

Suresh Chandra Satapathy

Vikrant Bhateja

Margarita N. Favorskaya

T. Adilakshmi *Editors*



# Smart Computing Techniques and Applications

Proceedings of the Fourth International  
Conference on Smart Computing and  
Informatics, Volume 2



# **Smart Innovation, Systems and Technologies**

**Volume 224**

## **Series Editors**

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

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

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

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

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

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

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

More information about this series at <http://www.springer.com/series/8767>

Suresh Chandra Satapathy · Vikrant Bhateja ·  
Margarita N. Favorskaya · T. Adilakshmi  
Editors

# Smart Computing Techniques and Applications

Proceedings of the Fourth International  
Conference on Smart Computing  
and Informatics, Volume 2



Springer

*Editors*

Suresh Chandra Satapathy  
School of Computer Engineering  
KIIT University  
Bhubaneswar, Odisha, India

Margarita N. Favorskaya  
Informatics and Computer Techniques  
Reshetnev Siberian State University  
of Science and Technologies  
Krasnoyarsk, Russia

Vikrant Bhateja  
Department of Electronics  
and Communication Engineering  
Shri Ramswaroop Memorial Group  
of Professional Colleges (SRMGP)  
Lucknow, Uttar Pradesh, India

Dr. A.P.J. Abdul Kalam Technical  
University  
Lucknow, Uttar Pradesh, India

T. Adilakshmi  
Department of Computer Science  
and Engineering  
Vasavi College of Engineering  
Hyderabad, India

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-981-16-1501-6

ISBN 978-981-16-1502-3 (eBook)

<https://doi.org/10.1007/978-981-16-1502-3>

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

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

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

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

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

# **Conference Committees**

## **Chief Patrons**

Sri. M. Krishna Murthy, Secretary, VAE  
Sri. P. Balaji, CEO, VCE

## **Patron**

Dr. S. V. Ramana, Principal, VCE

## **Honorary Chair**

Dr. Lakhmi Jain, Australia

## **General Chair**

Dr. Suresh Chandra Satapathy, KIIT DU, Bhubaneswar

## **Organizing Chair**

Dr. T. Adilakshmi, Professor and HOD, CSE, VCE

## **Publication Chairs**

Dr. Nagaratna P. Hegde, Professor, CSE, VCE  
Dr. Vikrant Bhateja, SRMGPC, Lucknow, UP, India

## **Program Committee**

Dr. S. Ramachandram, Former Vice Chancellor, OU  
Dr. Banshidhar Majhi, Director, IITDM, Kancheepuram  
Dr. Siba K. Udgata, Professor, HCU  
Dr. Sourav Mukhopadhyay, Associate Professor, IIT Kharagpur  
Dr. P. Radha Krishna, Professor, CSE, NIT Warangal  
Dr. M. M. Gore, Professor, MNNIT, Allahabad  
Dr. S. M. Hegde, Professor, NIT Surathkal  
Dr. Bapi Raju S., Professor, IIIT Hyderabad  
Dr. Rajendra Hegadi, Associate Professor, IIIT Dharwad  
Dr. S. Sameen Fatima, Former Principal, OU  
Dr. K. Shyamala, Professor, OU  
Dr. Naveen Sivadasan, TCS Innovation Labs, Hyderabad  
Dr. Badrinath G. Srinivas, Research Scientist—III, Amazon Development Center, Hyderabad  
Dr. Ravindra S. Hegadi, Professor, PAH Solapur University  
Dr. S. P. Algur, Professor and Chairman, CSE, Rani Channamma University, Belgavi  
Dr. R. Sridevi, HOD, Department of CSE, JNTUH

## **International Advisory Committee/Technical Program Committee**

Dr. Rammohan, South Korea  
Dr. Kailash C. Patidar, South Africa  
Dr. Naeem Hanoon, Malaysia  
Dr. Vimal Kumar, the University of Waikato, New Zealand  
Dr. Akshay Sadananda Uppinakudru Pai, University of Copenhagen, Denmark  
Dr. K. C. Santosh, the University of South Dakota  
Dr. Ayush Goyal, Texas A&M University, Kingsville  
Dr. Sobhan Babu, Associate Professor, IIT Hyderabad  
Dr. D. V. L. N. Somayajulu, Director, IIIT, Kurnool  
Dr. Siba Udgata, Professor, HCU  
Dr. R. B. V. Subramaanyam, Professor, NITW  
Dr. S. G. Sanjeevi, Professor, NITW

Dr. Sanjay Sengupta, CSIR, New Delhi  
Dr. A. Govardhan, Rector, JNTU Hyderabad  
Prof. Chintan Bhatt, Chandubhai Patel Institute of Technology, Gujarat  
Dr. Munesh Chandra Trivedi, ABES Engineering College, Ghaziabad  
Dr. Alok Aggarwal, Professor  
Dr. Anuja Arora, Jaypee Institute of Information Technology, Noida, India  
Dr. Divakar Yadav, Associate Professor, MMMUT, Gorakhpur, India  
Dr. Kuda Nageswar Rao, Andhra University, Visakhapatnam  
Dr. M. Ramakrishna Murthy, ANITS, Visakhapatnam  
Dr. Suberna Kumar, MVGR, Vizianagaram  
Dr. J. V. R. Murthy, Director Incubation and IPR, JNTU Kakinada  
Dr. D. Ravi, IDRBT, Hyderabad  
Dr. Badrinath G. Srinivas, Research Scientist-III, Amazon Development Center, Hyderabad  
Dr. K. Shyamala, Professor, OU  
Dr. P. V. Sudha, Professor, OU  
Dr. M. A. Hameed, Assistant Professor, OU  
Dr. B. Sujatha, Assistant Professor, OU  
Dr. T. Adilakshmi, Professor and HOD, CSE, VCE  
Dr. Nagaratna P. Hegde, Professor, CSE, VCE  
Dr. V. Sireesha, Assistant Professor, CSE, VCE

## **Organizing Committee**

Dr. D. Baswaraj, Professor, CSE, VCE  
Dr. K. Srinivas, Associate Professor, CSE, VCE  
Dr. V. Sireesha, Assistant Professor, CSE, VCE  
Mr. S. Vinay Kumar, Assistant Professor, CSE, VCE  
Mr. M. Sashi Kumar, Assistant Professor, CSE, VCE  
M. Sunitha Reddy, Assistant Professor, CSE, VCE  
R. Sateesh Kumar, Assistant Professor, CSE, VCE  
Mr. T. Nishitha, Assistant Professor, CSE, VCE

## **Publicity Committee**

Dr. M. Shanmukhi, Professor, CSE, VCE  
Mr. C. Gireesh, Assistant Professor, CSE, VCE  
Ms. T. Jalaja, Assistant Professor, CSE, VCE  
Mr. I. Navakanth, Assistant Professor, CSE, VCE  
Ms. S. Komal Kaur, Assistant Professor, CSE, VCE  
Mr. T. Saikanth, Assistant Professor, CSE, VCE

Ms. K. Mamatha, Assistant Professor, CSE, VCE

Mr. P. Narasiah, Assistant Professor, CSE, VCE

## **Website Committee**

Mr. S. Vinay Kumar, Assistant Professor, CSE, VCE

Mr. M. S. V. Sashi Kumar, Assistant Professor, CSE, VCE

# Preface

This volume contains the selected papers presented at the 4th International Conference on Smart Computing and Informatics (SCI 2020) organized by the Department of Computer Science and Engineering, Vasavi College of Engineering (Autonomous), Ibrahimbagh, Hyderabad, Telangana, during October 9–10, 2020. It provided a great platform for researchers from across the world to report, deliberate, and review the latest progress in the cutting-edge research pertaining to smart computing and its applications to various engineering fields. The response to SCI 2020 was overwhelming with a good number of submissions from different areas relating to artificial intelligence, machine learning, cognitive computing, computational intelligence, and its applications in main tracks. After a rigorous peer review with the help of technical program committee members and external reviewers, only quality papers were accepted for presentation and subsequent publication in this volume of SIST series of Springer.

Several special sessions were floated by eminent professors in cutting-edge technologies such as blockchain, AI, ML, data engineering, computational intelligence, big data analytics and business analytics, and intelligent systems. Eminent researchers and academicians delivered talks addressing the participants in their respective fields of proficiency. Our thanks are due to Prof. Roman Senkerik, Head of AI Lab, Tomas Bata University in Zlin, Czech Republic; Shri. Shankarnarayan Bhat, Director Engineering, Intel Technologies India Pvt. Ltd.; Ms. Krupa Rajendran, Assoc. VP, HCL Technologies; and Mr. Aninda Bose, Springer, India, for delivering keynote addresses for the benefit of the participants. We would like to express our appreciation to the members of the technical program committee for their support and cooperation in this publication. We are also thankful to the team from Springer for providing a meticulous service for the timely production of this volume.

Our heartfelt thanks to Shri. M. Krishna Murthy, Secretary, VAE; Sri. P. Balaji, CEO, VCE; and Dr. S. V. Ramana, Principal, VCE, for extending support to conduct this conference in Vasavi College of Engineering. Profound thanks to Prof. Lakhmi C. Jain, Australia, for his continuous guidance and support from the beginning of the conference. Without his support, we could never have executed such a mega event. We are grateful to all the eminent guests, special chairs, track managers, and reviewers

for their excellent support. A special vote of thanks to numerous authors across the country as well as abroad for their valued submissions and to all the delegates for their fruitful discussions that made this conference a great success.

Editorial Board of SCI 2020

Bhubaneswar, India  
Lucknow, India  
Krasnoyarsk, Russia  
Hyderabad, India

Suresh Chandra Satapathy  
Vikrant Bhateja  
Margarita N. Favorskaya  
T. Adilakshmi

# **List of Special Sessions Collocated with SCI-2020**

## **SS\_01: Next-Generation Data Engineering and Communication Technology**

Dr. Suresh Limkar, AISSMS Institute of Information Technology, Pune

## **SS\_02: Artificial Intelligence and Machine Learning Applications (AIML)**

Dr. Sowmya V., CEN, Amrita Vishwa Vidyapeetham, Coimbatore

Dr. Anand Kumar M., NIT Karnataka

Dr. M. Venkatesan, NIT Karnataka

Prof. Soman K. P., Amrita Vishwa Vidyapeetham, Coimbatore

## **SS\_03: Advances in Computational Intelligence and Its Applications**

Dr. C. Kishor Kumar Reddy, Stanley College of Engineering and Technology for Women, Hyderabad

P. R. Anisha, Stanley College of Engineering and Technology for Women, Hyderabad

**SS\_04: Blockchain Technology: Foundations, Challenges, and Applications**

Prof. Sandeep Kumar Panda, Faculty of Science and Technology, ICFAI Foundation for Higher Education, Hyderabad

Prof. Santosh Kumar Swain, School of Computer Engineering, KIIT (Deemed to be) University, Bhubaneswar

**SS\_05: Application of Machine Learning for Intelligent System Design**

Dr. Minakhi Rout, KIIT (Deemed to be) University, Bhubaneswar

**SS\_06: Advances in Big Data Analytics and Business Intelligence**

Dr. Vijay B. Gadicha, G. H. Raisoni Academy of Engineering and Technology, Nagpur

Dr. Ajay B. Gadicha, P. R. Pote College of Engineering and Management, Amravati

**SS\_07: Recent Advances in Artificial intelligence-Applications, Challenges, and Future Trends**

Dr. S. Velliangiri, CMR Institute of Technology, Hyderabad

Dr. P. Karthikeyan, Presidency University, Bengaluru

Dr. Iwin Thanakumar Joseph, KITS, Coimbatore

# Contents

<b>An Intelligent Tracking Application for Post-pandemic .....</b>	1
V. Roopa, R. Vasikaran, M. Sriram Karthik, S. Sindhu, and N. Vaishnavi	
<b>Investigation on the Influence of English Expertise on Non-native English-Speaking Students' Scholastic Performance Using Data Mining .....</b>	9
Suhashini Sailesh Bhaskaran and Mansoor Al Aali	
<b>Machine Learning Algorithms for Modelling Agro-climatic Indices: A Review .....</b>	15
G. Edwin Prem Kumar and M. Lydia	
<b>Design of Metal-Insulator-Metal Based Stepped Impedance Square Ring Resonator Dual-Band Band Pass Filter .....</b>	25
Surendra Kumar Bitra and M. Sridhar	
<b>Covid-19 Spread Analysis .....</b>	31
Srinivas Kanakala and Vempaty Prashanthi	
<b>Social Media Anatomy of Text and Emoji in Expressions .....</b>	41
Shelley Gupta, Ojas Garg, Radhika Mehrotra, and Archana Singh	
<b>Development of Machine Learning Model Using Least Square-Support Vector Machine, Differential Evolution and Back Propagation Neural Network to Detect Breast Cancer .....</b>	51
Madhura D. Vankar and G. A. Patil	
<b>Distributed and Decentralized Attribute Based Access Control for Smart Health Care Data .....</b>	67
B. Ravinder Reddy and T. Adilakshmi	
<b>Dynamic Node Identification Management in Hadoop Cluster Using DNA .....</b>	75
J. Balaraju and P. V. R. D. Prasada Rao	

<b>A Scientometric Inspection of Research Based on WordNet Lexical During 1995–2019 .....</b>	87
Minni Jain, Gaurav Sharma, and Amita Jain	
<b>Sentiment Analysis of an Online Sentiment with Text and Slang Using Lexicon Approach .....</b>	95
Shelley Gupta, Shubhangi Bisht, and Shirin Gupta	
<b>Fuzzy Logic Technique for Evaluation of Performance of Load Balancing Algorithms in MCC .....</b>	107
Divya, Harish Mittal, Niyati Jain, Bijender Bansal, and Deepak Kr. Goyal	
<b>Impact of Bio-inspired Algorithms to Predict Heart Diseases .....</b>	121
N. Sree Sandhya and G. N. Beena Bethel	
<b>Structured Data Extraction Using Machine Learning from Image of Unstructured Bills/Invoices .....</b>	129
K. M. Yindumathi, Shilpa Shashikant Chaudhari, and R. Aparna	
<b>Parallel Enhanced Chaotic Model-Based Integrity to Improve Security and Privacy on HDFS .....</b>	141
B. Madhuravani, N. Chandra Sekhar Reddy, and Boggula Lakshmi	
<b>Exploring the Fog Computing Technology in Development of IoT Applications .....</b>	149
Chaitanya Nukala, Varagiri Shailaja, A. V. Lakshmi Prasuna, and B. Swetha	
<b>NavRobotVac: A Navigational Robotic Vacuum Cleaner Using Raspberry Pi and Python .....</b>	159
Shaik Abdul Nabi and Mettu Krishna Vardhan	
<b>A Hybrid Clinical Data Predication Approach Using Modified PSO .....</b>	169
P. S. V. Srinivasa Rao, Mekala Srinivasa Rao, and Ranga Swamy Sirisati	
<b>Software Defect Prediction Using Optimized Cuckoo Search Based Nature-Inspired Technique .....</b>	183
C. Srinivasa Kumar, Ranga Swamy Sirisati, and Srinivasulu Thonukunuri	
<b>Human Facial Expression Recognition Using Fusion of DRLDP and DCT Features .....</b>	193
M. Avanthi and P. Chandra Sekhar Reddy	
<b>Brain Tumor Classification and Segmentation Using Deep Learning .....</b>	201
Manohar Madgi, Shantala Giraddi, Geeta Bharamagoudar, and M. S. Madhur	
<b>A Hybrid Approach Using ACO-GA for Task Scheduling in Cloud .....</b>	209
Simran Shrivastava, Sonika Shrivastava, and Lalit Purohit	
<b>K-Means Algorithm-Based Text Extraction from Complex Video Images Using 2D Wavelet .....</b>	219
Divya Saxena and Anubhav Kumar	

Contents	xv
<b>Opinion Mining-Based Conjoint Analysis of Consumer Brands . . . . .</b>	227
Kumar Ravi, Aishwarya Priyadarshini, and Vadlamani Ravi	
<b>Task Scheduling in Cloud Using Improved Genetic Algorithm . . . . .</b>	241
Shyam Sunder Pabboju and T Adilakshmi	
<b>Sentiment Analysis for Telugu Text Using Cuckoo Search Algorithm . . . . .</b>	253
G. Janardana Naidu and M. Seshashayee	
<b>Automation of Change Impact Analysis for Python Applications . . . . .</b>	259
T. Jalaja, T. Adilakshmi, and P. S. R. Abhishek	
<b>Enhancing Item-Based Collaborative Filtering for Music Recommendation System . . . . .</b>	269
M. Sunitha, T. Adilakshmi, and Mir Zahed Ali	
<b>Deep Learning-Based Enhanced Classification Model for Pneumonia Disease . . . . .</b>	285
S. Jeba Priya, S. Joshua Jaistein, G. Naveen Sundar, and T. Raja Sundrapandiyaneebanon	
<b>Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches . . . . .</b>	295
J. Srinivas, K. Venkata Subba Reddy, G. J. Sunny Deol, and P. Varaprasada Rao	
<b>A Novel Method for Optimizing Data Consumption by Enabling a Custom Plug-In . . . . .</b>	307
Vijay A. Kanade	
<b>An Effective Mechanism for the Secure Transmission of Medical Images Using Compression and Public Key Encryption Mechanism . . . . .</b>	317
T. K. Ratheesh and Varghese Paul	
<b>A Systematic Survey on Radar Target Detection Techniques in Sea Clutter Background . . . . .</b>	327
R. Navya and R. Devaraju	
<b>An Ensemble Model for Predicting Chronic Diseases Using Machine Learning Algorithms . . . . .</b>	337
B. Manjulatha and Suresh Pabboju	
<b>COVID-19 Face Mask Live Detection Using OpenCV . . . . .</b>	347
Anveshini Dumala, Anusha Papasani, and Sireesha Vikkurty	
<b>Chest X-Ray Image Analysis of Convolutional Neural Network Models with Transfer Learning for Prediction of COVID Patients . . . . .</b>	353
M. Shyamala Devi, P. Swathi, N. Pavan Kumar, Ravi Varma Tungala, Saranya Vivekanandan, and Priyanka Moorthy	

<b>Predicting Customer Loyalty in Banking Sector with Mixed Ensemble Model and Hybrid Model .....</b>	363
Jesmi Latheef and S. Vineetha	
<b>Design Patterns and Microservices for Reengineering of Legacy Web Applications .....</b>	373
V. Dattatreya, K. V. Chalapati Rao, and M. Raghava	
<b>A Comparative Study on Single Image Dehazing Using Convolutional Neural Network .....</b>	383
Poornima Shrivastava, Roopam Gupta, Asmita A. Moghe, and Rakesh Arya	
<b><i>Plasmodium falciparum</i> Detection in Cell Images Using Convolutional Neural Network .....</b>	395
Smaranjit Ghose, Suhrid Datta, C. Malathy, and M. Gayathri	
<b>An Online Path Planning with Modified Autonomous Parallel Parking Controller for Collision Avoidance .....</b>	403
Naitik M. Nakrani and Maulin M. Joshi	
<b>Real-Time Proximity Sensing Module for Social Distancing and Disease Spread Tracking .....</b>	415
Sreeja Rajesh, Varghese Paul, Abdul Adil Basheer, and Jibin Lukose	
<b>Automatic Depression Level Analysis Using Audiovisual Modality .....</b>	425
Aishwarya Chordia, Mihir Kale, Mukta Mayee, Preksha Yadav, and Suhasini Itkar	
<b>A Notification Alert System with Heartbeat and Temperature Sensors for Abnormal Health Conditions .....</b>	441
V. Sireesha, M. S. V. Sashi Kumar, S. Vinay Kumar, and R. M. Shiva Krishna	
<b>Recommender System for Resolving the Cold Start Challenges Using Classification .....</b>	451
Chandrima Roy, Siddharth Swarup Rautray, and Manjusha Pandey	
<b>A Skyline Based Technique for Web Service Selection .....</b>	461
Yamini Barge, Lalit Purohit, and Soma Saha	
<b>Novel Trust Model to Enhance Availability in Private Cloud .....</b>	473
Vijay Kumar Damera, A. Nagesh, and M. Nagaratna	
<b>Feature Impact on Sentiment Extraction of TEnglish Code-Mixed Movie Tweets .....</b>	487
S. Padmaja, M. Nikitha, Sasidhar Bandu, and S. Sameen Fatima	
<b>Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning .....</b>	495
M. Shyamala Devi, P. Swathi, M. Purushotham Reddy, V. Deepak Varma, A. Praveen Kumar Reddy, Saranya Vivekanandan, and Priyanka Moorthy	

<b>An Adaptive Correlation Clustering-Based Recommender System for the Long-Tail Items .....</b>	505
Soanpet Sree Lakshmi, T. AdiLakshmi, and Bakshi Abhinitth	
<b>Plant Leaf Identification Using HOG and Random Forest Regressor ...</b>	515
Jyotisagar Bal, Manas Kumar Rath, and Prasanta Kumar Swain	
<b>Deep Learning Based Facial Feature Detection for Ethnicity Recognition .....</b>	527
Sujitha Juliet Devaraj, R. Catherine Joy, I. Santhosh, and I. C. Kevin	
<b>Scanning Array Antenna Radiation Pattern Design Containing Asymmetric Null Steering Based on L-ASBO .....</b>	535
Anitha Suresh, C. Puttamadappa, and Manoj Kumar Singh	
<b>A Modified Novel Signal Flow Graph and Memory-Based Radix-8 FFT Processor Design .....</b>	545
A. Anitha, B. Triveni, Pinninti Kishore, and Makkenna Madhavi Latha	
<b>Vouch augmented Program Courses Recommendation System for E-Learning .....</b>	555
K. B. V. Rama Narasimham, C. V. P. R. Prasad, J. Jyothirmai, and M. Raghava	
<b>Heart Disease Prediction Using Extended KNN (E-KNN) .....</b>	565
R. Sateesh Kumar and S. Sameen Fatima	
<b>Prediction Analysis of Diabetes Using Machine Learning .....</b>	573
Srikanth Bethu, G. Charles Babu, B. Sankara Babu, and V. Anusha	
<b>Enhanced Goodput and Energy-Efficient Geo-Opportunistic Routing Protocol for Underwater Wireless Sensor Networks .....</b>	585
V. Baranidharan, B. Moulieshwaran, V. Karthik, R. Sanjay, and V. Thangabalaji	
<b>Early Detection of Pneumonia from Chest X-Ray Images Using Deep Learning Approach .....</b>	595
Prateek Sarangi, Pradosh Priyadarshan, Swagatika Mishra, Adyasha Rath, and Ganapati Panda	
<b>Detection of Network Anomaly Sequences Using Deep Recurrent Neural Networks .....</b>	605
R. Ravinder Reddy, K. Ayyappa Reddy, C. Madan Kumar, and Y. Ramadevi	
<b>Driver Drowsiness Detection Using Convolution Neural Networks .....</b>	617
P. Ravi Teja, G. Anjana Gowri, G. Preethi Lalithya, R. Ajay, T. Anuradha, and C. S. Pavan Kumar	
<b>Glaucoma Detection Using Morphological Filters and GLCM Features .....</b>	627
Babita Pal, Vikrant Bhatija, Archita Johri, Deepika Pal, and Suresh Chandra Satapathy	

<b>Analysis of Encryption Algorithm for Data Security in Cloud Computing .....</b>	637
Arijit Dutta, Akash Bhattacharyya, Chinmaya Misra, and Sudhangshu Sekhar Patra	
<b>A Machine Learning Approach in Data Perturbation for Privacy-Preserving Data Mining .....</b>	645
Jayanti Dansana and Adarsh Singh	
<b>IoT Service-Based Crowdsourcing Ecosystem in Smart Cities .....</b>	655
Arijit Dutta, Ruben Roy, Chinmaya Misra, and Kamakhy Singh	
<b>On Interior, Exterior, and Boundary of Fuzzy Soft Multi-Set Topology .....</b>	663
S. A. Naisal and K. Reji Kumar	
<b>Early Prediction of Pneumonia Using Convolutional Neural Network and X-Ray Images .....</b>	673
C. Kishor Kumar Reddy, P. R. Anisha, and K. Apoorva	
<b>Predicting the Energy Output of Wind Turbine Based on Weather Condition .....</b>	683
P. R. Anisha, C. Kishor Kumar Reddy, and Nuzhat Yasmeen	
<b>A Study and Early Identification of Leaf Diseases in Plants Using Convolutional Neural Network .....</b>	693
R. Madana Mohana, C. Kishor Kumar Reddy, and P. R. Anisha	
<b>Distributed and Energy Balanced Routing for Heterogeneous Wireless Sensor Network .....</b>	711
Shivani S. Bhasgi and Sujatha Terdal	
<b>ESRRAK-Efficient Self-Route Recovery in Wireless Sensor Networks Using ACO Aggregation and K-Means Algorithm .....</b>	719
Abhijit Halkai and Sujatha Terdal	
<b>An Explanation of Personal Variations on the Basis of Model Theory or RKT .....</b>	729
K. Reji Kumar	
<b>Fingerprint Enhancement Using Fuzzy Logic and Deep Neural Network .....</b>	735
Sridevi Sarraju and Franklin Bein	
<b>Gaussian Filter-Based Speech Segmentation Algorithm for Gujarati Language .....</b>	747
Priyanka Vishwas Gujarathi and Sandip Raosaheb Patil	
<b>Smart Farming Technology with AI &amp; Block Chain: A Review .....</b>	757
Deepali Jawale and Sandeep Malik	

Contents	xix
<b>Design and Development of Electronic System for Predicting Nutrient Deficiency in Plants .....</b>	765
Amruta Chore and Dolly Thankachan	
<b>Classification of Hyperspectral Images with Various Spatial Features .....</b>	773
Sandhya Shinde and Hemant Patidar	
<b>Detecting and Classifying Various Diseases in Plants .....</b>	781
Rashmi Deshpande and Hemant Patidar	
<b>Offline Handwritten Dogra Script Recognition Using Convolutional Neural Network .....</b>	789
Reya Sharma, Baijnath Kaushik, and Naveen Kumar Gondhi	
<b>“Device Design of 30 and 10 nm Triple Gate Single Finger Fin-FET for on Current (<math>I_{ON}</math>) and off Current (<math>I_{OFF}</math>) Measurement” .....</b>	799
Sarika M. Jagtap and Vitthal J. Gond	
<b>Fact Check Using Multinomial Naive Bayes .....</b>	813
Madhavi Ajay Pradhan, Ankita Shinde, Rohan Dhiman, Shreyas Ghorpade, and Swapnil Jawale	
<b>Author Index .....</b>	825

## About the Editors

**Suresh Chandra Satapathy** is currently working as Professor, KIIT Deemed to be University, Odisha, India. He obtained his Ph.D. in Computer Science Engineering from JNTUH, Hyderabad, and master's degree in Computer Science and Engineering from National Institute of Technology (NIT), Rourkela, Odisha. He has more than 27 years of teaching and research experience. His research interest includes machine learning, data mining, swarm intelligence studies and their applications to engineering. He has more than 98 publications to his credit in various reputed international journals and conference proceedings. He has edited many volumes from Springer AISC, LNEE, SIST and LNCS in the past, and he is also the editorial board member in few international journals. He is a senior member of IEEE and a life member of Computer Society of India. Currently, he is National Chairman of Division-V (Education and Research) of Computer Society of India.

**Vikrant Bhateja** is Associate Professor, Department of ECE in SRMGPC, Lucknow. His areas of research include digital image and video processing, computer vision, medical imaging, machine learning, pattern analysis and recognition. He has around 160 quality publications in various international journals and conference proceedings. He is associate editor of IJSE and IJACI. He has edited more than 30 volumes of conference proceedings with Springer Nature and is presently EiC of IGI Global: IJNCR journal.

**Dr. Margarita N. Favorskaya** is Professor and Head of the Department of Informatics and Computer Techniques at Reshetnev Siberian State University of Science and Technology, Russian Federation. Professor Favorskaya is a member of KES organization since 2010, the IPC member and Chair of invited sessions of over 30 international conferences. She serves as Reviewer in international journals (*Neurocomputing*, *Knowledge Engineering and Soft Data Paradigms*, *Pattern Recognition Letters*, *Engineering Applications of Artificial Intelligence*), Associate Editor of *Intelligent Decision Technologies Journal*, *International Journal of Knowledge-Based and Intelligent Engineering Systems* and *International Journal of Reasoning-based Intelligent Systems*, Honorary Editor of the *International Journal of Knowledge Engineering and Soft Data Paradigms*, Reviewer, Guest Editor and Book Editor

(Springer). She is the author or the co-author of 200 publications and 20 educational manuals in computer science. She co-authored/co-edited seven books for Springer recently. She supervised nine Ph.D. candidates and is presently supervising four Ph.D. students. Her main research interests are digital image and videos processing, remote sensing, pattern recognition, fractal image processing, artificial intelligence and information technologies.

**Dr. T. Adilakshmi** is currently working as Professor and Head of the Department, Vasavi College of Engineering. She completed her Bachelor of Engineering from Vasavi College of Engineering, Osmania University, in the year 1986, and did her Master of Technology in CSE from Manipal Institute of Technology, Mangalore, in 1993. She received Ph.D. from Hyderabad Central University (HCU) in 2006 in the area of Artificial Intelligence. Her research interests include data mining, image processing, artificial intelligence, machine learning, computer networks and cloud computing. She has 23 journal publications to her credit and presented 28 papers at international and national conferences. She has been recognized as a research supervisor by Osmania University (OU) and Jawaharlal Nehru Technological University (JNTU). Two research scholars were awarded Ph.D. under her supervision, and she is currently supervising 11 Ph.D. students.

# An Intelligent Tracking Application for Post-pandemic



V. Roopa, R. Vasikaran, M. Sriram Karthik, S. Sindhu, and N. Vaishnavi

**Abstract** The Global Pandemic has created a huge impact in the country at a tremendous rate affecting 3,794,314 people, snatching the lives of 66,678 people. This has created a huge impact on the county's economy and socioeconomic and psychological status of the citizens of the country. This health also created a situation which altered the lifestyle of the people. Due to the prevailing conditions of India the government has been liberalizing the lockdown and implementing unlock provisions in the country as a result it is difficult to track people movements and migrations which is the only preferable means to control the spread of the disease and to find the suspectable positive cases. A tracking system with a mobile-based application can be implemented which tracks the users of the application's travel history and movements on a daily basis in the society which keeps the records of the user along with physiological parameters (Temperature) so that the changes, suspects can be easily tracked.

## 1 Introduction

In September 17, 2019, a strange man admitted with an unknown disease in Wuhan, Hubei province, in China, which turns out to be a reason for global pandemic called COVID-19 which drastically changed and affected the whole human lifestyle in the world. It had a direct impact on people's health made them fall sick, increase the suffering rate, ultimately end up in the number of the causality, according to WHO reports COVID-19 affects in different ways for different people, dry cough, fever and tiredness are the most common symptoms of COVID-19. Based on their immune

---

V. Roopa (✉) · R. Vasikaran · M. Sriram Karthik · S. Sindhu · N. Vaishnavi  
Department of Information Technology, Sri Krishna College of Technology, Coimbatore, India  
e-mail: [v.roopa@skct.edu.in](mailto:v.roopa@skct.edu.in)

R. Vasikaran  
e-mail: [18uit155@skct.edu.in](mailto:18uit155@skct.edu.in)

M. Sriram Karthik  
e-mail: [18uit144@skct.edu.in](mailto:18uit144@skct.edu.in)

system strength, a person recovers from the disease with or without hospitalization, the pandemic has not only affected people's health but also devasted every country's economy, made socialization a forbidden word and had indirect effect on people's mental health too. Unfortunately, the world has to run amid of this pandemic problem. India has gone through lockdown from past April 2020 which put the whole country into universal lockdown to ensure the reduction of spreading of coronavirus, yet people will suffer even huge if they put to stay in lockdown for long time as we said the world has to run for all the people's basic needs and to ensure their economy, so month by month the grip has been loosened on this lockdown scenario, since then the COVID-19 cases are reaching its peak which is inevitable.

The COVID-19 tracking charts boosted exponentially after the stage by unlock programs taken by the government of India. The government has taken huge efforts to reduce the spread of COVID-19 and educating people about social distancing and asking them to maintain it. The government has closed every other economic business other than grocery shops for basic needs and hospitals for the health care in the beginning of the country's lockdown and it also closed the borders of the states, districts and cities and made transportation only allowed for daily needs of people making them to stagnant over a particular area but when the unlock of all this happened people are when made available to move from one place to another, the breakdown of virus also became unstoppable and the healthcare sector is struggling to track the patients movement because it helps to track and predict the people who are more prone to catch the disease because COVID-19 spreads through contact. And tracking of people is hard so. It is hard to predict the person who might be an active spreader of COVID-19 which will result in more patients without any contact history which ultimately result in social spread which is even more dangerous.

Currently in India, 3.77 M people are affected by COVID-19; at the same time, we cannot keep the people, so there has to be measure to ensure social distancing and that is just a promise which should be kept! but daily, we see news that it is not followed properly, so every shops, restaurant, supermarkets and places where people gather now asked to use a infra-red thermometer to check the customers and their temperatures are written down with their name and contact number, and this is the manual way the normal grocery, shopkeepers to big trading companies and other industries, but this is only useful for detecting if the person's temperature is high or low, but it cannot say its coronavirus or not and neither it seems to be useful for tracking only few that also in rare cases, so we see that there is no good solution to track the people who visit the places and trace them, so we came up with a solution that will be tremendously useful for the post lockdown scenario where the people came back to their new normal with facemasks and social distancing the project we took will be useful to track them and intimate them and also the shop owner and maintain the stats on who all come to their shops. For example, let us take a person is going to a coffee shop and then going to mall then to a restaurant after this post lockdown period then reaches home and this continues for about 2 weeks then he was diagnosed by COVID-19 at this time when the health department asks him where he been during all this period his memory cannot keep up all the details about where he visited and can barely remember if he misses anything it is going to end up as

social spread so many of people will affect by the disease and we cannot able to track people who are more prone to catch the disease so here comes in the game changer is our product which can be useful for tracking during post-pandemic time. The system works based on global positioning system to locate and track the users! first and foremost we are creating a mobile app so it will reach people easy and will have a good engagement. The system designed will have two provision: one for the users and another for the commercial user aka the shop owners and business. First comes the registration phase where give your details like name, age and address for basic information and you will be logged in to the application so whenever you enter any commercial center you have to show the device a QR code it is created by the work of RSA algorithm which appears in the app and the setup of IR thermometer which is interfaced with application using Bluetooth so that data will be automatically available in database through the commercial province of application it also shows up in your application for your own personal tracking for example if you have gone to a coffee shop you will show up the QR code then IR thermometer will detect your body temperature sends it to both you and the shop keeper and shows the name, the place at present aka coffee shop, body temperature at that time and these data's will keep on stacking. In the application, so in case, if you unfortunately affected by COVID-19, we will have a track of visits and able to warn them who visited the same coffee shop and also it is easier to detect where to track people and where not to track people and it can also show up areas which is more prone the disease thus helping other people to stay safe and lead a well-being post-pandemic life.

## 2 Literature Survey

Due to pandemic of COVID-19, all countries are looking toward mitigation plan to control the spread with the help of some modeling techniques [1]. Modeling techniques include wearing N95 mask with valves and to sanitize properly and to maintain distance between the wards. COVID-19 can be spread through respiratory droplets or due to close contact with the infected patients. SARS-CoV-2 was isolated from fecal samples of infected patients, which supports the significance of fecal-oral route in the transmission of SARS-CoV-2, but a WHO-China joint commission report has denied this route of transmission [2]. When a person affected by coronavirus knowingly or unknowingly makes contact with another person who is not affected by the virus despite any age groups. There is at most chance for the virus to proliferate. In comparison with traditional physical or hard sensors, MCS is inexpensive, since there is no need for network deployment, and its spatio-temporal coverage is outstanding. Two different approaches of MCS have been distinguished; they are (i) mobile sensing, which leverages raw data generated from the hardware sensors that are embedded in mobile devices (e.g., accelerometer, GPS, network connectivity, camera, or microphone, among others); and (ii) social sensing (or social networking), which leverages user-contributed data from OSN. The latter considers participants as ‘social sensors’, i.e., agents that provide information about their environment

through social media services after the interaction with other agents [3]. Information provided by the current situation regarding the pandemic should be updated properly through social media because majority of the population are accessing social media sites.

Social lockdown and distancing measures are the only tools available to fight the COVID-19 outbreak [4]. Prevention is always better than cure. Instead of suffering from the virus is good to stay away from the virus that can be done by distancing ourselves from others as of now social distancing is the only way to control this virus spread between wards until the arrival of vaccine. Early identification of non-compliance with the measures decreed in law RD 463/2020 and its subsequent extensions, such as (i) limitation of the freedom of movement of persons, (ii) opening to the public of unauthorized premises, establishments, areas of worship, etc., and (iii) agglomerations, among others. Social networks are an increasingly common way of reporting such events, and their identification can be used by authorities for resource planning [5]. People who do not follow rules and regulations given by the government to bring this spread to control should be punished in the name of advice and make people understand the seriousness behind every rules.

For instance, to make people know about the seriousness of this virus attack and to educate them about the death rate and recovery rate and to make them self-prepared if suppose they are prone to coronavirus. Its mission is to help citizens self-assess the probability of suffering the infectious disease COVID-19, to reduce the volume of calls to the health emergency number, informing the population, allowing an initial triage of possible cases and a subsequent follow-up by the Health Authorities [6]. If suppose, a ward is not feeling good and has all symptoms like fever, dry cough, tiredness or difficulties in breathing. Then it is always advisable to consult the nearby clinic and take necessary requirements to control the coronavirus spread. Smartphone-based contact-tracing applications are shown as a promising technology to finish or reduce the lockdown and quarantine measures. The technology of these mobile apps is based on the results of several years of research efforts on Mobile Computing, and particularly on Opportunistic Networking (OppNet) and MCS [7]. We are put under certain circumstances like where we are not allowed to roam around and to self-quarantine ourselves, We also cannot be without knowing the outside situation and condition about this virus spread at those situations. Smartphone technologies come in hand. For instance, we have several applications that give us a clear depiction about instant scenarios about the nation with its pandemic. It exhibits the data generation rate which can be calculated in time or frequency domain.

Data like death rate recovery and rate provided by healthcare should be updated properly to make people aware. Developing a novel vaccine is very crucial to defending the rapid endless of global burden of the COVID-19 pandemic. Big data can gain insights for vaccine/drug discovery against the COVID-19 pandemic. Few attempts have been made to develop a suitable vaccine for COVID-19 using big data within this short period of time [8]. We do not have a suitable vaccine developed for this virus. So, Big data plays a vital role in collecting information and control the spread to the maximum count and to control the spread until vaccine arrives. Conventional medicine alternatively called as allopathic medicine, biomedicine,

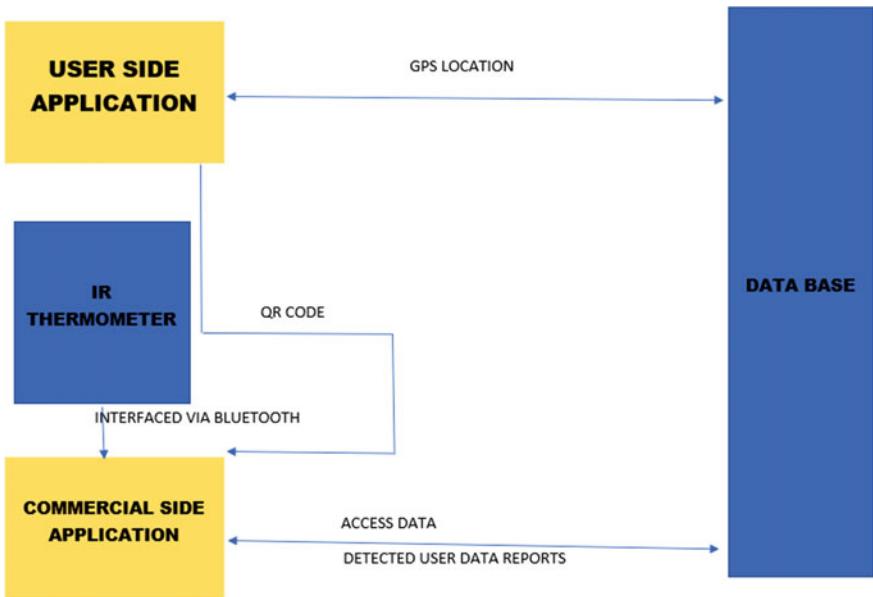
mainstream medicine, orthodox medicine and Western medicine, medical doctors and other professional healthcare providers such as nurses, therapists, and pharmacists use drugs, surgery or radiation to treat illnesses and eliminate symptoms [9]. Presently, there's no licensed medical medication for COVID-19 so people acquire some home remedies like drinking hot water, and having all rich nutritious food to improve immunity and perusing healthy lifestyle.

### 3 Proposed System

The system provides two user login: one for the common user and other for the business/commercial center; both are connected to the backend database. Both sides have a connection established between database and application. The user side application constantly tracks the location of the user using GPS (Global Positioning System) and updates in the database. This makes the continuous data tracking possible. The communication between the infra-red (IR) thermometer and the Commercial/business side application is made possible by interfacing the IR thermometer with the application via Bluetooth. At first when the user arrives at an any business/commercial center, the user's unique ID is scanned and detected by the commercial center side application as it can access user data from the database using a QR code. Then the user's body temperature is detected using the IR thermometer which is then sent to the business side user application via Bluetooth and it is automatically communicated to the database that gets updated for both the common user and business/commercial side user. Thus, the user activities and movements in such critical situation can be easily tracked and detected.

People have needs and duties which they had to fulfill which creates a necessity for them to move around in the society due to these indispensable needs the government issues social unlocks in the country which creates a difficulty in social distancing and tracking which is the source of blocking the spread of the disease and the only way to strategize the testing of susceptible cases. The proposed system focuses on tracking the movement of the user through a mobile-based application which is GPS enabled using unique ID and QR's for the users from one end and maintaining the user records at the business centers (restaurants, malls, supermarkets, garments, salon, etc.) at the other end so that the user and the backend database will be having the complete data of the user's activity and the business centers will be having the whole data and physiological parameter (temperature) of the customers which can be further used for tracking.

The system works by tracking the users of the application by tracking the user's activity using GPS (Global Positioning System), the users will be provided with a unique ID (QR in this case). The business centers will be provided separate login provisions to store the data of the customers including physiological data. When the users arrive at certain spot such as business centers the QR will be scanned by the business centers using the application for correct identification and their body



**Fig. 1** Working of intelligent AI tracking application

temperature will be checked using an infra-red (IR) thermometer which will be interfaced with the mobile app using the Bluetooth so that the data will be automatically sent to the database using the business center's application. The dataset will be safely stored in the databases which can be used for further tracking that can play a vital role during the unlocking process. This system has two separate provisions one for the common users and another for the business and commercial centers. The provision provides for the common users features QR code (which can be generated using RSA algorithm), tracking using GPS and data storage (Fig. 1).

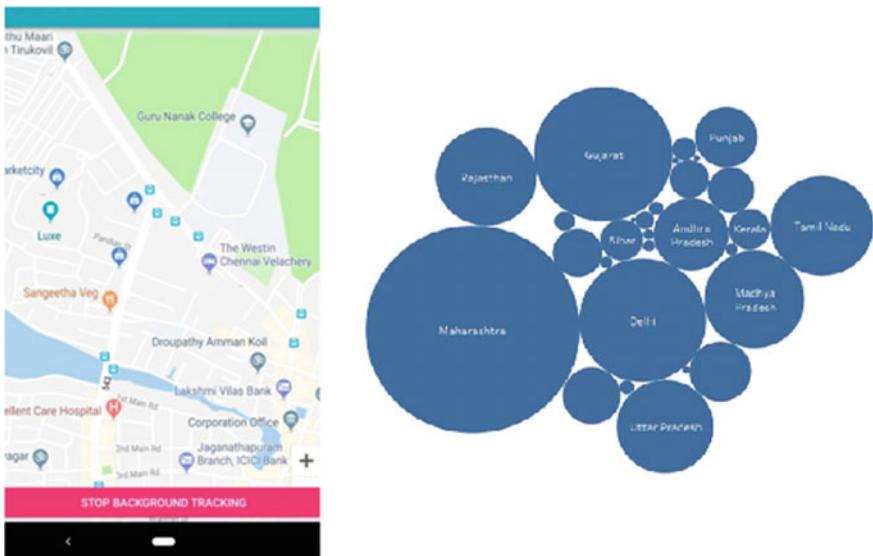
The provision provided for the commercial centers features Bluetooth interfacing between the IR Thermometer and the application, QR detection and data storage. Thus, these features make the tracking of the users at the social centers at ease and efficient. **RSA(Rivest-Shamir-Aldeman)** is an algorithm used to encrypt and decrypt messages. It is an asymmetric cryptographic algorithm. There are two different keys used in this algorithm. This method is called public-key cryptography. One of the keys can be given to anyone. The other key must be kept private. It is based on finding factors of a large composite number is difficult when the factors are prime numbers. It is called prime factorization and the key is also called key pair (public and private key) generator. The flow of application here first the customer enters a shop he will show up the QR code generated by the RSA algorithm thus the data transfers from user side to the commercial side which also includes the data of the individual's body temperature which is detected by the infra-red thermometer and all these data are stored in the database, thus this application will be enormously helpful for tracking the individual's movement in case of unfortunately getting affected by the pandemic

virus and also will play a major role in warning the user's who visited the same place as the patient. This application is simple, efficient and easy to deploy and we need not rely on any new system or hardware since it is available as an application, we can use it in mobile which most of them have in their hands.

## 4 Result and Discussion

As coronavirus is a contagious disease that spreads through social interaction between humans, Bluetooth tracing is vital for spread. Devices like mobile and tablets, it presents an ideal platform to introduce Bluetooth tracing software due to their ease of use and personalized usage. Therefore, several smartphone apps have been developed by governments, international agencies and other parties to mitigate the virus spread. In this paper, we analyze the large set of Bluetooth tracing with respect to different security and privacy metrics. We analyze Bluetooth tracing apps permission analysis, privacy analysis the security of the apps and review of the users. This approach has the benefit of not requiring network effects, because single individuals can track their locations without needing their contacts to have the app. The approach of logging location history is less private than direct tracing, but that may possibly be resolved with appropriate safeguards and redactions. Further, hybrid approaches involving both GPS data and Bluetooth proximity networks may prove to be useful to public health officials in modeling disease spread beyond just tracing. In Fig. 2 the Image 1 and 2(QR code) shows the QR code in the user side application being scanned by the commercial user application to access the user data and to record the body vital signs in the database to keep up to date records of the health conditions of the user this step is made automatic using the IR temperature interfaced via Bluetooth with the application and QR code to access the user database and directly storing them.

In Fig. 2, the Image 1 and 2(map) shows the person being tracked at each and every location they are traveling and every place they have visited at this situation, and this helps the government to track people in this post-pandemic situation to suspect the infectious places and infected persons so that we could avoid the spread of diseases. The application overall helps the government in keeping track of people who have been tested positive for the virus. It is also an excellent way to alert people about the number of infected cases in their area that have been identified as coronavirus positive or if they accidentally came in contact with a person suffering from COVID-19. The application requires being in running mode at all times to continue tracing individuals actively. The application can be used in such a way that it enables your smartphone to exchange the tracing keys periodically. This will help to locally store the unique ID of the people who have come into contact with the user. If later a user is tested positive for coronavirus, this method of cryptography will also ensure the privacy and the safety of your data. In addition to showing the data of the number of users who have been identified as positive, a map can be shown of the nearby area where people have been identified as positive for COVID-19.



**Fig. 2** Implementation of tracker on smart devices—cluster analysis

## References

1. Mahalle, P.N., Sable, N.P., Mahalle, N.P., Shinde, G.R.: Predictive analytics of COVID-19 using information, communication and technologies (2020)
2. Ying, S., et al.: Spread and control of COVID-19 in China and their associations with population movement, public health emergency measures, and medical resources, p. 24 (2020) [Online]. Available: <https://doi.org/10.1101/2020.02.24.20027623>
3. Li, R., Rivers, C., Tan, Q., et al.: The demand for inpatient and ICU beds for COVID-19 in the US: lessons from Chinese cities. medRxiv, 1–12 (2020). <https://doi.org/10.1101/2020.03.09.20033241>
4. Du, J., Vong, C.-M., Chen, C.L.P.: Novel efficient RNN and LSTM like architectures: recurrent and gated broad learning systems and their applications for text classification. IEEE Trans. Cybern. (2020). <https://doi.org/10.1109/TCYB.2020.2969705>
5. World Health Organization: Critical preparedness, readiness and response actions for COVID-19: interim guidance (2020)
6. Guo, B., Wang, Z., Yu, Z., et al.: Mobile crowd sensing and computing: the review of an emerging human-powered sensing paradigm. ACM Comput. Surv. (CSUR) **48**(1), 1–31 (2015)
7. International Labour Organization: The socioeconomic impact of COVID-19 in fragile settings: peace and social cohesion at risk. [https://www.ilo.org/global/topics/employment-promotion/recovers-and-reconstruction/WCMS\\_741158/langen/index.htm](https://www.ilo.org/global/topics/employment-promotion/recovers-and-reconstruction/WCMS_741158/langen/index.htm). Accessed 30 Apr 2020
8. Doran, D., Severin, K., Gokhale, S., et al.: Social media enabled human sensing for smart cities. AI Commun. (2015)
9. Adolph, C., Amano, K., Bang Jensen, B., et al.: Pandemic politics: timing state-level social distancing responses to COVID-19. medRxiv (2020)

# Investigation on the Influence of English Expertise on Non-native English-Speaking Students' Scholastic Performance Using Data Mining



Subhashini Sailesh Bhaskaran and Mansoor Al Aali

**Abstract** This investigation reports about an understanding of the connection between English skill and scholastic accomplishment of science students in Bahrain. Data from student information system were investigated by applying data mining techniques mainly decision tree algorithm. The results demonstrated a significant effect of English expertise on students' final cumulative grade point average (CGPA). These discoveries demonstrate that the English expertise of graduate students in a non-western polyglot scholastic background is significant for their scholarly accomplishment. Results from this investigation affirm the requirement for colleges in polyglot backgrounds to put resources into non-native English-speaking (L2) graduate students' English expertise toward the beginning of their scholastic projects. Instructional proposals are made, alongside recommendations for additional investigation.

## 1 Related Literature

### 1.1 *Language Expertise and Scholastic Accomplishment*

The acknowledgment that language helps in science and mathematics learning has prompted an expanded interest for substance territory expertise guidance [1–4]. Some research has created solid link between expertise and scholastic execution in science and arithmetic training (CCAAL 2010, p. 22). Disciplinary education, or substance territory explicit expertise, comprises of expertise aptitudes and learning that help graduate students' comprehension of ideas identified with a specific field of study, for example, science and arithmetic. Exceptional concern for disciplinary expertise in science and arithmetic instruction is significant for various reasons. Science writings are frequently testing to graduate students and require additional endeavors to

---

S. S. Bhaskaran (✉) · M. Al Aali  
Ahlia University, Manama, Bahrain  
e-mail: [sbhaskaran@ahlia.edu.bh](mailto:sbhaskaran@ahlia.edu.bh)

process. They are useful commonly and normally present thick and dynamic ideas, utilize new phrasing and language that graduate students are not likely to encounter in their day by day language use [5]. The expositive and specialized nature of science writings puts levels of popularity on graduate students' language aptitudes. The last incorporate knowing particular vocabulary; translating logical images and graphs; perceiving and understanding hierarchical examples regular to science writings; deriving principle thoughts, utilizing inductive and deductive thinking abilities; and perceiving circumstances and logical results connections [6]. Significant dominance of language skills and perusing expertise are in this way imperative for graduate students who are contemplating science and arithmetic, regardless of whether at essential, auxiliary or at tertiary level. This is considerably more so for graduate students in polyglot or bilingual backgrounds, who are not educated in their primary language but rather in a subsequent language (L2). Tooth (2006) contends that the particular semantic highlights that make science messages progressively thick and dynamic can cause perusing perception issues particularly for English L2 students. Because of a worldwide expanding relocation and development of individuals, bilingual and polyglot backgrounds are developing. It is assessed that half of the total populace utilizes more than one language or lingo in their regular day-to-day existence [7]. This extension of polyglots legitimizes more consideration for the expertise practices of L2 students in science and arithmetic training (cf. [8], Rhodes and Feder 2014). Most research into literacy practices in science and mathematics education comes from the richer western world. Unexpectedly, L2 examination into the rich polyglot backgrounds is scarce [9, 10]. The current paper is enlivened by the scarcity of research in the remainder of Middle East in regard to the job of language in science. It is an endeavor to fill this research gap, and specifically to all the more likely comprehend the job of graduate students' English expertise on scholarly accomplishment in a non-western, polyglot instructive background. It reports on a conceptual model that gives knowledge into the relationship between English expertise and scholastic performance of science college graduate students in Bahrain, and an experimental trial of that model.

This study analyzes the following questions:

- (a) Is there a direct relationship between non-native English-speaking students' scholastic English expertise and their scholastic accomplishment in science and mathematics instruction in Bahrain?
- (b) With respect to the level of English language expertise if there is noteworthy difference in the scholastic accomplishment of students in Bahrain?

## 2 Data Mining Process

The steps of Data Mining data preparation, data selection, transformation, modeling and evaluation are discussed below to unearth hidden information in the dataset that was used to relate English scores and CGPA of students.

### **3 Data Preparation**

Student dataset of a higher education offering different degrees at the undergraduate level was extracted from the student information system. The dataset pertains to graduated students belonging to 12 programs between 2003 and 2014. The size of the dataset is 646.

### **4 Data Selection**

Being the second step data selection encompasses the method of gathering necessary data supported by transformation of data which is the method of transforming data into the necessary layout necessary for modeling (Fayyad 1996). Data was obtained from the database using SQL queries related to English GPA and overall CGPA. Following the extraction several tables were clubbed into a single table. Cleaning was done by tackling missing values and variables were coded properly to allow the use of classification algorithms. A Few variables were obtained directly from the database. Certain features were computed or inferred based on other items present in various tables. Using feature selection, variables were extracted. These features were utilized in the following stage.

### **5 Modeling**

Modeling algorithms are used for the discovery of knowledge. Classification is used for categorizing data into different groups according to some conditions. A decision tree uses a tree of choices and their probable outcomes, involving opportunity incident outcomes, resource expenses, and efficiency. In order to find the relationship between English results and final GPA of science students decision tree was applied.

### The Algorithm

#### Stage 1:

The leaf is labeled with a similar class if the instances belong to similar class.

#### Stage 2:

For each attribute, the potential data will be figured and the gain in the data will be taken from the test on the attribute

#### Stage 3:

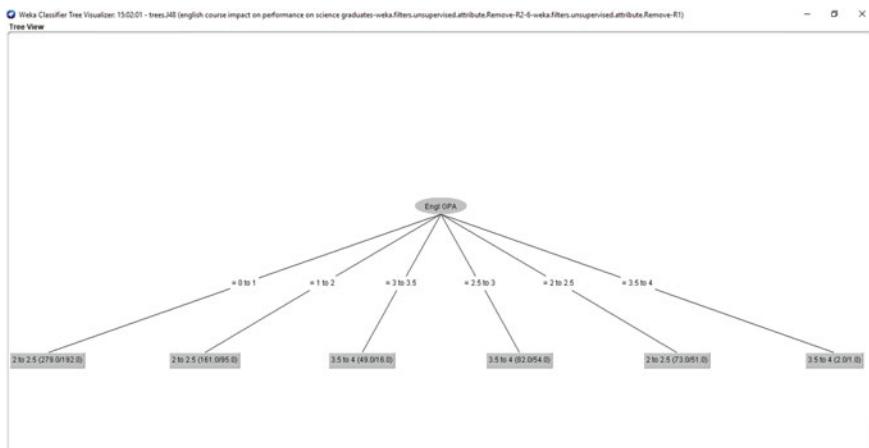
Finally the best attribute will be chosen depending upon the current selection parameter.

### == Detailed Accuracy By Class ==

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.364	0.205	0.394	0.364	0.378	0.163	0.665	0.410	3.5 to 4
	0.000	0.002	0.000	0.000	0.000	-0.018	0.586	0.215	2.5 to 3
	0.787	0.646	0.333	0.787	0.468	0.138	0.592	0.340	2 to 2.5
	0.035	0.074	0.146	0.035	0.056	-0.071	0.474	0.256	3 to 3.5
Weighted Avg.	0.336	0.263	0.241	0.336	0.253	0.062	0.579	0.314	

### == Confusion Matrix ==

```
a b c d <- classified as
63 1 97 12 | a = 3.5 to 4
18 0 83 12 | b = 2.5 to 3
29 0 148 11 | c = 2 to 2.5
50 0 116 6 | d = 3 to 3.5
```



The main aim of such a tree is to find the performance achieved by a student in terms of CGPA based on the performance in English courses. This tree helps in finding if English grades affect the overall GPA of the students. Such a knowledge could help the students and advisors to know if there is an impact of the English courses on the overall GPA and if there is an impact to coach the students on English courses. It can be seen from the decision tree output that when the GPA of English courses are lesser the final GPA of all courses is also lesser. Likewise when the GPA

of the English courses is higher, then it is the case of final GPA of all courses. When the English GPA is between 0 and 2.5, then the final GPA is between 2 and 2.5. When the English GPA is more than 2.5, then the final GPA falls between 3.5 and 4 showing a direct relationship between the performance of English courses and final GPA.

## 6 Discussion and Conclusion

This examination explored the job of English expertise on the scholastic accomplishment in an example of science and mathematics graduate students. Utilizing decision tree examinations, the investigation showed that students' scholastic English expertise is important for their scholarly accomplishment in a polyglot scholastic background. Consolidating two universally perceived perusing perception tests took into account an expansive estimation of graduate students' perusing abilities. A solid impact was found between graduate students' scholastic accomplishment and English results.

The added value of the research is threefold. Our research broadens the focus of L2 reading research in science and arithmetic instruction in a Bahraini polyglot background. As argued in the paper, essential research into perusing in developing nations is rare. Moreover, our investigation added considerably to our comprehension of the connection between L2 graduate students' English expertise and their science and arithmetic accomplishment at tertiary level. The examination substantiates past discoveries of the job of language in the science and arithmetic educational plan (displayed in the presentation), in light of information from a Middle Eastern background. Finally, by adopting a longitudinal strategy, it shows that even after 4 years of study, there is an indirect mediation connection between expertise and scholastic accomplishment. These discoveries have significant instructional ramifications. Our information affirms the requirement for colleges to put resources into L2 graduate students' expertise toward the beginning of their scholastic degrees. A Few universities in Bahrain have launched a mandatory Communicative program, orientation in English program where students are trained in language and interaction skills. This consideration ought to go past building vocabulary and dominance of word-perusing abilities, since this is not an assurance that graduate students can understand science writings (CCAAL 2010). Explicit thoughtfulness regarding preparing science content is along these lines required. A second instructional ramification has to do with the embodiment of creating perusing aptitudes: the more individuals read, the better they become at it [11–13]. Notwithstanding the CS courses, graduate students ought to take part in perusing widely as a feature of their scholarly courses. Despite the fact that college graduate students in Bahrain guarantee to esteem their reading material, they want to take in course content from different assets, for example, addresses and address notes [14, 15]. This conduct, joined with restricted access to perusing assets, confines graduate students' perusing improvement. It is in this way fundamental that college courses incorporate perusing assignments and give

adequate access to perusing materials. The finding that 10.3% of the graduate students revealed utilizing English in their home condition is in accordance with Bahrain being viewed as an ESL nation [16, 17]. The profoundly heterogeneous population, as far as age (from 17 to 35 years of age), and utilization of home language (speaking to 20 distinct dialects) demonstrates the requirement for establishments to cater to enormous disparities regarding age and language aptitudes inside their homerooms.

## References

1. Carnegie Council on Advancing Adolescent Literacy. Time to act: An agenda for advancing adolescent literacy for college and career success. New York, NY, Carnegie Corporation of New York (2010)
2. Fang, Z., Lamme, L., Pringle, R., Patrick, J., Sanders, J., Zmach, C., Henkel, M.: Integrating reading into middle school science: What we did, found and learned. *Int. J. Sci. Edu.* **30**(15), 2067–2089 (2008). <https://doi.org/10.1080/09500690701644266>
3. Hand, B.M., Alvermann, D.E., Gee, J., Guzzetti, B.J., Norris, S.P., Phillips, L.M., . . . Yore, L.D.: Guest editorial. Message from the Bisland group: what is literacy in science literacy? *J. Res. Sci. Teach.* **40**(7), 607–615 (2003)
4. Norris, S.P., Phillips, L.M.: How literacy in its fundamental sense is central to scientific literacy. *Sci. Edu.* **87**, 224–240 (2003)
5. Palinscar, A.: The next generation science standards and the common core state standards: Proposing a happy marriage. *Sci. Children* **51**(1), 10–15 (2013)
6. Barton, M.L., Jordan, D.L.: Teaching reading in science. A supplement to teaching reading in the content areas: If not me, then who? McREL, Aurora, CO (2001)
7. Ansaldi, A.I., Marcotte, K., Schererc, L., Raboyeaua, G.: Languagetherapy and bilingual aphasia: clinical implications of psycholinguistic and neuroimaging research. *J. Neurolinguistics* **21**, 539–557 (2008)
8. Barwell, R., Barton, B., Setati, M.: Multilingual issues in mathematics education: introduction. *Educ. Stud. Math.* **64**(2), 113–119 (2007)
9. Paran, A., Williams, E.: Editorial: reading and literacy in developing countries. *J. Res. Read.* **30**(1), 1–6 (2007)
10. Pretorius, E.J., Mampuru, D.M.: Playing football without a ball: language, reading and academic performance in a high-poverty school. *J. Res. Read.* **30**(1), 38–58 (2007)
11. Cox, K.E., Guthrie, J.T.: Motivational and cognitive contributions to students' amount of reading. *Contemp. Edu. Psychol.* **26**, 116–131 (2001)
12. Mullis, I. V. S., Martin, M. O., Kennedy, A. M., Foy, P.: PIRLS 2006 international report; IEA's progress in international reading literacy study in primary schools in 40 countries. Boston, MA: International Association for the Evaluation of Educational Achievement (IEA) (2007)
13. Organisation for Economic Co-operation and Development. PISA 2009 assessment framework, key competencies in reading, mathematics and science. Paris, France, Author (2009)
14. Owusu-Acheaw, M., Larson, A.G.: Reading habits among students and its effect on academic performance: A study of students of Koforidua polytechnic. *Lib. Phil. Practice, Paper* **1130**, 1–22 (2014)
15. Stoffelsma, L.: Short-term gains, long-term losses? A diary study on literacy practices in Ghana. *J. Res Read.* (2018). <https://doi-org.ru.idm.oclc.org/10.1111/1467-9817.12136>
16. Ahulu, S.: Hybridized English in Ghana. *Engl. Today: Int. Rev. Engl. Lang.* **11**(4), 31–36 (1995). <https://doi.org/10.1017/S0266078400008609>
17. Kachru, B.B.: Standards, codification and sociolinguistic realism: The English language in the outer circle. In: Quirk R., Widdowson H.G. (eds.) *English in the world: teaching and learning the language and literatures* pp. 11–30. Cambridge, England, Cambridge University Press (1985)

# Machine Learning Algorithms for Modelling Agro-climatic Indices: A Review



G. Edwin Prem Kumar and M. Lydia

**Abstract** Modelling lays a solid platform to assess the effects of climate variability on agricultural crop yield and management. It also aids in measuring the effectiveness of control measures planned and to design optimal strategies to enhance agricultural productivity and crop intensity. Models that aid in predicting drought, soil quality, crop yield, etc. in the light of climate variabilities can go a long way in enhancing global food security. Efficient modelling of agro-climatic indices will simplify the upscaling of experimental observations and aid in the implementation of climate-smart agriculture. This paper aims to present a comprehensive review of the use of machine learning algorithms for modelling agro-climatic indices. Such models find effective application in crop yield forecasting, crop monitoring, and management, soil quality prediction, modelling evapotranspiration, rainfall, drought, and pest outbreaks. The research challenges and future research directions in this area have also been outlined.

## 1 Introduction

The impact of weather and climate on agricultural yield has been significant. It has been proven that variability in climate-related parameters can have worse effects on food security on a global scale. The El-Nino Southern Oscillation (ENSO) being the lead player in causing inter-annual climate mode changes results in droughts and a significant reduction in crop yield. Variability in climatic conditions has significantly affected cropping area and intensity as well [1]. Hence, modelling of agro-climatic indices will be of great advantage to farmers to plan well in advance. The challenges involved in developing models for agriculture incorporating climate change involve the development of biophysical and economic models, discrete-event models, and models for dynamic change, interactions, management, and uncertainties [2].

Agro-climatic indices are defined based on the relationships between crop yield/management etc. to variation in climate. They are used to measure the optimal

---

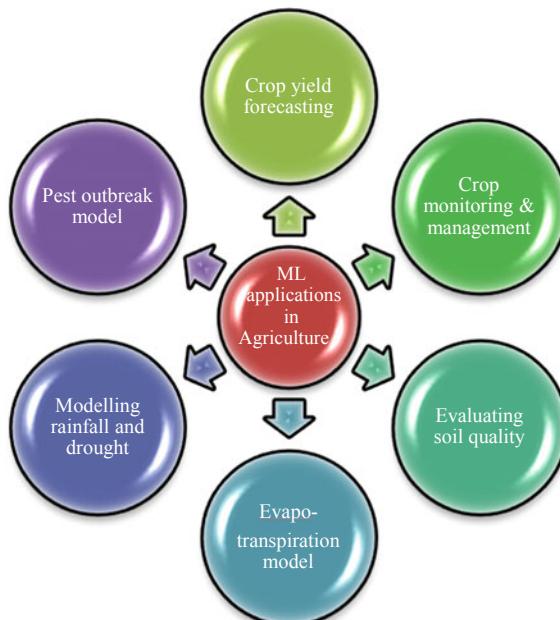
G. Edwin Prem Kumar ( · M. Lydia  
Sri Krishna College of Engineering and Technology, Coimbatore, India

climatic conditions required for desired agricultural performance in terms of yield, intensity, etc. The agro-climatic indices include parameters related to growing season, frost conditions, and multicriteria climate classification (MCC) system [3]. Parameters related to the growing season for viticulture include growing season temperature, precipitation, length, the number of dry and wet days. Frost conditions are depicted by mean data of first frost fall and last spring frost, number of days with frost and days with minimum temperature less than  $-15^{\circ}\text{C}$ . Furthermore, the MCC system includes parameters like heliothermal index, dryness index and cool night index.

Crop simulation models (CSMs) are process-based models that describe crop growth and development as a function of weather and soil conditions [4]. CSMs played a critical role in bioeconomic modelling, integrating the land use and agricultural productivity, both at regional and global scales. However, statistical models are known to scrutinize the relationship between agricultural crop yield and climate variables much better than CSMs as they capture the impact of many variables both directly and indirectly. These models outperformed CSMs by discovering relationships and modelling mechanisms of stressed biotic environments better. Though statistical models are severely limited by the availability of sufficient, quality data pertaining to weather, yield, agricultural management, etc., it has been found that hybrid statistical and CSMs can enhance the performance of statistical models [4].

In recent years, machine learning (ML) has emerged in a big way as a technique to revolutionize several areas of research including agriculture (Fig. 1). ML techniques along with big data, the Internet of Things (IoT), and several other advanced hybrid algorithms can handle complicated non-linear models and produce accurate

**Fig. 1** Machine learning applications in agriculture



results. ML models like regression, clustering, artificial neural networks (ANN), Bayesian models, support vector machines (SVM), ensemble learning, etc. find application in several agricultural processes like crop yield prediction, rainfall and drought prediction, crop management, pest outbreaks, etc. [5, 6].

This paper aims to provide an exhaustive review of the ML techniques used to model various agro-climatic indices, which are in turn useful in the estimation of key parameters in crop management. The ML techniques-based agro-climatic models used for several agricultural applications have been described in Sects. 2 and 3. The further research directions and challenges in this area have been suggested in Sect. 4 and the conclusions drawn from this survey are presented in Sect. 5.

## 2 Machine Learning for Crop Yield Forecasting

Machine learning-based agro-climatic models have been used for forecasting crop yield, monitoring, and management of crop. This section outlines the recent research work carried out in these areas. Annual crop yield is significantly determined by the climatic conditions. Regression-based models can be used to assess crop yield based on agro-climatic indices like Standardized Precipitation-Evapotranspiration Index (SPEI). Mathieu and Aires developed models to make annual estimation and seasonal prediction of crop yield using more than fifty agro-climatic indices [7]. Elavarasan and Vincent proposed a deep reinforcement learning model using agro-climatic indices for crop yield prediction and compared it with other ML models like ANN, Long Short-Term Memory (LSTM), gradient boosting, random forest (RF), etc. The accuracy of the model was evaluated using suitable evaluation metrics [8].

Mkhabela et al. predicted the crop yield of barley, canola, field peas and spring wheat using regression models in different agro-climatic zones, which included sub-humid, semi-arid and arid regions [10]. Johnson et al. used ML techniques and vegetation indices for crop yield forecasting. The vegetation indices used as predictors for crop yield include, the Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) derived from the Moderate-resolution Imaging Spectro-radiometer (MODIS), and NDVI derived from the Advanced Very High-Resolution Radiometer (AVHRR) [11]. ANN is the most common ML algorithm used and convolutional NN (CNN) is the most widely used DL algorithm for predicting crop yield. The most common features used for prediction of crop yield include temperature, rainfall and soil type [17]. Table 1 outlines the recent research carried out in crop yield forecasting.

**Table 1** Models for crop yield forecasting

Authors	Agro-climatic indices	ML technique
Mathieu and Aires [7]	SPEI and 49 other parameters	Regression model
Elavarasan and Vincent [8]	38 parameters	Deep Recurrent Q-Network (DRQN)
Mishra et al. [9]	Rainfall, temperature, humidity	Linear regression (LR), Ridge regression, Lasso regression, Support Vector Regression (SVR) with linear, polynomial and Radial Basis Function (RBF) kernels
Johnson et al. [11]	Vegetation indices (MODIS-NDVI, MODIS-EVI, AVHRR-NDVI)	Multiple linear regression (MLR), Bayesian neural networks, Model-based recursive partitioning
Chen et al. [12]	Photosynthetically Active Radiation (PAR), Radiation Use Efficiency (RUE), vegetation and meteorological indices	MLR
Folberth et al. [13]	Annual climate indices, growing season climate indices, monthly climate indices	Extreme gradient boosting, RF
Mathieu and Aires [14]	SPEI, average temperature	NN Classifier
Mupangwa et al. [15]	Big data from simulated cropping systems	Logistic regression, Linear discriminant analysis, K-nearest neighbour (K-NN), Classification and Regression trees, Gaussian naïve Bayes, SVM
Bai et al. [16]	Landsat 8 vegetation index and phenological length	SVM, NN, Mahalanobis Distance, Maximum Likelihood
Feng et al. [18]	Daily climate data	Hybrid modelling approach Agricultural Production Systems sIMulator (APSIM)-RF, APSIM-MLR
Feng et al. [19]	In-situ climate data, remote sensing data, wheat trial data, soil hydraulic properties	APSIM-RF, APSIM-MLR
Kamir et al. [20]	NDVI time series, climate time series, Yield maps, crop yield statistics	Regression, SVR with RBF
Cai et al. [21]	Satellite data, climate data	Least Absolute Shrinkage and Selection Operator (LASSO), NN, RF, SVM

(continued)

**Table 1** (continued)

Authors	Agro-climatic indices	ML technique
Zarei et al. [22]	United Nations Environment Programme (UNEP) aridity index, Modified De-Martonne index	Simple and multiple Generalizes Estimation Equation
Xu et al. [23]	Meteorological data	RF, SVM
Gumuscu et al. [24]	Air temperature, daily minimum air temperature, daily precipitation	k-NN, SVM, decision trees
Huy et al. [25]	12 climate indices	D-vine quantile regression model
Wang et al. [26]	Southern Oscillation Index (SOI), SOI phase, NINO3.4, Multivariate ENSO Index (MEI)	RF

### 3 Machine Learning for Crop Monitoring and Management

Machine learning-based agro-climatic models prove to be of great advantage in crop monitoring and management (Table 2). Agro-climatic indices along with high-resolution remote sensing data can be an effective tool for crop monitoring. Ballessteros et al. employed two agro-climatic indices namely reference evapotranspiration (ET<sub>0</sub>) and growing degree days (GDD) and few other vegetable indices for monitoring crop [27]. ML algorithms and vegetation indices were used to characterize and map cropping patterns [28]. The robustness of Landsat-based fraction of absorbed photosynthetically active radiation (fAPAR) models was assessed using ML algorithms [29].

### 4 Research Directions and Challenges

Accuracy in agricultural modelling is a very critical parameter to be satisfied. Effective models need a large amount of quality data, which is indeed a big challenge in this research. The performance of predictive modelling improves significantly if there is a substantial increase in the sample size in both spatial and temporal distributions [21]. Climate variability directly and indirectly affects several agricultural processes and the yield, making the modelling process a challenging one. Since the cropping area and intensity keeps changing, the use of static or historic data is likely to have potential errors. Annual mappings of crop area obtained from the satellite will aid in improved accuracy. The use of passive or active remote sensing data improves the accuracy of models as it incorporates parameters like canopy biomass and water content. ML algorithms that model the non-linearity better and capture

**Table 2** Models for crop monitoring and management

Authors	Objective	ML technique
Ballesteros et al. [27]	Crop monitoring	MLR
Feyisa et al. [28]	Map cropping patterns	SVM, RF, C5.0
Muller et al. [29]	Assess the fidelity of models	MLR, Decision tree, RF
Vindya and Vedamurthy [30]	Crop selection	Naïve Bayes Classification
Kale and Patil [31]	Decision support system	Data mining, Fuzzy logic
Shi et al. [32]	Ascertain change in NDVI	RF Regression with residual analysis
Lee et al. [33]	Projection of life-cycle environmental impacts	Boosted regression tree
Macedo et al. [34]	Estimation of crop area	Convolutional LSTM models
Young et al. [35]	Seasonal forecasting of daily mean air temperatures	Regularized Extreme Learning Machine
Sharma et al. [36]	Sustainable agriculture supply chain performance	ANN, Bayesian network, clustering, DL, Ensemble learning, regression, SVM
Jakariya et al. [37]	Assessment of vulnerability index	LR, Bayesian ridge regression, XGB regression, RF regression, Extremely randomized trees regression

well the relationship between the input and output variables are the need of the hour. The algorithms need to respond to variabilities at all levels and should ensure quicker convergence with higher computational speed.

## 5 Conclusion

Climate is both a resource and a restraint for agriculture. Early and consistent forecasting models play a critical role in farmer's decision-making pertaining to crop selection, yield, pest occurrence, irrigation needs, etc. Agro-climatic indices are indicators of climate characteristic which has definite agricultural significance. The spatial characteristics and temporal distribution of agro-climatic indices can be investigated and modelled to understand the growing seasonal parameters of different crops like wheat, maize, etc. This paper presented an exhaustive review on ML algorithms for crop yield forecasting, crop monitoring and forecasting based on agro-climatic indices. Research challenges and directions have also been presented.

## References

1. Iizumi, T., Ramankutty, N.: How do weather and climate influencing cropping area and intensity? *Glob. Food Sec.* **4**, 46–50 (2015)
2. Kipling, R.P., Topp, C.F.E., Bannink, A., Bartley, D.J., Penedo, I.B., Cortignani, R., del Prado, A., Dono, G., Faverdin, P., Graux, A.I., Hutchings, N.J., Lauwers, L., Gulzari, S.O., Reidsma, P., Rolinski, S., Ramos, M.R., Sandars, D.L., Sandor, R., Schonhart, M., Seddaiu, G., Middelkoop, J.V., Shrestha, S., Weindl, I., Eory, V.: To what extent is climate change adaptation a novel challenge for agricultural modellers? *Environ. Model. Softw.* **120**, 104492 (2019)
3. Ruml, M., Vukovic, A., Vujadinovic, M., Djurdjevic, V., Vasic, Z.R., Atanackovic, Z., Sivcev, B., Markovic, N., Matijasevic, S., Petrovic, N.: On the use of regional climate models: implications of climate change for viticulture in Serbia. *Agric. For. Meteorol.* **158–159**, 53–62 (2012)
4. Rotter, R.P., Hoffman, M.P., Koch, M., Muller, C.: Progress in modelling agricultural impacts of and adaptations to climate change. *Curr. Opin. Plant Biol.* **45(B)**, 255–261 (2018)
5. Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D.: *Mach. Learn. Agric. Rev.* **18**, 2674 (2018)
6. Priya, R., Ramesh, D.: ML based sustainable precision agriculture: a future generation perspective. *Sustain. Comput. Inf. Syst.* **28**, 100439 (2020)
7. Mathieu, J.A., Aires, F.: Assessment of the agro-climatic indices to improve crop yield forecasting. *Agric. For. Meteorol.* **253–254**, 15–30 (2018)
8. Elavarasan, D., Vincent, D.: Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access* **8**, 86886–86901 (2020)
9. Mishra, S., Mishra, D., Santra, G.H.: Adaptive boosting of weak regressors for forecasting of crop production considering climatic variability: an empirical assessment. *J. King Saud Univ. Comput. Inf. Sci.* (2017)
10. Mkhabela, M.S., Bullock, P., Raj, S., Wang, S., Yang, Y.: Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. *Agric. For. Meteorol.* **151**, 385–393 (2011)
11. Johnson, M.D., Hsieh, W.W., Cannon, A.J., Davidson, A., Bedard, F.: Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agric. For. Meteorol.* **218–219**, 74–84 (2016)
12. Chen, Y., Donohue, R.J., McVicar, T.R., Waldner, F., Mata, G., Ota, N., Houshamdar, A., Dayal, K., Lawes, R.A.: Nationwide crop yield estimation based on photosynthesis and meteorological stress indices. *Agric. For. Meteorol.* **284**, 107872 (2020)
13. Folberth, C., Baklanov, A., Balkovic, J., Skalsky, R., Khabarov, N., Obersteiner, M.: Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. For. Meteorol.* **264**, 1–15 (2019)
14. Mathieu, J.A., Aires, F.: Using neural network classifier approach for statistically forecasting extreme corn yield losses in Eastern United States. *Earth Space Sci.* **5**, 622–639 (2018)
15. Mupangwa, W., Chipindu, L., Nyagumbo, I., Mukuhani, S., Sisito, G.: Evaluating machine learning algorithms for predicting maize yield under conservation agriculture in Eastern and Southern Africa. *SN Appl. Sci.* **2**, 952 (2020)
16. Bai, T., Zhang, N., Mercatoris, B., Chen, Y.: Jujube yield prediction method combining Landsat 8 Vegetation Index and the phenological length. *Comput. Electron. Agric.* **162**, 1011–1027 (2019)
17. Klompenburg, T.V., Kassahun, A., Catal, C.: Crop yield prediction using machine learning: a systematic literature review. *Comput. Electron. Agric.* **177**, 105709 (2020)
18. Feng, P., Wang, B., Liu, D.L., Waters, C., Yu, Q.: Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agric. For. Meteorol.* **275**, 100–113 (2019)
19. Feng, P., Wang, B., Liu, D.L., Waters, C., Xiao, D., Shi, L., Yu, Q.: Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. *Agric. For. Meteorol.* **285–286**, 107922 (2020)

20. Kamir, E., Waldner, F., Hochman, Z.: Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS J. Photogram. Remote Sens.* **160**, 124–135 (2020)
21. Cai, Y., Guan, K., Lobell, D., Potgieter, A.B., Wang, S., Peng, J., Xu, T., Asseng, S., Zhang, Y., You, L., Peng, B.: Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* **274**, 144–159 (2019)
22. Zarei, A.R., Shabani, A., Mahmoudi, M.R.: Comparison of the climate indices based on the relationship between yield loss of rain-fed winter wheat and changes of climate indices using GEE model. *Sci. Total Environ.* **661**, 711–722 (2019)
23. Xu, X., Gao, P., Zhu, X., Guo, W., Ding, J., Li, C., Zhu, M., Wu, X.: Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. *Ecol. Indic.* **101**, 943–953 (2019)
24. Gumuscu, A., Tenekeci, M.E., Bilgili, A.V.: Estimation of wheat planting date using machine learning algorithms based on available climate data. *Sustain. Comput. Inf. Syst.* 100308 (2019)
25. Huy, T.H., Deo, R.C., Mushtaq, S., An-Vo, D.A., Khan, S.: Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. *Eur. J. Agron.* **98**, 65–81 (2018)
26. Wang, B., Feng, P., Waters, C., Cleverly, J., Liu, D.L., Yu, Q.: Quantifying the impacts of pre-occurred ENSO signals on wheat yield variation using machine learning in Australia. *Agric. For. Meteorol.* **291**, 108043 (2020)
27. Ballesteros, R., Ortega, J.F., Hernandez, D., Campo, A.D., Moreno, M.A.: Combined use of agro-climatic and very high-resolution remote sensing information for crop monitoring. *Int. J. Appl. Earth Obs. Geoinf.* **72**, 66–75 (2018)
28. Feyisa, G.L., Palao, L.K., Nelson, A., Gumma, M.K., Paliwal, A., Win, K.T., Nge, K.H., Johnson, D.E.: Characterizing and mapping cropping patterns in a complex agro-ecosystem: an iterative participatory mapping procedure using machine learning algorithms and MODIS vegetation indices. *Comput. Electron. Agric.* **175**, 105595 (2020)
29. Muller, S.J., Sithole, P., Singels, A., Niekerk, A.V.: Assessing the fidelity of Landsat-based fAPAR models in two diverse sugarcane growing regions. *Comput. Electron. Agric.* **170**, 105248 (2020)
30. Vindya N.D., Vedamurthy H.K.: Machine learning algorithm in smart farming for crop identification. In: Smys, S., Tavares, J., Balas, V., Iliyasu A. (eds.) Computational vision and bio-inspired computing, ICCVBIC 2019. Advances in Intelligent Systems and Computing, vol. 1108. Springer, Cham (2020)
31. Kale, S.S., Patil, P.S.: Data mining technology with Fuzzy Logic, neural networks and machine learning for agriculture. In: Balas, V., Sharma, N., Chakrabarti, A. (eds.) Data management, analytics and innovation. Advances in Intelligent Systems and Computing, vol. 839. Springer, Singapore (2019)
32. Shi, Y., Jin, N., Ma, X., Wu, B., He, Q., Yue, C., Yu, Q.: Attribution of climate and human activities to vegetation change in China using machine learning techniques. *Agric. For. Meteorol.* **294**, 108146 (2020)
33. Lee, E.K., Zhang, W.J., Zhang, X., Adler, P.R., Lin, S., Feingold, B.J., Khwaja, H.A., Romeiko, X.X.: Projecting life-cycle environmental impacts of corn production in the U.S. Midwest under future climate scenarios using a machine learning approach. *Sci. Total Environ.* **714**, 136697 (2020)
34. Macedo, M.M.G., Mattos, A.B., Oliveira, D.A.B.: Generalization of convolutional LSTM models for crop area estimation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **13**, 1134–1142 (2020)
35. Young, S.J., Rang, K.K., Chul, H.J.: Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management. *Agric. For. Meteorol.* **281**, 107858 (2020)
36. Sharma, R., Kamble, S.S., Gunasekaran, A., Kumar, V., Kumar, A.: A systematic literature review on machine learning applications for sustainable agriculture supply chain performance. *Comput. Oper. Res.* **119**, 104926 (2020)

37. Jakariya, Md., Alam, Md.S., Rahman, Md.A., Ahmed, S., Elahi, M.M.L., Khan, A.M.S., Saad, S., Tamim, H.M., Ishtiaq, T., Sayem, S.M., Ali, M.S., Akter, D.: Assessing climate-induced agricultural vulnerable coastal communities of Bangladesh using machine learning techniques. *Sci. Total Environ.* **742**, 140255 (2020)

# Design of Metal-Insulator-Metal Based Stepped Impedance Square Ring Resonator Dual-Band Band Pass Filter



Surendra Kumar Bitra and M. Sridhar

**Abstract** In this paper, a metal–insulator metal (MIM) based plasmonic stepped impedance square ring resonator (SI-SRR) band-pass filter (BPF) is designed and analyzed for dual-band applications. The MIM-based SI-SRR is investigated using commercially available CST studio suite. The proposed SI-SRR is compact and low power requirements suitable for Photonic Integrated Circuits (PICs). The SI-SRR is operated in the wavelengths of 1317 nm (227.6 THz) and 1640 nm (182.8 THz) with appropriate reflection and transmission parameters. The stepped impedance stubs are used in the ring resonator for tunable operating bands. The proposed SI-SRR has wide applications in PICs

## 1 Introduction

The waves that are produced at Metal and Insulator region when light is interacting are the Surface Plasmon Polarities (SPPs) [1]. These are high-speed electromagnetic waves traveling on the metal–insulator regions. MIM is one of the popular waveguides used for designing most of the optical devices [2]. MIM-based components are suitable for PICs. Recently, MIM-based components like stub ring resonators [3, 4], triangular resonators [5], rectangular ring resonator [6], square ring resonators [7], and circular ring resonators [8]. Different excitation schemes of ring resonators are briefly analyzed [9]. The ring resonators provide the dual-band operating wavelengths discussed in [8, 9].

In this work, first investigate the MIM-based plasmonic SI-SRR dual-band BPF for the O and L optical bands. The transmission performance of the filter is analyzed using CST studio suite. The SI-SRR improves the bandwidth and produces a more confinement nature. The concurrent plasmonic ring resonators and filters are operating in two or more frequency bands simultaneously. Several multi-band components are proposed in [9].

---

S. K. Bitra (✉) · M. Sridhar

Department of ECE, KLEF, Guntur, Andhra Pradesh, India

**Table 1** Design parameters of SI-SRR

S. No.	Parameter	Value in nm
1	L1	1000
2	L2	900
3	L3	200
4	W3	50
5	L4	200
6	W4	60
7	$d$	50
8	$g$	10

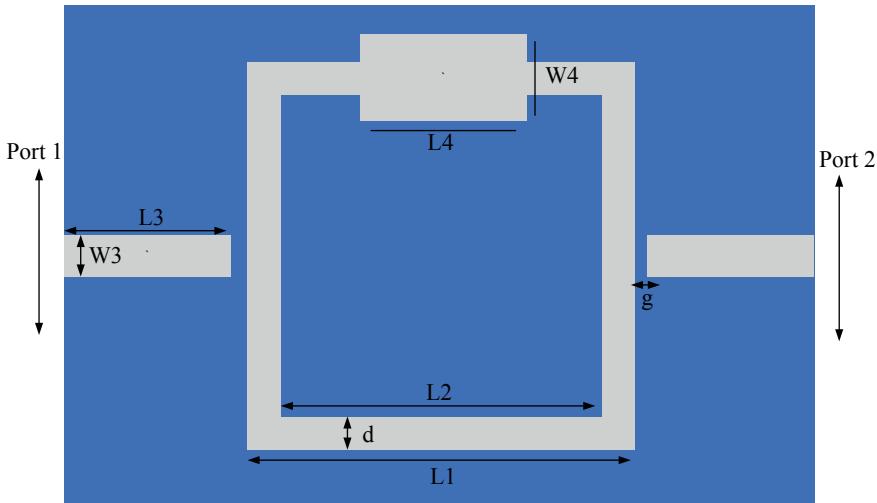
The paper organization is as follows: Sect. 2 describes the SI-SRR dual-band BPF design procedure and optimized parameters using MIM waveguide. The simulation results and field distributions are included in Sect. 3. Finally, the paper ends with the conclusion.

## 2 Stepped Impedance SRR Design

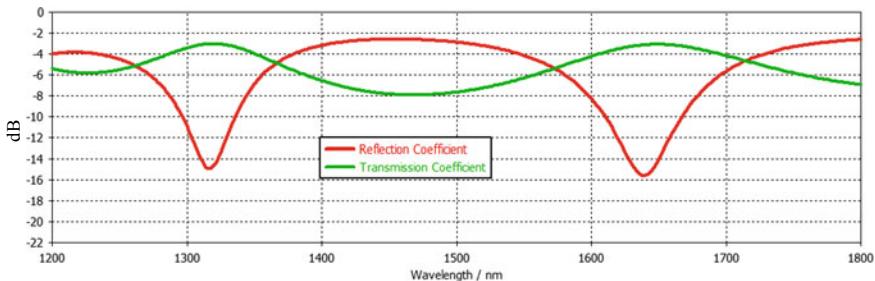
The proposed filter consists of coupled line feed and square ring resonator with stub loaded on one side of the ring forms the stepped impedance square ring resonator (SI-SRR) is shown in Fig. 1. The basic MIM characteristics and SRR dual-band characteristics are investigated in [10]. The enhancement of previous work is included in the stub to form the SI-SRR to give the better notching characteristics. The dimensions of the proposed SI-SRR dual-band BPF are represented in Table 1. The filter is designed and simulated in commercially available CST studio suite. The mesh sizes were taken as  $5 \text{ nm} \times 5 \text{ nm}$ . The dimensions of the SRR are calculated using [11]. The SI-SRR is suitable for optical O band (1260–1360 nm) and U band (1625–1675 nm). The filter is easily fabricated using lithographic techniques. For designing purpose, the metal is taken as silver and insulator is silica.

## 3 Simulation Results

The SI-SRR is designed using the CST studio suite using the above-optimized dimensions. By applying the coupled mode excitations, the reflection and transmission coefficients are observed and represented in Fig. 2. The reflection coefficient is highlighted in the red color, operated in dual bands are 1317 and 1640 nm with the power of approximately  $-15 \text{ dB}$  for both the bands. The transmission coefficient is represented in green color with the power of  $-3 \text{ dB}$  approximately. Due to the stub in



**Fig. 1** Proposed SI-SRR dual-band BPF



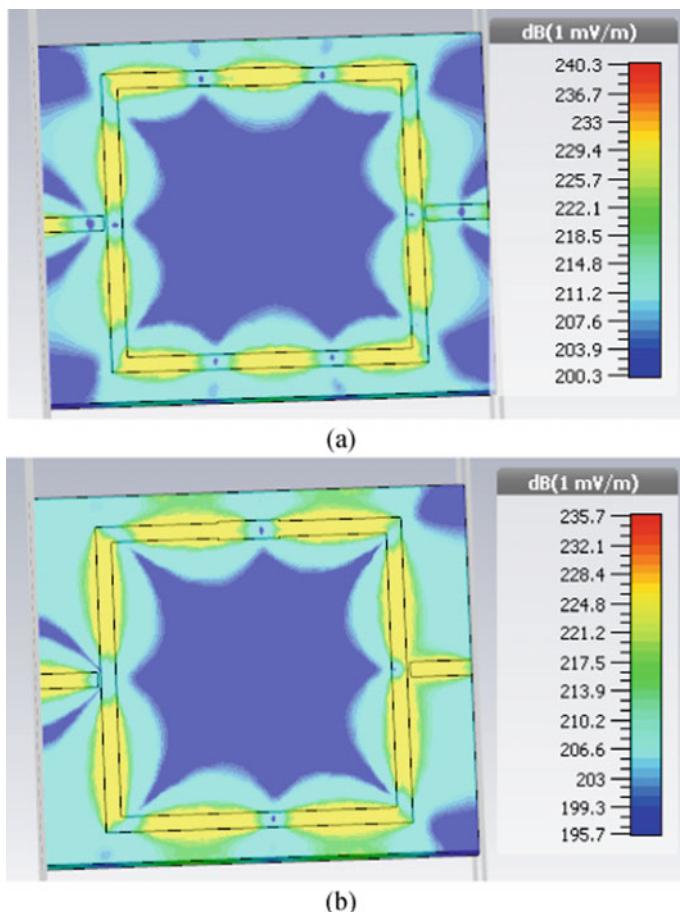
**Fig. 2** Reflection and transmission coefficient of SI-SRR dual-band BPF

the ring resonator, improving the bandwidth at O and U bands is observed. PML boundary conditions are used to simulate the SI-SRR filter.

Figure 3 represents the field distributions of the SI-SRR filter at 1317 and 1640 nm. The field distributions show the power confinement of the given power at the metal and insulator region.

## 4 Conclusion

The MIM-based plasmonic SI-SRR is suitable for optical O (1260–1360 nm) and U band (1625–1675 nm) dual-band applications. The Reflection and transmission coefficient of SI-SRR with resonant behavior is numerically analyzed. The center



**Fig. 3** Field Distributions at **a** 1317 nm (227.6 THz) and **b** 1640 nm (182.8 THz)

operated wavelength of the SI-SRR is 1317 and 1640 nm with  $-15$  dB reflection coefficient. The SI-SRR is easily fabricated using semiconductor fabrication procedure techniques. SI-SRR filter is best suitable for photonic integrated circuit applications.

## References

1. Barnes, W.L., Dereux, A., Ebbesen, T.W.: Surface plasmon subwavelength optics. *Nature* **424**, 824–830 (2003)
2. Ozbay, E.: Plasmonics: merging photonics and electronics at nanoscale dimensions. *Science* (80) **311**, 189–193 (2006)
3. Taylor, P., Li, C., Qi, D., Xin, J., Hao, F.: Metal insulator metal plasmonic waveguide for low distortion slow light at telecom frequencies. *J. Mod. Opt.* **61**, 37–41 (2014)

4. Zafar, R., Salim, M.: Analysis of asymmetry of fano resonance in plasmonic metal insulator metal waveguide. *J. Photonics* **23**, 1–6 (2016)
5. Oh, G., Kim, D., Kim, S.H., Ki, H.C., Kim, T.U., Choi, T.: Integrated refractometric sensor utilizing a triangular ring resonator combined with SPR. *IEEE Photonics Tech. Lett.* **26**, 2189–2192 (2014)
6. Yun, B., Hu, G., Cui, Y.: Theoretical analysis of a nanoscale plasmonic filter based on rectangular metal insulator metal waveguide. *J. Phys.* **43**, 385102 (1–8) (2010)
7. Liu, J., Fang, G., Zhao, H., Zhang, Y.: Plasmon flow control at gap waveguide junctions using square ring resonators. *J. Phys.* **43**, 055103 (1–6) (2009)
8. Setayesh, A., Miranaziry, S.R., Abrishamian, M.S.: Numerical investigation of tunable band pass\band stop plasmonic filters with hollow core circular ring resonators. *J. Opt. Soc. Korea* **43**, 82–89 (2011)
9. Vishwanath, M., Khan, H.: Excitation schemes of plasmonic angular ring resonator based band pass filters using MIM waveguide. *Photonics* **6**, 1–9 (2019)
10. Bitra, S.K., Sridhar, M.: Design of nanoscale square ring resonator band pass filter using metal insulator metal. In: Chowdary, P., Charkravarthy, V. (eds.) ICMEET Conference 2020, vol. 655. Springer, Singapore (2020)
11. Yun, B., Hu, G., Cui, Y.: Theoretical analysis of nanoscale plasmonic filter based on a rectangular metal insulator metal waveguide. *J. Phys. D Appl. Phys.* **43**, 385120 (2010)

# Covid-19 Spread Analysis



Srinivas Kanakala and Vempaty Prashanthi

**Abstract** Based on the public datasets afforded by John Hopkins University and Canadian health authorities, we developed a forecasting model of Covid-19 after analyzing the spread. Data related to the cumulative amount of definite cases, per day, in each country and another dataset consisting of various life factors, scored by the people living in each country around the globe. We are going to merge these two datasets to see if there is any relationship between the spread of the virus in a country by preprocessing, merging and finding correlation between datasets we will calculate needed measures and prepare them for an analysis, then we will try to predict the spread of cases by using various methods. Time series data tracking the number of people affected by the coronavirus globally, including confirmed cases of the coronavirus, the number of people who have died due to the coronavirus and the number of people who have recovered from the deadly infection. Data science can give accurate pictures of coronavirus outcomes and also helps in tracking the spread. Secondly using Covid-19 data, we can make supply chain logistics decisions in spreadsheets supplies of personal protective equipment and ventilators to hospitals and clinics across the world. An analysis of the country, by state and region, identifying locations of highest need for supplies and ventilators according to the dataset collected. This is called a supply plan. Finally, create a set of visualizations and then add these visualizations to a presentation so that we can report on findings.

## 1 Introduction

The tale Covid that began in Wuhan, China, has spread to practically all nations and was announced as pandemic [1]. The degree of this flare-up is fast. It is difficult to precisely survey the lethality of this infection and it has all the earmarks of being

---

S. Kanakala (✉)

VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India

V. Prashanthi

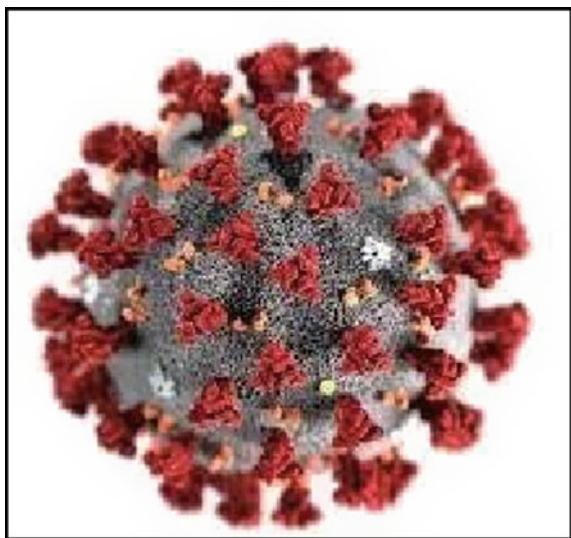
Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, India

unmistakably more deadly than the Covid that caused SARS and MERS. Researchers [2] have recognized two new strains of the Covid, demonstrating it is now been changed at any rate once. The greatest test is an obscure number of individuals have been contaminated by the infection without getting indicative. These individuals are transporters of the infection without themselves giving any indications. At first, individuals who gave no indications of disease were not isolated and this prompts the spread of the infection at a gigantic rate. The infection is additionally appeared to influence its hosts lopsidedly. Youngsters appear to be less inclined to be tainted while the moderately aged and more seasoned grown-ups are mysteriously contaminated. Men are bound to kick the bucket from the contamination contrasted with ladies, and furthermore individuals with a more fragile safe framework, Type 2 diabetes and hypertension. Be that as it may, as of late numerous alive and well youthful people have kicked the bucket from the contamination making it much harder to comprehend the impact of Covid-19 [3, 4].

## 2 Literature Survey

Coronavirus Fig. 1 is the overwhelming illness brought about by and was called Covid. This new virus was vague before and started in Wuhan, China, in December 2019. Coronavirus is a pandemic affecting many countries.

**Fig. 1** Coronavirus



## 2.1 Symptoms

The indications of Covid-19 [5] are general influenza like and a few patients increase an great kind of pneumonia. Patients have fever, muscle pain and body throbs, hacks and sore throat about following six days of getting the contamination. The huge people feel truly desperate and frail and improve all alone, yet a minority of patients will deteriorate following 5–7 days of ailment and the patients have windedness and exacerbating hack. The hack is dry and not wet. It is even observed that patients have solid cerebral pains. Furthermore, the side effects are somewhat unique in relation to influenza. People tainted from the infection might not have a virus. However, a few people do not become ill while being contaminated and are spreading the infection to new has. These individuals ought not to be all over town spreading the illness. Individuals who got tainted and have been effectively restored have likewise got contaminated by the Covid-19 once more. Making it much harder to contain the episode. There is no anti-microbial to treat the Covid-19 and it may not be accessible until the spring of 2021. This makes it much more imperative to take preventive activities.

## 2.2 No Symptoms

Coronavirus is basically extended with respiratory beads detached by somebody who is hacking or has other effects, like fever or sluggishness [6]. Many people who have coronavirus experience just mellow indications. These are the symptoms in the beginning. Most as of late it has been indicated that elevated levels of the infection are available in respiratory discharges during the “presymptomatic period that can a days ago to over seven days” before the fever and hack normal for Covid-19. This capacity of the infection to be communicated by individuals without side effects is a significant purpose behind the pandemic. Some news show that persons who are not having any manifestations can communicate the virus.

## 2.3 Coronavirus Modes of Spread

The Covid spreads principally from one person to other [7]. This occurs among peoples who are close to each other. Beads which are created when a contaminated individual hacks or wheezes may land in the mouths or noses of individuals who are close by, or potentially be breathed in into their lungs. An individual contaminated with Covid—even one without any side effects may emanate vaporizers when they talk or relax. Mist concentrates are irresistible viral particles that can buoy or float

around noticeable all around for as long as three hours. One can be affected by Covid-19 when he touches an item which has virus and then touches their own mouth, nose, or conceivably eyes.

## ***2.4 Importance of Social Distance and Self-isolation***

Every person should maintain social distancing of about 6 ft or more from others. Schools, gatherings, occasions, malls, etc. do not maintain any social distancing. Therefore, these are closed during Covid. This will help the society from virus, as the spread will be controlled. Self-isolation is an important measure that should be taken by the people who is affected by coronavirus. He or she should be isolated in separate room, even from family members. These people should not go to crowded places like schools, etc. Take clinical assistance. If you do not live in a region with intestinal sickness or dengue fever kindly do the accompanying.

## **3 Existing System**

**Aarogya Setu:** It is an Indian Covid-19 APP. “Contact following, Syndromic planning and Self-evaluation” computerized administration, fundamentally a portable application, created by the National Informatics Center under the Ministry of Electronics and Information Technology (MeitY). The cause for this application is to make people familiar with Covid-19 for well-being of people. It is an app which needs the mobile phone’s GPS and Bluetooth to follow the Covid contamination. The app can be accessed from Android and iOS versatile frameworks. With the help of Bluetooth, it indicates danger when one comes close (inside six feet of) to you who is with coronavirus, by viewing the database of cases around India. With the help of GPS, it can detect whether the region is a place with contaminated zone.

### **Drawbacks of Existing System:**

- It is forced through leader request with no legitimization.
- Recently, Robert Baptiste has tweeted that safety weaknesses in Aarogya Setu permitted programmers to realize who is tainted or not well in their preferred region. He additionally gave subtleties of what number of individuals were unwell and contaminated at the PM’s Office, the India Parliament and the Home Office.
- The application’s Terms of Service (TOS) gives restricted obligation to the administration. In this way, there is no administration responsibility in the event of information burglary of clients.

## 4 Proposed System

We developed a new model of Covid-19 after analyzing the spread. Data related to the cumulative no of confirmed cases, per day, in each Country and another dataset consisting of various life factors, scored by the people living in each country around the globe. We are going to merge these two datasets to see if there is any relationship between the spread of the virus in a country by preprocessing [8, 9], merging and finding correlation between datasets we will calculate needed measures and prepare them for an Analysis, then we will try to predict the spread of cases by using various methods. Time series data tracking the number of people affected by the coronavirus globally, including confirmed cases of the coronavirus, the number of people who have died due to the coronavirus and the number of people who have recovered from the deadly infection. Data preprocessing [10, 11] is a data mining technique which is used to transform the raw data in a useful and efficient format.

### Steps Involved in Data Preprocessing:

1. **Data Cleaning:** The information may have numerous insignificant and missed parts. To deal with such things, information cleaning is finished. This includes treatment of missed information, boisterous information and so on.
  - (a) **Missing Data:** This condition comes when some information is not present. This can be obtained from different techniques such as:
    - (i) **Ignore the Tuples:** This method is appropriate when the dataset is huge and different qualities are not present in a tuple.
    - (ii) **Fill the Missing Qualities:** There are many methods to fill this. One can do this physically, by characteristic mean or the most likely worth.
  - (b) **Uproarious Data:** Noisy information is a negligible information that cannot be deciphered by machines. It tends to be produced because of flawed information assortment, information section mistakes and so on. It very well may be dealt with in following manners.
    - (i) **Binning Method:** This strategy chips away at arranged information so as to smooth it. The entire information is isolated into portions of equivalent size and afterward different strategies are performed to finish the errand. Each fragmented is dealt with independently. We can supplant information of section by the mean or limit esteems could be utilized to finish the errand.
    - (ii) **Relapse:** Here information is turned smooth by fitting it into relapse function. The relapse utilized might be straight or various
    - (iii) **Bunching:** This methodology bunches the comparative information in a group. The anomalies might be not detected or else it comes under outside the bunches.

2. **Data Transformation:** This progression is taken so as to change the information in fitting structures appropriate for mining measure. This includes Normalization, Attribute Selection, Discretization, Concept Hierarchy Generation.
3. **Data Reduction:** Since information mining is a method that is utilized to deal with colossal measure of information. While working with gigantic volume of information, examination got more diligently in such cases. So as to dispose of this, we utilize information decrease method. It means to expand the capacity effectiveness and decrease information stockpiling and investigation costs. The different strides to information decrease are Data Cube aggregation, Attribute Subset Selection, Numerosity Reduction, Dimensionality Reduction.

## 5 Proposed System

The supply logistics from the USA are considered. This set is cleaned by removing the unnecessary columns. It narrows down the data to information we need the most. This prevents distraction and enables a clear idea of what needs to be analyzed. Calculating the vent creating a pivot table in python. Since the values are extremely clumped, we can sort by a required filter such as region like MidWest. We use a legend to depict this. We have shown a pie chart representation to give a clear picture of the distribution in Fig. 2 since we have many states and matplotlib does a rudimentary view of the analysis. Calculating the ventilator requirements for each state in the US.

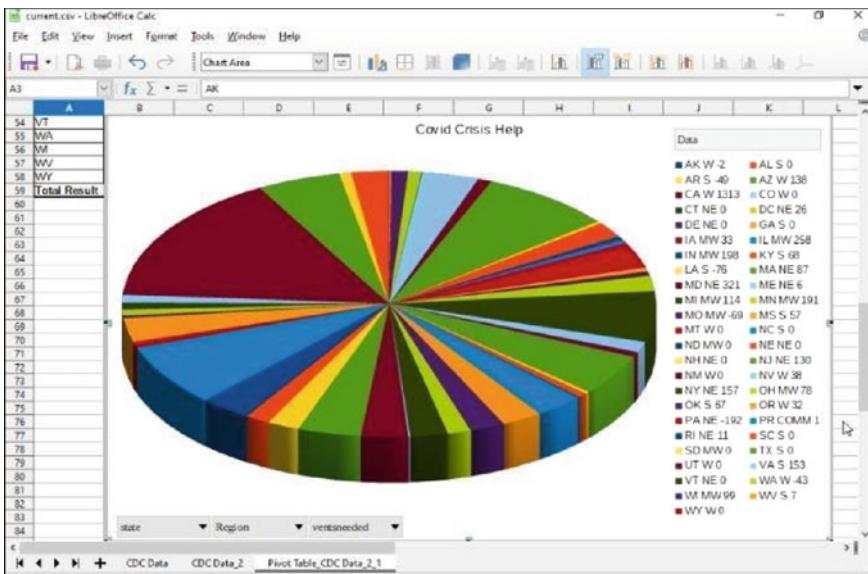
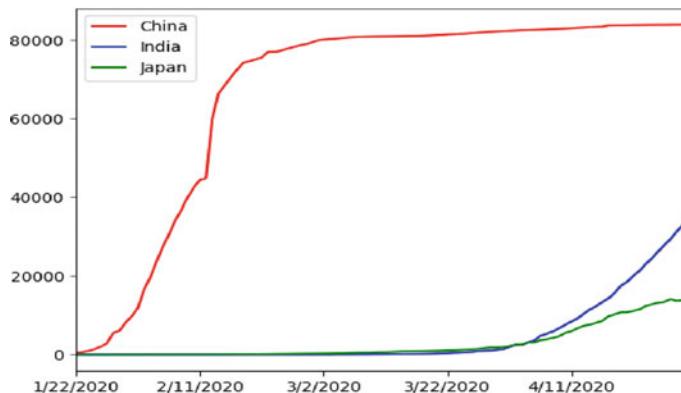


Fig. 2 Covid crisis pie chart

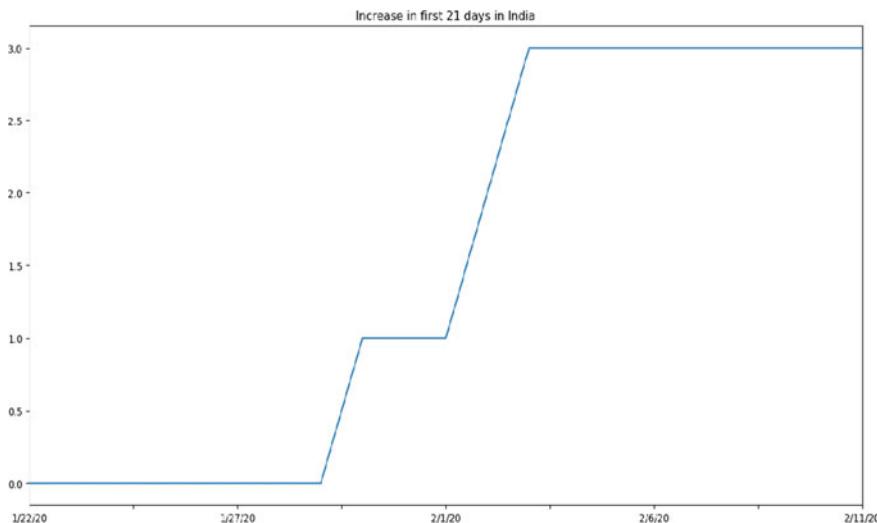
We take the number of patients hospitalized and the number of cumulative ventilators available to get the required number.

A specific case study to observe the number of cases in China, India and Japan show in Fig. 3. We can see that the rate of cases in the origin country of the virus has the maximum cases and since the other two countries are in close proximity, the cases have spread and also on a similar level in India and Japan.

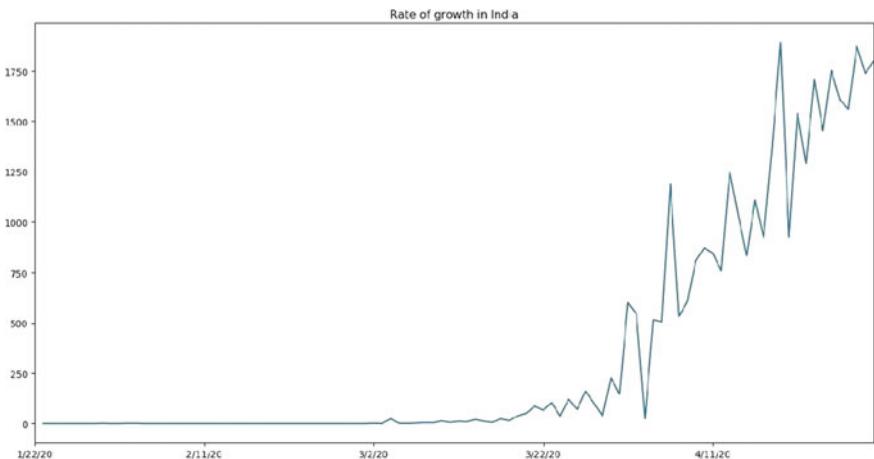
The spread of the virus in India can be seen by the 21st day where the sudden spike is evident shown in Fig. 4. The government issued lockdown to flatten this spike of increase shown in Fig. 5.



**Fig. 3** Specific case study to observe the number of cases in China, India and Japan

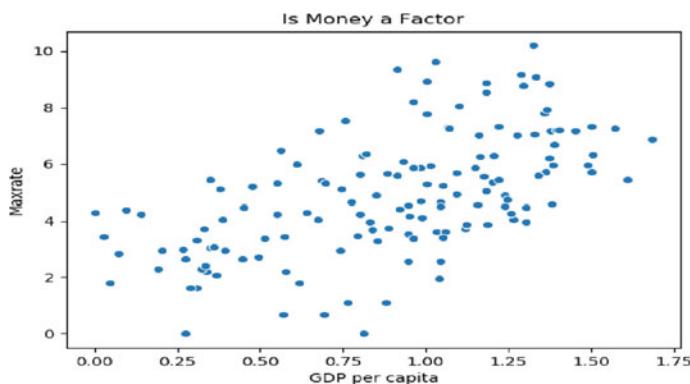


**Fig. 4** Spread of the virus in India can be seen by the 21st day

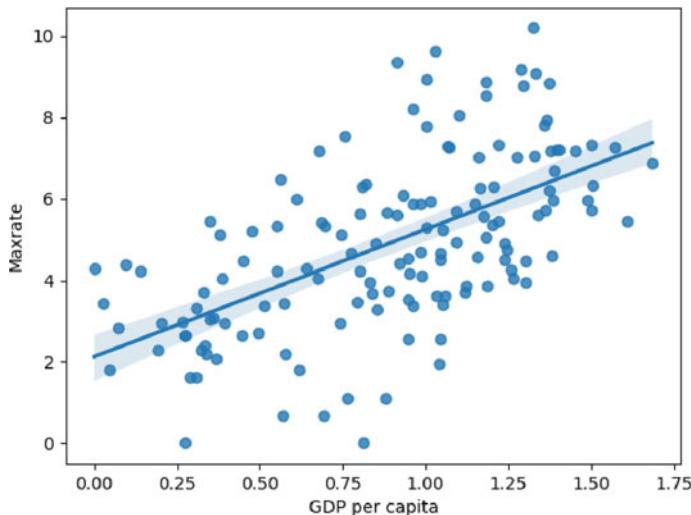


**Fig. 5** For first 21 days coronavirus spread

Corona cases versus GDP: Does the wealth of a country mean less corona cases? It does not seem so. In fact, developing countries have a lesser risk than developed countries as per our studies shown in Fig. 7. We have tried to correlate the number of cases and the development of a country. This can be due to various reasons like climatic conditions, etc. However, this is not due to lack of testing kits on the contrary in developing countries (Fig. 6).



**Fig. 6** Corona cases versus GDP



**Fig. 7** Future spread based on cases and GDP per capita

## 6 Conclusion

In this digital world, new data and information on the coronavirus and the progress of the occurrence have become accessible at an exceptional pace. Even though, tough questions stay without answer and exact answers to predict the dynamics of the situation will not receive in such stage. Analyzing and predicting the spread of viruses with the existing data will help us to have a better understanding to prevent the spread and to take preventive measures. To fight with coronavirus, we have to take care of ourselves and follow all the safety measures and rules that have been given by the government. Everyone can play a part in helping scientists to fight the coronavirus.

## References

1. Novel, Coronavirus Pneumonia Emergency Response Epidemiology: The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua liu xing bing xue za zhi= Zhonghua liuxingbingxue zazhi **41.2**, 145 (2020)
2. Perlman, S.: Another decade, another coronavirus, 760–762 (2020)
3. Abroug, F., et al.: Family cluster of Middle East respiratory syndrome coronavirus infections, Tunisia, 2013. Emerg. Infect. Dis. **20.9**, 1527 (2014)
4. Van Der Hoek, L., et al.: Identification of a new human coronavirus. Nat. Med. **10.4**, 368–373 (2004)
5. Guan, W.-j., et al.: Clinical characteristics of coronavirus disease 2019 in China. New England J. Med. **382.18**, 1708–1720 (2020)

6. Schoeman, D., Fielding, B.C.: Coronavirus envelope protein: current knowledge. *Virol. J.* **16**(1), 1–22 (2019)
7. Song, F., et al.: Emerging 2019 novel coronavirus (2019-nCoV) pneumonia. *Radiology* **295**.1, 210–217 (2020)
8. Kamiran, F., Calders, T.: Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.* **33**(1), 1–33 (2012)
9. Prashanthi, V., Kanakala, S.: Plant disease detection using Convolutional neural networks. *Int. J. Adv. Trends Comput. Sci. Eng.* **9**(3), 2632–2637
10. García, S., Luengo, J., Herrera, F.: Data preprocessing in data mining. Springer International Publishing, Cham, Switzerland (2015)
11. Prashanthi, V., Kanakala, S.: Generating analytics from web log. *Int. J. Engi. Adv. Technol.* **9**(4), 161–165

# Social Media Anatomy of Text and Emoji in Expressions



Shelley Gupta, Ojas Garg, Radhika Mehrotra, and Archana Singh

**Abstract** Social Media is a burgeoning platform where a person or a group can interact to collaborate and share their ideas through online mediums. Nowadays, social media content of Twitter, Facebook, Instagram, WhatsApp, online blogs, forums, etc. consists of text and emoji. Emoji are pictorial representation used in electronic media, which a user can or cannot combine with a text to express their sentiments. However, there is still too little work related to emoji in sentiment analysis. These emoji assist in extensive analysis of the opinions and sentiments of the public by means of Sentiment Analysis. Due to availability of large quantity of opinion-rich digital data, much of the current research is focusing on the area of the sentiment analysis. In this paper, the anatomy of social online media content is analyzed. Here, nearly 1.7 lac comments are analyzed. Also, the sentiments expressed by the end users on various popular products, delight events, etc., are analyzed to determine the usage of text and emoji on social online media like Facebook.

## 1 Introduction

In recent years, the number of people using social online media to share their opinions and emotions have increased enormously [1, 2]. Social online media like Facebook, Twitter and Instagram plays an important role in day to day life to communicate and to share opinion of the user without any restrain [3]. Facebook is one of the most popular social online media. It encloses the posts uploaded by the users which

---

S. Gupta (✉) · O. Garg · R. Mehrotra  
ABES Engineering College, Ghaziabad, Uttar Pradesh, India  
e-mail: [shelley.gupta@abes.ac.in](mailto:shelley.gupta@abes.ac.in)

O. Garg  
e-mail: [ojas.17bit1122@abes.ac.in](mailto:ojas.17bit1122@abes.ac.in)

A. Singh  
Department of Information Technology, Amity School of Engineering and Technology, Noida, India  
e-mail: [asingh27@amity.edu](mailto:asingh27@amity.edu)

consist of comments having both text and emoji. By this means the analysis can be done more extensively. Now the challenge is to frame a technology to identify and outline the sentiments.

Sentiment analysis is the automated process of visualizing and categorizing the opinions or sentiments expressed about a given subject or spoken language. In a world where 2.5 quintillion [4] bytes of data is generated every day; sentiment analysis has become a vital tool for making the data legitimate. Sentiment analysis can be useful in several ways. For example, in marketing, it helps in determining the popularity of the product, success rate of the product and helps in improving the quality of the product [5, 6]. Text and Emoji combination has provided the world a new way to express their emotions in a colorful manner.

The main objectives of the paper are:

- To evaluate online social media pages of multiple popular categories involving events (sport and movie), automobiles, brands, OTT platform, ecom platforms, etc.
- To evaluate and determine the total count of comments having both text and emoji along with the total count of emoji used.
- Comparative study among categories to determine among them the usage of Emoji.
- Comparative study among categories to determine among them the usage of Comments with Text and Emoji both.
- Study the aggregated anatomy of categories.

We use particularly Facebook because of the following reasons:

- Facebook is used by different people to express their opinion about different subjects; thus, it is a prominent source of people's sentiments.
- Facebook contains an enormous number of comments having both text and emoji and it grows every day.

Facebook's audience varies from regular users to celebrities, company representatives, political leaders, and even top brands have their profiles on Facebook. Therefore, it is possible to collect sentiments and opinions from different categories.

## 2 Literature Review

In recent years, Sentiment Analysis has become a prominent research topic in the field of Natural Language Processing [2]. With the rapid development of online networking sites and digital technologies, sentiment analysis is becoming more reputable than before. Sentiment Analysis has two approaches Machine Learning approach [6, 7] and Lexicon-based approach [5].

Machine Learning approach, is a technique to categorize the contextual data into predetermined categories without being explicitly programmed [7]. Machine

Learning approach can be divided into two methods [8, 9]. First, Supervised Learning which contains two sets of documents which are training set and a test set. Training set is used to study the data by the classifier. Test set is used for validating the data. Second, Unsupervised Learning, which involves the unlabeled data, for this, the algorithm recognizes the learning patterns in input and gives the specific output values. It also not requires prior training to mine data [10].

Lexicon-based approach, label the polarity (can be positive, negative, or neutral) of a textual content by aggregating the polarity of each word or phrase [11]. It is divided into two methods-Dictionary-based approach, which uses WordNet [12, 13], SentiWordNet [14] or any other dictionary to find suitable words of the sentiment word to determine the polarity. Second is Corpus-based approach, this helps in finding context-specific orientated sentiment words from a huge corpus [15].

Emoji have become competent aid of sentiment analysis as they are widely used in expressing feelings and emotions [2]. The meticulousness of recognizing emotions can increase and improve with the analysis of emoji. Emoji first originate in Japanese mobile phones in 1997 [16], and it became progressively popular worldwide in 2010s after being added to several mobile operating systems.

Most of the research are done with text [6, 7] only or with emoji [16] only but not involving the both. The objective of the paper is to determine the utilization of both text and emoji on social platform to express sentiments.

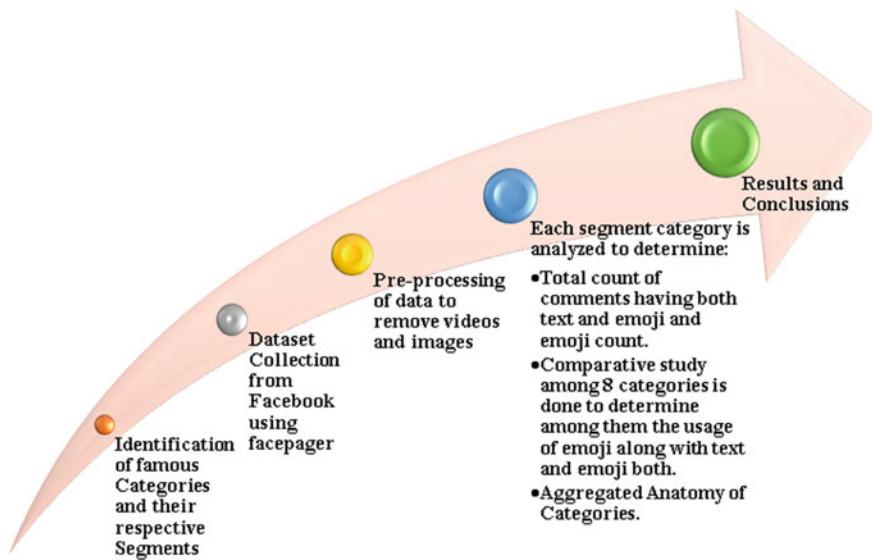
### 3 Proposed Approach

User's sentiments and opinions are the major criterion for the improvement of the quality and services. Overall, 117,000 online sentiment of 8 popular categories across the world have been downloaded as shown in Table 1. Sentiments or opinions of a person cannot be determined by only investigating the text. For definite and meticulous results, evaluation of both text and emoji should be done. The steps of the approach adopted are shown in Fig. 1.

- Step 1: Identify some of the famous categories like Events (Sport and Movie), Ecommerce Platforms, Automobiles, etc. These categories are further divided into segments like Events (Sport and Movie) into IPL, FIFA, Filmfare, etc.
- Step 2: The posts and comments of most popular categories on Facebook are downloaded using Facepager [17].
- Step 3: The data set is pre-processed to remove videos and images.
- Step 4: Each segment category is analyzed to determine:
  - Total count of comments having both text and emoji and emoji count.
  - Comparative study among 8 categories is done to determine among them the usage of emoji along with text and emoji both.
  - Aggregated Anatomy of Categories.

**Table 1** Statistics of post and comments under each category

S No.	Categories	Total segments in each category	Total posts in each segment	Total posts in each category	Total comments
1	Events (Sport and Movie)	7	200	1400	17,510
2	Ecommerce Platforms	7	200	1400	22,420
3	Automobiles	7	200	1400	21,198
4	Online Platforms	7	200	1400	23,775
5	OTT Platforms	7	200	1400	18,029
6	Gadgets	7	200	1400	18,701
7	Broadcasters	7	200	1400	29,319
8	Skin Care Brands	7	200	1400	20,669
	TOTAL	56	1600	11,200	171,621

**Fig. 1** Proposed approach

Step 5: The result of each segment is exhibited using charts.

Based on this analysis we determine the popularity of emoji among online users, Table 1 shows the statistics of post and comments under each category.

## 4 Implementation and Results

The dataset has been collected from Facebook and downloaded using Facepager. The module of Python called emoji is used to analyze emoji in dataset. Some of Python library used involves: Openpyxl to read and write in excel, Pathlib to work with file paths, re, workbook, etc. The results of (1) category-segment (2) category comparative study (3) aggregated anatomy of categories are listed below:

### 4.1 Category-Segment Analysis

Figure 2, shows the graphical representation of comment count, emoji count, and comments with text and emoji of various segments analyzed under different 8 categories.

### 4.2 Category Comparative Analysis

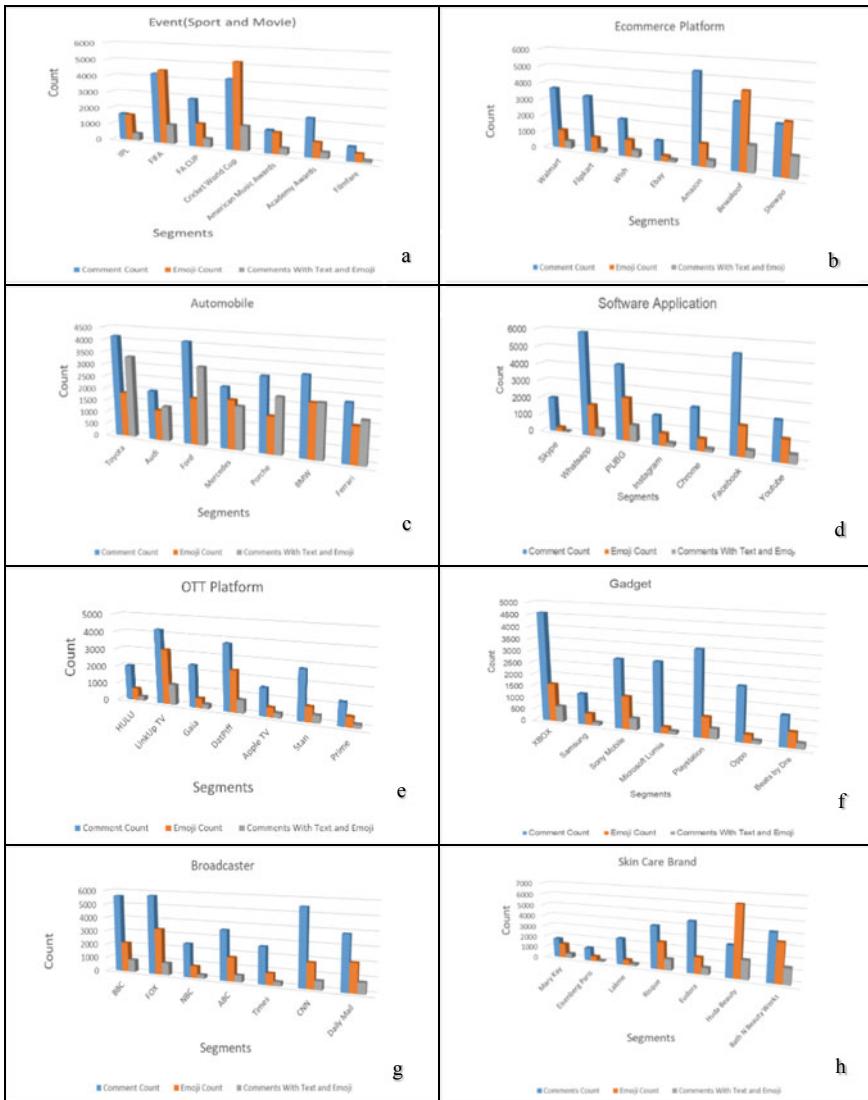
Table 2, shows that 6 categories out of 8 i.e. 75% of categories have total number of emoji, i.e., emoji count, greater than 40% of the total number of their respective comments considered i.e. comment count. Therefore, the total number of comments with text and emoji cannot be ignored. Thus, clearly reflecting the greater use of emoji by online users to express their sentiments.

Figure 3 shows that category of automobile incorporates maximum use of comments with text and emoji. Figure 4 shows that the usage of emoji is maximum in skin care brand and events (sports and movies).

### 4.3 Aggregated Anatomy of Categories

Figure 5 shows the following results:

- The emoji count is more than the 50% of the total number of comments considered.
- The count of comments with text and emoji both is nearly 25% of the total number of comments considered.
- The count of comments with text and emoji both is nearly 47% of the total number of emoji used.



**Fig. 2** **a** Events (Sport and Movie), **b** Ecommerce Platform, **c** Automobile, **d** Software Application, **e** OTT Platform, **f** Gadget, **g** Broadcaster, **h** Skin Care Brand

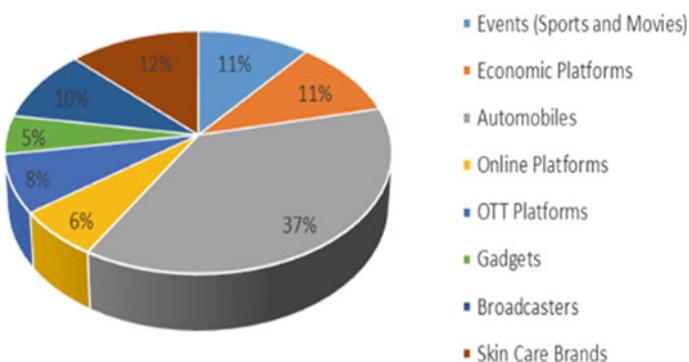
## 5 Conclusion

Our findings are based on the popular pages of Facebook and analyzing the total count of emoji and total count of comments with both text and emoji. From the analysis, we observe that these days, people are utilizing more emoji as a means of expressing their sentiments and opinions on social media. Thus, the sentiment analysis can be

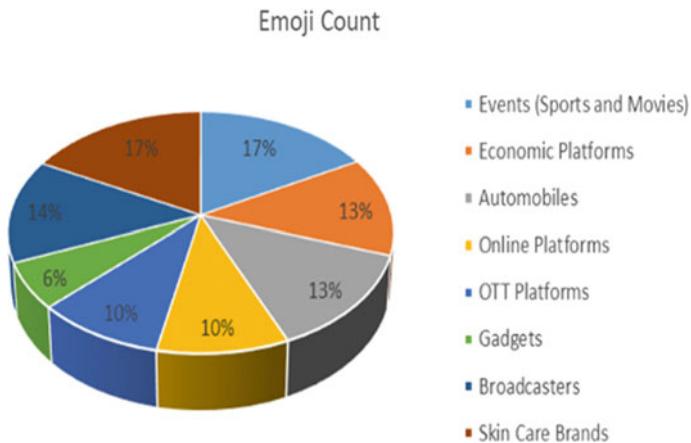
**Table 2** Category comment count, emoji count, and comments with text and emoji

S. No.	Category	Comment count	Emoji count	Comments with text and emoji
1	Event (Sport and Movie)	17,510	15,456	4604
2	Economic Platform	22,420	12,126	4511
3	Automobile	21,198	12,070	15,833
4	Online Platform	23,775	8835	2709
5	OTT Platform	18,029	8645	3277
6	Gadget	18,701	5573	2212
7	Broadcaster	29,319	12,663	4185
8	Skin Care Brand	20,669	15,772	5362
	Total	171,621	91,140	42,693

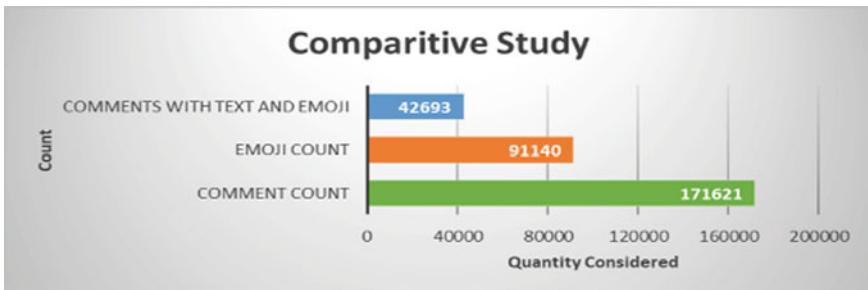
Comments with Text and Emoji

**Fig. 3** Categories comparison of comments with text and emoji

done more extensively when both emoji and text are considered. It provides us the direction for future work.



**Fig. 4** Categories emoji count



**Fig. 5** Aggregated anatomy of categories

## References

1. Suman, C., Saha, S., Bhattacharyya, P., Chaudhari, R. S.: Emoji helps! A multi-modal Siamese architecture for Tweet user verification. *Cogn. Comput.*, 1–16 (2020)
2. Gupta, S., Singh, A., Ranjan, J.: Sentiment analysis: usage of text and emoji for expressing sentiments. In: *Advances in Data and Information Sciences*, pp. 477–486. Springer, Singapore (2020)
3. Anjaria, M., Gudetti, R.M.R.: A novel sentiment analysis of social networks using supervised learning. *Soc. Netw. Anal. Min.* **4**(1), 181 (2014)
4. Marr, B.: How much data do we create every day? The mind blowing stats everyone should read (2020). Retrieved 21 June 2020, from <https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/>
5. Dey, A., Jenamani, M., Thakkar, J.J.: Senti-N-Gram: an n-gram lexicon for sentiment analysis. *Expert Syst. Appl.* **103**, 92–105 (2018)
6. Tripathy, A., Agrawal, A., Rath, S.K.: Classification of sentiment reviews using n-gram machine learning approach. *Expert Syst. Appl.* **57**, 117–126 (2016)

7. Tripathy, A., Anand, A., Rath, S.K.: Document-level sentiment classification using hybrid machine learning approach. *Knowl. Inf. Syst.* **53**(3), 805–831 (2017). <https://doi.org/10.1007/s10115-017-1055-z>
8. Ayyadevara, V. K.: Word2vec. In: *Pro Machine Learning Algorithms*, pp. 167–178. Apress, Berkeley, CA (2018)
9. Asghar, M.Z., Khan, A., Bibi, A., Kundu, F.M., Ahmad, H.: Sentence-level emotion detection framework using rule-based classification. *Cogn. Comput.* **9**(6), 868–894 (2017). <https://doi.org/10.1007/s12559-017-9503-3>
10. Schrauwen, S.: Machine learning approaches to sentiment analysis using the Dutch Netlog Corpus. Computational Linguistics and Psycholinguistics Research Center, pp. 30–34 (2010)
11. Da Silva, N.F.F., Hruschka, E.R., Hruschka, E.R.: Tweet sentiment analysis with classifier ensembles. *Decis. Support Syst.* **66**, 170–179 (2014). <https://doi.org/10.1016/j.dss.2014.07.003>
12. Miller, G.A.: Nouns in WordNet. WordNet: an electronic lexical database, pp. 23–46 (1998)
13. Lee, Y., Ke, H., Yen, T., Huang, H., Chen, H.: Combining and learning word embedding with WordNet for semantic relatedness and similarity measurement. *J. Am. Soc. Inf. Sci.* **71**(6), 657–670 (2019). <https://doi.org/10.1002/asi.24289>
14. Denecke, K.: Using SentiWordNet for multilingual sentiment analysis. In: 2008 IEEE 24th International Conference on Data Engineering Workshop (2008). <https://doi.org/10.1109/icdew.2008.4498370>
15. Hardeniya, T., Borikar, D.A.: Dictionary based approach to sentiment analysis—a review. *Int. J. Adv. Eng. Manage. Sci.* **2**(5) (2016)
16. Fernández-Gavilanes, M., Juncal-Martínez, J., García-Méndez, S., Costa-Montenegro, E., González-Castaño, F.J.: Creating emoji lexica from unsupervised sentiment analysis of their descriptions. *Expert Syst. Appl.* **103**, 74–91 (2018)
17. Facepager 3.6 Download (Free)—Facepager.exe (2020). <https://facepager.software.informer.com/3.6/>

# Development of Machine Learning Model Using Least Square-Support Vector Machine, Differential Evolution and Back Propagation Neural Network to Detect Breast Cancer



Madhura D. Vankar and G. A. Patil

**Abstract** Recently death rate among women increases due to Breast Cancer. Breast Cancer diagnosis system has been implemented with the help of Different Machine learning algorithms. Out of those, in our present work, we have used DE, LSSVM, and BPNN algorithms for developing Machine learning models. Here the training dataset is taken from UCI repository. In this system, we compare results of LS-SVM and backpropagation neural networks to generate accurate results at an early stage. Early detection increases the chances of treatment which helps to save women Life. It will be helpful to medical field because it avoids loss of data.

## 1 Introduction

Machine learning provides capability to computers for learning without being programmed. Machine learning algorithms are used to develop models with the help of training data. It includes two phases i.e. Training and Testing phase. Mostly Medical field utilizes different machine learning tools because it uses concepts such as classification and recognition systems.

This paper focuses on the breast cancer diagnosis at early stage, conducted using least square support vector machine (LS-SVM) classifier algorithm and back-propagation neural network algorithm (BPNN). Early diagnosis is necessary to increase survivability of women. Here we perform demonstration on Wisconsin Breast Cancer Dataset (WBCD) taken from UCI repository. System will use a Differential Evolution algorithm to optimize parameters of LS-SVM classifier which helps to enhance the accuracy. Here we use Supervised Learning mode of Neural Network which helps to train the neural network with a training dataset. Subsequently, it helps in the backpropagation algorithm to minimize the errors by adjusting interconnections weights. Here, the performance of the system is evaluated by using a tenfold cross validation method.

---

M. D. Vankar (✉) · G. A. Patil

Computer Science and Engineering Department, D. Y. Patil College of Engineering and Technology, Shivaji University, Kasaba Bawada, Kolhapur, Maharashtra, India

## 2 Literature Survey

Singh and Parveen [1] have proposed a hybrid methodology for classification with the help of support vector machine (SVM) and fuzzy c-means clustering. It provides accurate results to detect the brain tumor. It requires large memory space.

Lin et al. [2] have developed CAD system for characterizing breast nodules as either benign or malignant by using ultrasonic images. This paper uses classifier such as fuzzy cerebellar model neural network (FCMNN). It follows trial and error criteria to decide its all analysis parameters.

Vijay and Saini [3] have proposed system for detection of breast cancer using Image Registration Techniques. System describes features of Feed-forward back-propagation Artificial Neural Network (ANN) model. Mean Square Error (MSE) is used as analysis criteria to evaluate system performance. It is more time-consuming, because it follows trial and error method to compute number of neurons.

Utomo et al. [4] have used Extreme Learning Machine Neural Networks (ELM ANN) for developing decision support systems. ELM algorithm helps to generate accurate results and high sensitivity rate. This system requires more time to train the network than other methods. It provides low specificity rate.

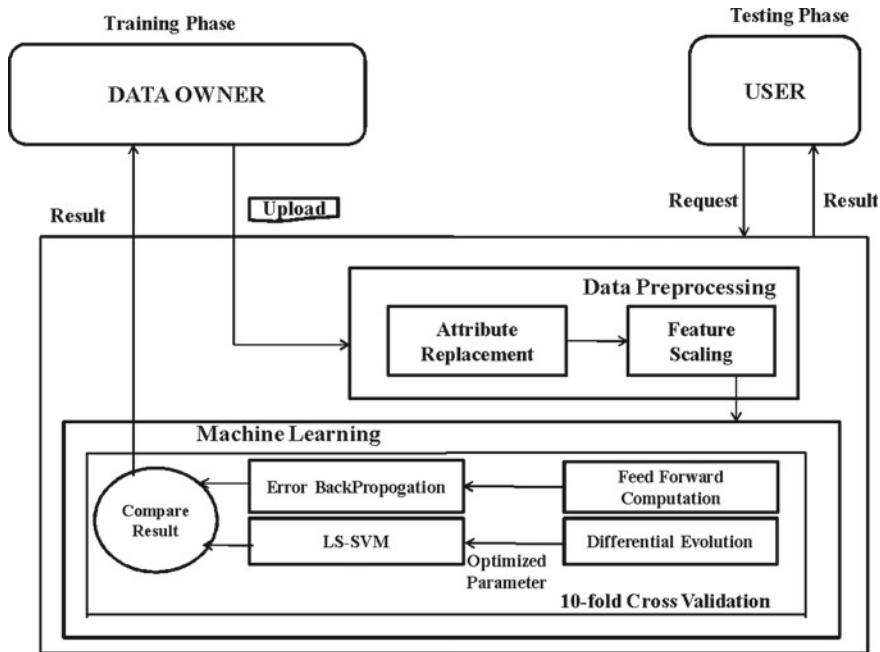
Saini et al. [5] have developed a model using an adaptive neural network to detect cancer. It requires more training time. It provides better accuracy as compared to the fuzzy system. It uses clustering features which allows to create number of clusters depends upon the problem size.

There was a need felt to develop a Machine Learning Model to optimize training time and enhance performance by using Least Square-Support Vector Machine, Differential Evolution and Back Propagation Neural network, so as to help detect more accurately breast cancer in the patients. The followings are the main objectives of this research work:

- To develop Machine Learning Model.
- To replace missing value in the dataset using a suitable algorithm.
- To optimize training time by Merging Differential Evolution with LS-SVM.
- To obtain Accurate detection results by comparing LS-SVM and BPN.
- To evaluate performance of the system with the help of a tenfold cross validation method.

## 3 System Architecture

Our proposed system will be used for developing machine learning model with the help of big data. System will process Dataset by using missing value replacement method to improve the performance. Then ‘Feature Scaling’ is used to convert data into binary format. After completion of conversion, the dataset will be transferred towards the machine learning model to train neural network by using two classifiers such as LS-SVM and BPN. Further, the focus is to compare the results of two



**Fig. 1** System architecture

classifiers and provide accurate results to the end-user with the help of analysis parameters (Fig. 1).

## 4 Methodology

### 4.1 Dataset Attributes

The given algorithms are implemented on the Wisconsin Breast Cancer Dataset (WBCD) taken from UCI repository of machine learning databases [13, 14]. Total Dataset consists of 10 Attributes. Out of those, nine attributes indicate as input variables and remaining one as output (2, 4). The value 2 indicates benignancy and 4 indicates that it is of malignant type of breast cancer. The dataset also contains another attribute such as ‘sample code number’ which is discarded, because it is not required for the classification process.

Table 1 provides information about attributes along with its range.

The functioning of the system has been divided into the following modules.

**Table 1** Attributes to be considered from the WBCD

S. no.	Attributes	Domain
1	Clump thickness	1 ... 10
2	Cell size—uniformity	1 ... 10
3	Shape of cell—uniformity	1 ... 10
4	Marginal adhesion	1 ... 10
5	Cell size—single epithelial	1 ... 10
6	Bare nuclei	1 ... 10
7	Bland chromatin	1 ... 10
8	Normal nucleoli	1 ... 10
9	Mitoses	1 ... 10
10	Class	2 shows benignancy, 4 shows malignancy

## 5 Module 1: Missing Value Replacement and Feature Scaling of Attributes

This module is used to make a dataset ready for training neural network using the following methods. Here, Data owner authentication, User authentication, and Missing value replacement mechanisms have been implemented.

### (A) Data Owner Authentication:

In Data owner authentication, the owner will sign in into the system by providing his/her username and password. If a new user wants to upload a testing database file into the system, he has to register giving his username, password, mobile number, address, and other details required. Once the data owner is authenticated, he can browse and load the Wisconsin Breast Cancer Training Dataset into the system for processing.

### (B) Litwise Deletion Algorithm to Ignore the Tuples Containing Missing Data:

This method helps to remove missing values within a given dataset. Dataset consists of attributes as mentioned in Table 1. To delete the subjects that have missing value is the simplest way of handling missing data. If any row contains a missing value for one attribute which represents ‘?’ symbol, then that entire row is deleted.

### (C) Feature Scaling of Attribute:

After completion of litwise deletion method, dataset will be forwarded towards Feature Scaling. The Feature Scaling of attribute is carried out to convert dataset values into binary format except last attribute i.e. class (2-benign, 4-malignant). The following formula is used to obtain scaling.

$$X' = [x - \min(x)] / [\max(x) - \min(x)]$$

Use the following considerations to obtain new values:

- (i) 0–5 values are converted to 0
- (ii) 6–10 are converted to 1.

## 6 Module 2: Training Neural Network using LS-SVM and DE

This module is focusing on the training of neural network using Least-Square Support Vector Machine and Differential Evolution merged with LS-SVM to improve the system's accuracy.

### (A) Least Squares Support Vector Machine (LS-SVM):

Figure 2 shows the role of LS-SVM to classify all training vectors in two classes with the help of hyperplane and to select the best hyperplane that leaves the maximum margin from both classes.

#### LS-SVM Concept:

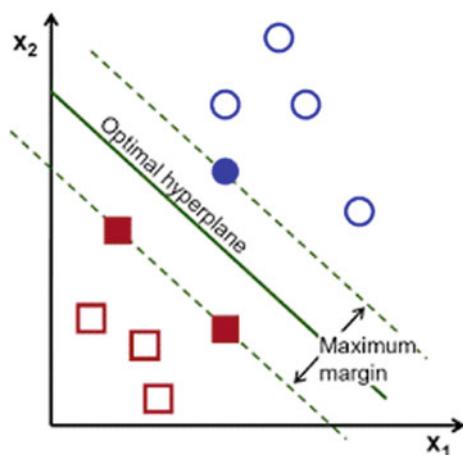
LS-SVM classifier is used to obtain suitable a hyper plane, which separates various classes. It is an extended version of SVM. As compared to other techniques, LS-SVM is very simple and faster.

#### LS-SVM Algorithm:

- Step 1: Load the training vector of  $n$  data points, where  $X_i$  represents the input vector and  $Y_i$  represents the corresponding  $i$ th target with values {2,4}.
- Step 2: For each input data point find out random weights

$$g(\vec{x}) = \vec{x} \cdot \vec{w} + w_o$$

**Fig. 2** Concept of LS-SVM



Step 3: Obtain the value of the bias term  $b$  and initialize the error  $e$  for each point randomly.

$$\begin{aligned} g(\vec{x}) &\geq 1 \quad \forall \vec{x} \in \text{Class 1} \\ g(\vec{x}) &\leq -1 \quad \forall \vec{x} \in \text{Class 2} \end{aligned}$$

Step 4: Minimize the objective function with the help of  $e, w$  and  $b$  values. Calculate the value of total margin.

Minimize  $\vec{w}$  term which helps to maximize separability between two classes.

Step 5: To satisfy the Karush-Kuhn-Tucker (KKT) conditions in the set of equations by developing the Lagrangian function with the solution

$$\vec{w} = \sum_{i=0}^N \lambda_i y_i \vec{x}_i$$

$$\sum_{i=0}^N \lambda_i y_i = 0$$

Step 6: Find out the number of support vectors.

Step 7: Use RBD kernel functions for classification of Training data. Use RBD kernel functions for classification of Training data.

#### (B) Differential Evolution (DE):

It is simple optimization Technique.

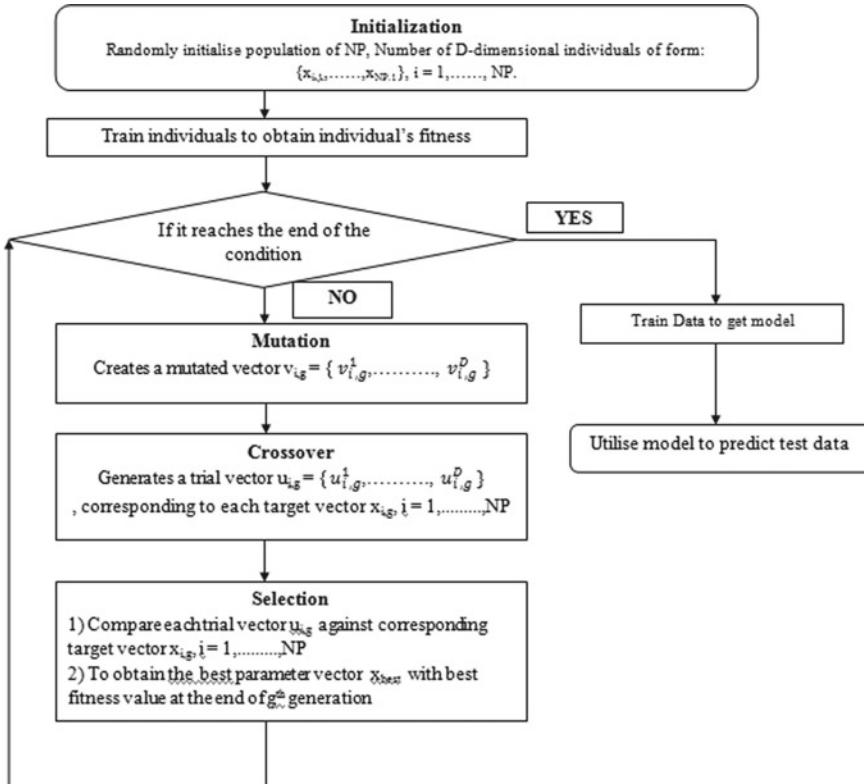
#### **DE-Based Optimization Technique:**

Figure 3 shows idea behind DE-LSSVM algorithm which helps to utilize developed model to predict test data.

DE has standard phases like initialization, mutation, crossover, and selection. DE improves its accuracy by merged it with LS-SVM classifier. Differential Evolution is used to optimize kernel parameters of LS-SVM. It will create a new fitness function which act as a classifier evaluation criteria, i.e., accuracy, mean absolute error, and root mean squared error.

## **7 Module 3: Training Neural Network Using Back Propagation Neural Network (BPNN)**

The Back Propagation algorithm is quite simple and easy to program. It is used for performing many tasks such as Classification, Clustering. Here we use a Supervised Learning model to train the neural network. It includes two phases such as



**Fig. 3** The flowchart of the DE-LSSVM algorithm

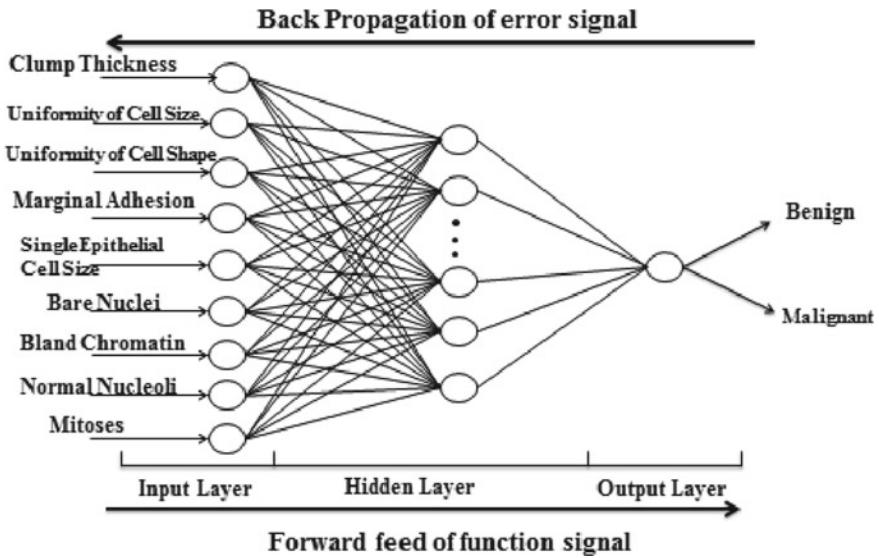
Feed Forward Computation and Backpropagation. Feed Forward Neural Network phase helps to calculate output then it will be checked against actual output. If calculated outputs are not satisfactory then connections between layers are modified by using Backpropagation algorithm. This procedure is performed again and again to minimize error which is calculated at Feed forward neural network phase (Fig. 4).

#### Four Phases of the Training Process:

1. **Initialization of weights:** It is used to assign some small random values.
2. **Feed forward approach:** In this phase, network has fixed synaptic weights. It is used to calculate error by using the following formula.

$$\text{Error Signal} = \text{Calculated Output} - \text{Desired Output}$$

3. **To calculate the propagated error:** It focuses on minimization of error by adjusting synaptic weights of network. Forward and backward passes are repeated until error is minimized.



**Fig. 4** Working of BPNN

$$W_{AB(\text{new})} = W_{AB(\text{old})} - \eta \frac{\partial E^2}{\partial W_{AB}}$$

where  $\eta$  represents the learning rate parameter,  $\frac{\partial E^2}{\partial W_{AB}}$  represents the sensitivity.

#### 4. To update weights in the neural network.

### 8 Module 4: Tenfold Cross Validation Method:

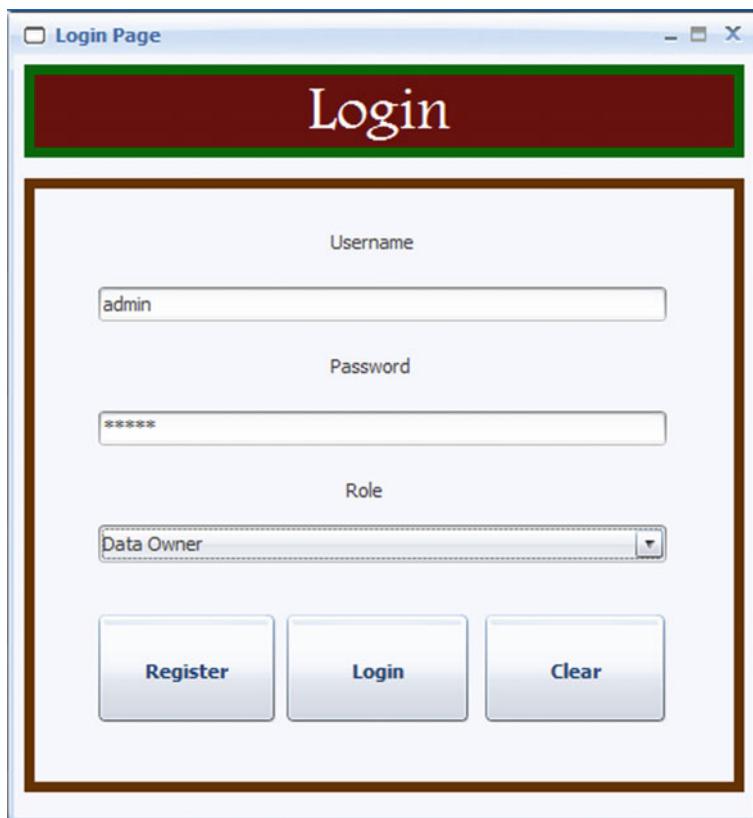
It mainly focuses on evaluation of system performance with the help of LS-SVM and BPN. Steps of tenfold cross validation performance evaluator are:

- First divide the given dataset into 10 sets.
- Then out of 10 sets, Training is performed on 9 sets and testing is performed on remaining 1 set.
- Take its mean accuracy by repeating this procedure 10 times.

### 9 Module 5: Experimental Results and Analysis

For experimentation net beans IDE has been installed.

Figure 5 shows Data owner authentication page, where owner will sign in into the system by providing his/her username and password.



**Fig. 5** Data owner authentication

After data owner authentication process, data owner can browse and load the Wisconsin Breast Cancer Training Dataset into system for processing. Here we will perform demonstration on 600 records of WBCD.

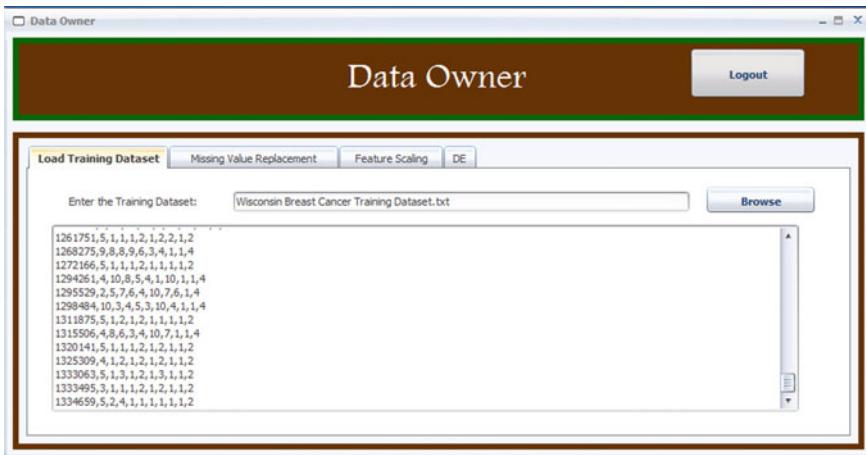
Figure 6 shows the Data Owner screen which is used to load training dataset for development of machine learning model.

Figure 7 shows the dataset after replacing missing values by using Litwise deletion method.

Figure 8 shows the dataset values in binary format except the last attribute that is being obtained after performing feature scaling.

Figure 9 shows a Data Owner screen with DE-LS-SVM Training results that include correctly and incorrectly classified instances which will help to calculate evaluation parameters of the system.

1. **Accuracy** = It represents percentage of correctly classified data.



**Fig. 6** Load dataset for training



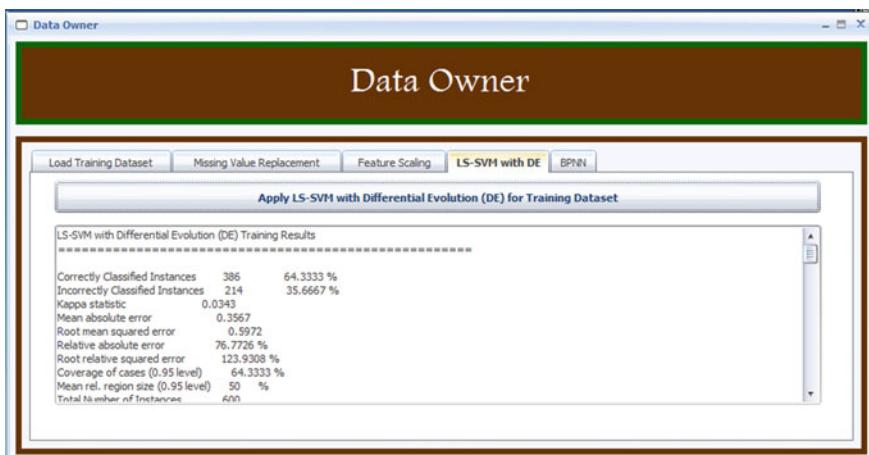
**Fig. 7** Litwise deletion method

$$\begin{aligned}
 &= \sum_{m=1}^c (\text{No. of correctly classified data in class } m) / \sum_{m=1}^c (\text{Total No. of data in class } m) \\
 &= (386/600) * 100 = 64.3333\%
 \end{aligned}$$

2. **Mean Absolute Error** = It measures error between observations using the following formula.



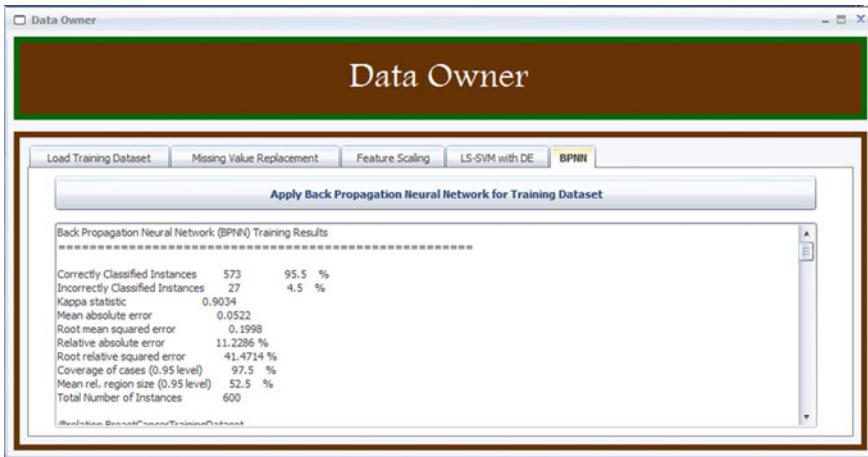
**Fig. 8** Feature scaling of attributes



**Fig. 9** Train dataset using DE and LS-SVM

$$\begin{aligned}
 \text{MAE} &= \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \\
 &= \frac{1}{n} \{\text{prediction} - \text{Actual observation}\} \\
 &= \frac{1}{600} \{600 - 386\} = 0.3566
 \end{aligned}$$

where,  $n$  = total number of instances within the dataset.



**Fig. 10** Train dataset using BPNN

3. **Root Mean Squared Error (RMSE)** = It shows the value i.e. square root of the MAE.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|} = \sqrt{0.3566} = 0.5971$$

Figure 10 shows a Data Owner screen with BPNN Training results that include correctly and incorrectly classified instances.

1. **Accuracy** =  $(573/600)*100 = 95.5\%$ .
2. **Mean Absolute Error (MAE)** =  $\frac{1}{600} \{600-573\} = 0.0522$
3. **Root Mean Squared Error** =  $\sqrt{0.0522} = 0.1998$ .

Tables 2, 4 and 5 contain calculated values of Accuracy, Mean Absolute Error, and Root Mean Squared Error of DE-LSSVM and BPNN algorithms with respect to number of records given by end-user.

From Table 2, we can easily calculate the average of accuracy for DE-LSSVM and BPNN algorithms with respect to the number of records within the dataset.

Table 3 gives clear idea that BPNN algorithm provides more accurate results as compared to DE-LSSVM algorithm.

Figure 11 shows that accuracy of BPNN is higher than DE-LSSVM algorithm. It is being observed that when number of records within dataset increases then accuracy of DE-LSSVM algorithm goes on decreasing.

Figure 12 shows that all analysis parameters of DE-LSSVM algorithm with respect to number of records within dataset.

Figure 13 shows that all analysis parameters of BPNN algorithm with respect to number of records within the dataset. From Fig. 11, it is being observed that when the

**Table 2** Accuracy between DE-LSSVM and BPNN algorithm

Number of records	Accuracy (%)	
	DE-LSSVM	BPNN
10	90	90
20	75	95
30	70	90
40	67	97.5
50	62	94
100	56	94
200	58	94.5
300	56.33	94.667
400	58.25	93.75
500	61.8	95
600	64.33	95.5

**Table 3** Average accuracy of DE-LSSVM and BPNN algorithm

Algorithm	DE-LSSVM	BPNN
Average accuracy (%)	65.34	93.99

**Table 4** Mean absolute error, Root mean squared error values of DE-LSSVM algorithm

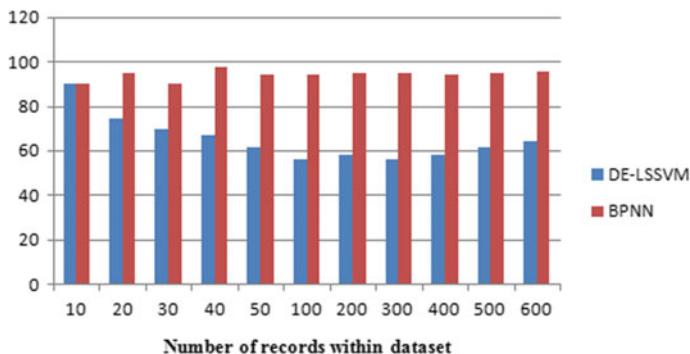
Number of records	Mean absolute error	Root mean squared error
10	0.1	0.3162
20	0.25	0.5
30	0.3	0.5477
40	0.325	0.5701
50	0.38	0.6164
100	0.44	0.6633
200	0.42	0.6481
300	0.4367	0.6608
400	0.4175	0.6461
500	0.382	0.6181
600	0.3566	0.0522

number of records within a dataset increases then accuracy of BPNN algorithm goes on increasing which means number of records within dataset is directly proportional to the accuracy of BPNN algorithm.

**Table 5** Mean absolute error, Root mean squared error values of BPNN algorithm

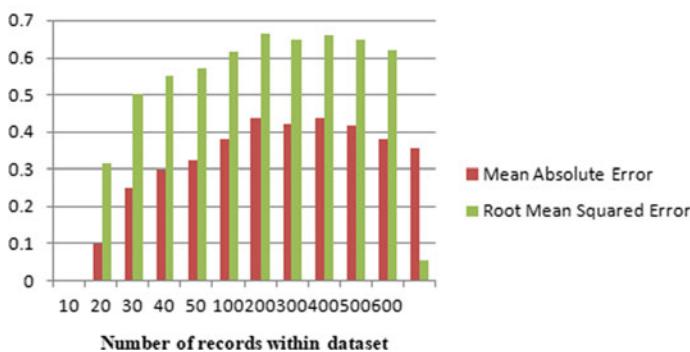
Number of records	Mean absolute error	Root mean squared error
10	0.1503	0.3222
20	0.1419	0.265
30	0.1079	0.2416
40	0.0668	0.1738
50	0.0944	0.2401
100	0.0728	0.2159
200	0.0609	0.2104
300	0.0651	0.2156
400	0.0679	0.2295
500	0.0603	0.2152
600	0.5971	0.1998

#### Accuracy between DE-LSSVM and BPNN Algorithm

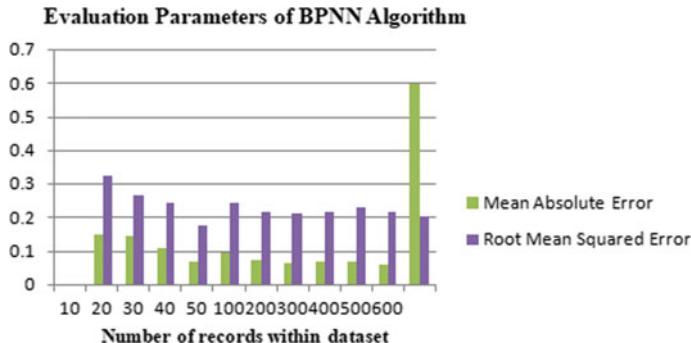


**Fig. 11** Accuracy between DE-LSSVM and BPNN algorithm

#### Evaluation Parameters of DE-LSSVM Algorithm



**Fig. 12** Evaluation parameters of DE-LSSVM algorithm



**Fig. 13** Evaluation parameters of BPNN algorithm

## 10 Conclusion

In this paper, Machine learning model has been developed for breast cancer patients with the help of DE-LSSVM and BPNN classification algorithms. Here missing values are handled by using Litwise deletion method under Dataset Preprocessing module which enhances system performance. DE is combined with LS-SVM which helps to improve system's accuracy and minimize the training time. Therefore it helps to generate results at early stage which increases chance of survivability of women. As mentioned in Table 3: BPNN gives 93.99% of average accuracy and LS-SVM gives 65.34% of average accuracy with respect to the available records in the given dataset. Therefore from the results of all analysis parameters, it is observed that BPNN algorithm provides more accurate results as compared to DE-LSSVM algorithm. This accuracy probably will go on increasing for large dataset. There is a further scope to go for implementation with solo classification techniques to improve the efficiency. Adding other features will help the system to handle and test more than one dataset at a time.

## References

1. Parveen and Singh, A.: Detection of brain tumor in MRI images, using combination of Fuzzy C-Means and SVM. In: IEEE Paper on Signal Processing and Integrated Networks (SPIN) (2015)
2. Hou, Y.-L., Lin, C.-M., Chen, K.-H., Chen, T.-Y.: Breast nodules computer-aided diagnostic system design using Fuzzy Cerebellar model neural networks. IEEE Trans. Fuzzy Syst. **22**(3) (2014)
3. Saini, S., Vijay, R.: Optimization of artificial neural network breast cancer detection system based on image registration techniques. Int. J. Comput. Appl. **105**(14) (2014)
4. Utomo, C.P., Kardiana, A., Yuliwulandari, R.: Breast cancer diagnosis using artificial neural networks with extreme learning techniques. Int. J. Adv. Res. Artif. Intel. **3**(7) (2014)
5. Singh, S., Saini, S., Singh, M.: Cancer detection using adaptive neural network. Int. J. Adv. Res. Technol. **1**(4) (2012)

6. Al-Anezi, M.M., Mohammed, M.J., Hammadi, D.S.: Artificial immunity and features reduction for effective breast cancer diagnosis and prognosis. IJCSI Int. J. Comput. Sci. Issues **10**(3) (2013)
7. Jouni, H., Issa, M., Harb, A., Jacquemod, G., Leduc, Y.: Neural network architecture for breast cancer detection and classification. In: IEEE International Multidisciplinary Conference on Engineering Technology (IMCET) (2016)
8. Menaka, K., Karpagavalli, S.: Breast cancer classification using support vector machine and genetic programming. Int. J. Innovative Res. Comput. Commun. Eng. **1**(7) (2013)
9. Fiiji, H.H., Almasi, B.N., Mehdikhan, Z., Bibak, B., Pilevar, M., Almasi, O.N.: Automated diagnostic system for breast cancer using least square support vector machine. Am. J. Biomed. Eng. (2013)
10. Usha Rani, K.: Parallel approach for diagnosis of breast cancer using neural network technique. Int. J. Comput. Appl. **10** (2010)
11. Sharma, A., Mehta, N., Sharma, I.: Reasoning with missing values in multi attribute datasets. Int. J. Adv. Res. Comput. Sci. Softw. Eng. **3**(5) (2013)
12. Kaushik, K., Arora, A.: Breast cancer diagnosis using artificial neural network. Int. J. Latest Trends Eng. Technol. (IJLTET) **7**(2) (2016)

### ***Web Links for Accessing Information About Breast Cancer Dataset***

13. <https://archive.ics.uci.edu/ml/index.php>
14. [https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))

# Distributed and Decentralized Attribute Based Access Control for Smart Health Care Data



B. Ravinder Reddy and T. Adilakshmi

**Abstract** Most access control solutions today provide the ability for centralized authorities, whether governments, manufacturers, or service providers to gain unauthorized access to and control devices by collecting and analyzing user's data. Data owners (DO) experience a necessity to concentrate on their own medical data and manage them. As an alternative, Blockchain promotes healthcare, including health care data, which actually may cause legal and privacy issues and solves the problem of sharing with third parties. Blockchain technology provides data owners with comprehensive, immutable records, and access to Smart Health Care (SHC) data free from service providers and websites. This paper presents a scheme, in which the data owner endorses the message based on attributes without leaking data other than the appended proof. The proposed model is Distributed and Decentralized scheme for Access Control.

## 1 Introduction

Smartwatches are ultimate medical sensors well suited for bringing essential medical monitoring into home which are easy to use, always running, and continuously in contact with our bodies. The major issue is when someone updates or modifies data without knowledge that lead to a major damage. Access Control of Smart Health Care (SHC) data can be an alternative solution for securing and providing privacy for data owners. SHC data stores the health-related personal data gathered from wearable devices like Fitbit. Some incorrect alteration of any such sensitive data

---

B. Ravinder Reddy (✉)

Department of CSE, UCE (A), OU, Hyderabad, India

e-mail: [ravinderreddycse@cvsr.ac.in](mailto:ravinderreddycse@cvsr.ac.in)

Department of CSE, Anurag Group of Institutions (A), Hyderabad 500088, India

T. Adilakshmi

Department of CSE, Vasavi College of Engineering (A), Hyderabad 500089, India

e-mail: [t\\_adilakshmi@staff.vce.ac.in](mailto:t_adilakshmi@staff.vce.ac.in)

can be harmful. Thus, privacy becomes an important aspect for a SHC system. The proposed model designates the patients to generate, share and manage data with e-healthcare providers by storing access rights of data collected on blockchain.

## 2 Literature Survey

### 2.1 Access Control

According to Zyskind et al. [1] blockchain can handle many real-time issues at large like. Currency and Transaction support, cloud storage, supply chains, and public charity.

Generally, in an organization, evaluation of access control policy to determine whether the requested access to a warehouse can be performed is done by a party that cannot be trusted at times. Attribute-Based Access Control (ABAC) policies are one of the common approaches to express access policy.

It combines a set of rules with conditions over a set of attributes assigned to subjects or resources. For instance, consider the resource,  $A$ , an object that controls the defined policy to resources, number of subjects,  $P_i$ , and number of resources  $X_j$ .  $P_i$  holds the rights to perform on actions like transfer of rights specified by defined policies on  $X_j$ . Hence, the approach needs  $A$  &  $P_i$  to perform actions which are independent from one another. The policy issuer and subject have no role to play while in the process of exchange. As coined by Di Francesco Maesa et al. [2] a policy creation transaction (PCT) is to be performed by  $A$  to transfer the access rights to a new subject in  $P_i$ . At last, after the policy is created, which is updated by resource owner  $A$ , is revoked.

Any Access control policy is defined based on Access Rights. An access right defined as permission for access of resource that is granted and which can allow a program or person to identify the ideal data. Digital access rights play crucial role in data security and compliance. These permissions control the ability of the users to view, read, write, change, navigate, and execute the contents of the file system. Table 1 represents the access rights.

**Table 1** Representation of access rights for different user types

User	User type	Type of access	Access control	File
Demo	Admin	Create different Users	Read, write execute	All files
Demo 1	Data owner (DO)	Create users and share access rights	Read, write execute	Uploaded file
Demo 2	User	As allowed by DO	As allowed by DO	As allowed by DO

## 2.2 *Blockchain*

On the other side, the invention of Bitcoin [3] by a person (or group of people) using the name Satoshi Nakamoto in 2008 has inspired blockchain and many other applications. Block chain contains three basic elements. The first is a distributed ledger, but it is centralized. The second element consensus algorithm is a way to make sure that all the ledger copies are the same for everyone and the process is called mining. The third, encryption and distribution of block, which contains data stored, and the result is a Hash code, which indicates the location of the block and its contents, and then links it with rest of the chain [4].

The blockchain-based Access Control is beneficial for, Easy of transfer access rights from one user to other, anytime validation can be done by any user to know who is having currently the access, and Proof of Existence of access rights is possible.

Certain properties or capabilities of blockchain ensure the highest security to transaction. Some of the terminology and core components of Blockchain are: Peer-to-Peer Network (P2P) [5], in which the users are connected to one another with no central control. Hash Pointer [6], which points to the next block of transaction, Digital Signatures [7], used for authenticating users. The two digital signature schemes which work on Elliptic and Edward curve respectively are Elliptic Curve Digital Signature Algorithm (ECDSA) and Edwards-curve Digital Signature Algorithm (EdDSA) [8]. Consensus [9] avoids intermediaries, Merkle Tree [10] is a technique that separates data and applies hash functions till a single hash value is achieved. Proof of Work (PoW) [11], a consensus algorithm used in the network to provide security and decision making, Proof of Stake [12], in bitcoin it is a collection of coins to win.

Byzantine Fault Tolerance [1], system is used to identify issues related to various components in the network.

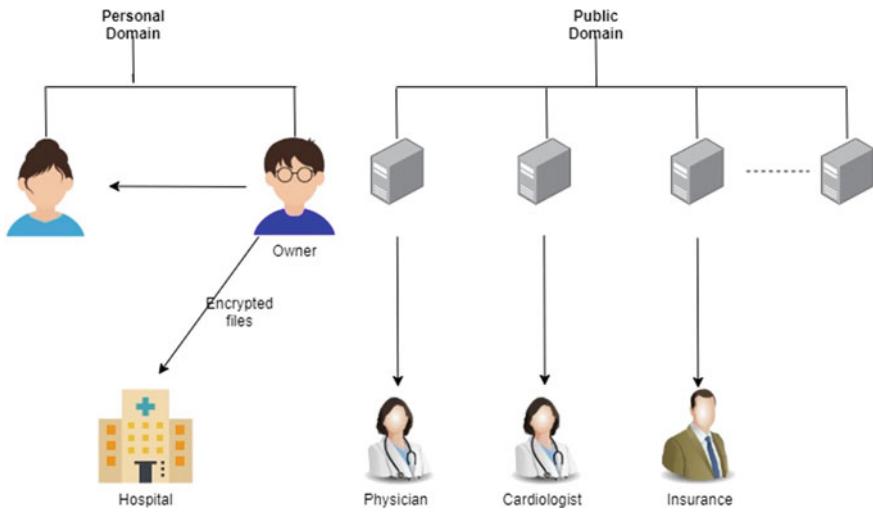
For example, Bob needs to upload a file and share it with someone like spark.

1. Bob: Uploads the file to cloud server.
2. Bob: gives access rights to spark for the file.
3. Spark: Initially can see the file but cannot read or download.
4. Spark: uses the private key generated by bob while giving attributes.
5. Spark: Finally, downloads the file.

In this using Blockchain and encryption can create two levels of security for the access control. Figure 1 represents the communication among patients, hospitals, and doctors.

## 2.3 *Smart Health Care on Cloud*

Deploying the SHC system on cloud ecosystem provides several opportunities like accessibility, computational elasticity, greater fault tolerance, and interoperability [13, 14]. As per HIPPA [15], the cloud providers are under non-covered entities. Thus,



**Fig. 1** Communication among patients, doctors, and hospital

the service providers have no obligation to ensure the integrity, confidentiality and to provide proper access to the consumers [16]. Consequently, the privacy becomes an important concern to adapt the medical system.

The health data is managed and controlled by the data owner or patient, unlike the other digital health records [17, 18]. Owners can selectively share their health-related data with third parties while hiding some data in private. The cloud can allow access the medical data anytime from anywhere. It can also support the system to prepare for appointments and maintain a more sophisticated view of personal health to share.

The cloud on the other side may have business interest in analyzing the health care data, and may also have malicious employees or sometimes even cloud may be hacked. As a result, the medical system will communicate with different users, and the given access control policy should support accountability and revocation features. Thus, the smart health care system must offer a tamper-proof feature and protect the data owner's privacy. In our proposed model, the underlying cloud infrastructure of the SHC system is considered as untrusted, and Blockchain used to provide security.

## 2.4 Limitations of Existing System

The existing system comes with various limitations like:

- ABAC is based on Central Authority and is not distributed.
- Creation or Configuring of ABAC is not so simple.
- Blockchain implementation of access control may be slow.

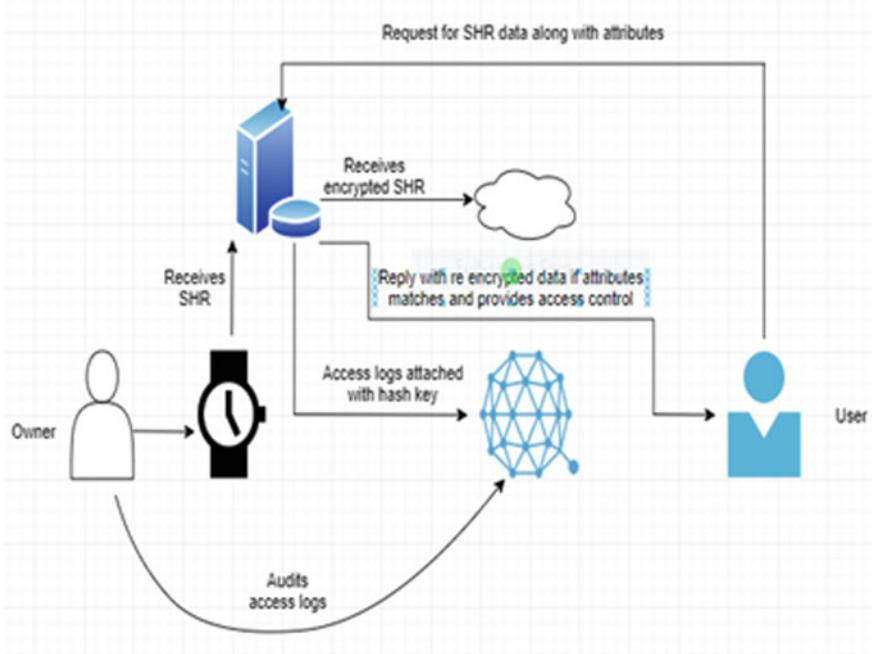
- Memory Utilization, Transaction time, Energy Consumption are high for blockchain.

### 3 Proposed Methodology

The proposed model in Fig. 2 works by collecting the data from wearable devices stores it on the cloud and assigns the data with access rights based on Attribute Based Access Control (ABAC) policies. The generated attributes and access rights or logs are stored on blockchain nodes which in turn can reduce the burden of overload, storage utilization on cloud, and improving performance making blockchain lightweight.

This system will provide high-quality preventive mechanisms for accessing smart health records. The Proposed model can also improve the efficiency in terms of power consumption, improves processing and execution time.

The benefits of this approach is secure sharing of patient information by providing decentralized access, improved patient and service provider interaction and safeguarding privacy. Both the general ABAC and Blockchain based ABAC schemes uses User, resource, object and environment as attributes. Table 2 depicts a comparison of Attribute based access control (ABAC) and blockchain ABAC.



**Fig. 2** Architecture of proposed model

**Table 2** Comparison of ABAC and blockchain based ABAC

S. no.	Parameter	ABAC	Blockchain based ABAC
1	Access rights	Granted based on policies framed using attributes	Granted by re-encryption using a hash key, based on attributes
2	Access control	Created by direct encryption	Created by specified attributes
3	Computational process	Expensive based on file size	Only access rights are stored
4	User tracing	Anonymous	Can be traced

**Table 3** Results of the proposed model

S. no.	Parameter	On-premise	Cloud	Cloud and blockchain
1	Energy consumption	40% less efficient than cloud and blockchain and 50% less efficient than cloud	10% higher	40% More efficient than without cloud and 10% less efficient than cloud
2	Execution time	3 s	2 s	40% More efficient than without cloud and 10% less efficient than cloud
3	File size	16 MB	16 MB	16 MB
4	Memory utilization	90%	96%	Improved
5	CPU load	93%	95%	Improved

## 4 Results

The proposed methodology can be implemented by considering the minimum hardware requirements that can be considered are 4 GB RAM, 100 GB Storage, and Processor-i7 and if tested on different environments like cloud, cloud and Blockchain and On-Premise, the expected results can be achieved as shown in Table 3.

## 5 Conclusion

The proposed model allows Data Owners get the control to generate, share and manage patient's data with healthcare providers by storing access rights of Cloud stored data on blockchain. Storing the access policies on blockchain can increase the security and avoid unauthorized access. The paper emphasized on shortfalls of existing system that can be avoided on further implementation of the proposed model. The proposed Architecture for SHC data can improve the Efficiency, Execution time by lowering the storage and finally improving the overall data access.

## 6 Future Scope

In the future we tend to implement the model using private blockchain. As an alternative solution Hyperledger Fabric, which reduces the overhead involved in mining and that supports 10,000 transactions/second can be considered for implementing. We feel that a comprehensive study is still a need for defining the fine grained Access Control using a light weight Blockchain.

## References

1. Zyskind, G., Nathan, O., Pentland, A.: Decentralizing privacy: using blockchain to protect personal data. In: 2015 IEEE Security and Privacy Workshops, San Jose, CA, pp. 180–184 (2015). <https://doi.org/10.1109/SPW.2015.27>
2. Di Francesco Maesa, D., Mori, P., Ricci, L.: Blockchain based access control. Lecture Notes in Computer Science, pp. 206–220 (2017). [https://doi.org/10.1007/978-3-319-59665-5\\_15](https://doi.org/10.1007/978-3-319-59665-5_15)
3. Nakamoto, S.: Bitcoin: a peer-to-peer electronic cash system (2008)
4. Nguyen, Q.K.: Blockchain—a financial technology for future sustainable development. In: Green Technology and Sustainable Development (GTSD), pp. 51–54 (2016)
5. Pappalardo, G., Di Matteo, T., Caldarelli, G., et al.: Blockchain inefficiency in the Bitcoin peers network. EPJ Data Sci. **7**, 30 (2018). <https://doi.org/10.1140/epjds/s13688-018-0159-3>
6. Ajao, L.A., Agajo, J., Adedokun, E.A., Kargong, L.: Crypto Hash Algorithm-based blockchain technology for managing decentralized ledger database in oil and gas industry. J. **2**(3), 300–325 (2019). <https://doi.org/10.3390/j2030021>
7. Martínez, V.G., Hernández-Álvarez, L., Hernandez Encinas, L.: Analysis of the cryptographic tools for Blockchain and Bitcoin. Mathematics **8**, 131 (2020). <https://doi.org/10.3390/math8010131>
8. Wang, L., Shen, X., Li, J., Shao, J., Yang, Y.: Cryptographic primitives in blockchains. J. Netw. Comput. Appl. (2018). <https://doi.org/10.1016/j.jnca.2018.11.003>
9. Ismail, L., Materwala, H.: A review of blockchain architecture and consensus protocols: use cases, challenges, and solutions. Symmetry **11**, 1198 (2019). <https://doi.org/10.3390/sym1101198>
10. Lee, D., Park, N.: Blockchain based privacy preserving multimedia intelligent video surveillance using secure Merkle tree. Multimedia Tools Appl. (2020). <https://doi.org/10.1007/s11042-020-08776-y>
11. Gervais, A., Karame, G.O., Wüst, K., Glykantzis, V., Ritzdorf, H., Capkun, S.: On the security and performance of proof of work blockchains. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security—CCS’16 (2016). <https://doi.org/10.1145/2976749.2978341>
12. King, S., Nadal, S.: Ppcoint: peer-to-peer crypto-currency with proof-of-stake, self-published paper, pp. 1–6 (2012)
13. Kaufman, L.M.: Data security in the world of cloud computing. IEEE Secur. Priv. **7**(4), 61–64 (2009)
14. Zhang, Y., He, D., Choo, K.R.: BaDS: blockchain-based architecture for data sharing with ABS and CP-ABE in IoT. Wirel. Commun. Mob. Comput. **2018**, 9 (2018)
15. What is HIPAA, <https://www.dhcs.ca.gov/formsandpubs/laws/hipaa/Pages/1.00WhatisHIPAA.aspx>, 2018.
16. Dwyer III, S.J., Weaver, A.C., Knight Hughes, K.: Health insurance portability and accountability act. Security Issues in the Digital Medical Enterprise, Society for Computer Applications in Radiology, 2nd edn (2004)

17. Baird, A., North, F., Raghu, T.S.: Personal Health Records (PHR) and the future of the physician-patient relationship. In Proceedings of the 2011 iConference, New York, NY, USA, pp. 281–288 (2011)
18. Wangthammang, M., Vasupongayya, S.: Distributed storage design for encrypted personal health record data. In: Proceedings of the 8th International Conference on Knowledge and Smart Technology, KST 2016, Thailand, pp. 184–189 (2016)

# Dynamic Node Identification Management in Hadoop Cluster Using DNA



J. Balaraju and P. V. R. D. Prasada Rao

**Abstract** The distributed system is gambling a vital position in storing and processing big data and information era is speedily increasing from various resources every minute. Hadoop has a scalable, and efficient disbursed machine supporting commodity hardware with the aid of combining specific networks within the topographical locality. Node support inside the Hadoop cluster is unexpectedly growing in one-of-a-kind versions which can be dealing with issues to control clusters. Hadoop does no longer offers node management, including and deletion Node futures. Node identification in a cluster absolutely depends on DHCP servers that handles IP addresses; hostname is based totally at the physical address (MAC) address of every node. There is a scope for the hacker to theft the information using IP or hostname and developing a disturbance in a dispensed gadget by using adding a malicious node by assigning replica IP or hostname. This paper proposing novel node management for the disbursed machine the usage of DNA hiding and producing a completely unique key by way of combing a completely unique Physical address (MAC) of node and hostname. This mechanism is supplying higher node control for the Hadoop cluster offering adding and deletion node mechanism through the use of constrained computations and providing better node security from hackers. the main objective of this paper is to design a set of rules to put in force node touchy records hiding using DNA sequences and supplying safety to the node and its data from hackers.

## 1 Introduction

Distributed Computing (DC) [1] furnishes a sensible area work with productive execution of a reply on a number PCs related with a system. For conveyed Computing DC, large undertakings are partitioned into littler troubles which would then be capable to be carried out on more than a few PCs concurrently free of one another. In the route of current years, the intermixing of software program engineering and

---

J. Balaraju (✉) · P. V. R. D. Prasada Rao

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

the multifaceted nature of science has lead to the affluent subject of bioinformatics. Bioinformatics [2] is one of the greater up-to-date territories and has made us totally conscious of a definitely special universe of science. Deoxyribonucleic Acid (DNA) [3] association investigation can be a protracted cycle going from a few hours to several days. This paper manufactures a dispersed framework that offers the reply for some bioinformatics-associated applications. The widely widespread goal of this paper is to bring together a Distributed Bioinformatics Computing System for hereditary succession examination of DNA. To process, we put away a big no. of DNA succession using different lengths of DNA preparations have been utilized for the sequential and nonconsecutive instance search to suppose about the framework's response time received using single and several PCs. Furthermore, quite a number lengths of DNA preparations had been likewise utilized for the instance ID to assume about its response time watched utilizing a solitary PC and several PCs. The specific goal of the proposed disseminated calculation for investigation of DNA successions are (1) Build up a potential disseminated DNA association examination calculations for diagram coordinating of DNA Gene grouping and sub-successions recognizable proof. (2) Execute them on an about coupled appropriated system, for example, widely wide-spread local and broad territory prepare utilizing fashionable programming language. The principal focus of this paper is to propose a calculation to actualize information covering up in DNA groupings to expand the multifaceted nature and making disarray to the programmers.

## 2 Hadoop Cluster

Hadoop [4] is an Apache open supply machine written in java that lets in dispersed making ready of massive datasets throughout bunches of PCs utilizing easy programming models. The Hadoop shape utility works in a scenario that offers dispersed potential and calculation throughout businesses of PCs. Hadoop is meant to scale up from single employee to a top-notch many machines, every supplying nearby calculation and capacity. In small Hadoop Cluster (HC) [5] have a solitary ace Node Server and a range of patron hubs. In a larger HC, HDFS hubs are overseen via a dedicated NameNode employee to have the record framework record, and an optionally available NameNode that can produce depictions of the namenode's reminiscence structures, as a consequence forestalling record framework debasement and loss of data. The dimension of HC's are shortly extended from 2005, the inventors is constrained to simply 20 to forty nodes in a groups. At that factor, they understood two issues, they are no longer accomplish its manageable till it ran dependably on the better groups. In the second stage, Yahoo efficaciously tried Hadoop on a one thousand hub bunch and start using it later yippee and Apache Software Foundation correctly tried a 4000 hub crew with Hadoop. HC bunch improved 4000–10,000+ in quite a number of deliveries.

### 3 DHCP

DHCP [6] is a conference that offers snappy, programmed, and focal administration for the dispersion of IP addresses [7] internal a system. DHCP is moreover used to organize the subnet cover, default entryway, and DNS employee statistics on the device. Media Access Control (MAC) Address is a novel identifier of the Network Interface Controller (NIC). A gadget hub can have distinctive NIC but every with novel MAC. A device chairman saves a scope of IP addresses for DHCP, and each DHCP patron on the LAN is organized to demand an IP tackle from the DHCP employee at some stage in gadget introduction. The solicitation and-award measure makes use of a lease thinking with a controllable timeframe, allowing the DHCP employee to get better and later on reallocate IP tends to that are now not recharged. The DHCP employee continually appoints an IP tackle to a bringing up purchaser from the vary characterized via the executive. The DHCP employee offers a personal IP address subordinate upon each's consumer identity established on predefined planning via the overseer.

### 4 DNA

Deoxyribonucleic Acid (DNA) cryptography [8], the utilization of the characteristics of DNA affords new probabilities and heading to statistics stowing away. This work will use the herbal traits of DNA successions. The units of DNA capacity, for example, imperative blending, and DNA document provide any other layer of safety to the proposed technique. So as to tightly close refined statistics thru unstable structures like the Internet, utilizing special kinds of facts insurance plan is fundamental. One of the famous strategies to invulnerable data thru the Internet is data protecting up. In view of the increasing quantity of Internet clients, the use of data stowing away or Stenographic strategies [9] is unavoidable. Before utilizing natural houses of DNA arrangements, generally implanting a thriller message into the host pics used to be the traditional technique of facts protecting up. The most enormous ones had been the attention of the mutilations of the photograph when the host image was modified to positive degrees. The key bit of their work is, the usage of natural attributes of DNA arrangements.

### 5 Literature Survey

**Alabady et al.** [10] is carried out a Network Security Model for Cooperative Network delivered a machine protection model. The creator has examined weaknesses, dangers, assaults, association shortcomings, and protection methods with device assurance.

**Balaraju et al.** [11] have examined large information advances and their advances for improved massive information. Information safety is a chief difficulty in the administration part, science, exploration, and commercial enterprise ventures. They likewise examined statistics stockpiling, handling, and safety territory and find out the challenges via making use of normal protection units for Hadoop. They encouraged a solitary DNA-based totally impenetrable hub for verification and metadata of the board for Hadoop which is the high-quality reply for strengthening statistics and dispose of NNSE blocks for protection metadata in the Namenode in Hadoop.

**Bin Ali et al.** [12] developed a Secure Campus Network have configured. The developed hierarchical structure of the campus community thinking about one-of-a-kind sorts of protection problems that make sure the exceptional of service.

**Pandya et al.** [13] are developed a Network Structure and mentioned 5 primary community topologies like Bus, ring, Star, Tree, and Mesh.

**Balaraju et al.** [14] are constructed up an Algorithm Built-in Authentication Based on Access (BABA) as a safety incidence coordinated as Hadoop hub for making positive about statistics in HDFS and straight away metadata safety for evading customers statistics in Hadoop. The instrument contributes a made certain about Hadoop Cluster barring making use of different safety sport plans which likewise lessens operational cost, calculations, expands data security, and giving steady safety solutions for Hadoop Cluster. The enhancement of this work is to reduce the computational weights of the proposed calculation.

**Kennedy et al.** [15] are carried out a Structured Network for a Small system. Creators have reenacted prepare configuration making use of the Cisco Packet Tracer programming and Wire Shark conference analyzer.

**Balaraju et al.** [16] had accomplished a solely new protection component, a Secure Authentication Interface (SAI) layer over the Hadoop Cluster. As a Single Security convention, this interface offers consumer verification, metadata security, and get admission to control. Contrasted the modern instruments, SAI can provide safety a much less computational weight. Creators focused on protection challenges and tended to for making certain about Big Data in a Hadoop Cluster thru a solitary, restrictive protection gadget known as Secure Authentication Interface. SAI made a confided in situation inner HC via confirming the two customers and their cycles.

All the above papers that are surveyed have proposed a number parts of allotted gadget structure, geographies, and execution but they have now not examined troubles seemed in possible usage. Numerous Authors are concentrating a made certain about the appropriation framework in Hadoop and finally which is treasured to data in the Hadoop crew have now not examined DHCP and MAC officers in element.

## 6 Problem Statement

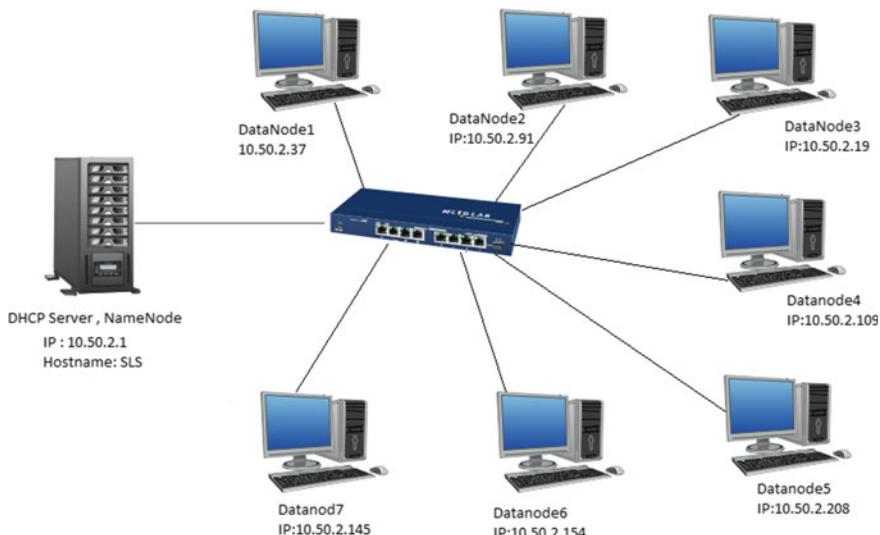
In any distribution system including Hadoop distribution structures linked a centralized community switch. each node in the network has a unique IP cope with this is dynamically assigned through the DHCP server by accumulating the bodily cope

with every related node and assigning static IP each node is hard for administrator in larger networks. IP cope with and hostname can seem any consumer working in a dispensed device and they'll access any node with suitable permissions. The problem with acting IP and hostname, the hacker may additionally vicinity reproduction hostname and replica IP deal with for disturbing or malfunctioning network by using placing a malicious node within the community. On the grounds that there may be scope dropping essential facts from a dispensed system by using IP cope with and it could be a safety hazard.

## 7 Related Work

The present invention relates to distributed computing systems and is more particularly directed to architecture and implementation of a scalable distributed computing environment which facilitates communication between independently operating nodes on a single network.

The primary objective of the research work is to create a layer on top of the distributed system especially for the Hadoop distributed system, so every node appears in the layer. For setting the environment we configured a centralized DHCP server and 250 nodes in the network with a different configuration. Multinode HC [17] configured, the master node configured within the DHCP server as Namenode,



**Fig. 1** Existing Hadoop cluster by configuring DHCP server

and the remaining nodes are data nodes. Figure 1 shows default distribution environment. The developed security layer is also configured within the DHCP Server for collecting the hostname and IP addresses which are stored itself.

The DNA algorithm within the proposed layer is converting the IP address, the hostname in different level and producing a unique key which appears for the user including the physical address (MAC) of NIC. The key generation is generating by using a highly secured DNA hiding [10] methodology for creating confusion hackers to access any node from the network and finally it becomes a highly secured distributed system.

### **Key generating Process:**

#### ***Start:***

1. ***Gathering Host, MAC, IP.***
2. ***Merging Host, MAC, IP as UNIQ\_Key.***
3. ***Translating UNIQ\_Key into BINARY Form.***
4. ***Translating Binary form to DNA.***
5. ***Assign a digit to DNA into a Number // a= 0, c= 1, g= 2, t= 3.***
6. ***Translating Number to Hexadecimal.***
7. ***Making UNIQ\_Key from Hexadecimal.***

#### ***Stop.***

Table 1 is displaying special key technology processes in extraordinary ranges for growing confusion to the hackers. The hostname of the node is eight characters, it is dynamically assigning from generated special key and altering every 7 days, it can be up to date robotically central table. Table 2 is displaying the hostname and fame of node information. So, the hackers get confusions to get admission to a unique node the usage of hostname from the dispensed system. The essential gain of the proposed technique is no longer having everlasting hostname to get entry to node and it shared data.

The secured layer additionally includes nodes facts with the aid of preserving central desk alongside hostname. Internally each and every node linked with different nodes with the aid of performing hostname managed via the protection layer. All these hostnames in the community are maintained by means of a invulnerable layer which include the storage status, processing configuration by way of periodically updating in the central table. Figure 2 is showings the impenetrable layer within DHCP, Namenode server alongside with special hostnames for every node. The impervious layer is updating the central desk when a new node is introduced or putting off of a node from the allotted system.

## **8 Result Analysis**

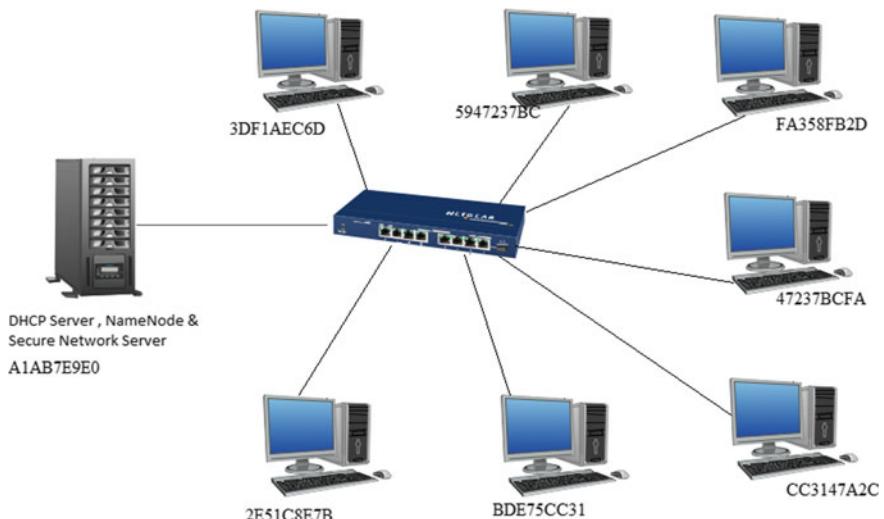
Overall performance analysis is analyzed with small disbursed system with proposed comfortable layer all the prevailing security techniques aren't concentrated to cover

**Table 1** Procedure for hiding system information using DNA

Steps	Description	MAC Address	IP Address	Hostname
01	Node sensitive data	10:78:D2:55:95:A8		
02	Merged data	10:78:D2:55:95:A8-10:50:4:8- HD_DN01		
03	Binary form	00,110,001,00,110,000,00,111,010,00,110, 111,00,111,000,00,111,010,01,000,100,00,110, 010,00,111,010,00,110,010,00,110,01,00,111,010,00, 111,001,00,110,010,00,110,001,010,01,000,00,100, 111,000,00,101,010,00,110,001,00,110,000,00,101, 110,00,110,01,00,110,000,00,101,010,00,110,00,100,00, 101,110,00,111,000,00,101,101,00,100,000,01,001,000,01, 000,100,01,011,111,01,000,100,01,001,110,00,110,000,00,110,000,01,	10.50.4.8	HD_DN01
04	DNA form	ATACATAAATGCATGTATGAAATGCCACAATAGATGGATCCATTCATGGATGCAATCCAT GGAACATGAAGTCAATACATAAAGAGATCCATAAAAGAGATCAAGAGATGAAGACAGAA CAGACACACCTTCACACATGATAAAATAC		
05	A = 0, C = 1, G = 2, T = 3	0301 0300 0321 0323 0320 0322 1010 0302 0322 0311 0311 0322 0321 0311 0322 1001 0320 0231 0301 0300 0202 0311 0300 0202 0310 0202 0320 0201 0200 1020 1010 1133 1010 1030 0300 0301		
06	Decimal to Hexa	5CC3148053C4906F901A54B4DE8225FD8071B87BA02E51C8E7B2C2BDE75CC314 7A2CCCD09B85CEA1AB7E9E03DF1AEC6D5947237BCFA358FB2DED		
05	Unique Key	5CC3148053C4906F901A54B4DE8225FD8071B87BA02E51C8E7B2C2BDE75CC3147A2CCC D09B85CEA1AB7E9E03DF1AEC6D5947237BCFA358FB2DED		

**Table 2** Nodes status information in central server

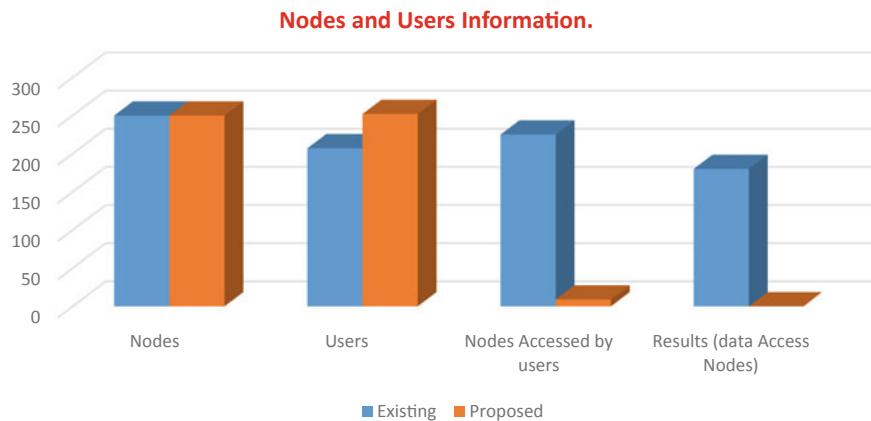
MAC	Node_Hostname	Status	Joined/removed
A4-1F-72-58-BB-01	A1AB7E9E0	Active	14-MAR-2020
F5-10-72-58-BB-01	3DF1AEC6D	Active	14-MAR-2020
C4-1F-B2-58-BB-01	5947237BC	Active	14-MAR-2020
A4-1F-72-58-BB-01	FA358FB2D	Removed	23-JUN-2020
A4-1F-72-58-BB-01	47237BCFA	Active	14-MAR-2020
A4-1F-72-58-BB-01	CC3147A2C	Active	14-MAR-2020
A4-1F-72-58-BB-01	BDE75CC31	Active	14-MAR-2020
A4-1F-72-58-BB-01	2E51C8E7B	Removed	05-JUL-2020

**Fig. 2** Proposed Hadoop cluster by dynamic hostnames

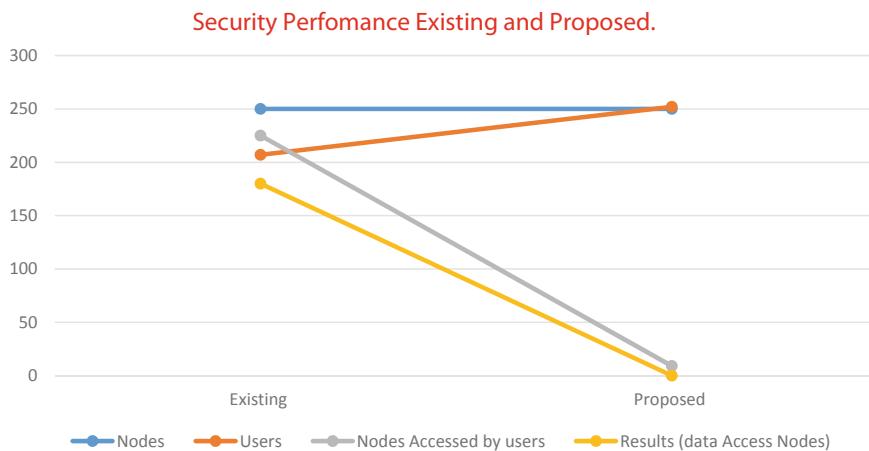
nodes data. Table 3 is showing the nodes user get right of entry to information in configured environment. Parent three is about assessment of nodes and person facts in present and evolved gadget (Figs. 3 and 4).

**Table 3** Nodes and users information

	Existing	Proposed (Security layer)
Nodes	250	250
Users	207	252
Nodes accessed	225	9
Results (data access nodes)	180	0



**Fig. 3** Nodes and user data existing and proposed



**Fig. 4** Performance evolution of existing and proposed

This safety layer possibly offers  $24 \times 7$  securities for HC which may be very beneficial for small allotted system to keep their statistics securely. Determine 4 is indicates the overall performance extended safety for above-configured surroundings. This will increase the facts protection, unique operational, and decrease preservation hassle.

## 9 Conclusion

The research studies of this work, we've got carried out a secured distributing system with the help of DHCP server, IP, Host, and MAC aggregate. The unique allotted machine network allotted a dedicated unique ID to each node for securing community at the side of information. the administrator best has entire privileges to get entry to all of the nodes such as the server and the others cannot get right of entry to any node in the network without knowing IP or Host. The performance of the community management and distributed System is improved in the long run data security is superior. The complete community is deliberate to be automatic which entails minimal person intervention. The destiny scope of this work is to beautify the same security layer to use to the wan community placed in special places.

## References

1. Yang, H., Lee, J.: Secure distributed computing with straggling servers using polynomial codes. *IEEE Trans. Inf. Forensics Secur.* **14**(1), 141–150 (2019). <https://doi.org/10.1109/TIFS.2018.2846601>
2. Khanan, A., Abdullah, S., Mohamed, A.H.H.M., Mehmood, A., Ariffin, K.A.Z.: Big data security and privacy concerns: a review. In: Al-Masri, A., Curran, K. (eds.) Smart technologies and innovation for a sustainable future. Advances in Science, Technology & Innovation (IEREK Interdisciplinary Series for Sustainable Development). Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-01659-3\\_8](https://doi.org/10.1007/978-3-030-01659-3_8)
3. Roy, M., et al.: Data security techniques based on DNA encryption. In: Chakraborty, M., Chakrabarti, S., Balas, V. (eds.) Proceedings of International Ethical Hacking Conference 2019. eHaCON 2019. Advances in Intelligent Systems and Computing, vol. 1065. Springer, Singapore (2020). [https://doi.org/10.1007/978-981-15-0361-0\\_19](https://doi.org/10.1007/978-981-15-0361-0_19)
4. Samet, R., Aydin, A., Toy, F.: Big data security problem based on Hadoop framework. In: 2019 4th International Conference on Computer Science and Engineering (UBMK), Samsun, Turkey, pp. 1–6 (2019). <https://doi.org/10.1109/UBMK.2019.8907074>
5. Akhgarnush, E., Broeckers, L., Jakoby, T.: Hadoop: A standard framework for computer cluster. In: Liermann, V., Stegmann, C. (eds.) The Impact of Digital Transformation and FinTech on the Finance Professional. Palgrave Macmillan, Cham (2019). [https://doi.org/10.1007/978-3-030-23719-6\\_18](https://doi.org/10.1007/978-3-030-23719-6_18)
6. Rajput, A.K., Tewani, R., Dubey, A.: The helping protocol “DHCP”. In: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, pp. 634–637 (2016)
7. Mohsin, M., Prakash, R.: IP address assignment in a mobile ad hoc network. In: MILCOM 2002. Proceedings, Anaheim, CA, USA, vol. 2, pp. 856–861 (2002). <https://doi.org/10.1109/MILCOM.2002.1179586>
8. Sajisha, K.S., Mathew, S.: An encryption based on DNA cryptography and steganography. In: 2017 International Conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, pp. 162–167 (2017). <https://doi.org/10.1109/ICECA.2017.8212786>
9. Kar, N., Mandal, K., Bhattacharya, B.: Improved Chaos-based video steganography using DNA alphabets. *ICT Express* **4**(1), 6–13 (2018). ISSN 2405-9595. <https://doi.org/10.1016/j.icte.2018.01.003>
10. Alabady, S.: Design and implementation of a network security model for cooperative network. *Int. Arab J. Technol.* **1**(2) (2009)

11. Balaraju, J., Rao, P.V.V.P.: Recent advances in big data storage and security schemas of HDFS: a survey (2018)
12. Bin Ali, M.N., Hossain, M.E., Masud Parvez, Md.: Design and implementation of a secure campus network. *Int. J. Emerg. Technol. Adv. Eng.* Website **5**(7) (2015). [www.ijetae.com](http://www.ijetae.com) (ISSN 2250-2459. ISO 9001:2008 Certified Journal)
13. Pandya, K.: Network structure or topology. *Int. J. Adv. Res. Comput. Sci. Manage. Stud.* **1**(2), 6 (2013)
14. Balaraju, J., Prasada Rao, P.V.R.D.: Designing authentication for Hadoop cluster using DNA algorithm. *Int. J. Recent. Technol. Eng. (IJRTE)* **8**(3) (2019). ISSN: 2277-3878. <https://doi.org/10.35940/ijrte.C5895.0983>
15. Offor, Kennedy, J., Obi, Patrick, I., Nwadike Kenny, T., Okonkwo II: *Int. J. Eng. Res. Technol. (IJERT)* **2**(8) (2013)
16. Balaraju, J., Rao, P.: Innovative secure authentication interface for Hadoop cluster using DNA cryptography: a practical study (2020).[https://doi.org/10.1007/978-981-15-2475-2\\_3](https://doi.org/10.1007/978-981-15-2475-2_3)
17. Gugnani, S., Khanolkar, D., Bihany, T., Khadilkar, N.: Rule based classification on a multi node scalable Hadoop cluster. In: Fortino, G., Di Fatta, G., Li, W., Ochoa, S., Cuzzocrea, A., Pathan, M. (eds.) *Internet and distributed computing systems. IDCS 2014. Lecture Notes in Computer Science*, vol. 8729. Springer, Cham (2014). [https://doi.org/10.1007/978-3-319-11692-1\\_15](https://doi.org/10.1007/978-3-319-11692-1_15)
18. Demidov, V.V.: Hiding and storing messages and data in DNA. In: *DNA beyond Genes*. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-36434-2\\_2](https://doi.org/10.1007/978-3-030-36434-2_2)

# A Scientometric Inspection of Research Based on WordNet Lexical During 1995–2019



Minni Jain, Gaurav Sharma, and Amita Jain

**Abstract** The purpose of this study is to conduct a scientometric inspection for the research based on WordNet lexicon. WordNet is a lexical database for the English Language developed by George A. Miller and his team at Princeton University in 1985. This study reviews WordNet-based research mainly published in the Web of Science (WoS) database from 1995 to 2019. The publication data has been analyzed computationally to present year-wise publications, publications growth rate, country-wise research publications, authors who published a large number of papers in this field. The results have been shown in the form of figures. The present study will be useful for a better understanding of the trajectory of the research work done in the field of WordNet during the last 25 years and shall highlight the prominent research areas and categories in the field of WordNet.

## 1 Introduction

WordNet is a lexical resource for the English language. It groups English phrases into units of synonyms known as synsets. WordNet has been widely used in the field of Natural Language Processing (NLP), Artificial Intelligence (AI), Text Analysis, Machine Translations, Information Retrieval, etc. in the recent past [1]. A common use of WordNet is to determine the similarities between words. It was developed in the Princeton University under the direction of prof. George A. Miller in 1985. Since then, a lot of research has been done to strengthen the WordNet using semantic relations between different components. It is a combination of dictionary and thesaurus

---

M. Jain · G. Sharma (✉)

Department of Computer Science and Engineering, Delhi Technological University, Delhi 110042, India

M. Jain

e-mail: [minnijain@dtu.ac.in](mailto:minnijain@dtu.ac.in)

A. Jain

Department of Computer Science and Engineering, Ambedkar Institute of Advanced Communication Technologies and Research, Delhi 110031, India

because unlike dictionary it also found the meaningful relations between words. WordNet allows the different lexical categories such as noun, verbs, adverbs, and adjectives however it ignores the propositions and determiners [2]. It will generate the valid meaningful sentence for the use of a set of synsets to link with the semantic relations. WordNet includes mainly two terms word forms and sense pairs. Word forms is a string over the finite alphabets and sense pair is the set of meanings. This word forms with a particular sense is known as “word” in the English language [1]. So more than 166,000 word forms and sense pairs are used in the WordNet [3].

Semantic relations are the relations between meanings or relations between two sentences or relations between two words [4]. Some semantics relations are given below:

- Synonymy is the set of synonyms (synsets) to represent word senses.
- Antonymy is the set of opposite names of word forms.
- Hyponymy is the inverse of hypernymy (super-name), so the hyponymy is (sub-name). It is a transitive relation.
- Meronymy is the inverse of holonymy (whole-name), so the meronymy is the (part-name). It shows the member parts.
- Troponymy (manner-name) is used for verbs. Entailment is showing the relations between verbs.

These semantics relations are represented by pointers in between word forms or synsets. More than 116,000 pointers are used in WordNet. Semantic relations in WordNet is an open-source and freely available for everyone on the Internet. WordNet 3.1<sup>1</sup> is the latest version. The increasing growth of research work on WordNet since the period of inception inspires us to analyze the progress of research work on WordNet. In this paper, a scientometric analysis has been made computationally to highlight the role of WordNet in the development of language. The present study shall be helpful to understand the trajectory of the research work done on WordNet during a large span that is from 1995 to 2019. The resource data on the WordNet research publications has been taken from the Web of Science (WoS). We have analyzed the data so, obtained from the Web of Science (WoS) computationally to identify the year-wise research publications on WordNet and represented graphically for better understanding. The WordNet research publications data have also been analyzed to identify author-wise publications.

This information we are representing in the form of graphs for better understanding. Thus, the aim of the present study is to provide the analytic account of the progress of the research work for example the major areas of research work, major concepts, and major approaches to research on WordNet during the period of time 1995–2019 [5–19]. This study shall be useful to understand how and why the research work on WordNet has grown over time and which countries, organizations/Institutions, authors have contributed more substantially. It will also help to understand the major potential research areas and applications of WordNet. Section 2

---

<sup>1</sup><https://wordnetweb.princeton.edu/perl/webwn>.

describes the data and methodology; Sect. 3 highlights the results of the study with corresponding visualizations, and conclusions are given in Sect. 4.

## 2 Data and Methodology

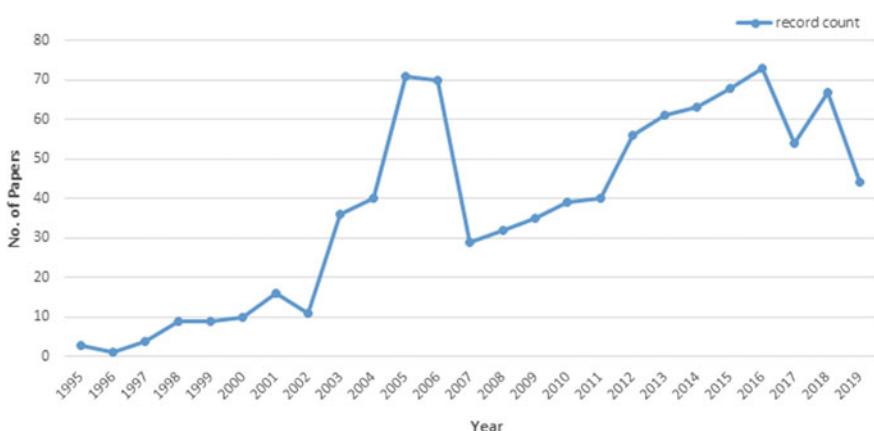
For the scientometric analysis of WordNet the data is retrieved from Web of Science (WoS). The total number of papers retrieved is 947 but out of these 104 papers are filtered. The collected data was based on the time span 1995–2019 [5–19]. Web of Science (WoS) is the large database for the different documents, articles, reviews, proceeding papers, editorial materials, etc. [5]. So, the study in this paper is about 25 years of work on WordNet to be analyzed using tables and graphical representations.

## 3 Analytical Study

This section describes the details of various important indicators computed through the analysis of data.

### 3.1 Year-Wise Publications

First of all, we have measured the number of published papers on WordNet for each of the years 1995–2019. Figure 1 shows the number of published papers in WordNet



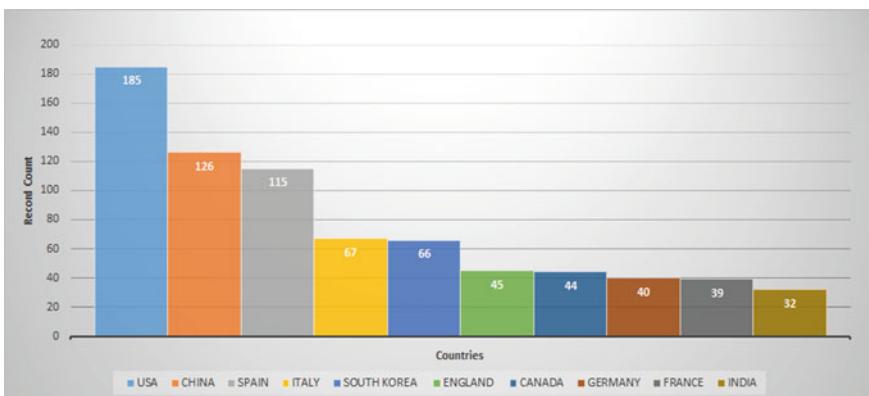
**Fig. 1** Year-wise publications graph

on a year-wise plot. We have a column about NOP (No. of publications) and total citations and average citations per year. We can observe that this graph has been more or less flat till 2002, after which there is a steep rise. From 2006 to 2007, there is a drastic fall in the number of papers in this time period. After that, the graph is consistently increasing and slightly up and down was happened in between year by year.

### 3.2 Country-Wise Publications

We have analyzed the country-wise contributions of WordNet research during 25 years of the period from 1995 to 2019. The research paper publications can be seen in the terms of record count in Web of Science (WoS). The topmost countries according to the research paper record count, as listed above are illustrated in the graph in Fig. 2. The graph is explained by different colors, these colors are indicating the “*Countries*”, and the values are “*Record Count*”. The highest number of published papers or record count about WordNet is on the account of the USA.

After that many countries come, China is the second topmost country who has a high number of published papers on WordNet. Spain is just comparatively low than China. After that Spain, there is a huge difference between the other countries. The other major countries from where WordNet work is reported to include Italy, South Korea, England, Canada, Germany, France, India, etc. These are the only top 10 countries which we included in this paper.



**Fig. 2** Country-wise publication graph

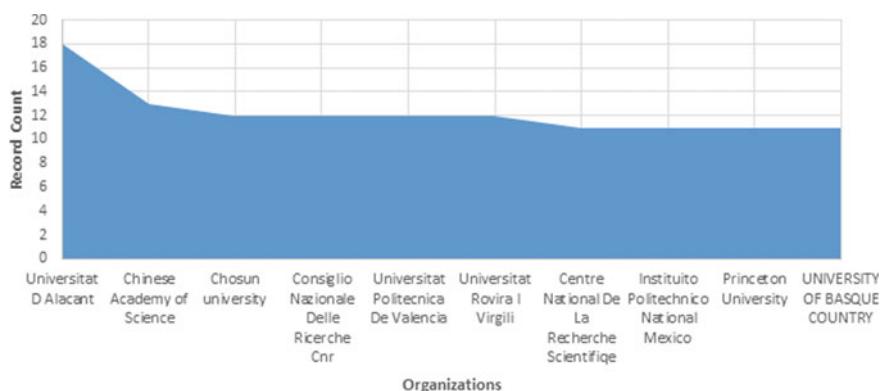
### 3.3 Top Organization Publications

After analyzing WordNet research publications to extract the county-level results, we tried to understand the organization-level research output. First, we have to identify the top organization's contributions significantly to WordNet work during the 1995–2019 period. We list the top 10 organizations in descending order of the number of publications originating from them. The table indicates four different indicators, namely NOP (Number of Papers), TC (Total Citations), ACPP (Average Citations Per Paper) and h-index for the WordNet research output originating from various organizations. The values of NOP, TC, ACPP, and h-index are computed from the Web of Science (WoS). ACPP values is calculated as:

$$\text{ACPP} = \frac{\text{TC}}{\text{NOP}} \quad (1)$$

The h-index metric is the author-level metric which is measured both the productivity and the impact of the citations published work of a scientist or scholar. The h-index is calculated for individuals, institutions, journals, etc. We can observe that the Universitat D Alacant has the highest number of research papers on the topic of WordNet. This is following by the Chinese Academy of Science which is in the second position by few numbers of papers (Fig. 3).

We have selected only the top 10 organizations, in which Universitat D Alacant having the largest number of papers (18 papers) after that Chinese Academy of Science has the second-highest papers (13 papers) there have 4 organizations [Chosun university, Consiglio Nazionale Delle Ricerche Cnr, Universitat Politecnica De Valencia, Universitat Rovira I Virgili] who have the same number of papers (12 papers) Now the last 4 organizations [Centre National De La Recherche Scientifique, Instituto Politecnico National Mexico, Princeton University, University of Basque Country] also have the same number of papers (11 papers).



**Fig. 3** Area graph for organization-wise publications



**Fig. 4** Treemap of author-wise publications

### 3.4 Author-Wise Publications

We have also analyzed the WordNet research publication data to identify the most productive and most cited authors. We are defining here highly productive authors who produce a high amount of research papers published during the period of 1995–2019. We present the list of top 15 most productive authors, in this table we indicate the four indicators again are NOP (Number of papers), TC (Total Citations), ACPP (Average Citations Per Paper), and h-index. We can observe that Kim P is the most productive author on WordNet research during 1995–2019. He published 12 papers on WordNet, his total citations are 88, his ACPP is 7.33 and h-index is 6. This is followed by the Rosso P he published 11 papers, his total citations are 77, ACPP is 7, and h-index is 4. But if the most productive author is also the most cited author is not compulsory. Sanchez D. published 10 papers, his total citations are 488, ACPP is 48.8, and h-index is 10. According to the Web of Science (WoS) using this table the most productive author is Kim P. with 12 papers published on WordNet, highest total citations having author is Weikum G. with 521 citations, the most average citations per paper author is Weikum G. with 74.43 and the highest h-index author is Sanchez D. with h-index 10. Here for the better understanding, we make the treemap chart in Fig. 4, which shows the top 15 most productive authors with their NOP.

## 4 Conclusions

The paper presents an analytical study in the field of WordNet research. We analyzed the research publications on WordNet research and extract the data from Web of Science (WoS) in the time span 1995–2019. The analytical study helped to identify the year-wise publications, country-wise contributions, top organization publications,

and top author-wise publications. In this paper, we used the figures which result for better understanding of the research work done in the field of WordNet. From the figures, we conclude that the USA, China, and Spain are the most productive countries with the record count of more than 100 in the field of WordNet research while India has been placed on number 10 with 32 record count. It representing the year-wise publications that the highest number of research papers have been published in the year 2016. But the remarkable growth in the research publications has been found from 2003. It appears that extensive use of WordNet as a tool has attracted computer science researchers and software developers. Author-wise treemap shows that the authors Kim P., Rosso P., and Sachez D. are the only authors who published the research papers in double digits. Universitat D Alacant is the topmost organization who published the highest number of research papers in the field of WordNet while Chinese Academy of Science and Chosun University are the topmost second and third organizations in the production of research papers on WordNet. organization of the WordNet is at the 9th place in the list of top 10 organizations of the research on WordNet. The analysis has predicted to provide an analytical study to researchers working in this domain in exploring the discipline. This paper will also be helpful for the beginners who are new in this field and wants to do research in the WordNet. In the upcoming time, the country-wise contributions, year-wise publications, and author-wise publication will change when the research will continue in this field with time because this paper analyzed only for 25 years (1995–2019) of time span.

## References

1. Miller, G.A.: WordNet: a lexical database for English. *Commun. ACM* **38**, 39–41 (1995). <https://doi.org/10.1145/219717.219748>
2. Miller, G.A., Hristea, F.: WordNet nouns: classes and instances. *J. Comput. Linguist.* **32**, 1–3 (2006). <https://doi.org/10.1162/coli.2006.32.1.1>
3. Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: WordNet: an on-line lexical database. *Int. J. Lexicogr.* **3**, 235–312 (1993)
4. Wei, T., Lu, Y., Chang, H., Zhou, Q., Bao, X.: A Semantic approach for text clustering using WordNet and lexical chains. *Expert Syst. Appl. Int. J.* **42**, 2264–2275 (2015). <https://doi.org/10.1016/j.eswa.2014.10.023>
5. Mah, T., Ben Aouicha, M., Ben Hamadou, A.: A new semantics relatedness measurement using WordNet features. *J. Knowl. Inf. Syst.* **41**, 467–497 (2014). <https://doi.org/10.1007/s10115-013-0672-4>
6. Montoyo, A., Palomar, M., Rigau, G.: Interface for WordNet enrichment with classification systems. *Database Expert Syst. Appl.* **2113**, 122–130 (2001)
7. Gomes, P., Pereira, F.C., Paiva, P., Seco, N., Carreiro, P., Ferreira, J.L., Bento, C.: Advances in Artificial Intelligence, vol. 2671, pp. 537–543 (2003)
8. Miller, G.A., Fellbaum, C.: WordNet then and now. *Lang. Resour. Eval.* **41**, 209–2214 (2007). <https://doi.org/10.1007/s10579-007-9044-6>
9. Christiane Fellbaum Vossen, P.: Challenges for multilingual WordNet. *Lang. Resour. Eval.* **46**, 313–326 (2012). <https://doi.org/10.1007/s10579-012-9186-z>
10. Otegi, A., Arregi, X., Ansa, O., Agirre, E.: Using knowledge based relatedness for information retrieval. *Knowl. Inf. Syst.* **44**, 689–718 (2015). <https://doi.org/10.1007/s10115-014-0785-4>

11. Gomes, P., Pereira, F.C., Paiva, P., Seco, N., Carriero, P., Ferreira, J.L., Bento, C.: Noun sense disambiguation with WordNet for software design retrieval. *Advances in Artificial Intelligence*, vol. 2671, pp. 537–543 (2003)
12. Sonakshi, V.I.J., Tayal, D., Jain, A.: A machine learning approach for automated evaluation of short answers using text similarity based on WordNet graphs. *Wirel. Pers. Commun.* (2019). <https://doi.org/10.1007/s11277-019-06913-x>
13. Rudnicka, E., Piasecki, M., Bond, F., Grabowski, L., Piotrowski, T.: Sense equivalence in PLWordNet to Princeton WordNet mapping. *Int. J. Lexicogr.* **32**, 296–325 (2019). <https://doi.org/10.1093/ijl/ecz004>
14. Jiang, Y.C., Yang, M.X., Qu, R.: Semantic similarity measures for formal concept analysis using linked data and WordNet. *Multimedia Tools Appl.* **78**, 19807–19837 (2019). <https://doi.org/10.1007/s11042-019-7150-2>
15. Zhu, N.F., Wang, S.Y., He, J.S., Tang, D., He, P., Zhang, Y.Q.: On the suitability of WordNet to privacy management. *Wirel. Pers. Commun.* **103**, 359–378 (2018). <https://doi.org/10.1007/s11277-018-5447-5>
16. Cai, Y.Y., Zhang, Q.C., Lu, W., Che, X.P.: A hybrid approach for measuring semantic similarity based on IC-weighted path distance in WordNet. *J. Intel. Inf. Syst.* **51**, 23–47 (2018). <https://doi.org/10.1007/s10844-017-0479-y>
17. Ehsani, R., Solak, E., Yıldız, O.T.: Constructing a WordNet for Turkish using manual and automatic annotation. *ACM Translations on Asian and Low-Resource Language Information Processing*, vol. 17 (2018). <https://doi.org/10.1145/3185664>
18. Guinovert, X.G., Portela, M.A.S.: Building the Galician WordNet: methods and applications. *Lang. Resour. Eval.* **52**, 317–339 (2018). <https://doi.org/10.1007/s10579-017-9408-5>
19. Vij, S., Jain, A., Tayal, D., Castillo, O.: Fuzzy logic for inculcating significance of semantic relations in word sense disambiguation using a WordNet graph. *Int. J. Fuzzy Syst.* **20**, 444–459 (2018). <https://doi.org/10.1007/s40815-017-0433-8>

# Sentiment Analysis of an Online Sentiment with Text and Slang Using Lexicon Approach



Shelley Gupta, Shubhangi Bisht, and Shirin Gupta

**Abstract** Sentiment analysis is a technique that helps data analysts to analyze the various online users' opinions about a particular product or service. There are various approaches for performing sentiment analysis. Lexicon-based approach uses sentiment lexicons to calculate the polarity of the sentences. It is observed that nowadays online lexicons consist of text and online slang as well. The proposed approach calculates the polarity of the sentence by evaluating a polarity score of sentiments with text and slang. The sentiment polarity of tweets is also evaluated using the machine learning classification techniques like SVM, Random Forest, and Linear regression with accuracy, recall, precision, and F-score parameters. The accuracy of the proposed approach has been evaluated to 96% for Twitter dataset containing 17,452 tweets and 97% for other social media sentiments.

## 1 Introduction

Sentiment analysis is an excellent way to get to know the reaction and feelings of people (particularly consumers) about any product, topic or idea [1, 4] expressed on review sites, e-commerce sites, online opinion sites, or social media like Facebook, Twitter, etc. Sentiment analysis is employed using three broad techniques: Lexical based [3, 9, 10, 17], machine learning based [9, 10, 17] and hybrid/combined [2].

Lexicon-based analysis is governed by matching the new tokens with pre-defined dictionaries that have existing tagged lexicons as positive, negative, or neutral [11] polarity and score. Machine Learning approach as stated in [2, 9, 17] is the most popular approach for sentiment analysis due to its accuracy and ease of adaptability. Generally labeled and sizable datasets are used which requires human annotators that are expensive and time-consuming. Hybrid/combined [2] approach combines

---

S. Gupta (✉) · S. Bisht · S. Gupta  
ABES Engineering College, Ghaziabad, Uttar Pradesh, India  
e-mail: [shelley.gupta@abes.ac.in](mailto:shelley.gupta@abes.ac.in)

S. Bisht  
e-mail: [shubhangi.16bit1027@abes.ac.in](mailto:shubhangi.16bit1027@abes.ac.in)

the accuracy of machine learning approach with the speed of lexicon-based approach with the aim to make it more accurate.

The use of slangs for expressing one's opinion saves time and space within the word limit using slangs and abbreviations. Most of the existing sentiment analysis approaches remove the slangs during the pre-processing of the dataset. Thus, removing the impact of online slangs in sentiment score evaluation. Although our belief is inclusion of slang in determining the polarity and score of sentiment at sentence and document level, can enhance the sentiment score of different approaches.

Our distinct contributions for the paper can be summarized as follows: (i) Our approach is a rule-based approach that deals with slang by replacing their meaning in an online sentiment and then calculating the sentiment score at the sentence level. (ii) The approach determines the average score of the total number of positive and negative sentences to evaluate the sentiment score and polarity of an online document. (iii) Our approach outperforms the existing approaches as it calculates the sentiment score using online slang dictionary at sentence and document level.

The paper is organized as follows: Sect. 2 describes the literature review to Sentiment analysis and the background of this work. Section 3 demonstrates the proposed approach. Section 4 demonstrates the score calculation and result evaluation. Sections 5 and 6 represent the conclusion and limitations with future work.

## 2 Literature Review

A large proportion of the population uses the fast medium of English vocabulary of slang and abbreviations while expressing their emotions on a social media platform. Very few existing sentiment analysis techniques have considered commonly used slangs while calculating the polarity of the sentences.

General Inquirer (GI) [12] is a text mining approach that uses oldest manually constructed lexicons. This approach has classified a large number of words (11 k) into a few categories (183 or more). GI has been widely used to automatically determine the properties of text.

LIWC discovers the feature of speaking and writing language. It works by revealing the patterns in the speech by calculating recurrence with which words of different categories are used with respect to the total number of words in the sentence [13].

SentiWordnet 3.0 [5] is an enriched version of SentiWordNet 1.0 [8] which supports sentiment classification. The whole process of SentiWordNet 3.0 [5] uses semi-supervised and random walk processes instead of manual glosses.

Vader [6] has successfully calculated the overall polarity of sentences by taking into consideration the score of each word of the sentence and performing a qualitative analysis to identify those properties of text which affect the intensity of the sentence.

It tells about how positive or negative the sentiment is. However, the only shortcoming of Vader [6] was the use of a very limited number of slangs.

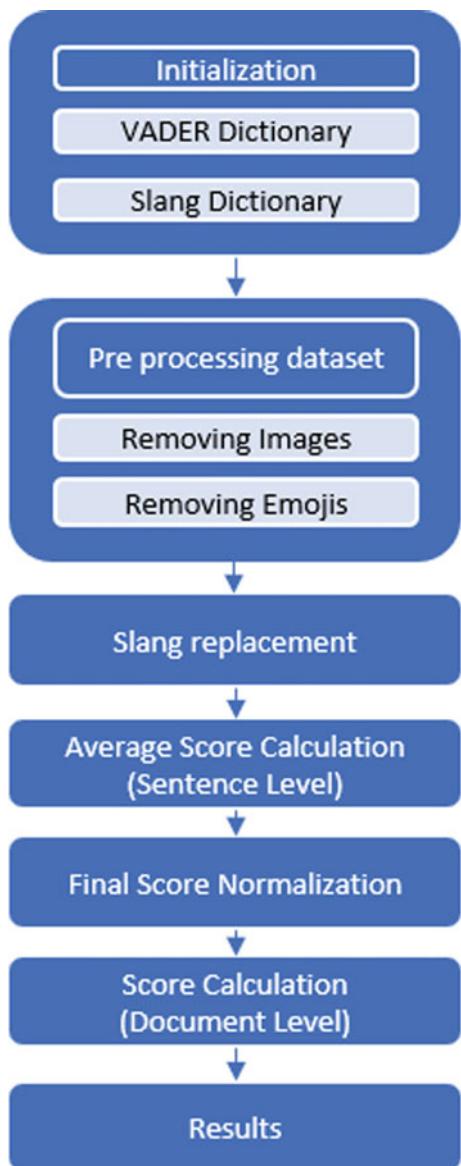
Senti-N-Gram [11] extracts the n-gram scores from a customer-generated data using a ratio-based approach which depends on the number of positive and negative sentences in a given dataset. It worked on n-gram lexicons with successfully handling the cases of intensifiers and negation.

However, the above approaches have not mentioned anything about slangs. Thus, this paper aims at suggesting an approach that deals with sentiment analysis suggesting with the abilities of Vader [6] and slang at sentence and document level as well.

### 3 Proposed Approach

The steps of proposed approach are elaborated below and shown in Fig. 1. Figure 3 elaborates the definitions of symbols used in algorithm Fig. 2.

1. *Initialization:* A slang dictionary is initialized considering near about 300 most commonly used slangs. The keys correspond to the slang and the values correspond to the meaning of the respective slang. Every slang is considered in lowercase and uppercase letters as well, e.g. lol: laugh out loud, LOL: laugh out loud. The existing Vader [6] dictionary of common lexicons created by using well-established word-banks, LIWC [13], ANEW [18] and GI [12] is also initialized.
2. *Pre-processing of Dataset:* The links of various images and emojis are removed from the dataset. The tweets with only image links were replaced by a numeric data type NaN (Not a Number).
3. *Slang Replacement:* The above pre-processed dataset as any slang is determined it is replaced with its meaning. E.g.: The sentence “The joke you cracked was funny!!!ahaha lol.” will become “The joke you cracked was funny!!!ahaha laugh out loud.”
4. *Score Calculation (Sentence Level):* The score of each sentence is calculated on the basis of the lexicons [6]. This provided a polarity to each sentence as positive, negative, or neutral considering the following five cases of punctuation, capitalization, degree modifiers, contrastive conjunction, and negation [6].
5. *Final Score Calculation Normalisation:* The sentiment score of each sentence lies between  $-4$  and  $+4$ . Then a Normalization function is applied to map each score to a value between  $-1$  and  $+1$ . The normalization function used is Hutto Normalization function [6].
6. *Score Calculation (Document Level):* Suppose an online document consists of ‘ $n$ ’ number of tweets. The score and polarity of the document are determined on the basis of the average score of positive sentences or negative sentences. If the average score of positive sentences is greater than the average score of negative sentences, the polarity of the document is positive or vice versa.

**Fig. 1** Proposed approach

#### 4 Demonstration for Score Calculation

Table 1 contains online sentiments covering all five cases of quantitative analysis, i.e., punctuation, negation, capitalization, intensifiers, and contrastive conjunction. The polarity score of these comments are the same for our proposed approach and Vader

<i>Symbols</i>	<i>Remarks</i>
dict	Dictionary of 1 slangs and their meanings
df	Data frame containing the dataset
Tweets	Column containing Tweets
sheet	Document containing dataset
i	counter variable
l	Number of slangs in dictionary
n	Number of tweets in a document
text	Tweets with images and emojis
text1	Tweets without images
strip_image	Function for removing images
Tweet without image	Data frame column containing tweets without images
text2	Tweets without images
text3	Tweets without emojis
strip_emoji	Function for removing emojis
Tweet without emoji	Data frame column containing tweets without emojis
sentence	Tweets whose score is calculated
word	Access each word of sentence
replace_slangs	Function for replacing slangs with their meanings
text4	Texts with replaced slangs
score	Stores avg score of each sentence
average_score	Function for calculating the average score of sentences
alpha	Normalization parameter
norm_score	Stores the normalized score
asns	Average score of negative sentences
aspss	Average score of positive sentences
avg_neg	Function calculates average score of all negative sentences.
avg_pos	Function calculates average score of all positive sentences
document_score	Final score of entire documents

**Fig. 2** Symbols used in algorithm

[6] for online sentiments ‘without slang’. Although, the polarity score of proposed approach and Vader [6] is considerably different for online sentiments ‘with slangs’.

```

# Step 1: Initialization
dict := {slango: meaning0, slang1:meaning1, slang2:meaning2, ..... slangi: meaningi};
df := sheet[“Tweets”];

# Step 2: Eliminating images and emojis
for i:= 0 to n-1 do:
    text:= df. iloc[i];   text1 :=strip_image(text);
    df [“Tweets without image”] :=text1;
end

for i=0 to n-1 do:
    text2 = df. iloc [i, “Tweets without image”];   text3 = strip_emoji(text2);
    df [“Tweets without emoji”]:= text3;
end

# Step 3: Replacing Slangs
for i =0 to n-1 do:
    sentence = df. iloc [i, ” Tweets without emoji”]
    if word in sentence == dict[slang] then:
        sentence =replace_slangs ();
    else do:
        sentence =sentence;
    end
    df[“Tweets with replaced slangs”]=sentence;
end

# Step 4: Score Calculation (Sentence Level)
for i =0 to n-1 do:
    text4 = df. iloc [i,” Tweets with replaced slangs”];
    score=average_score(text4);
end

# Step 5: Applying Normalization Function
for i =0 to n-1 do:
    norm_score=score / math. sqrt ((score*score) +alpha);
end

# Step 6: Score calculation (Document Level)
asns=avg_neg ();      asps=avg_pos ();
if (ASPS > ASNS):   document_score = ASPS;
else                  document_score = ASNS;
end

```

**Fig. 3** Proposed algorithm

**Table 1** Demonstration of score calculation

Without slangs	Proposed/Vader approach scores	With slangs	Proposed approach score	Vader approach score
LOVED your work	0.6841	TBH LOVED your work gw	0.8964	0.6841
The joke you cracked was funny!!!	0.7163	The joke you cracked was funny!!! hahaha lol	0.9077	0.8617
I love you!!	0.6988	I Love you!! xoxo	0.9059	0.8684
You were my friend but after the stunt you pulled i do not like you	-0.1444	You were my friend but after the stunt you pulled I do not like you gth	-0.8387	0.6542
I hope both of you are having best time	0.7964	I hope both of you are having best time ILYF	0.9062	0.7964
The service here is marginally good	0.3832	SRSLY The service here is marginally good	0.2280	0.3832
I do not like your tone	-0.2755	I do not like your tone, it is vb	-0.7098	0.2755
Kobe Bryant was a GREAT basketball player	0.729	Kobe Bryant was a GREAT and awsm basketball player	0.8730	0.7034
Food was horrible!!!	-0.6571	The food was ABS horrible !!!	-0.6877	-0.6571
Your outfit was good!	0.4926	Your outfit was good amz !	0.6239	0.4926

#### 4.1 Experimental Setup

To implement the proposed approach, we use Python and Vader a lexicon and rule-based sentiment analysis tool [6, 15]. (i) Tweets of various most followed personalities of Twitter are downloaded to conduct various experiments [14, 16]. The dataset consists of 16,000 tweets of 80 personalities across the world, 40 males and 40 females. The tweets of different personalities are stored as a document. (2) Comments on Facebook consists of 1452 online sentiments downloaded using facepager [7].

## 4.2 Results Evaluation

The tweets of each personality are stored as a document in a single excel. The scores of each tweet are evaluated at sentence level. At document level, the average score of all the tweets of personalities is taken to determine polarity and sentiment score of the tweets of various personalities. The experiment results at sentence level are illustrated in Tables 2 and 3. Table 2 clearly shows our proposed approach outperforms existing [6] for Twitter dataset, for SVM, Random Forest, and Linear Regression. Table 3 demonstrates that our approach shows accuracy of 97% for SVM and 98% for Random Forest for Facebook dataset. Thus, it is observed that our proposed approach gives better accuracy when text dataset with different slangs are evaluated.

## 5 Conclusion

The proposed approach attempts to calculate the polarity of online sentiments by considering the slangs used by the users in their online sentiments. It evaluates the sentence as negative, positive, or neutral by successfully replacing the slangs with their meanings to produce the correct score. The proposed approach is a further extension of the existing approach Vader [6] by overcoming its limitation i.e. usage of only common slangs.

## 6 Limitations and Future Work

Limitations of this approach are: (1) Exaggerated slangs such as “loooooool”, “rofl!!!!” etc. are not captured. (2) Use of emoticons such as 😊 😅 🤣, etc. in sentiments are very common but our approach doesn’t consider them. (3) We have considered only 300 slangs however more could be considered. This paves the way for future work of proposing a framework for sentiment analysis that considers exaggerated slangs, emojis, much more slangs than 300 for better and accurate results.

**Table 2** Comparative results on Twitter dataset (Sentence level)

ML techniques	SVM			Random forest			Linear regression					
	P	R	F	A	P	R	F	A	P	R	F	A
Evaluation metrics												
Without Slang (Vader)	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.99	0.99	0.93	0.93	0.93
With Slang (Proposed Approach)	0.95	0.965	0.96	0.95	0.95	0.94	0.95	0.96	0.95	0.95	0.95	0.95

**Table 3** Comparative results on Facebook dataset (Sentence level)

ML techniques	SVM			Random forest			Linear regression		
	P	R	F	A	P	R	F	A	P
Evaluation metrics									
Without Slang (Vader)	0.96	0.96	0.969	0.96	0.95	0.95	0.95	0.97	0.96
With Slang (Proposed Approach)	0.98	0.98	0.979	0.975	0.96	0.96	0.96	0.98	0.98

## References

1. Liu, B., Zhang, L.: A survey of opinion mining and sentiment analysis. In: Mining Text Data, pp. 415–463. Springer, Boston, MA (2012)
2. Nagamanjula, R., Pethalakshmi, A.: A novel framework based on bi-objective optimization and LAN2FIS for Twitter sentiment analysis. *Soc. Netw. Anal. Min.* **10**(34), 34 (2020)
3. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **37**, 267–307 (2011). [https://doi.org/10.1162/COLI\\_a\\_00049](https://doi.org/10.1162/COLI_a_00049)
4. Anjaria, M., Guddeti, R.M.R.: A novel sentiment analysis of social networks using supervised learning. *Soc. Netw. Anal. Min.* **4**(1), 181 (2014)
5. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: Proceedings of LREC, 10 (2010)
6. Gilbert, C.H.E., Hutto, E.: Vader: a parsimonious rule-based model for sentiment analysis of social media text. In: Eighth International Conference on Weblogs and Social Media (ICWSM-14), vol. 81, p. 82. (2014). Available at (20/04/16). <https://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>
7. Facepager. Get the software safe and easy. (2020). Retrieved 8 June 2020, from <https://facepager.software.informer.com/3.6/>
8. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: LREC, vol. 6, pp. 417–422 (2006)
9. Korayem, M., Aljadda, K., Crandall, D.: Sentiment/subjectivity analysis survey for languages other than English. *Soc. Netw. Anal. Min.* **6**(1), 75 (2016)
10. Wang, Z., Lin, Z.: Optimal feature selection for learning-based algorithms for sentiment classification. *Cogn. Comput.* **12**(1), 238–248 (2020)
11. Dey, A., Jenamani, M., Thakkar, J.J.: Senti-N-Gram: an n-gram lexicon for sentiment analysis. *Expert Syst. Appl.* **103**, 92–105 (2018)
12. Stone, P.J., Bales, R.F., Namenwirth, J.Z., Ogilvie, D.M.: The general inquirer: a computer system for content analysis and retrieval based on the sentence as a unit of information. *Behav. Sci.* **7**(4), 484–498 (1962)
13. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: LIWC 2001, 71(2001) (2001). Lawrence Erlbaum Associates, Mahway
14. Gupta, S., Singh, A., Ranjan, J.: Sentiment analysis: usage of text and emoji for expressing sentiments. In: Advances in Data and Information Sciences (2020)
15. VaderSentiment (2020). Retrieved 7 June 2020, from <https://pypi.org/project/vaderSentiment/>
16. Find out who's not following you back on Twitter, Tumblr, & Pinterest (2020). Retrieved 7 June 2020, from <https://friendorfollow.com/twitter/most-followers/>
17. Han, H., Zhang, J., Yang, J., Shen, Y., Zhang, Y.: Generate domain-specific sentiment lexicon for review sentiment analysis. *Multimedia Tools Appl.* **77**(16), 21265–21280 (2018)
18. Nielsen, F.Å.: A new ANEW: evaluation of a word list for sentiment analysis in microblogs. arXiv preprint [arXiv:1103.2903](https://arxiv.org/abs/1103.2903) (2011)

# Fuzzy Logic Technique for Evaluation of Performance of Load Balancing Algorithms in MCC



Divya, Harish Mittal, Niyati Jain, Bijender Bansal, and Deepak Kr. Goyal

**Abstract** Mobile Cloud Computing (MCC) is considered next-generation applications. One most important issue in MCC is Load Balancing. Uniform distribution of load among all the virtual servers can provide a better response time. Users demand more services with better results. Many algorithms are proposed to address this issue. Analysis of prevalent load balancing algorithms in cloud computing is done using various parameters like throughput, resource utilization, response time, etc., To compare the existing algorithms in a much better way and to compare their performance, a fuzzy logic technique is evolved in this paper. The technique is illustrated using a suitable numerical example. Evaluation of performance of 8 Algorithms is carried out on the basis of four metrics throughput, response time, migration time, and overhead. Using this technique Gradation of prevalent algorithms may be done easily in terms of their performance on the basis of desired metrics.

## 1 Introduction

Cloud computing is pay-per-use computing model. Mobile Cloud Computing is considered next-generation applications. One most important issue in Mobile Cloud Computing is Load Balancing. Uniform distribution of load among all the virtual servers can provide a better response time. It has gained much attention these days. Users demand more services with better results. Many algorithms are proposed to address this issue. Analysis of prevalent load balancing algorithms in cloud computing is done using various parameters like throughput, resource utilization, response time, etc., To compare the existing algorithms in a much better way and to compare their efficiency, a fuzzy logic technique is evolved in this paper. In future experimental study will be done so that gradation of prevalent algorithms may be

---

Divya · N. Jain · B. Bansal · D. Kr. Goyal  
Department of CSE, Vaish College of Engineering, Rohtak, India

H. Mittal (✉)  
BM Institute of Engineering and Technology, Sonepat, India

done in terms of their efficiency. Literature Review and Analysis of some prevalent algorithms of Load Balancing Algorithms are described in Sect. 2. The proposed model is described in Sect. 3 followed by concluding remarks and future work in Sect. 4.

## 2 Literature Review

### 2.1 *Software Quality Assessment Based on Fuzzy Logic*

Mittal Harish et al. “Software Quality Assessment Based on Fuzzy Logic Technique”, [5]. Provides fuzzy logic based precise approach to quantify quality of software modules on the basis of inspection rate and error density. TFNs were used to represent inspection rate and error density of the software. Software modules are given quality grades using fuzzy logic.

### 2.2 *Deriving Quality Metris for Cloud Computing Systems*

Matthias Beker, Sebastian Lehrig, Steffen Becker, proposed “Systematically Deriving Quality Metrics for Cloud Computing Systems” [1], 2015. They derived and classified metrics according the goal question in a top-down fashion by defining the goal to analyze cloud systems and questions that help achieving goals.

### 2.3 *Dynamic Round Robin (DRR)*

Lin et al. [3] proposed for energy-aware virtual machine scheduling and consolidation. They analyzed the problem of power consumption problems in data centers. DRR reduce a significant amount of power consumption. Performance Parameters considered Throughput, Overhead, Fault tolerance, Migration time, and Resource Utilization.

### 2.4 *Load Balancing in Cloud Computing*

**Power-Based Load Balancing for Cloud Computing PALB.** Galloway et al. [2] presented a load balancing approach to IaaS cloud Architecture. The approach maintains the state and decides the number of compute nodes that should be operating.

They considered Throughput, Overhead, Fault tolerance, Migration time, Response Time, and Resource utilization.

**Max–Min Task Scheduling Algorithm for Load Balance [4], Mao et al. (2014)** in Cloud Computing is required to distribute the dynamic local workload evenly across all the nodes to achieve high user satisfaction and resource utilization. Performance Parameters Throughput, Overhead, Response Time, and Resource utilization were studied.

**Mishra et al.** [19] described various load balancing techniques in homogeneous and heterogeneous cloud computing environments. A system architecture, with distinct models for the host, VM is described. They proposed taxonomy for the load balancing algorithm in the cloud environment. To analyze the performance of heuristic-based algorithms, simulation is carried out in CloudSim simulator.

**Afzal and Kavitha** [20] presents a comparative study on load balancing approaches. They framed a set of problem-related questions and discussed them in the work. The data collected for this study had been gathered from five potential databases. A multilevel taxonomy-based classification was proposed by considering five criteria. The study revealed that task scheduling is important in both proactive and reactive approaches.

## 2.5 Quality Metrics

Table 1 illustrates Scalability Metrics, Elasticity Metrics, and Efficiency Metrics.

**Table 1** Quality metrics

Metric		Unit	Example
Scalability metrics	Scalability Range (ScR)	Max	The system scales upto 100 req./min
	Scalability Speed (ScS)	Max, rate	The system scales upto 100 req./min, with linear increase rate 1 req/month
Elasticity metrics	Number of Service Level Objectives (SLO) violations (NSLOV)	1/time unit	40 SLO violations/hour
	Mean time to Quality repair (MTTQR)	Time unit	30 s for an additional 10 requests/hour
Efficiency metrics	Resource Provisioning Efficiency (RPE)	[0,∞]	5 more resources than actual resource demand
	Marginal Cost (MC)	Monetary unit	\$2 for an additional 100 requests/hour

### 3 Proposed Model for the Evaluation and Comparison of Load Balancing Algorithms

Load Balancing is reassigning the total load to the individual nodes of the collective system that facilitates networks to provide maximum throughput with minimum response time. Load Balancing Algorithms are of two types static and dynamic. Static algorithm divides the load equally on servers and is called round-robin algorithm, while dynamic algorithm uses weights on servers. Static algorithm creates imbalanced traffic while dynamic algorithm tries to balance the traffic.

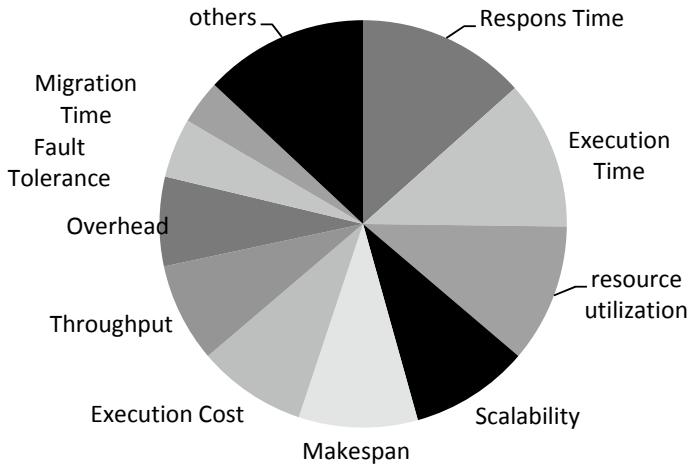
In 2019 Afzal and Kavihta, Journal of Cloud Computing Advances Systems and Applications, investigated that in existing literature there are 16 major Metric in Load Balancing, which are given in Table 2.

In order to evaluate performance of load balancing algorithms, one must first identify performance parameters/Metrics that strongly influence the performance. Such metrics are

- **Throughput-High** throughput is necessary for overall system performance. Generally, it is measured in MB/s.
- **Overhead**—It should be **low**.
- **Fault Tolerance**—It should be **high**.
- **Response Time**—is the time interval between sending a request and receiving its response. It should be **low**. It is measured in ms.

**Table 2** Metrics in load balancing

S. no.	Metric	Contribution%
1	Throughput	7.87
2	Overhead	7.09
3	Migration time	3.93
4	Response time	13.39
5	Execution times	11.81
6	Resource utilization	11.02
7	Scalability	9.45
8	Waiting time	2.36
9	Execution cost	8.66
10	Makespan	9.45
11	Degree of balance	4.72
12	Power consumption	3.14
13	Service level violation	0.78
14	Task rejection ratio	1.50
15	Fault tolerance	4.72
16	Migration cost	0.11



**Fig. 1** Metrics that strongly influence the performance of Load Balancing Algorithms

- **Resource Utilization**—is the proper utilization of the resources. It should be **high**. It is usually measured in milliseconds (ms).
- **Scalability**—is the ability to perform uniform load balancing. It should be **high**.
- **Migration Time**—is the time required in migrating the jobs or resources from one node to another. It should be **low** (Fig. 1).

#### Assumption: Number of concurrent users should be the same for each Algorithm

The metrics are of two types—Metric for which performance increases with increase in the value of metric e.g., throughput, fault tolerance, resource utilization, scalability, and metric for which performance decreases with decrease in the value of the metric e.g., response time, migration time, and overhead. We take the metrics as Triangular Fuzzy Number  $(a, m, b)$  and suppose that  $m$  divides the line joining the points  $a$  and  $b$  in the ratio 1:1, so that

$$m = \frac{a + b}{2} \quad F = \frac{b - a}{2m} = \frac{b - a}{b + a}$$

For  $b > a$

$$\mu(x) = \begin{cases} 0, & x \leq a \\ \frac{m-x}{m-a}, & a \leq x \leq m \\ \frac{b-x}{b-m}, & m \leq x \leq b \\ 0, & x \geq b \end{cases} \quad (1)$$

For  $b < a$

$$\mu(x) = \begin{cases} 0, & x \geq a \\ \frac{x-m}{a-m}, & a \leq x \leq m \\ \frac{x-b}{m-b}, & m \leq x \leq b \\ 0, & x \leq b \end{cases} \quad (2)$$

## Fuzzification

- (a) For metric whose performance is directly proportional to performance (Table 3)

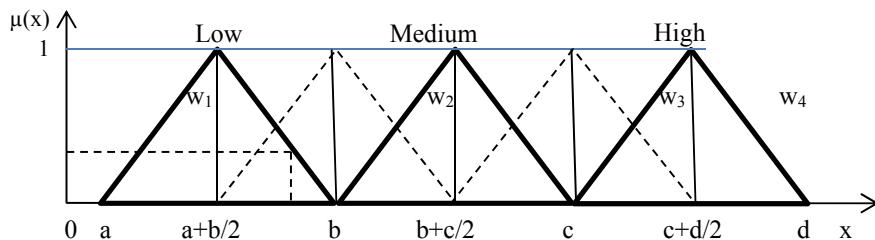
where,  $a, b, c, d, w_1, w_2, w_3$ , and  $w_4$  are real constants,  $w_1 < w_2 < w_3 < w_4$ . A value of  $x$  may have two membership functions. For  $\frac{a+b}{2} \leq x \leq b$ , it has membership function  $\mu$  as low and  $1 - \mu$  for medium. For  $b \leq x \leq \frac{b+c}{2}$ , it has membership function  $\mu$  as medium and  $1 - \mu$  for low (Fig. 2).

- (b) For metric whose performance is indirectly proportional to performance (Table 4).

where  $a, b, c, d, w_1, w_2, w_3$ , and  $w_4$  are real constants and  $w_1 < w_2 < w_3 < w_4$  (Fig. 3).

**Table 3** Linguistic Variables for Metric for which performance is directly proportional to the value of metric

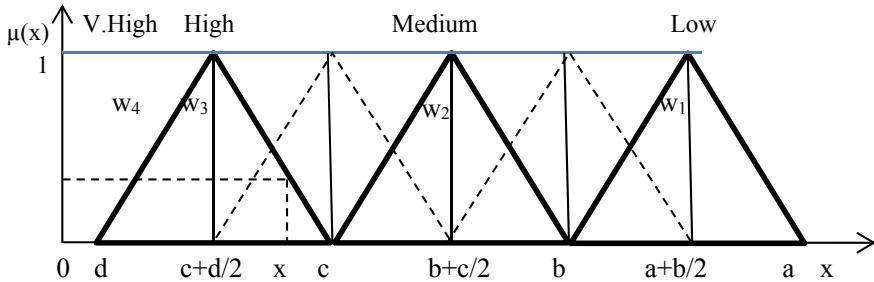
Information interval	Linguistic variable	Weights
$[a, b]$	Low	$w_1$
$[b, c]$	Medium	$w_2$
$[c, d]$	High	$w_3$
$[d, \infty]$	Very high	$w_4$



**Fig. 2** TFNs when performance is directly proportional to the value of the metric

**Table 4** Linguistic Variables for Metric for which performance is indirectly proportional to the value of metric

Information interval	Linguistic variable	Weights
$[0, d]$	Very high	$w_4$
$[d, c]$	High	$w_3$
$[c, b]$	Medium	$w_2$
$[b, a]$	Low	$w_1$



**Fig. 3** TFNs when performance is indirectly proportional to the value of the metric

A value of  $x$  may have two membership functions. For  $\frac{a+b}{2} \geq x \geq b$ , it has membership function  $\mu$  as high and  $1 - \mu$  for medium. For  $b \geq x \geq \frac{b+c}{2}$ , it has membership function  $\mu$  as medium and  $1 - \mu$  for high.

### Defuzzification

After evaluating membership for the value of metric, find its contribution in the performance as marks for this metric, using the following rules.

Equation 3 for Metric for which performance increases with increase in the value

$$m = \begin{cases} \mu_i * w_1, & a \geq x \geq \frac{a+b}{2} \\ \mu_i * w_1 + (1 - \mu_i) * w_2, & \frac{a+b}{2} \geq x \geq b \\ \mu_i * w_2 + (1 - \mu_i) * w_1, & b \geq x \geq \frac{b+c}{2} \\ \mu_i * w_2 + (1 - \mu_i) * w_3, & \frac{b+c}{2} \geq x \geq c \\ \mu_i * w_3 + (1 - \mu_i) * w_2, & c \geq x \geq \frac{c+d}{2} \\ \mu_i * w_3 + (1 - \mu_i) * w_4, & \frac{c+d}{2} \geq x \geq d \\ \mu_i * w_4, & x < d \end{cases} \quad (3)$$

Equation 4 for Metric for which performance decreases with increase in the value.

$$m_i = \begin{cases} \mu_i * w_1, & a \geq x \geq \frac{a+b}{2} \\ \mu_i * w_1 + (1 - \mu_i) * w_2, & \frac{a+b}{2} \geq x \geq b \\ \mu_i * w_2 + (1 - \mu_i) * w_1, & b \geq x \geq \frac{b+c}{2} \\ \mu_i * w_2 + (1 - \mu_i) * w_3, & \frac{b+c}{2} \geq x \geq c \\ \mu_i * w_3 + (1 - \mu_i) * w_2, & c \geq x \geq \frac{c+d}{2} \\ \mu_i * w_3 + (1 - \mu_i) * w_4, & \frac{c+d}{2} \geq x \geq d \\ \mu_i * w_4, & x < d \end{cases} \quad (4)$$

If we take  $w_1 = 10, w_2 = 20, w_3 = 40, w_4 = 50, Mi$  are marks out of 50.

Let Total Marks is Sum of marks of all the metrics (taking MAX value as 100). Then Grade of Algorithm is calculated as (Table 5).

**Table 5** Classification of grades

Total marks (100)	Grade
$0 \leq M_{\text{total}} < 5$	1
$5 \leq M_{\text{total}} < 10$	2
$10 \leq M_{\text{total}} < 20$	3
$20 \leq M_{\text{total}} < 30$	4
$30 \leq M_{\text{total}} < 40$	5
$40 \leq M_{\text{total}} < 50$	6
$50 \leq M_{\text{total}} < 60$	7
$60 \leq M_{\text{total}} < 70$	8
$70 \leq M_{\text{total}} < 80$	9
$M_{\text{total}} \geq 80$	10

**Table 6** Values of throughput and response time of various algorithms

Algorithm	Response time in ms	Throughput in MB/s
A	2600	17
B	1700	26
C	800	37
D	2700	38
E	2650	27
F	1800	44
G	850	18
H	3400	17

**Table 7** Weights for throughput

Throughput in MB/s	Complexity	Weights
[15,25]	Low	10
[25,35]	Medium	20
[35,45]	High	40
[45,∞]	Very high	50

**Illustrative Example:** We take a simple case by taking into consideration only four metrics.

**Inputs:** Throughput, Response time, Migration Time and Overhead. But the result can be extended to any desired number of Metrics.

**Output:** Grade of Algorithm.

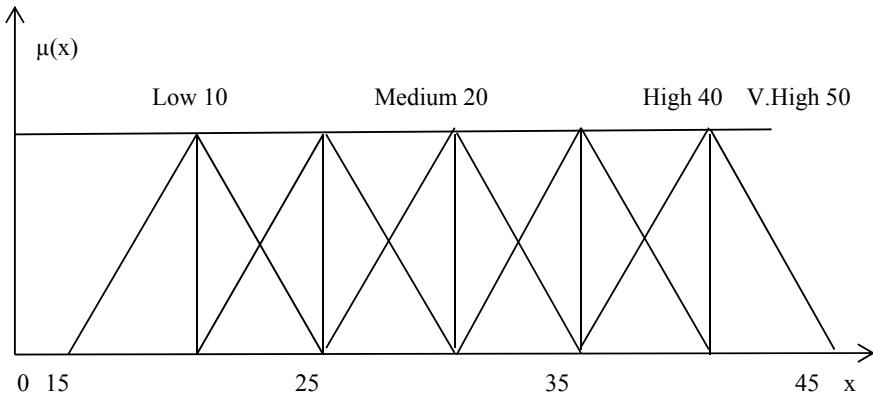
The values of throughput and response time of various algorithms are given in Table 6.

## Fuzzification

Taking,  $n = \text{no. of concurrent users}$ , Complex Matrix for Throughput (Table 7 and Fig. 4).

$$M_{\text{thr}} = \begin{cases} \mu_{\text{thr}} * 10, & 15 \leq x \leq 20 \\ \mu_{\text{thr}} * 10 + (1 - \mu_{\text{thr}}) * 20, & 20 \leq x \leq 25 \\ \mu_{\text{thr}} * 20 + (1 - \mu_{\text{thr}}) * 10, & 25 \leq x \leq 30 \\ \mu_{\text{thr}} * 20 + (1 - \mu_{\text{thr}}) * 40, & 30 \leq x \leq 35 \\ \mu_{\text{thr}} * 40 + (1 - \mu_{\text{thr}}) * 20, & 35 \leq x \leq 40 \\ \mu_{\text{thr}} * 40 + (1 - \mu_{\text{thr}}) * 50, & 40 \leq x \leq 45 \\ \mu_{\text{thr}} * 50, & x > 45 \end{cases} \quad (5)$$

Value of weights for low, medium and high, and very high are 10, 20, 40, and 50, respectively. The calculated values of  $\mu_{\text{thr}}$  are given in Table 8



**Fig. 4** TFNs for throughput

**Table 8** Calculated values of marks for throughput

Algorithm	Throughput in MB/s	$\mu_{\text{thr}}$	$1 - \mu_{\text{thr}}$	Marks out of 50
A	17	0.4	0.6	4
B	26	0.2	0.8	12
C	37	0.4	0.6	28
D	38	0.6	0.4	32
E	27	0.4	0.6	14
F	44	0.2	0.8	48
G	18	0.6	0.4	6
H	17	0.4	0.6	4

## Response Time

$$M_{\text{res}} = \begin{cases} \mu_{\text{res}} * 10, & 3500 \geq x \geq 3000 \\ \mu_{\text{res}} * 10 + (1 - \mu_{\text{res}}) * 20, & 3000 \geq x \geq 2500 \\ \mu_{\text{res}} * 20 + (1 - \mu_{\text{res}}) * 10, & 2500 \geq x \geq 2000 \\ \mu_{\text{res}} * 20 + (1 - \mu_{\text{res}}) * 40, & 2000 \geq x \geq 1500 \\ \mu_{\text{res}} * 40 + (1 - \mu_{\text{res}}) * 20, & 1500 \geq x \geq 1000 \\ \mu_{\text{res}} * 40 + (1 - \mu_{\text{res}}) * 50, & 1000 \geq x \geq 500 \\ \mu_{\text{res}} * 50, & x < 500 \end{cases} \quad (6)$$

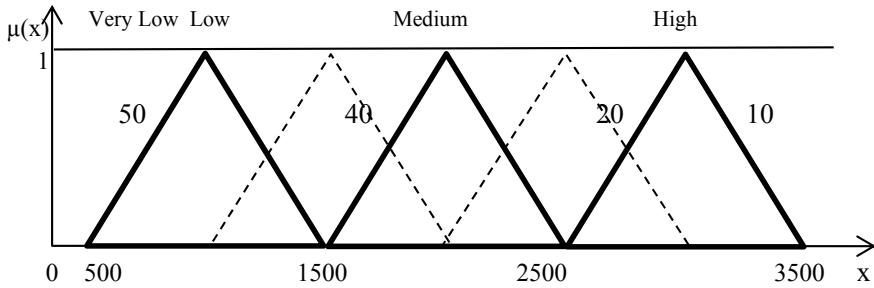
Value of weights for high, medium and low and very low are 10, 20, 40, and 50, respectively (Table 9). The calculated values of  $\mu_{\text{res}}$  are given in Table 10 (Fig. 5).

**Table 9** Calculated values of marks of response time

Algorithm	Response time (ms)	$\mu_{\text{res}}$	$1 - \mu_{\text{res}}$	$M_{\text{res}}$ (Out of 50)
A	2600	0.2	0.8	18
B	1700	0.4	0.6	32
C	800	0.6	0.4	44
D	2700	0.4	0.6	16
E	2650	0.3	0.7	17
F	1800	0.6	0.4	28
G	850	0.7	0.3	43
H	3400	0.2	0.8	2

**Table 10** Calculated values of marks of migration time

Algorithm	Migration time in ms	$\mu_{\text{mgr}}$	$1 - \mu_{\text{mgr}}$	$M_{\text{mgr}}$ (Marks are out of 50)
A	2400	0.2	0.8	12
B	1600	0.2	0.8	36
C	900	0.2	0.8	48
D	2900	0.2	0.8	18
E	2850	0.7	0.3	13
F	3400	0.2	0.8	2
G	550	0.1	0.9	49
H	3300	0.4	0.6	4



**Fig. 5** TFNs representing response time using Fig. 3

$$M_{mgr} = \begin{cases} \mu_{mig} * 10, & 3500 \geq x \geq 3000 \\ \mu_{mig} * 10 + (1 - \mu_{mig}) * 20, & 3000 \geq x \geq 2500 \\ \mu_{mig} * 20 + (1 - \mu_{mig}) * 10, & 2500 \geq x \geq 2000 \\ \mu_{mig} * 20 + (1 - \mu_{mig}) * 40, & 2000 \geq x \geq 1500 \\ \mu_{mig} * 40 + (1 - \mu_{mig}) * 20, & 1500 \geq x \geq 1000 \\ \mu_{mig} * 40 + (1 - \mu_{mig}) * 50, & 1000 \geq x \geq 500 \\ \mu_{mig} * 50, & x < 500 \end{cases} \quad (7)$$

$$M_{ovr} = \begin{cases} \mu_{ovr} * 10, & 10 \geq x \geq 9 \\ \mu_{ovr} * 10 + (1 - \mu_{ovr}) * 20, & 9 \geq x \geq 8 \\ \mu_{ovr} * 20 + (1 - \mu_{ovr}) * 10, & 8 \geq x \geq 7 \\ \mu_{ovr} * 20 + (1 - \mu_{ovr}) * 40, & 7 \geq x \geq 6 \\ \mu_{ovr} * 40 + (1 - \mu_{ovr}) * 20, & 6 \geq x \geq 5 \\ \mu_{ovr} * 40 + (1 - \mu_{ovr}) * 50, & 5 \geq x \geq 4 \\ \mu_{ovr} * 50, & x < 4 \end{cases} \quad (8)$$

Value of weights for high, medium and low, and very low are taken 10, 20, 40, and 50, respectively. The calculated values of  $\mu_{ovr}$  are given in Table 11

Grades of Algorithms are calculated in Table 12 using Tables 8, 9, 10 and 11 (Fig. 6).

## 4 Conclusion and Future Work

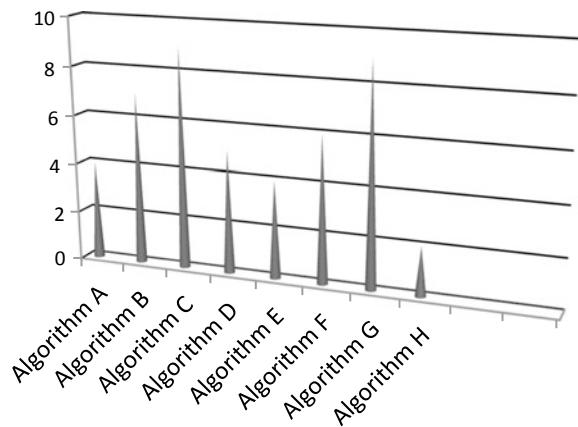
Analysis of prevalent load balancing algorithms in cloud computing is done using various parameters like throughput, resource utilization, response time etc., To compare the existing algorithms in a much better way and to compare their performance, a Fuzzy Logic Technique based Model is proposed. The Model is explained by a suitable Numerical by taking 4 metrics Throughput, Response Time, Migration

**Table 11** Calculations for overhead

Algorithm	Overhead in million Rs	$\mu_{ovr}$	$1 - \mu_{ovr}$	$M_{ovr}$ (Marks are out of 50)
A	7.5	0.5	0.5	15
B	6.5	0.5	0.5	30
C	4.5	0.5	0.5	45
D	8.5	0.5	0.5	15
E	8.50	0.5	0.5	15
F	9.5	0.5	0.5	5
G	4.5	0.5	0.5	45
H	9.5	0.5	0.5	5

**Table 12** Grades of algorithms

Algorithm	$M_{res}$	$M_{thr}$	$M_{mgr}$	$M_{ovr}$	Total marks (100)	Grade
	<b>Out of 50</b>					
A	18	4	12	15	24.5	4
B	32	12	36	30	55	7
C	44	28	48	45	82.5	9
D	16	32	18	15	41	5
E	17	14	13	15	29.5	4
F	28	48	2	5	41.5	6
G	43	6	49	45	71.5	9
H	2	4	4	5	7.5	2

**Fig. 6** Comparison of performance of various algorithms

Time and Overhead. In future experimental study of the existing realistic algorithms can be done using this model so that gradation of prevalent algorithms may be done in terms of their performance on the basis of desired metrics. Present MCC Architectures are not up to the mark for the present requirements. There is an immense need to tackle the issues of the MCC environment. There is vast scope for this study in the fast-emerging field of Healthcare Monitoring. Fuzzy logic has vast capabilities to address the challenges; hence effort is to evolve fuzzy-based Models.

## References

1. Becker, M., Lehrig, S., Becker, S.: Systematically deriving quality metrics for cloud computing systems. In: Proceedings of the 6th ACM/SPEC international conference on performance engineering, pp. 169–174 (2015)
2. Galloway, J.M., Smith, K.L., Vrbsky, S.S.: Power aware load balancing for cloud computing. In: Proceedings of the World Congress on Engineering and Computer Science, vol. 1, pp. 19–21 (2011)
3. Lin, C.C., Liu, P., Wu, J.J.: Energy aware virtual machine dynamic provision and scheduling for cloud computing. In: 2011 IEEE International Conference on Cloud Computing, pp. 736–737. IEEE (2011)
4. Mao, Y., Chen, X., Li, X.: Max–Min task scheduling algorithm for load balance in cloud computing, pp. 457–465
5. Harish, M., Pardeep, B., Puneet, G.: Software quality assessment based on fuzzy logic technique. *Int. J. Soft Comput. Appl.* (3), 105–112 (2008). ISSN: 1453-2277
6. Fernando, N., Loke, S.W., Rahayu, W.: Mobile cloud computing: a survey. *Future Generat. Comput. Syst.* **29**(1), 84–106 (2013)
7. Jia, W., Zhu, H., Cao, Z., Wei, L., Lin, X.: SDSM: a secure data service mechanism in mobile cloud computing. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM), pp. 1060–1065 (2011)
8. Kosta, S., Aucinas, A., Hui, P., Mortier, R., Zhang, X.: Unleashing the power of mobile cloud computing using Thinkair, pp. 1–17. CoRR abs/1105.3232 (2011)
9. Liang, H., Huang, D., Cai, L.X., Shen, X., Peng, D.: Resource allocation for security services in mobile cloud computing. In: Proceedings of IEEE Conference on Computer Communications Workshops (INFOCOM), pp. 191–195 (2011)
10. Naz, S.N., Abbas, S., et al.: Efficient load balancing in cloud computing using multi-layered Mamdani fuzzy inference expert system. *(IJACSA) Int. J. Adv. Comput. Sci. Appl.* **10**(3) (2019)
11. Ragmani, A., et al.: An improved Hybrid Fuzzy-Ant Colony Algorithm applied to load balancing in cloud computing environment. In: The 10th International Conference on Ambient Systems, Networks and Technologies (ANT) April 29–May 2, LEUVEN, Belgium (2019)
12. Shiraz, M., Gani, A., Khokhar, R., Buyya, R.: A review on distributed application processing frameworks in smart mobile devices for mobile cloud computing. *IEEE Commun. Surv. Tutorials* **15**(3), 1294–1313 (2013)
13. Sanaei, Z., Abolfazli, S., Gani, A., Buyya, R.: Heterogeneity in mobile cloud computing: taxonomy and open challenges. *IEEE Commun. Surv. Tutorials* **16**(1), 369–392 (2014)
14. Sethi, S., et al.: Efficient load Balancing in Cloud Computing using Fuzzy Logic. *IOSR J. Eng. (IOSRJEN)* **2**(7), pp. 65–71 (2012). ISSN: 2250-3021
15. Zadeh, L.A.: From computing with numbers to computing with words-from manipulation of measurements to manipulation of perceptions. *Int. J. Appl. Math. Comput. Sci.* **12**(3), 307–324 (2002)
16. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965)

17. Zadeh., L.A.: The concept of a linguistic variable and its applications to approximate reasoning—part I. *Inf. Sci.* **8**, 199–249 (1975)
18. Zhou, B., Dastjerdi, A.V., Calheiros, R.N., Srivama, S.N., Buyya, R.: A context sensitive offloading scheme for mobile cloud computing service. In: Proceedings of the IEEE 8th International Conference on Cloud Computing (CLOUD), pp. 869–876 (2015)
19. Mishra, S.K., Sahoo, B., Parida, P.P.: Load balancing in cloud computing: a big picture. *J. King Saud Univ. Comput. Inf. Sci.* **32**, 2, 149–158 (2020)
20. Afzal, S., Kavitha, G.: Load balancing in cloud computing—a hierarchical taxonomical classification. *J. Cloud Comput.* **8**, 1, 22 (2019)

# Impact of Bio-inspired Algorithms to Predict Heart Diseases



N. Sree Sandhya and G. N. Beena Bethel

**Abstract** Optimization techniques are employed to deal with dynamic, difficult, and robust problems. Most of the Machine learning algorithms are implemented to predict heart diseases. Classification techniques are one of the methods that is highly used in machine learning for prediction. Some classification methods predict accuracy with acceptable range, but others may not. In this paper, we streamline two different bio inspired algorithms, Ant and Bat are used for heart disease prediction. Here, we extracting the key features from heart disease attributes using these two bio-inspired algorithms. Then these extracted features are implemented to the different classifiers. In this research, we examine the bio inspired algorithms optimized with Random Forest and SVM classifiers and compared the results. Ant colony optimization and Bat colony optimization give better results with SVM classifier than Random Forest classifier. When comparing the results in this research, Bat algorithm is better-optimized algorithm than ant algorithm.

## 1 Introduction

Most of the people in the world are suffering from heart diseases, which leads to death frequently. According to mortality statistics, it is proven that the main origin of death in the world is because of heart diseases. As per WHO (World Health Organization) four out of 5 cardiovascular diseases (CVD) deaths are due to heart strokes and heart attacks [1]. Heart plays a major role in functioning of our body and it is a very important organ. It pumps blood throughout the body. Whenever heart disease occurs, all the functionalities of heart is not properly worked. So, it is important to predict heart diseases with an intelligent approach for better results. In this research, we implemented smart optimization algorithms like bio inspired algorithms for heart disease prediction.

Bio inspired algorithms are nature-inspired algorithms developed to resolve complex complications. With huge data, it became more challenging to provide

---

N. Sree Sandhya (✉) · G. N. Beena Bethel  
CSE Department, GRIET, Hyderabad, Telangana, India

optimum solution [2]. At these times, bio inspired algorithms are recognized to solve complex problems with the novel approaches. The objective of this paper is to enhance the attributes of dataset for better prediction. By extracting the features of bio inspired algorithms, the attribute class labels are analyzed for further process. These class label values are generated by considering the remaining attribute values with the extracted bio inspired algorithm features. Then respective classifier is applied on this dataset for predicting the accuracy.

## 2 Related Work

In healthcare, the patient's data is increasing day-by-day with extra medical information. To predict any disease, it is important to collect related information of various scenarios of patient's data. Different techniques and tools are depending on this data for prediction of a disease. The key aim of this paper is to analyze the patient's records and extract the main features and implementing the bio inspired algorithms to predict the heart disease. Kora and Ramakrishna [3] presented a method for predicting myocardial infarction based on the changes in ECG signals. Myocardial infarction is predicted by using the proposed method called improved bat algorithm. But for implementation of improved bat algorithm they just taken 13 patient records. Out of 13, 7 are myocardial infarction patients records and remaining 6 are normal individual records. Four methods like Support Vector Machine, K-Nearest Neighbors, Levenberg–Marquardt Neural Network, and Conjugate Gradient Neural Network are implemented in both normal bat and improved bat algorithms. They concluded that improved bat algorithm gives better results when compared with normal bat algorithm.

Dubey et al. [4] diagnosed heart diseases in the early stage by combining the Ant optimization algorithm with data mining techniques (DMACO). They considered the pheromone value of ant and recognized the risk level. Whenever pheromone value increases the risk also increases. Then Ant algorithm with data mining techniques are applied to generate the detection rate of the heart disease and concluded that DMACO develops the pheromone intensity and improves the detection rate of the disease.

From the above-discussed works, it is clear that all the techniques used for heart disease prediction are hybrid classification methods. But the work in this research is different from the above-discussed works in a way that, here we implemented two different optimization techniques with two different classifiers. In this study, we analyze the efficiency by combining optimization techniques with classifiers.

### 3 Methodology

The system architecture used in this research for predicting the heart disease is shown in Fig. 1.

Here, Heart data means dataset which contains related data of heart disease patients. Heart Disease Dataset used in this research is taken from UCI Machine Learning Repository. UCI Repository contains different datasets from different domains. David Aha developed this UCI repository by in 1987 [5]. In this research, we used Cleveland (Cleveland clinic Foundation) data, which contains 303 instances with 76 raw attributes. Out of 76, we used only 14 attribute values. Out of 14, the last value is the predicted attribute based on these 13 attribute values. We are considering the Cleveland dataset because it has less amount of missing data.

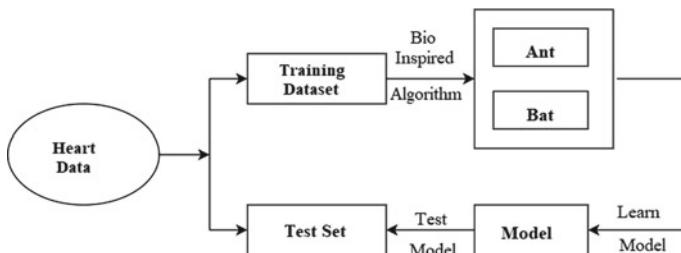
#### 3.1 Optimization Techniques

##### 3.1.1 Ant Colony Optimization

Marco Dorigo was presented Ant Algorithm in 1992. This Ant algorithm was designed based on the real ants' inspiration. It is highly used to resolve complex difficulties. Ant algorithm finds optimal solution by performing iterations [6]. The main goal is to select the finest features with lowest redundancy.

Let us assume ' $m$ ' ants develop solutions from a finite set of components ' $C$ '. Consider an empty solution  $s^x = \phi$ , at each iteration it is extended by adding new possible solution from the set of neighbors  $N(s^x) \subseteq C$ . The path of the graph can be designed by constructing these partial solutions [7]. At construction step, these solution components are formed by using probabilistic approach. Then

$$P(C|s^x) = \frac{T_{ij}^\alpha \cdot \eta_{ij}^{\beta}}{\sum_{C \in N(S^x)} T_{ij}^\alpha \cdot \eta_{ij}^{\beta}} \quad (1)$$



**Fig. 1** Architecture

where  $C = \{c_{ij}\}$ ,  $i = 1, 2, 3, \dots, m$  and  $j = 1, 2, 3, \dots, n$ ; and  $\forall c_{ij} \in N(S^x)$ .

$T_{ij}$  is a pheromone value and  $\eta'_{ij}$  is heuristic value. Both  $T_{ij}$  and  $\eta'_{ij}$  are accompanying with the component  $c_{ij}$ .  $\alpha$  and  $\beta$  are determining pheromone and heuristic information and these are real positive parameters. By using this probability rule ants are constructing solutions is called a tour [8]. To find the best solutions, we have to update the pheromone values. Here we are gathering good solutions associated with pheromones and avoid bad ones. This is achieved through pheromone evaporation.

$$T_{ijs} \leftarrow (1 - \rho) \cdot T_{ij} + \rho \cdot \sum_{s \in S | c_{ij} \in s} F(s) \quad (2)$$

where  $S$  is the solutions used for update,  $\rho$  is evaporation rate and it lies between  $(0, 1]$ ,  $F(s)$  is fitness function and  $F: S \rightarrow R^+$  is a function then  $f(s) < f(s^c) \Rightarrow F(s) \geq F(s^c), \forall s \neq s^c \in S$ .

The algorithm is as follows:

### Ant Algorithm

- Step 1 create the initial parameters like  $N$  nodes and  $M$  arcs. Then constant amount of pheromone is assigned to all arcs.
- Step 2 Ant  $k$  uses the pheromone trail at node  $i$  to compute the next node  $j$  by using probabilistic approach. It is calculated using (1).
- Step 3 When ants are traversed in between the arc  $(i, j)$  the pheromone value is updated, which are called local trails. These changes are done using the following equation.

$$T_{ij} \leftarrow T_{ij} + \Delta T^k$$

- Step 4 When ant  $k$  moved to the next node, the pheromone evaporation is done using the following equation.

$$T_{ij} \leftarrow (1 - \rho) \cdot T_{ij}, \forall (i, j) \in A$$

- Step 5 Repeat step 2 until ant target point reached. This is called iteration cycle and it involves ant's movements, pheromone evaporation and deposits.
- Step 6 Whenever ants reach the target, they will update the pheromone using global trails and finds the optimal solution. These global trails are done using (2)
- Step 7 This process repeats until termination condition is satisfied. If so, it will generate the output otherwise repeat above steps once again.
- Step 8 End.

### 3.1.2 Bat Colony Optimization Algorithm

In 2010, Xin-She Yang developed bat algorithm. It is developed based on the behavior of micro bats communication [9] using echolocation. Generally, bats using echo-based location to search food and travel from one place to another. During this search of food bats, change their velocity, frequency, and sound accordingly. The bat algorithm is presented as follows.

#### Bat Algorithm

- Step 1 Define the objective function or fitness function of a bat.
- Step 2 Initialize the bat parameters like frequency ( $f_i$ ), Loudness (A) and pulse emission rate ( $r_i$ ).
- Step 3 Randomly generate the bat population.
- Step 4 Sort the current population values (preferably in descending order).
- Step 5 Generate new frequency, velocity and position values using (3), (4), (5) respectively until maximum iteration criteria is done.
- Step 6 Generate the best solution as a local solution in step 5.
- Step 7 New solutions should be stored in a resource log.
- Step 8 Update the values of loudness and pulse emission rate using the (6) and (7).
- Step 9 Fitness of new solution is tested w.r.t. A and r then rank the bats and find the current best position.
- Step 10 Repeat the process until termination criteria is satisfied.
- Step 11 End.

Bats fly randomly with the velocity  $v_i$  at position  $x_i$  with different frequency ranges  $f[\min, \max]$ , varying wavelength  $\lambda$  and loudness  $A_0$  to search for prey. The wavelength is adjusted automatically and pulse emission rate can be varied in between  $[0,1]$  and it depends on the target of proximity. Loudness value lies between  $[A_0, A_{\min}]$ . Here the bat is randomly assigned the frequency between  $[f_{\min}, f_{\max}]$ , hence it is called frequency-tuning algorithm. Each bat is associated with velocity  $v_i^t$  and position  $x_i^t$  in search space at each iteration  $t$ , with respect to assigned frequency  $f_i$ . Hence at each iteration, we need to update  $f_i$ ,  $v_i$  and  $x_i$  and along with these parameters loudness and pulse emission rate also be updated. To do this, we use the following equations

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \quad (3)$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x^*)f_i \quad (4)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad (5)$$

$$A_i(t+1) = \alpha A_i(t) \quad (6)$$

$$r_i(t+1) = r_i^0 [1 - e^{-\gamma t}] \quad (7)$$

where  $f_i$  is the existing frequency,  $v_i^t$  is the existing velocity and  $X_i$  is the existing position.  $A_i$  is the loudness and  $r_i$  is the pulse emission rate.

$v_i^t - 1$  is the previous velocity,  $x_i^t - 1$  is the previous position.

$\alpha, \beta$  are random values and lies between  $[0,1]$  and  $\gamma$  is a constant value and  $\gamma > 0$ .  $x^*$  is the present best position.

### 3.2 Classification

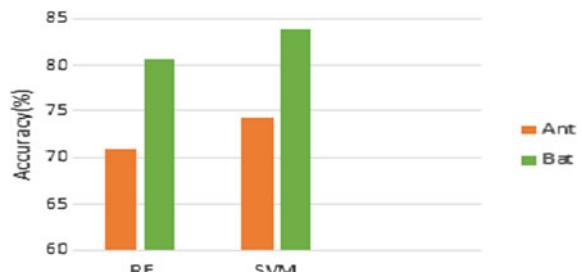
Heart Disease Prediction can be perceived as a clustering problem or classification problem [10]. In contrast, we have a tendency to fashioned a model on the immense set of presence and absence data. So, we can turn it to classification once we extract the features from the respective bio inspired algorithm. In this research, we comparing the updated dataset using two classifiers namely random forest and support vector machines.

## 4 Experiments and Results

Here, we confer the heart disease dataset researches and evaluations in python environment. The objective of this project was to test which optimization technique organizes the heart disease the best with a specific classifier. We use tenfold validation to evaluate the performance of classification methods for predicting the heart disease. To avoid uneven operation consequences, each trial was run 10 times, and the optimized classification accuracy was chosen for assessment. Then the accuracy comparison of ant and bat algorithms with both the classifiers is as described in Fig. 2.

From Fig. 2, it is observed that irrespective of the classifier the order of the optimization algorithms which gives the best accuracy is Bat and Ant algorithms respectively. When we compare these two algorithms in the view of feature extraction,

**Fig. 2** Accuracy comparison graph of Ant and Bat algorithm using Random Forest (RF) and SVM classifiers



Ant algorithm takes the least priority and Bat Algorithm gives better results in both the cases. In this research, With Random Forest classifier, Ant gives 70.96% and bat algorithm gives 80.64% accuracy. With SVM Classifier, Ant gives 74.2% and Bat algorithm gives 83.87%. Both the algorithms give better results with SVM classifier.

## 5 Conclusion and Future Scope

In this paper, we mainly focus the two bio inspired algorithms feature extraction using two classifiers Random forest and SVM. Optimized SVM classifier gives better results in each case than Optimized Random Forest classifier. Bat algorithm gives better accuracy with both classifiers. So, we conclude that Bat algorithm is better than Ant algorithm implemented in this research. Similarly, in the case of classifiers it is also clear that SVM gives better prediction results than Random Forest. Because both classifiers applied on same bio algorithm but classification results vary. In Future works, there is a large scope for more bio algorithms with multiple classifiers. Whenever we implement multiple classifiers on a single bio algorithm will give better clarity of that respective feature extraction quality and increases the scope for better results.

## Reference

1. Khourdifi, Y., Bahaj, M.: Heart disease prediction and classification using Machine Learning Algorithms Optimized by particle swarm optimization and Ant Colony Optimization. *Int. J. Intell. Eng. Syst.* (2019)
2. Darwish, A.: Bio-inspired computing: algorithms review, deep analysis, and the scope of applications. *Future Comput. Inf. J.* (2018)
3. Kora, P., Ramakrishna, K.S.: Improved Bat Algorithm for the detection of myocardial infarction. *Springer Plus* **4**(1), 666 (2015)
4. Dubey, A., Patel, R., Choure, K.: An efficient data mining and Ant Colony Optimization Technique (DMACO) for heart disease prediction. *Int. J. Adv. Technol. Eng. Explor. (IJATEE)* **1**(1), 1–6 (2014)
5. Dataset, UCI Machine learning Repository [online]; <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>
6. Rao, P., et al.: An efficient approach for detection of heart attack using Noble Ant Colony Optimization concept of data. *IJESRT* (2018)
7. Dorigo, M.: Ant Colony optimization. *Scholarpedia* **2**(3), 1461 (2007)
8. Nasiruddin, I., Ansari, A.Q., Katiyar, S.: Ant Colony Optimization: a tutorial review, Conference Paper (2015)
9. Hinduja, R., Mettildha Mary, I., Ilakkya, M., Kavya, S.: CAD diagnosis using PSO, BAT, MLR and SVM. *Int. J. Adv. Res. Ideas Innovations Technol.* (2017)
10. Kotsiantis, S., Pintelas, P.E., Zaharakis, I.D.: Machine learning: a review of classification and combining techniques. *Artif. Intell. Rev.* **26**(3), 159–190 (2006)

# Structured Data Extraction Using Machine Learning from Image of Unstructured Bills/Invoices



K. M. Yindumathi, Shilpa Shashikant Chaudhari, and R. Aparna

**Abstract** The identification and extraction of unstructured data have always been one of the most difficult challenges of computer vision. Parsing this sort of large data is very challenging; however, recent advancements in computer vision technology help make this feasible. A ubiquitous commonplace item that many consumers receive is difficult to be transformed into raw data. Receipts contain a dense amount of data that can be useful for future analysis, but there exists no widely available solution for transforming receipts into structured data. Existing solutions are either very costly or inaccurate. This paper introduces a data pipeline for the identification, cropping, and extraction of unstructured data within healthcare bills/invoice images. This pipeline outperforms existing solutions by a large margin, and offers the ability to automatically pull out semantic data such as description and unit price from an image of a bill. It achieves this success by using Logistic Regression, KNeighbours, and OpenCV Scikit to crop the image. Optical Character Recognition (OCR) is applied to detect chunks of text and process the image. The accuracy observed is approximately 93% for Logistic Regression and 81% for KNeighbours.

## 1 Introduction

Recent years have seen a growing interest in harnessing advances Machine Learning (ML) and Optical Character Recognition (OCR) to translate physical and handwritten documents into digital copies. It appears to be difficult to create digital documents from scratch. Ultimately, a solution to get the simplicity of documents generation

---

K. M. Yindumathi (✉) · S. S. Chaudhari · R. Aparna

Department of Computer Science and Engineering, M.S. Ramaiah Institute of Technology, Bengaluru 560054, India

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

S. S. Chaudhari

e-mail: [shilpasc29@msrit.edu](mailto:shilpasc29@msrit.edu)

R. Aparna

e-mail: [aparna@msrit.edu](mailto:aparna@msrit.edu)

while ensuring the ease of digital documents usage is today's need. Identification of text characters generally deals with the identification of optically formed characters and is often called OCR. OCR's simple concept is to turn any hand-written or typed text into data files that can be interpreted and read by computer. Any document or book can be scanned with OCR, and the editable text file can then be translated from a computer. The OCR program has two main benefits which include the potential to improve efficiency by successfully reducing staff participation and processing data. Most broadly, the fields of implementation of this system are postal offices, banks, publishing sector, government agencies, education, banking, health care.

The universal OCR system consists of three main steps which are image acquisition and preprocessing, feature extraction, and classification [1]. Image preprocessing phase cleans up and enhances the image by noise removal, correction, binarization, dilation, color adjustment and text segmentation, etc. Feature extraction is a technique for extracting and capturing certain pieces of information from data. In the classification phase, the portion of the divided text in the document image will be mapped to the equivalent textual representation. The original invoice image is initially pre-processed by secondary rotation and edge cutting to eliminate the unnecessary background information. The region of the critical information in the normal image obtained is then derived by following the pattern, which is the focal point of the recognition of the material. OCR is used to translate image information into text and enable easy use of the interpreted content. The principle point is to analyze certain ML algorithms and process text using categorization and classification.

In particular, this paper is interested in the processing of unstructured bill image data and converting it into a simple-to-use, analyzable structured data format. It helps to resolve some of the fundamental difficulties and inconsistencies associated with parsing this sort of unstructured data. We use a combination of bill image datasets and custom receipt data collected over the past few months by a small group of people to train and evaluate the effectiveness of this system. Our aim is to be able to process and parse receipt data from most standard use cases, which involves steps such as correcting the input image orientation, cropping the receipt to remove the background, running OCR to pull text from the image, and using algorithm to determine relevant data from the OCR result. A good, satisfactory system handles the above characteristics and further edge cases by pulling the description and unit price from image inputs. The paper demonstrates how this result was achieved and provides both a quantitative and qualitative evaluation with respect to the success of this system.

## 2 Related Works

This section discusses related papers concerning the specific task of analyzing and classifying receipt-data using machine learning. Researchers mostly focused on developing techniques to improve the recognition and extraction of text from unstructured data whereas industry has focused on creating commercial systems to reduce

manual labor costs for inputting receipt image data for analysis or reporting. However, neither produces an optimal system due to degradations in either accuracy or cost.

One commonly used extraction mechanism for text detection is a Convolutional Neural Network (CNN) [1, 2]. This class of OCR utilizes a Long Short Term Memory (LSTM) to propose regions of interest where text may exist as well as a CNN to determine the likelihood of text appearing at that location. These systems all provide end-to-end for text identification—from localizing and recognizing text in images to retrieving data within such text. Deep learning based method for bank serial number recognition system consistently achieved state-of-the-art performance on bank serial number text detection benchmarks [1]. Some research work has also been done on graph neural networks to define the table structure for text extraction. Canny edge detector for image data identification [3] is proposed for utilizing all kinds of images, failed for handwritten bills. Table 1 presents summary of the text extraction techniques in various fields along with experimentation.

### 3 Proposed Structured Data Extraction

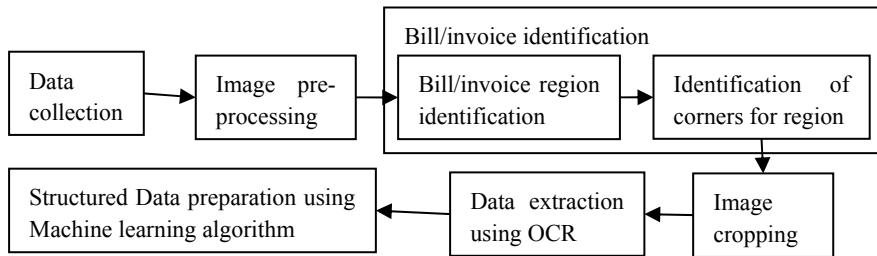
This section describes the process leading up to classify and extract information from receipts (typed bills and handwritten bills) using machine learning.

A text segmentation technique defines the first character in the line and then tries to interpret the whole line according to the position of the first character, and in the case of handwritten text, the flow is not within. And those characters aren't remembered well. If the picture includes an object that is not a bill or other rectangular piece of paper or entity, the job then classifies the entity and recognizes it as the bill required to retrieve the material. Further text analysis can be done to determine whether the collected information is from a bill.

The process of structured data extraction from the unstructured image data of the bill involves data pipeline that starts with collecting the receipts in the form of unstructured image data, converting the receipts into image format, correcting the image orientation, cropping the receipt to remove the background, running OCR to pull text from the image, and training machine learning algorithms to determine relevant data from the OCR result, writing a custom made algorithm to extract specific fields of data from the receipts that we are interested in and finally evaluating the classifiers performance. To eliminate the useless context information, the origin bill image is preprocessed to get grayscale image for to accelerate the processing. The image quality is calibrated to a better angle and view for easier retrieval of the data. Next, the region of the necessary information in the normal image obtained is extracted by matching the prototype, which is the center of identification of the information. Recognized optical character is used through subsequent use of the derived information to translate the image information into text. Initially, OpenCV is used to detect bills from the image, and to remove unnecessary noise from the image. Then, the Tesseract OCR engine is used to transfer intermediate images for further processing. Tesseract tries to apply Segmentation of Text to catch written text

**Table 1** Comparison of text extraction from bills

Paper	Metrics	Algorithm	Results
[3]	Convert binary text object into ASCII	OCR	Extract all kinds of blur images
[2]	Canny edge detector for image data identification	Open CV, Tesseract OCR	Extract all kinds of images, failed for handwritten bills
[1]	Number and letter segmentation	Tanimo to measure theory	Rate of numeral recognition 99% and letter recognition 92%
[4]	Detecting and filtering images with low quality	Definition evaluation algorithm based re-blur	Detect low-quality image and screen out undesirable image
[5]	Secondary rotation, rotate image, degree of rotation	OCR data is exerted as excel format	The accuracy of this method can be reached upto 95.92%
[6]	Various documents formats	Combination of heuristic filtering and OCR	Best result with an average accuracy of 92%
[7]	Block level image into individual characters	Softmax CNN classifier	Best result with an accuracy of 99.92%
[8]	Image accuracy for a selected text region	OCR based CNN	The accuracy rate is 95% for taken at a distance of 10 km
[9]	CNN to remove the visual portrayals, entity-aware network to decrypt EoIs	Entity –aware mechanism for feature extraction and LSTM as input	The accuracy for model trained over real data is about 95.8%
[10]	Graph Neural Network to define the table structure	graphlet discovery algorithm	Accuracy related to table detection is about 97%
[11]	Denomination based K-means clustering	Tesseract for text extraction framework	Accuracy rate is 93.3%
[12]	Bounding box is designed for specific elements	OCR	Accuracy after pre-processing is about 87.5%
[13]	FCNN and LSTM for text classification	Text detection Tesseract 4.0 and EAST detector	Best result with an accuracy 83.3%
[14]	Dense Nets and CTC for text conversion	Area extraction- YOLOv3	Average accuracy of recognition is about 0.96
[15]	Entity extraction by labelling Regexp patterns	OCR	Arabic text is cursive difficult to determine accuracy
[16]	BPN predict text or non-text area block wise	OCR	Recognition accuracy 99.24%
[17]	Column based character segmentation	Text recognition with OCR	Character recognition accuracy 89% and table recognition 98%



**Fig. 1** Data pipeline phases for extracting structured data from unstructured image data

in various fonts and languages. Figure 1 shows sequential phases of data pipeline for extracting structured data from unstructured image data. Each step can be executed in pipeline for multiple images as each phase can be independent.

**Phase 0: Data collection:** The data set we are using for this project is 40 healthcare bills. Those receipts are divided into 2 major categories, normal health care bills and covid health care bills. All the receipts are less than a year old. Healthcare bills are the biggest major categories with amount of variance while covid related healthcare bills are highest amount and hence the dataset have more mixed types of healthcare bills. We made sure that the backdrop wasn't too loud, and that the contrast between the receipt was good. The receipts were guided more or less vertically, without too much coercion than required. Nevertheless, most of the receipts were crumpled during transportation so that they had folds and creases and some had also washed out letters to make the job not so straightforward at all.

**Phase 1: Pre-processing Image:** We converted the images to grayscale that conveniently reduces the data by two-thirds and accelerates the subsequent processing process. We have normalized global illumination by using an old image processing trick to remove gradients of slow illumination: slow gradients lead to modulation of image intensity at low frequencies, so filtering the image with a high-pass filter render the global illumination clear. The filter is implemented effectively using the discrete cosine transform (DCT): convert the image into frequency space, cancel the low-frequency components and convert the image back into the space domain. Usage of DCT instead of the discrete Fourier transform avoids dealing with complex numbers.

**Phase 2: Healthcare bills identification:** The healthcare bills is usually a rectangular piece of paper, defining the area using a quadrilateral with the four-vertex  $\{p1, p2, p3, p4\}$ . The vertices are selected such that the polygon occupies as much of the receipt as possible and as little of the background as possible. The preprocessed images is to segment the image into pixels representing the receipt and those not. In particular, blur the image to eliminate noise and add a 60% resolution level for initial segmentation. There we apply binary closure to remove small false detections in the past and fill flaws along the contour. Finally, it discards all but the largest blob in the image. The process of bill identification is 6-step as follows. (1) Identify bill receipt based on

four-vertex polygon. (2) Convert image into grayscale using Gaussian blurring. (3) Remove noise and set threshold with 60% resolution. (4) Extract the features from the image using edge detection. (5) Measure the receipt end-points in the image. (6) Outline the end-points with green blobs.

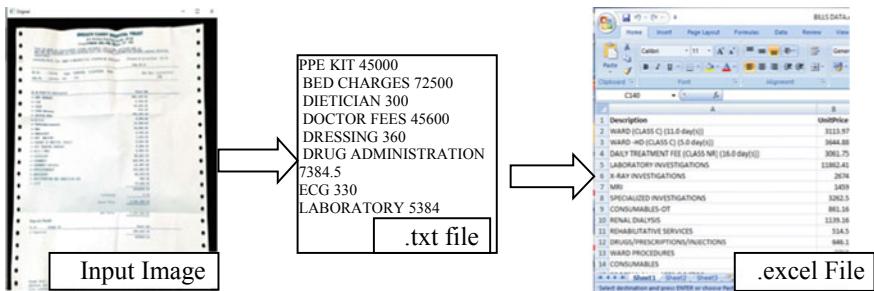
**Phase 3: Corner Identification:** Putting the vertices in the corners make the polygon clip. The lower left corner has a very obtuse angle which gives an impression that accurate corner detection cannot be as simple as internal angle calculation. In other cases, the receipt may be distorted, rounded or even corners caused by an irregular tear in odd locations. We depend on the receipt edges for a more robust solution. We measure the receipt outlines from the foreground mask (using binary morphology) and then apply a probabilistic transformation to get the start and end points of the line segments into the image. Instead, we calculate the intersection of each pair of horizontal and vertical segments (green blobs) to produce a list of corner candidates, which we reduce with average change (red crosses) to more than a reasonable number. We tried different methods of finding the receipts, each with varying success rates.

**Phase 4: Cropping, de-skewing and enhance:** Next steps are to strip it out of the picture and boost the contrast of what is written on the receipt. Both are standard image processing operations, and scikit-image provides transforming wrap using a mapping of pixel input and pixel output positions to do them in one step. We simply calculate the quadrilateral edge-lengths for the output shape and create a rectangle with width and height equal to the full length of the segments top and bottom and left and right, respectively. Gray thresholds and corresponding pixel filter, where we have removed all blobs that crossed the boundary of the pixel in the right direction. Threshold effectively suppresses many of the luminous variations while keeping the text intact. Nevertheless, several characters, particularly the small ones, are hard to distinguish in the binary image. Therefore by Gaussian blurring, we first feather the mask out and replace it with the original image (receipt). It essentially protects all hidden areas but with soft rather than sharp edges. The process of Cropping, de-skewing and enhance is 5-step as follows. (1) Load the necessary XML classifiers and load input images. (2) Image can be resized to 500 pixels, cropped, blurred, de-skew. (3) Convert image into grayscale using Gaussian blurring. (4) Image segmentation based on words, characters. (5) Extract the features from the image using edge detection.

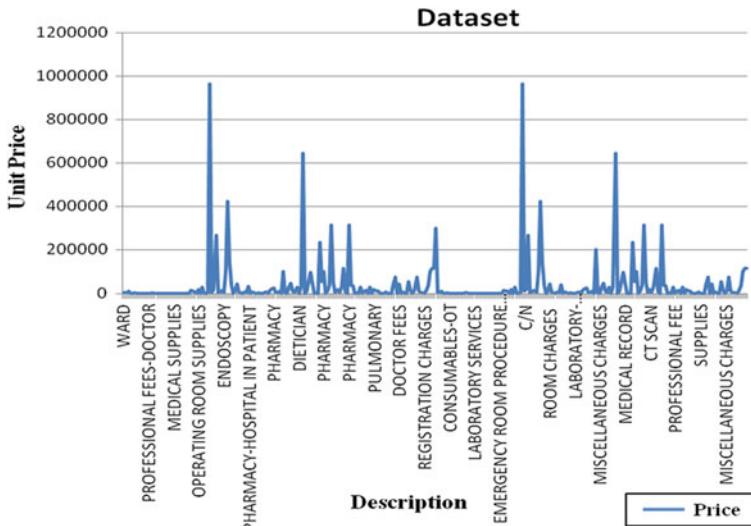
**Phase 5: Extraction of Data:** The most-free alternative is the Tesseract OCR engine. We used pytesseract fundamental, which is a simple wrapper around Tesseract: call `image_to_string` to transform the image to a single formatted string. Call `image_to_data` for assured recognition and other useful information for individual fragments of the text. The underlying OCR engine uses Long Short-Term Memories (LSTM) network. Tesseract architecture includes adaptive thresholding to convert image into binary images, component analysis to extract OCR image with a black background and white text, line/word finding using contours and blobs, recognize words using machine learning classifier two-pass process to extract text from image. We must strip out detections of low trust and those that are simply too poor to be text for

performance boosting. The default output is in a tab separate values (TSV) format. Pytesseract can automatically convert TSV into a data-frame using pandas. The data extraction process is 5-step process as follows. (1)Apply Tesseract OCR for image. (2) Differentiate word counters associated with image. (3) Differentiate letter counters associated with word counter image. (4) Preprocess letter images and TSV data is generated. (5) Consolidate associated data to text then to excel file as shown in Fig. 2.

**Phase 6: Data preparation:** Until feeding the data into any algorithm, there are several measures that need to be taken to prepare the data by transforming it in such structured sample data set used into this method is described as seen in Fig. 3. The vectorization and tokenization act is the principal steps of this process. We decided to differentiate these modules and left it this way to other preparation phases to make



**Fig. 2** OCR processing of proposed system



**Fig. 3** Dataset of healthcare bills

the testing and optimization as close as possible. We also tried to study whether we could use machine learning to retrieve the receipt from different data points. In this sense, when referring to a data point, it is assumed to be a particular form of token which contains a value that varies between transactions, in our case the unit price and definition. The token may contain different formatting or style in different documents in our solution, the only constraint that is required is that it is a special token that in most situations makes it ideal for extracting data points.

We used tools offered by scikit-learn libraries them to suit the classifiers' specific implementation advice. The process of extracting specific data points is very different from classifying the receipts. We could not use the same algorithms or classifiers for this task. We opted to use the following classification models for evaluation based on that text extraction. (1) Logistic Regression and (2) Kneighbors Classifier. We then continued with the best of those models for tuning and maximizing hyper-parameters before progressing with a second round of benchmarking. Evaluation to be able to achieve similar outcomes over different runs with the same random seed was used for the data shuffling process when comparing different algorithms. The benchmarking software was a basic python script that supplied the pipeline with a new classifier for a round of checking against the data collection. The classifier was then replaced by the next classifier, and it repeated the process.

To parse a given OCR output, words are first tokenized using sklearn libraries and then keywords for each given category identified. Once each keyword is identified, we ran a spatial search for all given text inputs nearest to the text input containing the keyword for insightful information. For example, for parsing pricing data, keywords such as "price", "total", "description" and "amount" are first identified from the OCR output. For each given positive keywords match, a nearest-neighbor search is conducted to look for text bounding boxes containing pricing information. The keyword-price pair is selected for the bounding boxes that are furthest down on the page. Once the unit-price and description information are found, this information is returned as output and a plot containing the individual images generated during the pipeline combined with the parsed information.

The extracted text data set is divided into a training set (80%) and a test set (20%). Both sets are divided into two parts, one part containing the OCR extracted receipt text and the other one containing the value for the data point referred to as the label. The model has access to the labels during training to know what it is looking for while the prediction function only has access to the document text. The labels for the test data is instead used after the predictions have been made by the model to compare against. The Fig. 3 is a graphical representation of the actual dataset.

## 4 Results Analysis

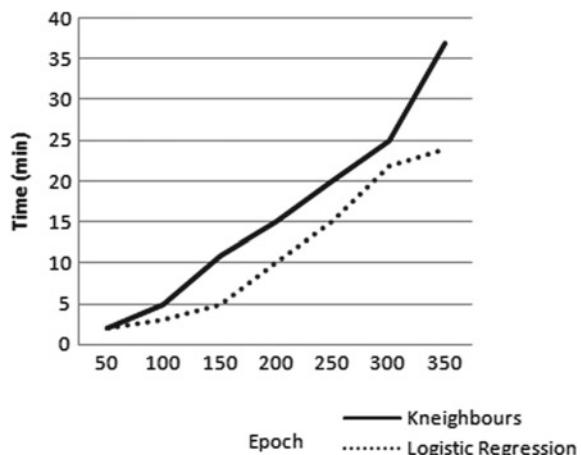
The development environment used for this project includes PYTHON version 3 as language to develop the software in order to meet the project requirements. For the initial development and exploration, we have used SPYDER as an IDE and then later

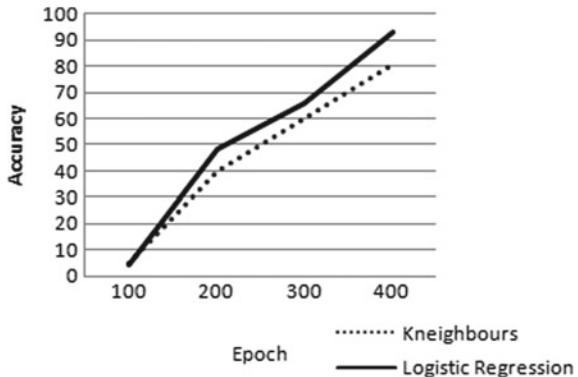
we have shifted to PYCHARM, which supports Python and makes the task completion ease. Various other libraries and pre-trained models used include Pytesseract, numpy, imutils, cv2, pandas, sklearn, etc. These Programming algorithms have well defined models that are trained with some samples to increase the accuracy of the result. The result depends on the amount of training given to the model. The receipts accuracy statistic is measured as the number of correct tokens divided by the overall tokens count. The cumulative number of tokens in the record generated by the OCR algorithm is the sum of tokens in it. This is because noise is seen incorrect tokens generated by the OCR algorithm. The classification efficiency metric is the mean of the product of carrying out a cross-validation test of three folds on the data collection. While relating to the consistency of going forward algorithms it is the mean value of this cross validity to which it applies. The accuracy increases by a small margin after running the optimization test. The default values were found to be strongly efficient, and accurately identified the gap in precision between the optimized algorithms. In the case of KNeighbours, the variance area corresponds to the difference in time being expressed in Fig. 4 against the Logistic Regression while Fig. 5 represents the accuracy. The final results for the accuracy and time is presented in Table 2.

The final results for the accuracy and time is presented in Table 2.

With approximately 93% precision, we find this to be very fair though bearing in mind the relatively limited data collection analyzed. With the assessed model, subcategories which are much smaller subsets with less variation between categories also have very high accuracy. There were several classifiers managed to score 81% or more in the cross-validation test, which may mean that classifying receipts is a very simple task that machine learning is very suitable to start with, particularly because the optimization of the hyper parameter did not see any major improvements and compared to the Kneighbors we outperformed them with great margins. We assume these findings are very positive, but as these findings derive only from principle solution verification, a wider analysis will be required to validate them.

**Fig. 4** Algorithm validation time



**Fig. 5** Algorithms accuracy**Table 2** Algorithm accuracy and time

Algorithm	Accuracy (%)	Time (min)
Logistic regression	93	24.3
KNeighbours	81	37.5

This is an experiment requiring estimation of a numerical value. Running the python script describes the data set structure first, then matches the pattern and tests it on the test dataset. Finally, on Logistic Regression we got 93% accuracy and on KNeighbours 81% accuracy. This was achieved with a basic custom-made approach that has tremendous room for development, and predicted to do even better with more complex solutions. As such, we find these findings to be very positive and they could require time and work to determine a more complicated approach.

## 5 Conclusion

The paper describes the analysis of different text extraction methodologies based on Machine learning algorithms. The research is to separate the content from imprinted invoice. The downside of the inspection is that, even though the picture consist a document that is not a bill, a rectangular slip to paper or an object, the inspection will recognize the object and find it to be necessary bill, the substance of which will be deleted. The applied algorithm is Logistic Regression and the KNeighbours for the extracted information to predict the accuracy. The sample images taken for implementation and the accuracy observed is approximately 93% for Logistic Regression and 81% for KNeighbours. The effort is to make people understand their bills and to bring about some level of transparency in the market. We think these results are very promising, but as these results only stem from a proof of concept solution, a larger study might be needed to verify them. The findings of this work indicate that there are many ways to enhance the efficiency of automated processing of receipts. The

program will be successful in detecting bills that are mutilated because experiment findings suggest that there are reasonably strong outcomes from local thresholding techniques. The binarization methods require adaptation of the brightness. This translation is particularly necessary for the identification of characters. We need to analyze the speed-up of goal character recognition.

**Acknowledgements** We are grateful to Vadivel Karuppantan and Shailesh Prabhu of VB Ideas Private Limited, Bangalore for their assistance in problem formulation and periodic reviews.

## References

1. Umam, A., Chuang, J.-H., Li, D.-L.: A Light Deep Learning Based Method for Bank Serial Number Recognition (2018)
2. Rahmat, R.F., Gunawan, D., Faza, S., Haloho, N., Nababan, E.B.: Android-Based Text Recognition on Receipt Bill for Tax Sampling System. Department of Information Technology, Universitas Sumatera Utara, Medan, Indonesia (2018)
3. Sidhwa, H., Kulshrestha, S., Malhotra, S., Virmani, S.: Text extraction from bills and invoices. In: International Conference on Advances in Computing, Communication Control and Networking, pp. 564–568 (2018)
4. Mizan, C.M., Chakraborty, T., Karmakar, S.: Text recognition using image processing. *Int. J. Adv. Res. Comput. Sci.* **8**(5), 765–768 (2017)
5. Song, W., Deng, S.: Bank bill recognition based on an image processing. In: 2009 Third International Conference on Genetic and Evolutionary Computing, pp. 569–573 (2009)
6. Jiang, F., Zhang, L.-J., Chen, H.: Automated image quality assessment for certificates and bills. In: 1st International Conference on Cognitive Computing, pp. 1–5 (2017)
7. Sun, Y., Mao, X., Hong, S., Xu, W., Gui, G.: Template matching-based method for intelligent invoice information identification. In: AI-Driven Big Data Processing: Theory, Methodology, and Applications, vol. 7, pp. 28392–28401 (2019)
8. Holt, X., Chisholm, A.: Extracting structured data from invoices. In: Proceedings of Australasian Language Technology Association Workshop, pp. 53–59 (2018)
9. Guo, H., Qin, X., Liu, J., Han, J., Liu, J., Ding, E.: EATEN: Entity-aware Attention for Single Shot Visual Text Extraction. Department of Computer Vision Technology (2019)
10. Riba, P., Dutta, A., Goldmann, L., Fornes, A., Ramos, O., Lladós, J.: Table Detection in Invoice Documents by Graph Neural Networks. Computer Vision Center (2018)
11. Abburu, V., Gupta, S., Rimitha, S.R., Mulumani, M., Koolagudi, S.G.: Currency recognition system using image processing. In: Tenth International Conference on Contemporary Computing, pp. 10–12 (2017)
12. Raka, P., Agrwal, S., Kolhe, K., Karad, A., Pujeri, R.V., Thengade, A., Pujeri, U.: OCR to read credit/debit card details to autofill forms on payment portals. *Int. J. Res. Eng. Sci. Manag.* **2**(4), 478–481 (2019)
13. Kopeykina, L., Savchenko, A.V.: Automatic privacy detection in scanned document images based on deep neural networks. In: International Russian Automation Conference (2019)
14. Meng, Y., Wang, R., Wang, J., Yang, J., Gui, G.: IRIS: smart phone aided intelligent reimbursement system using deep learning. In: College of Telecommunication and Information Engineering, vol. 7, pp. 165635–165645 (2019)
15. Rahal, N., Tounsi, M., Benjlaiel, M., Alimi, A.M.: Information extraction from Arabic and latin scanned invoices. In: IEEE 2nd International Workshop on Arabic and Derived Script Analysis and Recognition, pp. 145–150. University of Sfax (2018)

16. Umam, A., Chuang, J.-H., Li, D.-L.: A Light Deep Learning Based Method for Bank Serial Number Recognition. Department of Computer Science (2018)
17. Xu, L., Fan, W., Sun, J., Li, X.: A Knowledge-Based Table Recognition for Chinese Bank Statement Images, pp. 3279–3283. Santoshi Naoi Fujitsu Research and Development Center, Beijing (2016)

# Parallel Enhanced Chaotic Model-Based Integrity to Improve Security and Privacy on HDFS



B. Madhuravani, N. Chandra Sekhar Reddy, and Boggula Lakshmi

**Abstract** This article presents how to provide security to large chunk of data. Data be the crucial in all areas, is to be protected during storage and retrieval. The large volumes of data called Big Data will be stored in HADOOP in HDFS file systems. This research helps in understanding the Big Data and provides a model which improves security and efficient storage of data onto HDFS. This paper uses an enhanced Quadratic Chaotic Map in the process of generating keys to improve integrity. The model is implemented in parallel manner to reduces time in the process of verifying integrity. The model is implemented and tested with HADOOP Cluster which improved security and minimized time.

## 1 Introduction

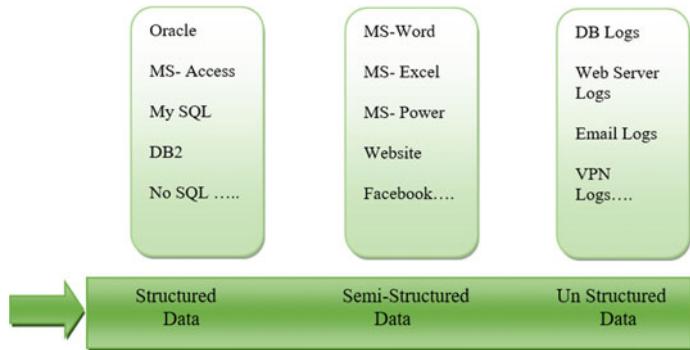
Data being fundamental in all areas be protective and manageable. As data is increasing day by day with emerging trends Internet of Things IoT, it should be processed and managed in efficient manner. The emerging technology in Computer Science field, the Big Data manage data sets whose size is impractical and unimageable. The data can be in various forms like structure, semi-structure and un structure (Fig. 1) [1].

### 1.1 *Data-Information-Knowledge-Wisdom-Decision (DIKWD) Pyramid*

The computer systems are processing machines for data which accepts input, stores data in disks, process and produces result in the form of output. The data can be of different representations, which is given as DIKWD Pyramid (Fig. 2) [2].

---

B. Madhuravani (✉) · N. Chandra Sekhar Reddy · B. Lakshmi  
Department of Computer Science and Engineering, MLR Institute of Technology, Dundigal,  
Hyderabad, Telangana, India



**Fig. 1** Types of data

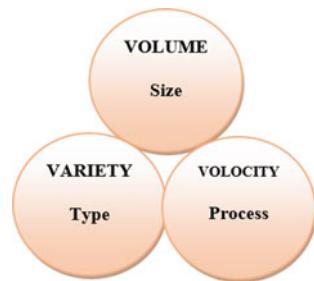
**Fig. 2** DIKWD pyramid



Data is un-processed, un-organized and un-structured data. Which is available as raw. Where information is processed and structured data. Knowledge is an information gained through study and analysis. The experience and knowledge helps to make wise decisions.

## 1.2 Big Data

Big Data—large volumes of data, an emerging trend in the field of computer science. In wireless sensor networks in real world are connected with large number of devices and storing and producing of large volume of data. Hence there is a large demand in storing and processing of large data. One solution to this is Big Data. Hence there is a need for parallel processing of data and also the security issues are boosted up for Bid Data.

**Fig. 3** Three V's

The Big Data characteristics are defined as Three V's (Fig. 3) [3]—V-Volume, V-Variety, and V-Velocity.

### 1.3 Hadoop

A solution to Big Data problem is Hadoop (Fig. 4) [4]. It's the combination of Map Reduce and HDFS. Where Map Reduce does processing and HDFS used for storage.

## 2 Enhanced Quadratic Map

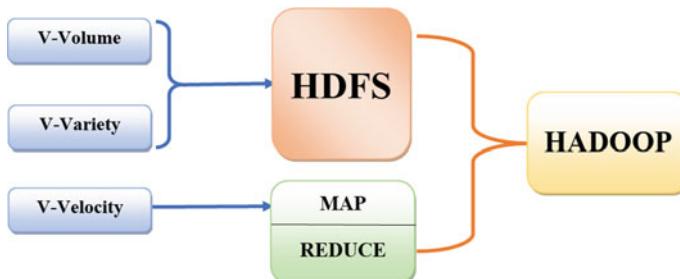
The general equation for Quadratic Map [5, 6] is given as

$$x_{n+1} = px_n^2 + qt$$

Let  $p = 1$  and  $q = 1$

$$x_{n+1} = x_n^2 + t$$

$$x^{(1)} = [x^{(1)}]^2 + t$$

**Fig. 4** HADOOP

$$\begin{aligned}
(x^{(1)})^2 - x^{(1)} + t &= 0 \\
x_2^{(1)} &= 0.5(1 \pm \sqrt{1 - 4t}) \\
x_{n+2} &= x_{n+1}^2 + t \\
x_{n+2} &= (x_n^2 + t)^2 + t \\
x_{n+2} &= x_n^4 + tx_n^2 + tx_n^2 + (t^2 + t) \\
x_{n+2} &= x_n \\
x^4 + 2x^2 - x + (tx^2 + t) &= (x^2 - x + t)(x^2 + x + 1 + t) = 0 \\
x_y^{(2)} &= 0.5 * (1 \pm \sqrt{-3 - 4t})
\end{aligned}$$

Let  $p = -x$   $q = 1$   $t = s$

$$x_{n+1} = px_n^2 + qt$$

$$\begin{aligned}
x_{n+1} &= s - tx_n^2 \\
s \in (0, 2) \text{ and } t \in (0, 1)
\end{aligned}$$

```

Algorithm Enhanced_Qudratic(upper_bound,lower_bound)
{
    for 1 to n do
    {
        Calculate point = (Random_Value * (upper_bound-
                                             lower_bound)) + lower_bound;
        s = point;
        t = Random_Value;
        x = Random_Value;
        plotting = s - power(x,2);
    }
}

```

Plotting data is presented in Table 1 and its corresponding plotting is given in Fig. 5.

### 3 Proposed Model

#### 3.1 Big Data Security System

The Big Data Security system [7] is depicted in Fig. 6. The model takes input from different big data sources. Process the input data in storing and retrieving from HDFS. The HDFS uses privacy and security services to overcome several security challenging issues.

**Table 1** Plotting data

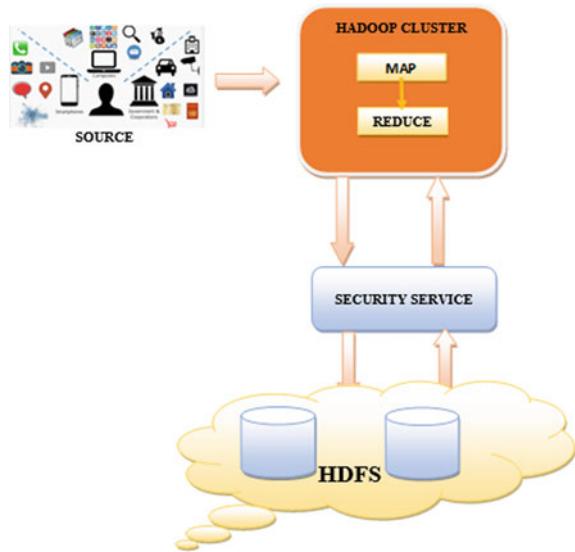
$X$	$Y$								
1	0.232875	21	0.058198	41	0.195583	61	0.024572	81	0.072284
2	0.166929	22	0.080503	42	0.068599	62	-0.06414	82	0.339254
3	0.025129	23	0.31357	43	0.096259	63	0.312608	83	0.414752
4	0.344497	24	0.294245	44	0.278975	64	0.23461	84	0.072936
5	0.215562	25	0.001356	45	-0.10992	65	0.021318	85	0.236616
6	0.162311	26	0.217931	46	0.361283	66	-0.22555	86	0.345707
7	0.047713	27	0.410416	47	0.057775	67	0.136593	87	0.177242
8	0.209773	28	0.040958	48	0.494532	68	0.362841	88	0.152431
9	0.26003	29	0.289463	49	-0.06218	69	0.124408	89	0.322687
10	0.099767	30	0.123042	50	0.291799	70	0.153504	90	0.298568
11	-0.18155	31	0.301667	51	-0.00484	71	0.108776	91	0.216872
12	0.290262	32	0.004594	52	0.13529	72	0.45248	92	0.110915
13	0.268044	33	-0.10173	53	0.275352	73	-0.17845	93	0.165477
14	-0.16444	34	-0.03964	54	0.162848	74	0.063284	94	0.335702
15	0.002755	35	0.470395	55	0.139831	75	0.301267	95	0.056919
16	0.389281	36	-0.08801	56	-0.05347	76	0.178351	96	0.275423
17	-0.02248	37	-0.03638	57	0.117242	77	0.36466	97	0.285088
18	0.259248	38	0.247029	58	-0.07157	78	-0.08031	98	0.065591
19	0.282307	39	0.301248	59	-0.13844	79	0.116092	99	0.102762
20	-0.05414	40	0.257316	60	0.367234	80	-0.08216	100	0.295698

**Fig. 5** Enhanced quadratic map plotting

### 3.2 Proposed Parallel Processing Authentication Model

The proposed authentication model uses the enhanced quadratic chaotic in the generation of pseudorandom numbers based on system/node identity. The data from Big Data Sources will be stored and retrieved from HDFS using proposed authentication model, where the storage and processing can be performed parallelly which in turn improves the security and storage efficiency [8] (Fig. 7).

**Fig. 6** Big data security system



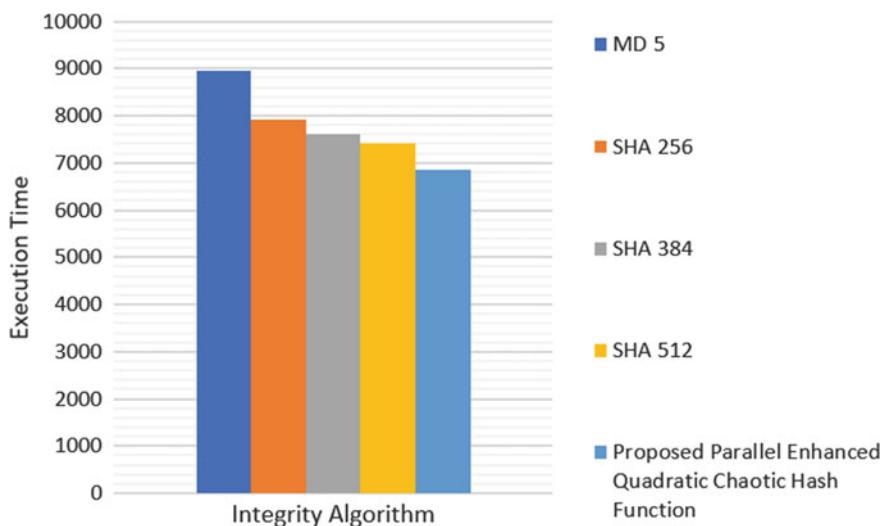
**Fig. 7** Proposed parallel processing authentication model

#### 4 Experimental Results

The system is implemented and analyzed in Java platform for comparative analysis in terms of time taken to perform and verify the integrity with traditional hash functions and proposed quadratic chaotic hash function. The results (Table 2, Fig. 8) proved the proposed model is taking less time for integrity verification, in turn, improves the life time of sensor nodes and improves energy.

**Table 2** Comparative analysis

Integrity algorithm	Execution time (ms)
MD 5	8963
SHA 256	7924
SHA 384	7612
SHA 512	7419
Proposed parallel enhanced quadratic chaotic hash function	6854



**Fig. 8** Comparative analysis in terms of execution time

## 5 Conclusion

This paper presented a model which improved security and lifetime through parallel enhanced quadratic chaotic hash model. Obviously, the execution time of the system is improved with parallel processing approach. The files should encode before storing data into HDFS. The experimental results proved that the system is efficient in terms of execution time.

## References

1. <https://www.datamation.com/big-data/structured-vs-unstructured-data.html>
2. [https://en.wikipedia.org/wiki/DIKW\\_pyramid](https://en.wikipedia.org/wiki/DIKW_pyramid)
3. <https://www.zdnet.com/article/volume-velocity-and-variety-understanding-the-three-vs-of-big-data/>

4. Hadoop, W.T.: The Definitive Guide: The Definitive Guide. O'Reilly Media (2009), 2. Borthakur, D.: HDFS architecture guide. HADOOP APACHE PROJECT <https://hadoop.Apach>
5. Kanso, A., Yahyaoui, H., Almulla, M.: Keyed hash function based on a chaotic map. *Inf. Sci.* **186**, 249–264 (2012)
6. Madhuravani, B., Murthy, D.S.R.: A hybrid parallel hash model based on multi-chaotic maps for mobile data security. *J. Theor. Appl. Inf. Technol.* **94**(2), (2017). ISSN: 1992-8645
7. Sai Prasad, K., Chandra Sekhar Reddy, N., Rama, B., Soujanya, A., Ganesh, D.: Analyzing and predicting academic performance of students using data mining techniques. *J. Adv. Res. Dyn. Control Syst.* **10**(7), 259–266 (2018)
8. Chandra Sekhar Reddy, N., Chandra Rao Vemuri, P., Govardhan, A., Navya, K.: An implementation of novel feature subset selection algorithm for IDS in mobile networks. *Int. J. Adv. Trends Comput. Sci. Eng.* **8**(5), 2132–2141 (2019). ISSN 2278-3091

# Exploring the Fog Computing Technology in Development of IoT Applications



Chaitanya Nukala, Varagiri Shailaja, A. V. Lakshmi Prasuna, and B. Swetha

**Abstract** In the 21st era, IoT is assuming a significant part in creating Smart urban communities. With the development of IoT, information is developing with increasing speed. As the information is developing the need to store information is likewise expanding. More the information, the dormancy will be high to store and recover information from the cloud. The idea of mist processing was started to reduce the inertness for getting to information to and from the cloud. Haze processing gives the capacity, figuring just as systems administration administrations toward the end purpose of the system. Haze hubs likewise have restricted computational abilities. Because of certain shortcomings, haze figuring and distributed computing can't continue alone, so both these advances are coordinated to fabricate keen IoT foundation for Smart city. Mist figuring have a significant job and preeminent duty being developed of a Smart city. This paper examines different utilization of mist registering and their usage in Smart urban areas. It additionally proposes a model for Waste administration framework in a city. Mist figuring can assist with overseeing the waste assortment of the city in a keen manner. Based on our survey, a few open concerns and difficulties of mist processing are examined, and the bearings for future analysts have additionally been talked about.

---

C. Nukala (✉)

Department of CSE, RGM College of Engineering and Technology (Autonomous), Nandyal, Andhra Pradesh 518501, India

V. Shailaja

Department of IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Telangana 500090, India

A. V. Lakshmi Prasuna · B. Swetha

Department of IT, Mahatma Gandhi Institute of Technology, Gandipet, Hyderabad, Telangana 500075, India

## 1 Introduction

A creating number of physical articles are being related with the IoT [1]. It is the interconnection of different physical elements that pass on and exchange data the sensors, shrewd meters, telephones and vehicles, radio-recurrence distinguishing proof (RFID) labels, and actuator [2]. The interconnection of these devices enables shrewd IoT applications like following on the web trucked merchandise, condition observing, medical services keen home and savvy framework, and so forth.

IoT devices make a great deal of data, which procure tremendous figuring office, stockpiling zone, and correspondence information move limit. Cisco said 50 billion gadgets could be associated by Internet in 2020 [3], and it will build 500 billion by 2025 [4].

The idea of mist figuring was presented by Cisco in the year 2012 [5]. The beginning objective of haze registering is to upgrade the profitability and to diminish volume of information which is moved to cloud for preparing. For ease, the board of all assets haze layer go about as a middle among gadgets and cloud server farms. Haze processing can offer types of assistance in different zones i.e. observation, transportation division, clever urban areas, medical services, and keen structures. Mist figuring can be utilized in various kinds of IoT administrations [5–7]. To start with, Smart E-Health Gateway can be utilized for patients to checking their wellbeing status [8]. Crisis caution can be enacted and send the alerts to the proprietor [9]. A haze based Electronic Data Interchange (EDI) is an exhaustive virtualized device outfitted with the capacity, transmission and registering limit [5].

Further segments of the article are clarified as follows: Sect. 2 contains Motivation of the investigation, Sect. 3 explains the Layered design of IoT and mist processing and how it can function for keen utilizations of IoT, Sect. 4 examines uses of IoT and haze registering, Sect. 5 clarifies load adjusting in mist figuring condition, Sect. 6 examine the proposed philosophy, Sect. 7 has been regarding open issues and supportive gestures lastly Sect. 8 finish up the paper and clarifies the opportunity of Future.

## 2 Motivation

This manuscript provides gives a short conversation of the application territories of IoT and FC. The fundamental ideas of IoT and FC are talked about that incorporates points of interest, hindrances, and engineering of FC.

Mist layer requires load adjusting to accomplish asset productivity, stay away from over-burden in the system, to improve framework execution, and furthermore to ensure the framework against disappointments. The inspiration of this paper is to investigating the keen waste administration framework which incorporates load adjusting on the haze layer. This paper likewise talks about different open issues and difficulties looked in mist conditions.

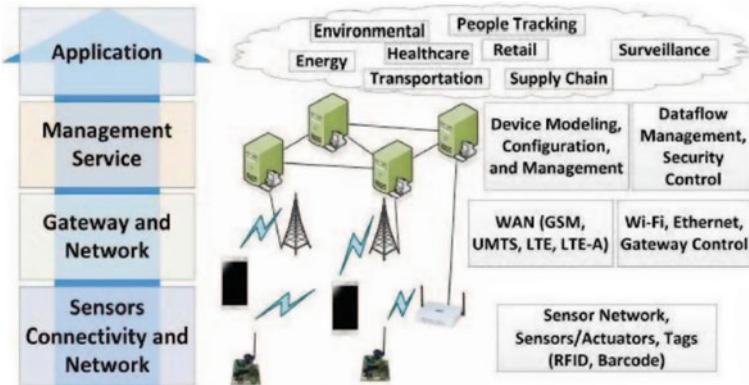


Fig. 1 Layers of IoT architecture [10]

### 3 Layered Framework of IOT and FC

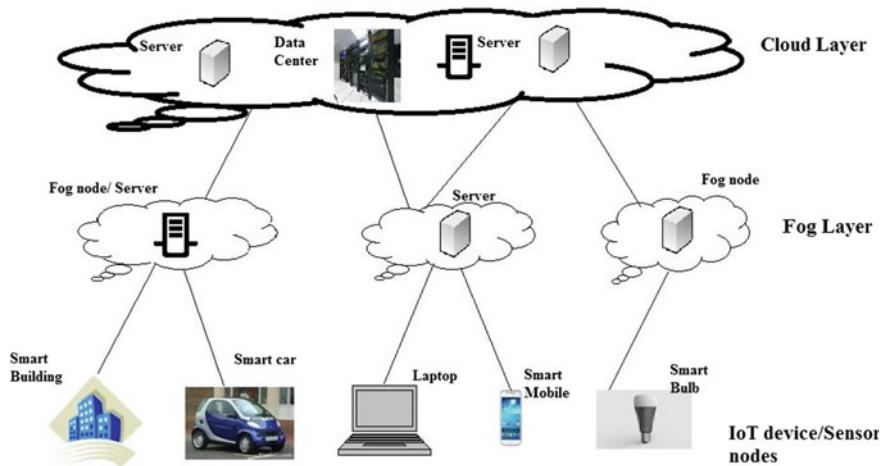
#### 3.1 Layered Framework of IOT

IoT has been arrangement of devices that send, share, and use data from the physical condition to offer types of assistance to individuals, endeavors, and society. The essential three-layer engineering is appeared in Fig. 1 [10].

- Layer of Sensors: This detect and collecting information from nature. Sensors, scanner tag marks, RFID labels, GPS, camera, and actuator are available in this layer
- Layer of Network: This utilized to assemble the information from sensor and sends to the web. Liable for organize layer is interfacing with other savvy things, arrange gadgets, and workers. Its features are moreover used for communicating and handling sensor data. Utilizing various advancements, different kinds of conventions and heterogeneous systems are accumulated.
- Layer of Middleware: This gets information from layer Network. Its inspiration is administration the board, information stream the executives, and security control. It moreover performs information taking care of and takes decisions normally taking into account results.
- Layer of Application: It gets information from the Middleware layer and gives overall administration of the application.

#### 3.2 Layered Architecture of FC

The class of CC for example FC have three layer engineering. The lower layer contains IoT gadgets. Mist layer is the center layer. IoT gadgets are coupled to cloud



**Fig. 2** Architecture of Fog computing

layer through mist layer. Entire gadgets store their information on the cloud. The haze layer channels information, and the information which isn't quickly needed is diverted to the cloud. The every now and again got to information is put away on the mist layer. Layered design of FC is clarified as beneath: Fig. 2 shows the engineering of FC.

- **Brilliant IoT gadgets:** The a great many sensors hubs and implanted frameworks having low transfer speed and low inertness are utilized at this layer. Savvy gadgets like brilliant structures, advanced cells, workstations, shrewd power bulbs, keen vehicles, and so forth can be considered as IoT gadgets which gather the information and send this information to the haze layer [14].
- **Haze Layer:** Network layer of haze is additionally partitioned into two sections: Fog system and Core organize.

**Network of Fog:** It incorporates 3G/4G/5G/LTE/Wi-Fi and so forth multi-edge benefits that are utilized to interface distinctive detecting gadgets with the haze hubs. Haze hubs are utilized to channel information assembled by method of IoT gadgets and not regularly utilized information is diverted to the upper layer for example cloud layer.

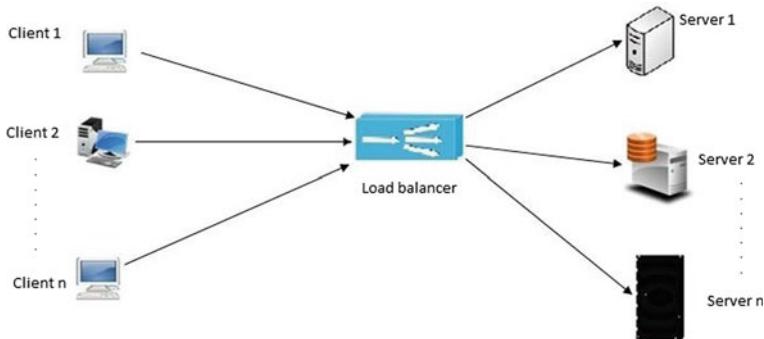
**Network of Core:** QoS, manifold protocol Label Switching (MPLS), manifoldcast, and security were deliberated at this phase [11].

- **Layer of Cloud:** It incorporates a great deal of server farms and cloud facilitating IoT investigation. The colossal information accumulated through various IoT gadgets are put away in the huge server farms situated at different areas on the planet.

## 4 Applications of FC and IOT

Haze works in dispersed condition. It offers types of assistance to the last client at the edge gadget. FC have different application regions where we can incorporate with IoT that can be examined as follows:

- a. **Meticulous Healthcare:** In view of dirtied condition different sorts of microscopic organisms have been being spread noticeable all around which causes different ailments.  
Each individual has occupied today as a result of quick ways of life. Shrewd healthcare has the brilliant IoT which monitors exercises of individuals and measures different boundaries of their body and continues transferring the information on the haze hubs, which have been being seen by the specialists. The information put away on the haze hubs have been being utilized by specialists to treat the patients inside time. The individuals have been wearing insightful gadgets and these have been additionally connected to mist hubs which have been ceaselessly sending the estimations of body boundaries (temperature, pulse, and so on.) so as to the mist hubs. These wise wearable gadgets help to monitor individuals' wellbeing [7].
- b. **Meticulous Parking:** Because of much increment in transportation in the urban communities, all the more parking spots have been required. Individuals need to meander to a great extent to locate the fitting parking spot for them. FC presented a novel thought of Meticulous stopping. With the utilization of FC the stopping spaces can be introduced with the sensors which continue following climate the parking have been a has vacant or full [12].
- c. **Meticulous Agriculture:** Agribusiness has the wide territory to be given consideration since has zone has from where all the urban communities have been getting food. Brilliant agribusiness idea has been introduced in most recent couple of years. Shrewd detecting and figuring have been playing a critical obligation in keen agribusiness. A couple of savvy agribusiness approaches have been imagined around there. Brilliant water system frameworks have been given. Shrewd sensors have been being repaired in the fields [13]. FC gives a stage to working the sensors in the fields, which have been ceaselessly watching the yields. The sensors identify necessities of the harvests and persistently store information in the haze hubs which send the cautions to the ranchers about the prerequisites of the yields. The shrewd farming assumes an essential function for building a brilliant city.
- d. **Meticulous Waste administration:** The earth has corrupting step by step, and to moderate the planet we require sharp consideration towards the normal assets. Presently days, day by day developing waste and water exhaustion from the earth looks for more consideration. Keen trash the board framework can be named the answers for improvement of condition in this time. Such keen frameworks will be created with the assistance of cloud just as FC to gather and deal with the waste all the more proficiently [8].



**Fig. 3** Balancing of load in network

## 5 Load Balancing in FC Environment

In a system, scarcely any frameworks stay under-stacked sooner or later stretch, while the others convey the whole heap of the system. To keep up the heap in a reasonable plan, “Burden Balancing” gets vital. “Burden Balancing endeavors to disperse the heap in indistinguishable extents all through assets relying upon response capacity all together that each valuable asset isn’t over-burdened or underutilized in a cloud device” [14]. Burden adjusting additionally needed to be done to dodge halt and diminish the worker flood issue. Figure 3 shows the heap adjusting.

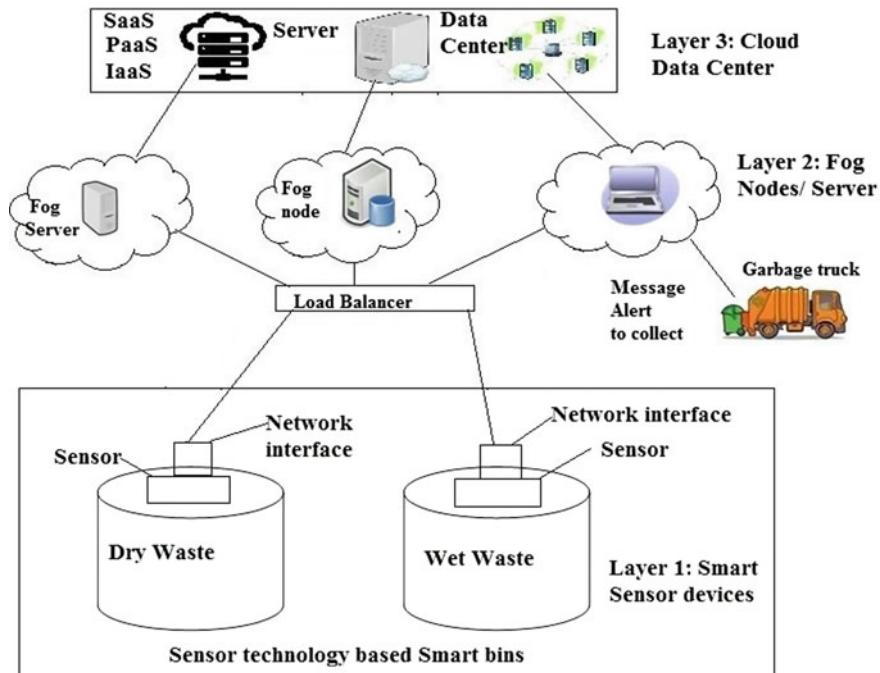
A portion of the objectives of mist based burden adjusting are talked about as underneath:

- In instance of failure of system, the balancing of load offers plans for backup.
- Consequently, performance could be enhanced.

## 6 Projected Methodology

The novel waste administration network has been projected in this manuscript by modifying the load at the layer of mist. Here, projected method comprises of three phases: mist, datacenters of cloud, gadgets related to sensor. It could be stated that comprise of sensors. Moreover, these canisters smart would tied up further over the layer of mist that data of channels have to be transferred towards cloud. The balancer of burden would adapt mist hubs heap. Moreover, it might distinct the similar heap on entire hubs.

These hubs of haze would readily create the messages for advising the transporters of trash for waster gathering. The sensors of security would established in canisters that lighten regarding receptacles. Further, in instance, anybody might try for canisters tempering, where these sensors would start signal clamoring that might spare by considering receptacles. Further, the containers smart would be linked by 3G or



**Fig. 4** Load balancing model

4G towards layer of mist. The application would be formed for contributing overall circumstance. The ensuing Fig. 4 shows the proposed model of burden adjusting framework.

## 7 Open Issues and Encouragements

There were divergent open issues, which could be worked in an addition. Here, accompanying could be examined further:

- Interaction among hubs: further, it might be examined in addition to explore the engineering by which hubs of haze might interact over one another.
- Recognition devices: the gadgets that are smart were exorbitant than another fundamental market components. Here, these could be deliberated deprived of any issue.
- It could be a significant issue in this contemporary world. FC pre-requisite more computations of security for real-time practice.
- It shall be in condition of mist. As it causes challenges when pair of workers have been loaded very much, while other workers would be stacked beneath.

- Effectiveness of energy: The effectiveness of energy could a prominent in FC test. The mist utilizing force has to be lessened.

## 8 Conclusion and Future Scope

The objective of this manuscript portrays joining of IoT and FC for assisting divergent implementations. What's more, load adjusting has been proposed. Next to, a couple of uses, including the brilliant farming, keen medical services, shrewd stopping, and savvy squander the executives showed. The basic purpose behind this survey is to give a significant perception and preferences of IoT and its fuse with mist/edge processing and how burden adjusting should be possible when FC incorporated with IoT. There are still more open regions for future analysts, for example, Haze organizing, asset provisioning, greater headway in transportation.

## References

1. Al-Fuqaha, A., Guizani, M., Mohammadi, M., Aledhari, M., Ayyash, M.: Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun. Surv. Tutor.* **17**(4), 2347–2376 (2015)
2. Gubbia, J., Buyya, R., Marusic, S., Palaniswamia, M.: Internet of things (IoT): a vision, architectural elements, and future directions. *Futur. Gener. Comp. Syst.* **29**(7), 1645–1660 (2013)
3. Evans, D.: The Internet of Things: How the Next Evolution of the Internet is Changing Everything. Cisco White Paper (2011)
4. Camhi, J.: Former Cisco CEO John Chambers Predicts 500 Billion Connected Devices by 2025. Business Insider (2015)
5. Luan, T.H., Gao, L., Xiang, Y., Li, Z., Sun, L.: Fog Computing: Focusing on Mobile Users at the Edge. arXiv preprint [arXiv:1502.01815](https://arxiv.org/abs/1502.01815) (2015)
6. Emilia, R., Naranjo, P.G.V., Shojafar, M., Vaca-cardenas, L.: Big Data Over SmartGrid—A Fog Computing Perspective Big Data Over SmartGrid—A Fog Computing Perspective (2016)
7. Khan, S., Parkinson, S., Qin, Y.: Fog computing security: a review of current applications and security solutions. *J. Cloud Comput.* **6**(1) (2017)
8. Mahmud, R., Kotagiri, R., Buyya, R.: Fog Computing: A Taxonomy, Survey and Future Directions, pp. 1–28 (2016)
9. Desikan, K.E.S., Srinivasan, M., Murthy, C.S.R.: A novel distributed latency-aware data processing in fog computing—enabled IoT networks. In: Proceedings of the ACM Workshop on Distributed Information Processing in Wireless Networks—DIPWN’17, pp. 1–6 (2017)
10. Chi, Q., Yan, H., Zhang, C., Pang, Z., Xu, L.D.: A reconfigurable smart sensor interface for industrial WSN in IoT environment. *IEEE Trans. Ind. Inform.* **10**(2) (2014)
11. Chiang, M., Zhang, T.: Fog an IoT: an overview of research opportunities. *IEEE Internet Things J.* **3**(6), 854–864 (2016)
12. Perera, C., Qin, Y., Estrella, J.C., Reiff-Marganiec, S., Vasilakos, A.V.: Fog Computing for Sustainable Smart Cities: A Survey (2017)

13. Varshney, P., Simmhan, Y.: Demystifying Fog Computing: Characterizing Architectures, Applications and Abstractions. In: Proceedings—2017 IEEE 1st International Conference of Fog Edge Computing ICFEC, pp. 115–124 (2017)
14. Dastjerdi, A.V., Gupta, H., Calheiros, R.N., Ghosh, S.K.: Fog Computing: Principles, Architectures, and Applications, pp. 1–26 (2016)

# NavRobotVac: A Navigational Robotic Vacuum Cleaner Using Raspberry Pi and Python



Shaik Abdul Nabi and Mettu Krishna Vardhan

**Abstract** Human life is becoming more advanced day by day and the use of technology is making our lives easier. By this consideration, one of our daily task is to keep our surroundings clean by natural way using a broomstick, as technology is increasing day by day we came with manual vacuum cleaners and then moved to robotic vacuum cleaners. In robotic vacuum cleaner there exist many different methods which are used to clean in home or hotels. In NavRobotVac (Navigational Robotic Vacuum cleaner) model, we consider the science of architecture (vastu) and implemented using python programming with Raspberry Pi. It automatically scans area around and starts cleaning. In this model, IoT sensors are used to read real-time data thus it gives technical efficiency and economic efficiency.

## 1 Introduction

Vacuum cleaners were developed in mid-1920 and then we came up with different designs and features [1]. In 1996 the 1st robotic vacuum cleaner was developed using programming. NavRobotVac cleans your home even when you are out of your home, reduces work, and saves more time for the household [2]. There are many companies who develops vacuum cleaners using different methods with very high cost [3]. For developing a robotic vacuum cleaner there are mainly two different tasks, firstly scanning the area which needs to be cleaned, we scan the area in a ground level using sensors surrounded by the robot then processed in a raspberry pi using python language, followed grid mechanism where each cube in a grid is based on the size and distance moved by robot, vacuum fan and two sweeping motors are turned off while scanning using a relay module. Secondly by using scanned data robots start the cleaning process from its initial position, remembers the area it already cleaned so

---

S. A. Nabi (✉)

Department of CSE, Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India  
e-mail: [dr.nabi@sreyas.ac.in](mailto:dr.nabi@sreyas.ac.in)

M. K. Vardhan

Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India

that no redundant cleaning takes place of same location, stores the dust in a removable bin. Scanning and cleaning both follow different algorithms depending upon input given by sensors. Developed a better algorithm for an efficient cleaning process as we are making it cheap by using ultrasonic sensor, raspberry pi, micro gear motors, line tracking sensor, compass sensor, relay module, and DC motor.

Robotic vacuum cleaners have reached a good amount of success within the domestic market. The iRobot Corporation (one of the most popular players) claims to sold 6 million units of their products within one decade [4].

According to the statistics of the International Federation of Robotics [5], about 2.5 million personal and service robots were sold in 2011, an increase of 15% in numbers (19% in value) compared to 2010 [6]. This statistical figures clearly highlights the increasing demand of domestic robots at our homes, which encourages to creates new interaction patterns. In addition to this, the demand for the operation of many new cleaning robots will follow the same tendency. Because of that, we have designed our robot with effectively in terms of price and time.

## 2 Related Work

An existing robotic [7] vacuum cleaner had some drawbacks like colliding with obstacles and stopped at a shorter distance from walls and other objects. It was not able to reach all corners and edges of the room and left those areas unclean [8].

A design and implementation of a smart floor cleaning robot uses a microcontroller [9] where data cannot be stored, it is difficult to find the place which is not cleaned and which is cleaned, action will be performed only based on the current data read by sensors.

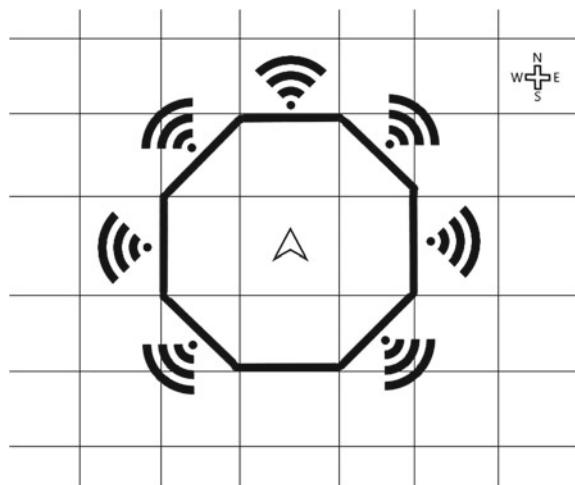
In existing vacuum cleaner, robot doesn't have any idea about its current position and direction in which it is moving, if any obstacle occurs robot turns to the direction with higher free space even though the robot already cleaned that location.

## 3 NavRobotVac Design

NavRobotVac uses a grid mechanism for the virtual view of a robot, the size of each grid depends upon dimensions of body and wheels, the robot length and width is  $30 \times 30$  cm and 1.5 cm radius of a wheel from which each grid is considered as  $10 \times 10$  cm. Grid mechanism of NavRobotVac is shown in Fig. 1. The robot uses a compass sensor to know the direction for easy scanning stores the area and performs the next operation based upon direction and data given by ultrasonic sensors.

The proposed system is designed in such a way that it is targeted to meet the user needs and also reaches more number of users. It consists of two wheels of radius 1.5 cm which are placed exactly in the center, for one complete rotation of wheels it moves 10 cm distance which is considered the size of a cube in a grid. A line

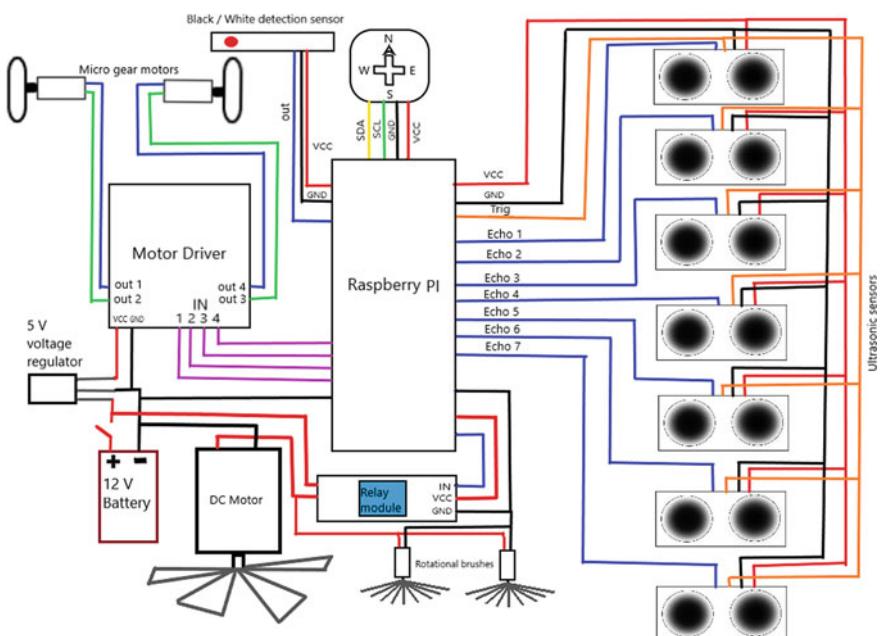
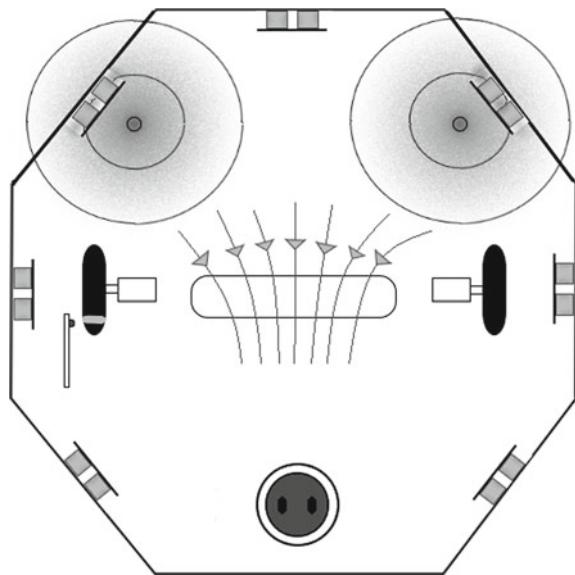
**Fig. 1** Grid mechanism of NavRobotVac



detecting sensor is used to identify one complete rotation of a wheel by sensing a white line on black wheel and based on the movement the current location of the robot is updated continuously. The robot rotates very smoothly for the turn in the same place. A 360° rotatable wheel is fixed at the back for balancing the robot. Speed of the wheel varies based on the function it performs. If the robot decides to turn left or right then wheels are rotated in the opposite direction to each other so that the robot stays in the same position and only direction gets changed. Robot stops only if it completes its task or if the pause button is pressed. The design of NavRobotVac is shown in Fig. 2.

The robot consists of seven ultrasonic sensors on each side except back, compass sensor is fixed facing the north direction, line detecting sensor is fixed beside a wheel, along with relay module all the sensors are connected to raspberry pi for the power supply and to transfer the data with which the data is processed to give instruction to wheels connected to micro gear motors along with the fan connected to DC motor to suck the dust and rotational brushes fixed to micro gear motors. A 12 V power supply is given to the relay and 5 V is supplied to the motor driver by reducing voltage using a voltage regulator. The detailed circuit diagram of NavRobotVac is shown in Fig. 3.

Robot contains start, pause buttons, and a switch for main power supply. When the switch is on all the sensors and processor gets turned on. When the start button is pressed by placing the robot at the center of the room for the first time, the robot starts scanning in a spiracle pattern in an anti-clockwise direction. Once the scanning is done scanned data is stored in raspberry pi memory. Immediately after the scanning robot starts the cleaning process, cleaning is done for every 24 h (fixed in a program). Pause button is used to stop the cleaning or scanning process until it is pressed again.

**Fig. 2** Robot design**Fig. 3** Circuit diagram of NavRobotVac

## 4 Functionality

It mainly consists of two functionalities. The first one is scanning and second is cleaning. If any obstacle appeared in the process of scanning or cleaning, the robot performs operation based on the current input and previous data if present (NPY file) and continues its operation. The detailed descriptions of these two functionalities are defined as follows.

### 4.1 Scanning Algorithm

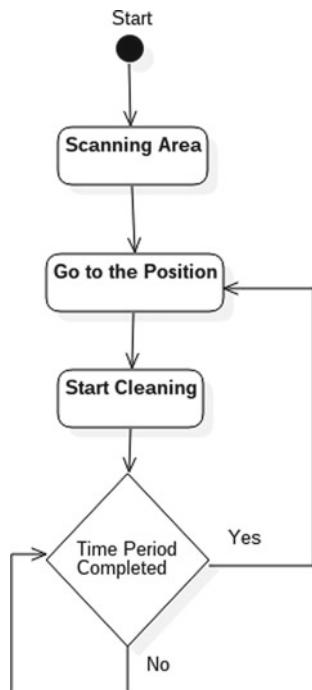
A matrix is created and saved in NPY file before the scanning process begins to record the area of the location, then the robot turns towards the north direction and starts moving until it reaches a wall, then it turns towards left i.e. west and begins scanning using ultrasonic sensors, process the data in raspberry pi and updates the file. Robot initially completes scanning borders of the location and scans inside the scanned area. We followed the spiral pattern for the scanning process. The algorithm reads the file if there is any area which is not scanned and moves towards the location if it finds any, if not the scanning process stops. Then all the extra rows and columns in a matrix which are useless are removed from the file and store only the scanned area.

### 4.2 Cleaning Algorithm

Cleaning algorithm uses the NPY file saved by the scanning algorithm of the location that needs to be cleaned, make a copy of the file to note the place which is already cleaned while cleaning. To start the cleaning process the robot needs to go to its initial position which is the south west corner. A method is executed to find the initial position and move to the location. Then the robot starts cleaning the west wall from south to north, after reaching north wall the robot turns east, moves front and turns towards south, using the same method cleaning process is done from west to east. Cleaning algorithm is done in a parallel pattern. The process begins after the scheduled time is completed.

NavRobotVac is implemented using python programming language; execution is done in raspberry pi containing processor and raspbian operating system. The program is kept in bash location for the auto start of execution as soon as the power turns on. The basic functionality of the product is shown in Fig. 4.

**Fig. 4** Functionality of NavRobotVac



### 4.3 Obstacles Avoidance

In the process of scanning when any obstacle comes, robot considers it as a non-cleaning area. As the scanning pattern is spiracle in an anti-clockwise direction the robot moves towards its left and continues scanning.

When the process of cleaning is going on, if any object appears in front of the robot, it will pass the obstacle from its right side and updates that area as a non cleaning area. In the next turn when the same area is free then the robot moves towards that area and updates to the empty space in NPY file which needs to be cleaned.

## 5 Results and Discussions

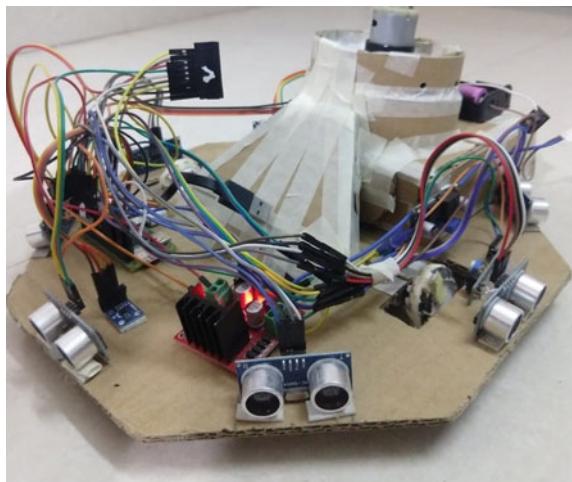
NavRobotVac model aim is to strive the advance homemade robotic vacuum cleaners for domestic needs with reference to numerous features: usability, apparent convenience, and style. These are the basic factors to meet the domestic needs for the adoption of advanced technology. Based on users need and time to time to meet their prospects and their methods of employing a robotic vacuum cleaner, we can reach their expectations and meet their true needs and take them under consideration

while the process of evolving these types of systems with advanced technologies. The resultant of our proposed system is shown in Fig. 5.

The proposed system considered a sample area with five corners to test the robot as shown in Fig. 6. It starts with north direction and ends with after scanning the entire blocked area. The result can be seen in a NPY file which contains 0, 1, and 2's, 0 says that area is out of border, 1 represents empty area to be cleaned and 2 represents obstacle in the field. An example for output of proposed model is shown in Fig. 7.

In an existing system robot works on a microcontroller as a central unit which only works on a present input it leads to a redundant cleaning of same location. Whereas in NavRobotVac, it uses Raspberry pi as a central unit which gives the instruction by taking present input and previous data given by sensors.

**Fig. 5** NavRobotVac



**Fig. 6** Sample area with 5 corners



**Fig. 7** Sample area output

While comparing with existing system, Robots store mapped area in a file which requires additional memory whereas in proposed system the mapped area is stored in a NPY format which consumes less memory and it is the fastest method of loading in data. This robot is designed in octagon shape to minimize the size and for easy movement of robot. Also it is designed in a minimum price as there is no extra equipment like camera etc.

## 6 Conclusion and Future Scope

Due to revolution in the technologies, robots which are used for domestic purpose are transformed from the straight forward “random-walk” methodology to further advanced navigation systems, comprising a domestic technology for a reasonable cost.

With this user-friendly approach, the developed product efficiently ensures Scanning of a new area which needs to be cleaned and stores the scanned area as a NPY file in a memory present in raspberry pi. Cleaning process executes after every time period completed, simultaneously updating the area file if any obstacles are found. By implementing this system a user is free from keeping his home clean as that work is done by a robotic vacuum cleaner. Users need to keep a time gap to activate the cleaning process. After cleaning is done the user takes a dustbin from the robot, cleans it and fixes it back for the next clean.

In future, the vacuum cleaner NavRobotVac can be added with many features like mopping, carpet cleaning, and an application can be designed for the robot to make track of it.

**Acknowledgements** We would like to thank everyone, who has motivated and supported us for preparing this Manuscript.

## References

1. Mitchel, N.: The Lazy Person's Guide to a Happy Home: Tips for People Who (Really) Hate Cleaning, 9 Jan 2016 [Online]. Available: <https://www.apartmenttherapy.com/thelazy-personsguide-to-a-happy-home-cleaning-tips-forpeople-who-reallyhate-cleaning-197266>. Accessed 22 June 2016
2. Prassler, E., Munich, M.E., Pirjanian, P., Kosuge, K.: Domestic robotics. In: Springer Handbook of Robotics, pp. 1729–1758. Springer International Publishing (2016)
3. Hong, Y., Sun, R., Lin, R., Shumei, Y., Sun, L.: Mopping module design and experiments of a multifunction floor cleaning robot. In: 2014 11th World Congress on Intelligent Control and Automation (WCICA), pp. 5097–5102. IEEE (2014)
4. C.o. iRobot: iRobot: Our History, 29 May 2015. [Online]. Available: <https://www.irobot.com/About-iRobot/CompanyInformation/History.aspx>
5. Adkins, J.: 'History of Robotic Vacuum Cleaner, 21 June 2014 [Online]. Available: <https://prezi.com/pmn30uytu74g/history-ofthe-robotic-vacuum-cleaner/>. Accessed 14. 04. 2015
6. Kapila, V.: Introduction to Robotics (2012). <https://engineering.nyu.edu/mechatronics/smart/pdf/Intro2Robotics.pdf>
7. Shashank, R., et al.: Shuddhi—Acleaning Agent. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **9**(2S) (2019). ISSN: 2278-3075
8. Frolizzi, J., Disalvo, C.: Service robots in the domestic environment: A study of Roomba vacuum in the home. In: International Conference on Human Robot Interaction HRI, pp. 258–265 (2006)
9. Appliance, C.: How to Choose the Best Vacuum Cleaner, 23 Feb 2016 [Online]. Available: <https://learn.compactappliance.com/vacuum-cleaner-guide/>. Accessed 22 June 2016

# A Hybrid Clinical Data Predication Approach Using Modified PSO



P. S. V. Srinivasa Rao, Mekala Srinivasa Rao, and Ranga Swamy Sirisati

**Abstract** Enhancement of diagnostic predictive mechanisms in Clinical Decision Support Systems (CDSS) is performed through improving disease staging predictions, illness progression prediction, and reducing the number of features in high dimensional clinical data sets. All the works in this paper uses benchmark datasets from University of California Irvine machine learning repository for verification. Clinical datasets are obtained through various diagnostics procedures, and the data from different sources was transformed into a single format before the start of the analysis. Since there are many diagnostics procedures, the transformed data usually has a large number of features. Preliminary data cleaning procedures like duplicate removal, noise dismissal, and filling up of missing data are performed initially. Data reduction techniques are applied over this cleaned data. A dataset is mainly due to two main reasons: too many instances or too many attributes. Once the noisy and duplicate instances are removed, having too many instances usually generates an accurate classifier. However, having too many features is a negative aspect of data analytics because the dataset's unimportant features tend to bring down the classifier's accuracy that has produced using the dataset. Reducing the number of features to an optimal level to enhance the classifier's accuracy is the aim of feature reduction algorithms. Every disease has various stages of severity. In this paper, an effective DSS support system for the Clinical Data predication Approach was proposed based on the combination of Hybrid technique using C4.5, decision tree with Particle Swarm Optimization (PSO) evaluated for various diseases.

---

P. S. V. Srinivasa Rao (✉)

Vignan's Institute of Management and Technology for Women, Kondapur, Ghatkesar Mandal, Telangana, India

M. S. Rao

Department of CSE, Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India

R. S. Sirisati

Department of CSE, Vignan's Institute of Management and Technology for Women, Kondapur, Ghatkesar Mandal, Telangana, India

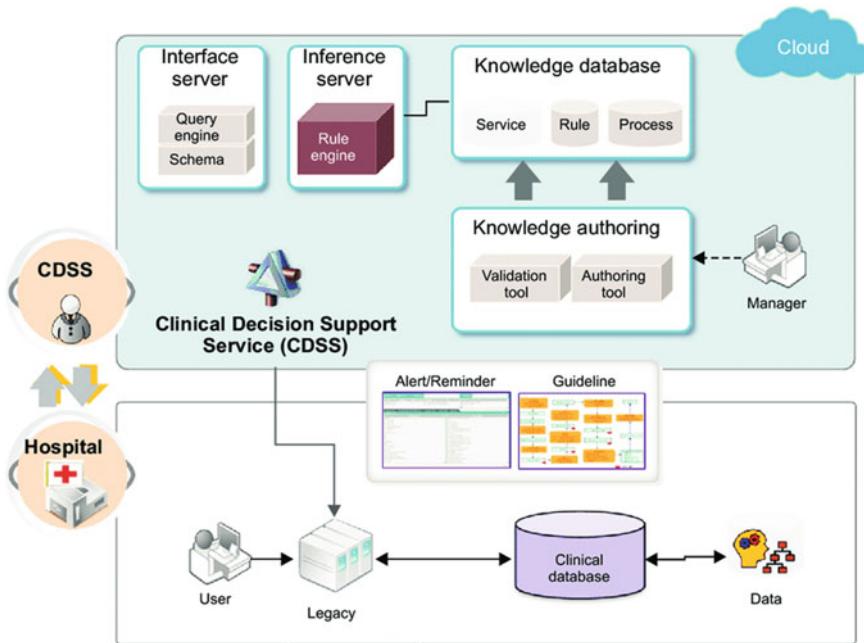
## 1 Introduction

Every country in the world needs healthy individuals in order to have an efficient human resource. Improving the health care field does not only requires personnel support. With the active support of software tools, hardware equipment, and rich information, healthcare domains could see an unprecedented change. In recent years, Data predication Approaches designed primarily to help during the treatment phase is gaining momentum, but the Data predication Approaches to support during the diagnosis phase is still in its rudimentary form. Since the sophistication of machine learning and other data analytics algorithms are getting real in the market place, Data predication Approaches research in clinical diagnosis is gaining ground. With that in mind, the problem statement objectives are formulated [1, 2]. The Clinical Data predication Approach (CDPA) assists physicians and other healthcare professionals in their decision-making process. According to Robert Hay ward, Clinical Data predication Approaches link with health care domain to effect health. CDPA uses Computational Intelligence (CI) techniques as a means to make decisions. CDPA uses various types of patient data during the process of decision making. CDPA prominently finds its usage mainly in the treatment phase. It is used as an alerting/alarming system. Integrating with the Internet of Things (IoT) technologies, CDPA supports treating personnel during the hospitalization period like rule-based drug delivery, emergency vital alert systems, and continuous sensing and monitoring systems. The CDPA is used to give correct information to the necessary personnel at the right time through technology supporting automation. Thus as of now, the role of CDPA is mostly pertained to sensing and alerting. Now the time is ripe enough to apply CDPA extensively during the diagnosis phase, as there are too many advancements in machine CI. Some of the prominent applications of CDPAes are CO [3–5] (Fig. 1).

- i. Tracking the progression of diseases.
- ii. Suggesting suitable diet needs for patients.
- iii. Suggesting an effective treatment procedure.
- iv. Staging of diseases.
- v. Predicting the causality of diseases.
- vi. Pharmacological decision support.
- vii. Estimating the outcome of a treatment procedure.
- viii. Estimating the outcome of a surgery.
- ix. Early warning systems in the ICU-intensive care unit.
- x. Diagnosis of diseases.

The effectiveness of CDPA is decided by way of handling the two main challenges, such as data handling and data analytics.

Clinical datasets are obtained through various diagnostics procedures, and the data from different sources got transformed into a single format before the start of the analysis. Since there are many diagnostics procedures, the transformed data usually has a large number of features. Preliminary data cleaning procedures like duplicate removal, noise dismissal, and filling up of missing data are performed initially. Data



**Fig. 1** DSS support system for clinical data predication approach

reduction techniques are applied over this cleaned data. A dataset is mainly due to two main reasons: too many instances or too many attributes. Once the noisy and duplicate instances are removed, having too many instances usually generates an accurate classifier. However, having too many features is a negative aspect of data analytics because the dataset's unimportant features tend to bring down the classifier's accuracy that has produced using the dataset. Reducing the number of features to an optimal level to enhance the classifier's accuracy is the aim of feature reduction algorithms [4–6].

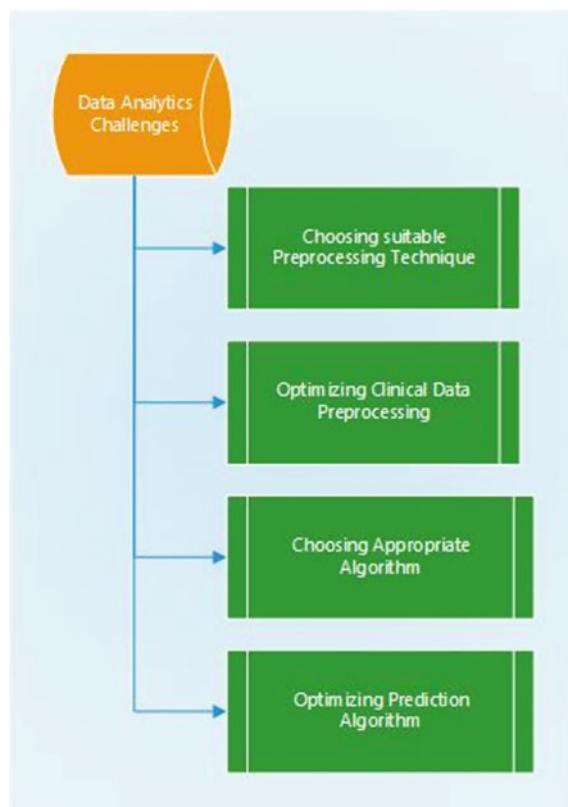
## 2 Related Work

Clinical datasets are obtained through various diagnostics procedures, and the data from different sources got transformed into a single format before the start of the analysis. Since there are many diagnostics procedures, the transformed data usually has a large number of features. Preliminary data cleaning procedures like duplicate removal, noise dismissal, and filling up of missing data are performed initially. Data reduction techniques are applied over this cleaned data. A dataset is mainly due to two main reasons: too many instances or too many attributes. Once the noisy and duplicate instances are removed, having too many instances usually generates

an accurate classifier. However, having too many features is a negative aspect of data analytics because the dataset's unimportant features tend to bring down the classifier's accuracy that has produced using the dataset. Reducing the number of features to an optimal level to enhance the classifier's accuracy is the aim of feature reduction algorithms. The main challenge is that the sequence's prediction is made with a trained entity that is not trained with the transition data. Continuous-time series clinical data corresponding to individual patients is not available in enormous quantity than the current data severity levels. So in the proposed work, we have used the clinical dataset having current severity levels to detect the transition sequence [7–9].

Patient-centric disease sequence prediction is performed in all of the previous researches use time-series data. No research work addresses the prediction of possible transition stage of disease for a patient without the availability of that patient's time-series data in Fig. 2. The availability of time-series data, such as ICU data, is comparatively low. So the previous works using time series data do not have many benchmark data sets to verify, so the reliability of the proposed Algorithm can not be verified to the fullest. Previous works that focus on time series data assume that the disease

**Fig. 2** Data preprocessing phases



Attribute	Type	Missing Value	Values
Department	polynomial	0	2
Chief Complaint	polynomial	0	912
History of Current Illness	polynomial	2	1031
History of Past illness	polynomial	1	867
Gender	polynomial	0	2
Age	polynomial	0	130
Name	polynomial	0	1012
Check in Date	polynomial	0	1034
Check out Date	polynomial	14	839
In-patient ID	Integer	0	1017
Check in Diagnosis	polynomial	5	773
Check out Diagnosis	polynomial	45	796
UHID	polynomial	0	1034
DATE_VALIDATEED	polynomial	0	1034

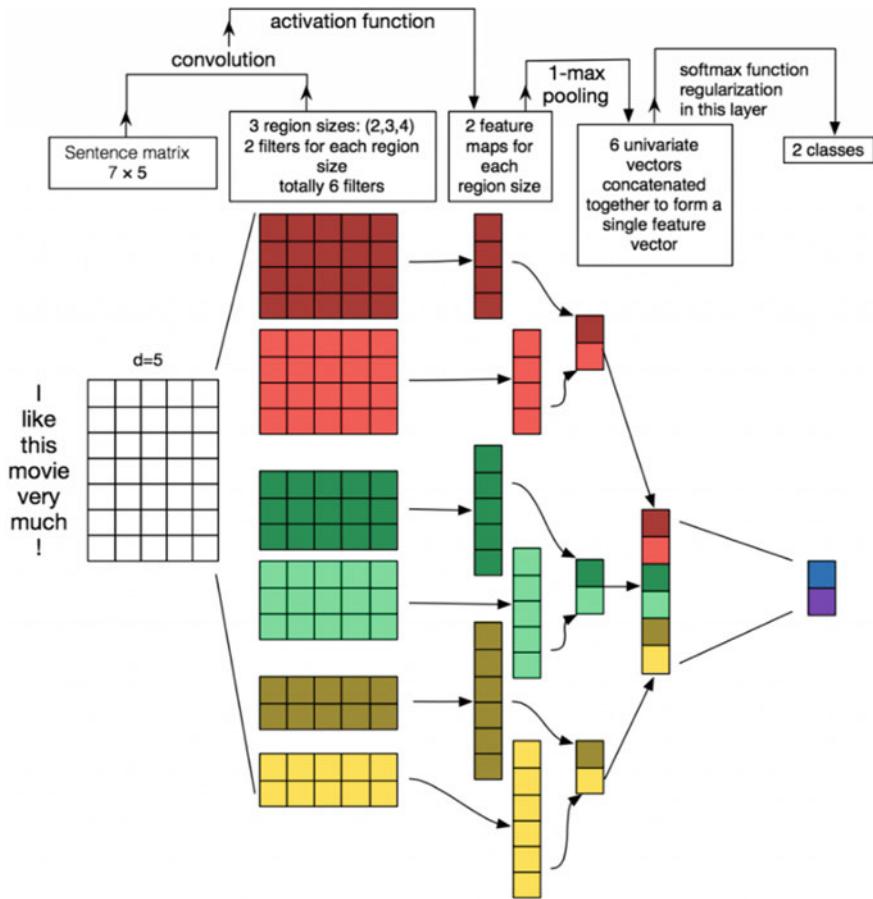
**Fig. 3** Attribute considered DSS in clinical data

progression takes a sequential route, which is not necessarily the default case. The disease can jump stages, and it does not refer to passing through the stages in a fast manner. There is a deficiency in research works that can redefine clinical test data mapping with new disease stages. All the research mentioned above gaps is addressed. The proposed solution uses nontime series data and predicts the patient-centric disease transition stage. The solution provided can even redefine the staging pattern of the disease [2, 10] (Fig. 3).

### 3 Proposed DSS Support System for Clinical Data Predication Approach Using C4.5 with PSO

After reducing the features form the equation formulated as part of this proposed work using C4.5 decision tree and PSO technique for Algorithm 1, are mentioned below [9, 11, 12]. The density of the cluster is found out using the Equations.  $\text{cluster}_i = 1$  Density of Cluster, P-cluster where  $p$  is the particles (Fig. 4).

$$p_k^t = \{p_k 1, p_k 2^t, \dots, p_k D^t\} \quad (1)$$



**Fig. 4** DSS phases of hybrid C4.5 decision tree with PSO

$$p_1^t = \{p_1 1, p_1 2^t, \dots, p_1 D^t\} \quad (2)$$

$$\begin{aligned} \text{Va}_i d^t &= w * \text{Va}_i d^{t-1} + c_1 r_1 (p_i d^t - x_i d^t) \\ &\quad + c_2 r_2 (p g d^t - x_i d^t), \quad m = 1, 2, \dots, M \end{aligned} \quad (3)$$

$$x_i d^{t+1} = x_i d^t + \text{Va}_-, \quad d = 1, 2, \dots, D \quad (4)$$

Here, each decision tree can be treated as a particle for Swarm Optimization. The Eq. 1 represents best solution tree particle, Eq. 2, with the population  $l$ , and Eqs. 3 and 4 deliberates velocity and updated dynamic position.

---

**Algorithm: 1: DSS support system for Clinical Data predication Approach**

---

- i. Calculate cluster =  $L_i = 1$ " Cluster – DPti cluster.and F\_cluster
  - ii. Assign medial diseases Data Set of all cluster  $M \times N$
  - iii. Training Set cluster nodes of diseases
  - iv. Test Samples Examples
  - v. Apply C4.5 and Decision Tree with using Eqs. 1 and 2
  - vi. Apply Current Decision Level using PSO with Eqs. 3 and 4
  - vii. Assign the final Particle as Decision parameter
- 

A unified model for feature selection is also proposed, significantly to reduce features in massive datasets. It does reduction by selecting features appearing in both the set of feature subsets selected through MBPSO and other existing Binary Genetic Algorithm based methods and the model. An evaluation parameter for wrapper-based feature selection methods is called the Trade-off factor (TF). The following are the highlights of the proposed work. Wrapper-based methods are used in this work, and feature selection is in every step of the Algorithm through the creation of the SVM classifier. The existing Binary Genetic Algorithm (BGA) is used along with in unified approach. A comparison of various feature selection models is performed using a neural network classifier.

Particle Swarm Optimization technique (PSO) is one of the evolutionary optimization algorithms. It is a heuristic algorithm. It has proved its mettle in optimization; PSO has been considered to provide a solution in this work. Binary Particle Swarm Optimization (BPSO) technique is a version of PSO, which uses particles having binary digits. An updated version called Mutated Binary Particle Swarm Optimization (MBPSO) is proposed in this. BPSO is used in feature selection problems. Every digit in a particle used in BPSO is either 0 or 1. If there is a dataset having n number of features, then the particles will be having n number of binary digits. Each feature corresponds to each digit in a particle's binary string. The features corresponding to 1's are all congregated, which is considered a derived data set. Once every particle had its fitness value found, the particle with the highest fitness value is the global best particle. Then every other particle will update its velocity and position towards the global best particle. During the movement, binary digits of particles will be updated. Again the process is repeated till the fitness of the global best particle is lesser than a threshold value, which is the preset or specific number of iterations are fixed, so that after completing the present iteration, the features of the dataset corresponding to the global best particle is considered as the reduced feature subset. The movement of particles in search space is said to be an exploration phase of PSO, and the movement of particles towards the global best particle is said to be the exploitation phase. PSO algorithm is proved to be good to have acceptable exploitation. That is, it tends to converge fast. However, it has a problem with exploration. The proposed method called MBPSO aims to alleviate the problem of exploration by introducing a mutation mechanism, which is usually associated with genetic algorithms. The workflow diagram is shown in Fig. 2. The idea is to selectively apply the mutation process, which is defined as a random update of particles' digits. Last k particles with

the least fitness values are subjected to mutation, thereby the process of exploitation is not disturbed, and the process of exploration is enhanced through making the last  $k$  fittest particles scrambled through the search space. Thus, the proposed method called MBPSO aims at enhancing the exploration phase of BPSO. The evaluation of finished selection using the neural network uses random sampling without replacement of both random samplings without replacement of random sampling without replacement during training and testing. There is only one hidden layer with ten neurons in evaluating neural networks. For every method, the count of input neurons varies with the outcome of feature selection, and the count of neurons in the input layer is equal to the count of features selected. All datasets used evaluate have only one class, so there is only one neuron in the neural network's output layer. Genetic Algorithms (GA) are evolutionary optimization algorithms. GA algorithms are heuristic algorithms. The GA components are population generation, selecting fit members from the population to perform a cross over. After cross-over operation, the members are subjected to the next process called a mutation. Members are called genetic strings. A string can be a sequence of integers or real numbers. If a string is composed of binary numbers, then it is called a binary string. GA, which uses binary strings, is said to be Binary Genetic Algorithm (BGA). The meaning of cross-over between two strings is that both of the two strings undergoing the process will lose certain parts of it, and the lost part will be filled with part from other strings. This process of marriage between two strings is said to be cross-over. The resultant modified strings are said to be offspring. A mutation is a process where a string's digits are randomly chosen and assigned with a random value (binary string 0 or 1). BGA finds its use in feature subset selection. The process starts with population generation, meaning, and creation of strings.  $N$  number of strings are created. Each string will have  $m$  number of binary digits, equal to the high dimensional dataset's total count of features. Once strings are generated, then the fitness of every string is computed. Features corresponding to 1 alone in strings are congregated, and the resultant dataset is called derived dataset. A derived dataset, a classifier is created, and that classifier is evaluated. The classifier's accuracy is set as the fitness value of that string. If the classifier's accuracy is more, then the fitness of the string will also be more. Thus every string of the entire generated population finds its fitness value. This kind of methodology where a classifier is generated to fix the fitness value is called wrapper-based methodology. Once every string had its fitness found, then the first  $k$  strings are chosen to be part of the reproductive pool. In the reproductive pool, pairs of strings are chosen randomly, and they are subjected to cross-over and mutation. BGA used in work uses a single point cross over.

## 4 Experiments and Result

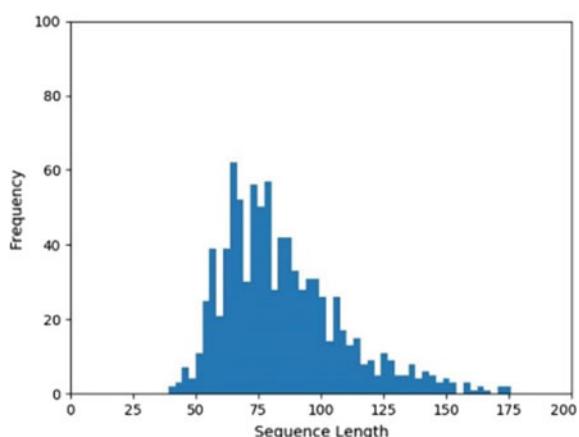
There are many dark areas in clinical data analytics, like genomic data analytics, which can solve genetic and hereditary diseases. These diseases data set is considered

from the UCI ML standard repository. We can now identify a genetic disease occurrence, but the cure can not be given. However, once the genomic data research/bio-informatics unfolds the secrets of genetic code, then wonders can be done in curing incurable diseases, and many unknown areas like the one that is mentioned in our research that is the randomness in disease progression, the key for this riddle might be in the genetics field. Nowadays, computation infrastructure is very much ready, and there is the availability of fast processing systems like supercomputing and cluster computing, and the platform to share the enormous data is also getting shape through the cloud and sample data storage. The advancement of big data analytics is in the budding phase, and once this reaches a matured phase, with high computing powered machines along with efficient big data analytics algorithms with enormous support data available, most of the health care problems can be solved which helps people lead a more happy life. Every symptom that is detected during suffering from the disease is due to some parameters going wrong. The flaw in a specific parameter is caused due to various reasons. The entity which brings flaw to a specific parameter is said to be a causal entity. All the entities which are causal for a specific parameter are called Concepts. Many flawed parameters characterize the disease, and each parameter has its own set of concepts. The concept sets of vital parameters may intersect. If we represent the concept dependencies, we end up getting a map data structure, which depicts interdependencies of concepts, and it is called a cognitive concept map. In order to set right the parameters, the concepts have to be provoked. The weigh of involvement of concepts in deciding the parameters is patient-centric. It can be taken as a problem statement, and a solution can be provided for accurate prediction of the percentage of involvement of concepts in tweaking the features of vital parameters. For example, if red blood cells count is said to be a parameter and this count can be altered through tweaking one or more concepts, how much to tweak is patient-centric. If this can be done, patient-centric targeted treatment can be provided. In today's scenario, time-series data is available in abundant quantities both for in and outpatients. Every hospital has electronic medical records for all of their patients. These data are now getting shared in cloud platforms. Even though there is no standardized mechanism to share the data is not available currently due to various privacy data laws, the data can be shared with patients' consent, But the time series data corresponding to fetal development is very scarce. Not much research has been done in capturing many vitals or parameters during the fetal development phase. Currently, the data being captured development phase is very minimal, and it is of non-invasive type. If there is a safe and commonly usable minimal invasive technology is developed for fetal analysis, many genetic disorders can be out. The fetus's presently morphological structure is set as a benchmark to analyze the genetic disorders, which has a very high error rate. If extensive data is collected during fetal development, through minimal safe invasive and non-invasive methods, a good repository of sequential data would be obtained to analyze using Computational Intelligence techniques. Many genetic disorders can be prevented or mutated upon if they are detected at the correct time, or the decision to abort the fetus can be taken based on the severity of the genetic disorder. Diseases such as down syndrome,

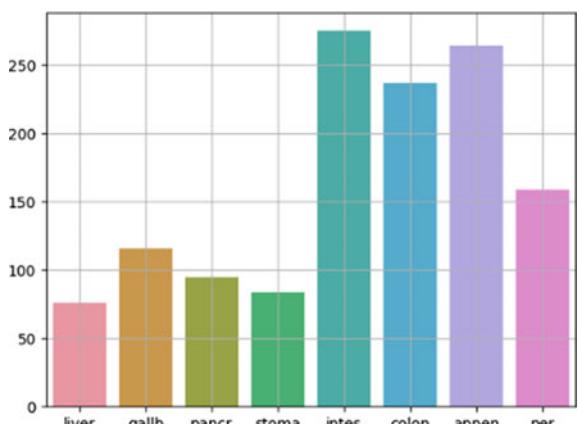
Huntington disease, fragile x syndrome, hearing loss, sickle cell disease, turner's syndrome, Fragile X syndrome, and many more can be detected more accurately.

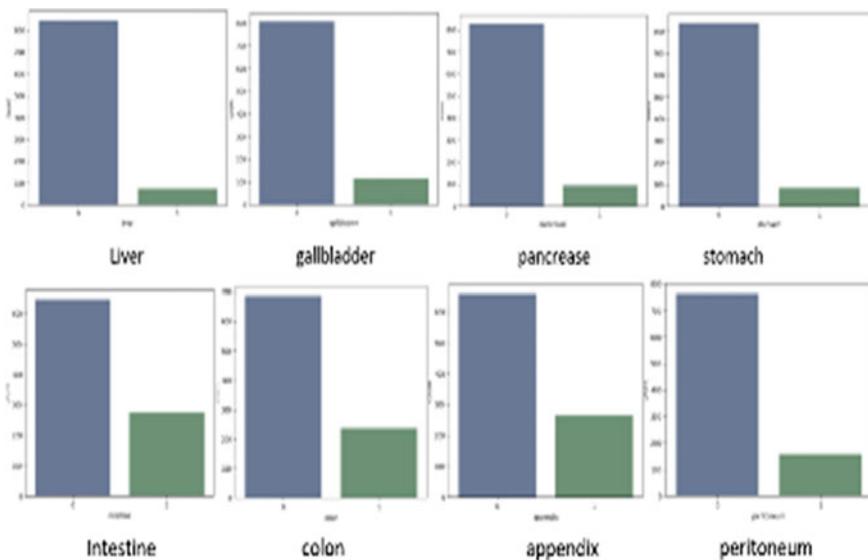
As per Figs. 5, 6, 7, 8, 9 and 10 describes various parameters predicted using proposed C4.5 and PSO. The behavioral and social communication pattern data and fetal development data can be correlated by analyzing autism and Asperger syndrome. In the previous idea, analyzing fetal data, a genetic disease of morphological disorders can be identified. However, in this case, intellectual ability can be predicted through fetal time series data if the current intellectual disability is successfully correlated with fetal clinical data. Intelligent prosthesis or learning limbs: Presently, much research has brought good results in designing sophisticated weightless artificial limbs or prosthesis. Nevertheless, still much more can be done to make a prosthetic limb more intelligent. For example, if the pattern of ordinary people's movements and activities are captured, having parameters such as pressure in the joints, angular velocity and stress-induced angle changes, vibrations, flexibility. The

**Fig. 5** Disease support with frequency



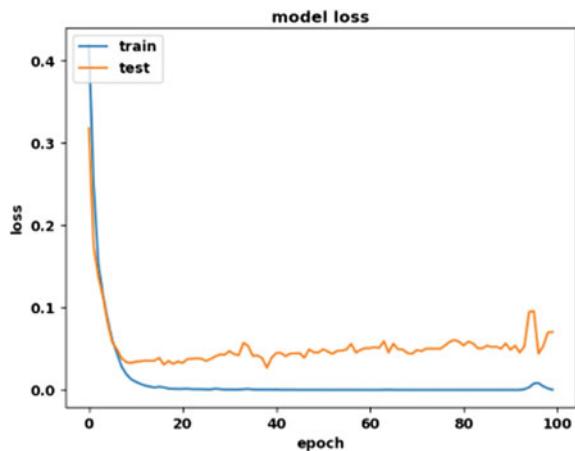
**Fig. 6** Various Disease with level parameter-1





**Fig. 7** Various Disease with level parameter-2

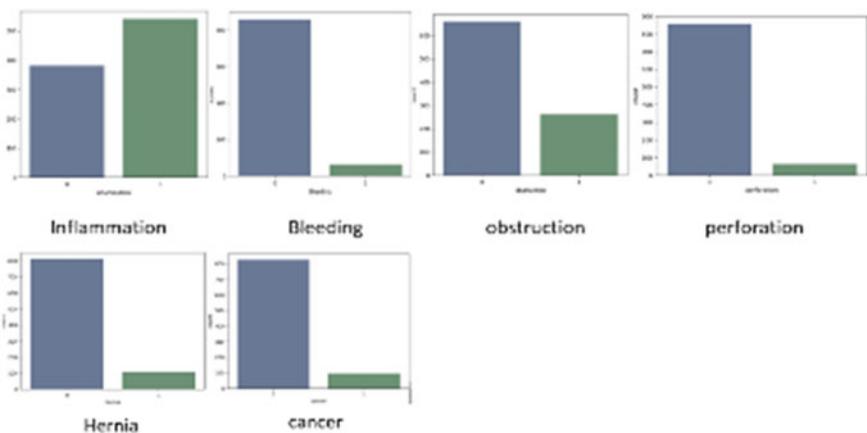
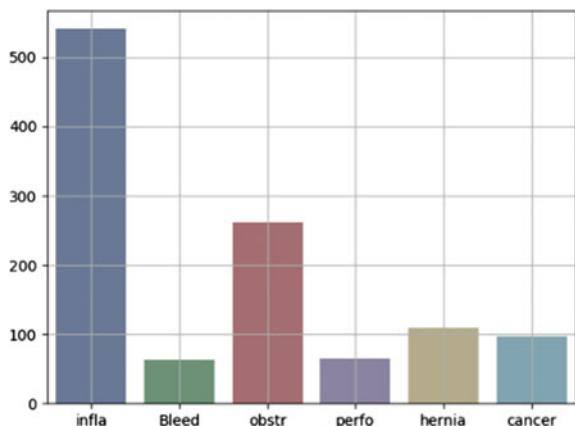
**Fig. 8** Various Disease with level parameter-3



relation between the patterns can be understood, and the prosthetics can also adjust the parameters, just like a normal limb. For example, the parameter changes during running are captured well, and then the training can come out to be fair and so the prosthetics leg can even be designed to run in the same manner with which an average person runs. This knowing can be utilized in developing humanoid robots too.

Sensory organ functionality is dependent on neurological structure. With the advancement in neural and deep learning methodologies, a sensory organ function can be emulated artificially. Artificial and efficient sensory organs are the theme

**Fig. 9** Various Disease with level parameter-4



**Fig. 10** Various Disease with level parameter-5

of the future. Imagine an artificial eye capable of seeing than a normal eye and ear, hearing more than its natural counterpart. Artificial sensory organs using neural networks need needs a long way to go. However, proper funding and much research coordination in this domain will revive many special people's lives. Drug suggestion is another area where computational intelligence algorithms can be applied. Drug sensitiveness varies from patient to patient.

Moreover, the effect of drugs to vary from patient to patient. Using cognitive maps and weight optimization algorithms, patient-centric drug-related decisions such as what type of drug, quantity, duration, and mode of delivery can be decided. The proposed objectives to improve diagnostic prediction in clinical Data predication Approaches are achieved through improving upon accurate staging, progression prediction, and efficient preprocessing of data through feature selection. The research work of the paper is summarized as follows. This work also used coronary heart

disease as a reference to verify the Algorithm. The proposed method's highlight is to arrive at a patient-centric possible disease stage transition sequence using nontime series data. Unlike most of the methods, which predicts the next stage for a patient having a specific disease using clinical time-series data, the proposed Algorithm uses nontime series data corresponding to different persons. The proposed method uses a novel approach rooted in vector algebra, gravitational force-based, and nearest neighbor property.

## 5 Conclusions

The proposed algorithm gives a solution to the problem statement's third objective: feature subset selection using a wrapper-based method with heuristic optimization algorithms such as modified particle swarm optimization algorithm and Genetic Algorithm. However, some previous works on this field use heuristic algorithms that did not address exploring and exploiting heuristic algorithms together. The proposed work hybrid methodology based on the C4.5 decision tree with PSO enhances both exploration and expletory processes through genetic Algorithm and modified particle swarm optimization algorithms. An evaluation parameter called the trade-off factor is proposed, which is used to evaluate the selected features.

It has been found that computational intelligence algorithms are indispensable in health care applications. In the further coming years, computational intelligence algorithms will rise only because of its efficient functioning mechanisms due to much energy is getting into research in these domains. The research on clinical data analytics holds the key to future diagnostics and treatment. Clinical data analytics is in the starting stage from where it will revolutionize the health care industry. Technology support in the health care industry pertains to clinical tests, and in the last five decades, the use of technology in testing has grown, but the application of technology in the diagnostics phase is still not reliable. Extensive research is needed to come out with concrete results to be trusted upon the technology in the decision. Some of the future research works that can be taken up to enhance clinical Data predication Approaches are discussed in the next section.

## References

1. Herland, M., Khoshgoftaar, T.M., Wald, R.: Survey of clinical data mining applications on big data in health informatics. In: Proceedings of ICMLA '13—vol. 02, pp. 465–472. Google Scholar Digital Library Physionet-MIMICIII [n.d.]. <https://archive.physionet.org/physiobank/database/mimic3cdb/>
2. Cai, X., Perez-Concha, O., Coiera, E., Martin-Sanchez, F., Day, R., Roffe, D., Gallego, B.: Real-time prediction of mortality, readmission, and length of stay using electronic health record data. J. Am. Med. Inf. Assoc. **23**(3), 553–561 (2016). <https://doi.org/10.1093/jamia/ocv110>

3. Li, L., Cheng, W.-Y., Glicksberg, B.S., Gottesman, O., Tamler, R., Chen, R., Bottinger, E.P., Dudley, J.T.: Identification of type 2 diabetes subgroups through topological analysis of patient similarity. *Sci. Trans. Med.* **7**, 311 (2015)
4. Sirisati, R.S.: Machine learning based diagnosis of diabetic retinopathy using digital fundus images with CLAHE along FPGA Methodology. *Int. J. Adv. Sci. Technol. (IJAST-2005-4238)* **29**(3), 9497–9508 (2020)
5. Purushotham, S., Meng, C., Che, Z., Liu, Y.: Benchmark of deep learning models on large healthcare MIMIC Datasets. *CoRR abs/1710.08531* (2017). <https://arxiv.org/abs/1710.08531>
6. Sirisati, R.S.: Dimensionality reduction using machine learning and big data technologies. *Int. J. Innov. Technol. Explor. Eng. (IJITEE-2278-3075)* **9**(2), pp 1740–1745 (2019)
7. Pai, S., Bader, G.D.: Patient similarity networks for precision medicine. *J. Mol. Biol.* (2018)
8. Pai, S., Hui, S., Isserlin, R., Shah, M.A., Kaka, H., Bader, G.D.: netDx: Interpretable patient classification using integrated patient similarity networks. *bioRxiv* (2018)
9. Kanungo, T., Mount, D., Netanyahu, N., Piatko, C., Silverman, R., Wu, A.: An efficient K-means clustering algorithm analysis and implementation. *IEEE TPMAI* **24**(07), 881–892 (2002)
10. Mao, Y., Chen, W., Chen, Y., Lu, C., Kollef, M., Bailey, T.: An integrated data mining approach to real-time clinical monitoring and deterioration warning. In: *Proceedings of SIGKDD’12*, pp. 1140–1148 [n.d.]
11. Ma, T., Zhang, A.: Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 7–10. IEEE (2017)
12. Yu, Shi.: Multiclass spectral clustering. In: *Proceedings Ninth IEEE International Conference on Computer Vision*, vol. 1, pp. 313–319 (2003). <https://doi.org/10.1109/ICCV.2003.1238361>

# Software Defect Prediction Using Optimized Cuckoo Search Based Nature-Inspired Technique



C. Srinivasa Kumar, Ranga Swamy Sirisati, and Srinivasulu Thonukunuri

**Abstract** These days, software systems are very complex and versatile. Therefore it is essential to identify and fix the software error. Software error assessment is one of the most active areas of research in software engineering. In this research, we are introducing soft computing methods to assess software errors. Our proposed technique to software gives errors and accurate results. In our proposed method, the error database is first extracted, which acts as an input. After that, the collected input (data) is clustered by the clustering technique. For this purpose, we use the modified C-Mean Algorithm. Therefore, the data is clustered. An efficient classification algorithm then groups clustered data. For this reason, we use a hybrid nervous system. Therefore, there are software bugs, and these errors are optimized using the MCS algorithm. Our proposed method for software error assessment is implemented on the Java platform. Performance measurement is measured by various parameters such as execution rate and execution time. Our proposed Cuckoo search based strategy is comparable to many existing strategies. Graphical representation of comparison results from our proposed strategy for identifying software proposals is one that effectively evaluates profitable strategy and reasonable reference rates.

## 1 Introduction

Software Defect Prediction (SDP) plays an essential part in reducing software development costs and maintaining Achilles' and others' high quality (2017). When there is a recurring software failure in the system, it automatically causes a software error. Software error is a bug introduced by software developers and shareholders. A software vulnerability assessment's primary purpose is to improve the quality,

---

C. Srinivasa Kumar (✉) · R. S. Sirisati

Department of CSE, Vignan's Institute of Management and Technology for Women, Kondapur, Ghatkesar Mandal, Telangana, India

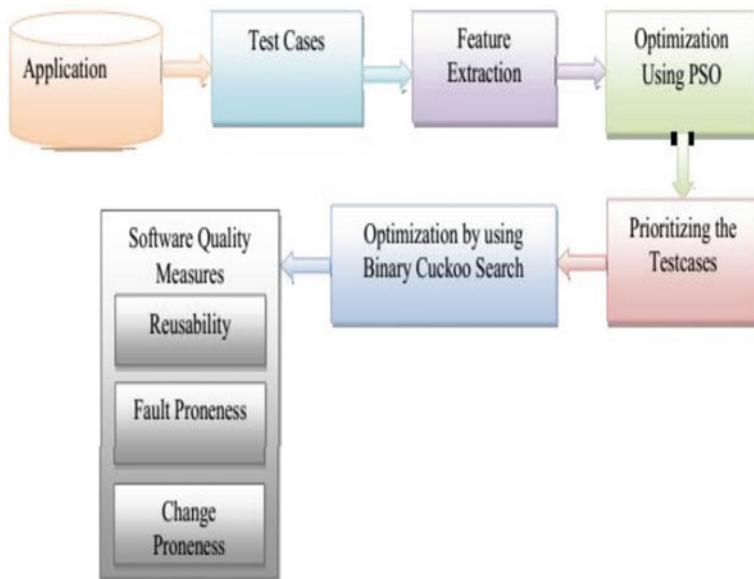
S. Thonukunuri

Department of Mathematics, Vignan's Institute of Management and Technology for Women, Kondapur, Ghatkesar Mandal, Telangana, India

marginal cost, and time of software products. For high-performance error assessment, researchers and others have been working on selecting consistent features and experimental learning algorithms since the 1990s (2014). To assess the error-ability of those who support software testing activities. Software errors are a significant cause of failure for large engineering projects, causing significant financial damage in the twenty-first century. In software quality, various error assessment methods are proposed. The strategy is used on various assumptions, including source code metrics (e.g., alignment, coordination, size) by. Software vulnerability assessment helps to test resources through software modules to address their vulnerability effectively. Current software error assessment models for estimating the number of errors. The software module may fail to provide an appropriate command because accurately estimating the number of errors in a software module due to noise data is challenging. Each software error arises in different conditions and environments and therefore differs in its specific characteristics. Software errors can have a substantial negative impact on software quality. Therefore, error assessment in software quality and software reliability is exceptional. Comparative Software Quality Engineering is a new area of research in the distorted assessment (2014). Current Error Assessment Function (1) Calculate the number of errors remaining in the software system, (2) Identify the error associations, and (3) Classify the error-characteristic features of the software components that are generally error-free and error-free. The reference result can be used as a necessary step by the software developer can control the software process (2009). Software Error Prediction (SDP) empirical studies are very biased with data quality suffer from widely limited generalizations (1986). False assumptions can help improve software quality and reduce the distribution cost of those software systems. In 2005, SDP research overgrew. It allows researchers to collect error assessment data sets from real-world projects for public use and create repetitive and comparable models throughout the study. To date, many SDP based works have done extensive research on matrix and learning algorithms that describe code modules for designing prediction models (2014). Therefore software vulnerability assessment plays an essential role in improving software quality. This software can help reduce testing time and cost. Therefore, it is used in many organizations to save time, improve software quality, software testing, and project resources. Assessing software vulnerabilities for errors in historical databases. In the real world, lumbering elephants are exposed by the aggression of speeding dwarfs. As software projects grow in size, defective assessment technology plays an essential role in supporting developers and speeding up the time to market with more reliable software products [2] (Fig. 1).

## 2 Parallel Works

The general software error assessment procedure follows the machine learning methods. The first step is to find examples from the software. An example of code, function, class or method. These examples arise from various problem tracking systems, version control systems, or e-mail archives. For example, the software has



**Fig. 1** Software defect prediction model block diagram

different dimensions. These examples can be classified as Bug B or Bug Number, Clean C or Clean. After identifying examples using ranges and dimensions, the first step of machine learning pre-processing methods is used to create new examples. Pre-processing is applied to capture features, measure data, and mute noise (2014) [3, 4]. It is not mandatory to apply to all types of error assessment models. After the pre-processing, examples were created to practice the error assessment model. The model of the model causes distorted events and pure events. The number of errors in the example is called regression. This event only gives false or altered results for both, so it is called binary classification. There are many applications for software bug prediction. Its primary purpose is to allocate resources to test software products effectively. Error Assessment Model produces error encounter software, and its scope. Once the reference model is built, its performance should be evaluated. Performance is usually measured in two ways: power presents energy and descriptive energy. Pride Dimensions Predictive power: power measures a model's accuracy for injecting software artifacts with absent power defects. Accuracy, recall, F-measurement, AUCRC, false-positive rates on X-axis, and reasonable favorable rates on Wi-axis are all used in the classification range and are commonly used in error assessment studies. Diffusion Dimensions Diffusion Energy: In addition to measuring power dissipation energy, descriptive energy is also used in distorted studies. The descriptive power model measures how well one understands the difference in data. The measurements described for R2 or standard deviation are commonly used to determine the descriptive strength [5–9].

### 3 Proposed Modified Cuckoo Search Technique.

Yang and Deb recently introduced all Cuckoo search optimization methods. Cuckoo has an aggressive breeding strategy. The female lays her fertile eggs in another species' cage, so surrogate parents inadvertently raise brother Atul Bisht and others (2012) [10]. Coconut eggs are sometimes found in the nest, and surrogate parents either throw it or throw it into the nest or start their flock elsewhere. Cuckoo Search Optimization Algorithm considers various design parameters and controls based on the three main compatibility rules of Azim et al. (2011) (1) Each Cuckoo lays one egg at a time and lays it randomly in the selected nest. (2) Good nests with high-quality eggs are passed on to the next generation; (3) The number of host nests available is determined, and the hockey bird's ability determines the number of Cuckoo eggs. In this case, the host bird can lay eggs or build a new nest. This straw finally selected can be calculated using the n furnace fraction of the current frequency for simplicity. It needs to replace with a new chamber (using new random solutions) in the next cycle. This method has been successfully demonstrated in some benchmarking tasks. The particle optimization method is better than other methods, including manual (2007). Cuckoo Search Algorithm is a metaheuristic algorithm inspired by Cuckoo's reproductive nature and is easy to implement. There are plenty of places for cocaine. Each egg represents a solution, and the nasal egg represents a new solution. The parasites of some species of chickens are unusually attractive. These birds can lay eggs in the host nest and mimic external features such as the host egg's color and color. If this strategy is not successful, the host may throw an egg out of the nose or leave the nest, leading Azimuth to others (2011) [11–13]. Based on this situation, the researchers developed an evolutionary optimization algorithm called Cuckoo Search (CS) and collected CS using their own rules:

Coockoo Search Optimization cost function

$$\text{Min } \text{CS}(S) = \text{Min}[\text{cs1}(S), \text{cs2}(S), \dots, \text{csM}(S)] \quad (1)$$

Inequality Constrains

$$g_b(S) \geq 0, \quad b = 1, 2, \dots, c \quad (2)$$

$$h_l(S) = 0, \quad l = 1, 2, \dots, a \quad (3)$$

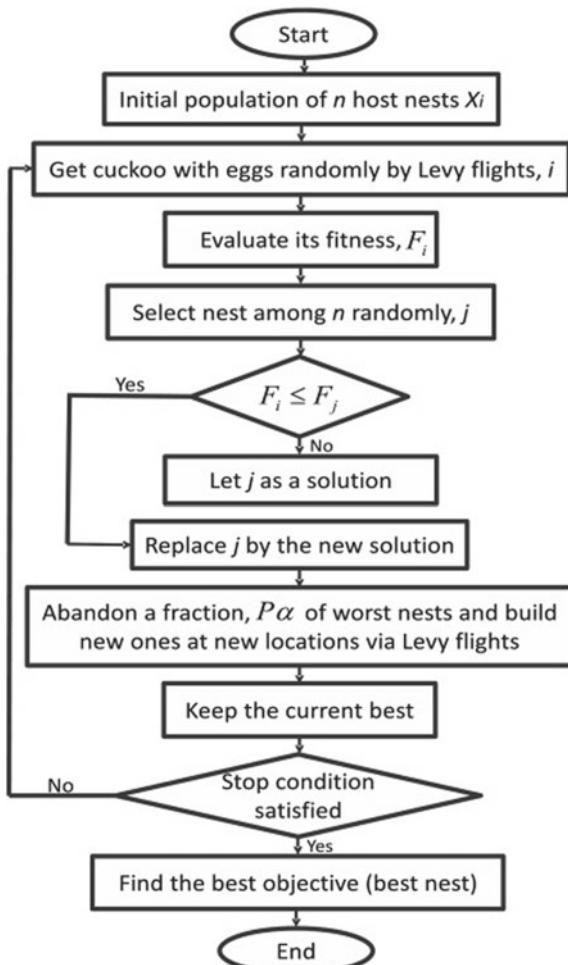
Here Eq. (1) represents Coockoo Search cost optimization function for software predictive parameters, Eqs. (2), and (3) represents constraints to make inequalities for searches. The new solution satisfies the  $N$ -Dimensional Boolean lattice. The solutions are updated in the corners of the hypercube. Additionally, suppose a given attribute is selected or not. In that case, the solution uses a binary vector, where an attribute is selected to create the new dataset and if not 0. In our particular way, the solution represents the value of the attribute. Weight optimization used in the

modified Cuckoo search algorithm. The Cuckoo search algorithm refers to a meta-heuristic algorithm that attributes its origin to Cuckoo's reproductive behavior and is easy to implement. There are many nests for the nose. Each egg signals a solution, and a Cuckooed egg matches the new solution. Novel and best solutions instead of the nest's terrible solutions. As a modified nervous system, we have modified the standard Cuckoo search algorithm to include the gas supply during the upgrade phase levy uses the flight equation.

$$Sb(t') = Sb(t) + 0.01 \times \alpha \times L(\beta) \times (P1(t) - P2(t)) \times m$$

Optimization in gas supply is better than usual. The methods of the optimization process are shown as follows (Fig. 2):

**Fig. 2** Flow of modified Cuckoo search algorithm



---

**Algorithm-1: Modified Cuckoo search algorithm (Phase-I)**


---

1. Each Cuckoo randomly selects a N host nests to lay eggs  $X_i$
  2. The number of host nests available is determined randomly hops i
  3. The nests containing the best quality eggs  $X_i$  are passed on to the next generation fitness ( $F_i$ )
  4. Select n randomly j
  - 5 If  $F_i <= F_j$  (the host finds an egg in the bird's nose, it may lay eggs or leave it.)
  4. Early stages j be solution
  5. Let and j new step
  - Replace j. Evaluate fitness performance  $P_\alpha$
  7. Modify using the Levy flight equation
  8. Find best the result
  9. Stop
- 

---

**Algorithm-2: Modified Cuckoo search algorithm (Phase-II)**


---

Step 1: Get started

The host nest population ( $M_i$ , where  $i = 1, 2, n$ ) begins unilaterally

Step 2: Creating a new cocaine step

With the help of Levi's aircraft, Cuckoon was randomly selected to create new solutions. It is assessed to determine the dominance of carved Cuckoo solutions

Step 3: Fitness Assessment iterations 9 and 10 best assess fitness here by choosing the following equations. Fitness Maximum popularity

SP - TP refers to the selected population - refers to the total population

Step 4: Restore step

---

Initially, the levy is applied by planes, resolving the cautionary transition. The quality of the novel solutions is assessed, and one of them is selected arbitrarily. Suppose the quality of the novel solution in the selected niche is better than the previous solution. It replaces using a new solution (Cuckoo). Otherwise, the previous solution is considered the best solution.

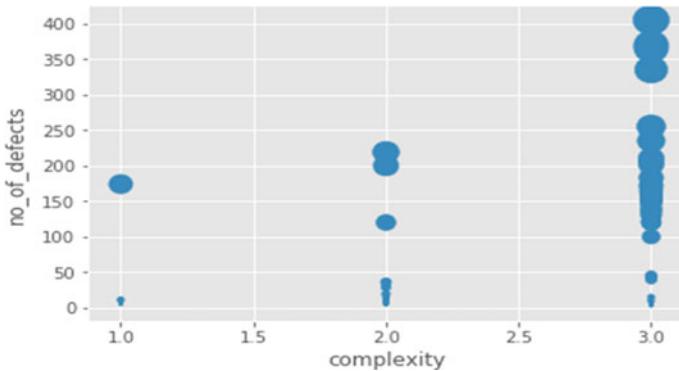
## 4 Results and Discussion

Software error assessment is a recent research topic; Many researchers have focused on providing efficient technology by providing software quality. Various technologies have been suggested and used before, but all have some limitations. Improving software quality is the primary goal of our specific software vulnerability assessment method (Table 1).

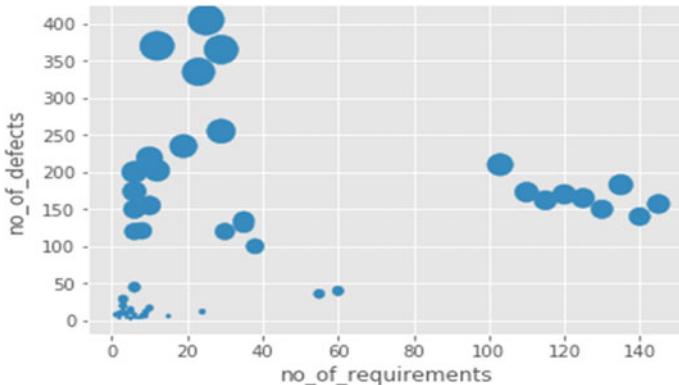
Improves software performance and efficiency. The targeted software is implemented in Java using the error prediction platform, Netbeans 8, and JDK 1.8. This section examines our performance based on various performance steps and provides sample results of the experiments. The reference method's effectiveness is evaluated in our specific software using evaluation metrics such as software execution time and reference rate. The evaluation metric evaluates our specific software, shows the evaluation methods' effectiveness, and justifies these systems' theoretical and practical

**Table 1** Evaluated parameters for software defect prediction in Cuckoo based search

Automation_scope	Brd_availability	Build_quality	Change_in_schedule	Complexity	Criticality	Dependencies_criticality
No	Yes	3	No	High	High	Medium
No	Yes	3	No	High	High	Medium
No	No	3	No	Medium	Medium	Medium
No	No	3	No	Medium	Medium	Medium
Development_methodology	Documentation_quality	Environment	Environment_downtime	Fs_availability	No_of_defects	No_of_requirements
Waterfall	5	Dev,SIT, UAT and Production	Yes	Yes	10	3
Waterfall	5	Dev,SIT, UAT and Production	Yes	Yes	100	38
Hybrid Agile	4	Integration, Regression, Production	Yes	Yes	12	24
Hybrid Agile	4	Integration, Regression, Production	Yes	Yes	6	15
No_of_test_cases	No_of_test_scenario	Release_duration	Requirement_validation	Technology	Type_of_requirement	Unit_test_defects
150	25	12	Yes	Java 8, Oracle 12 g, Thick client	Non Functional	5
700	70	24	Yes	Java 8, Oracle 12 g, Thick client	Functional	70
167	0	8	Yes	.NET	Functional	0
130	0	7	Yes	.NET	Functional	0



**Fig. 3** Number of complexity versus No. of defects



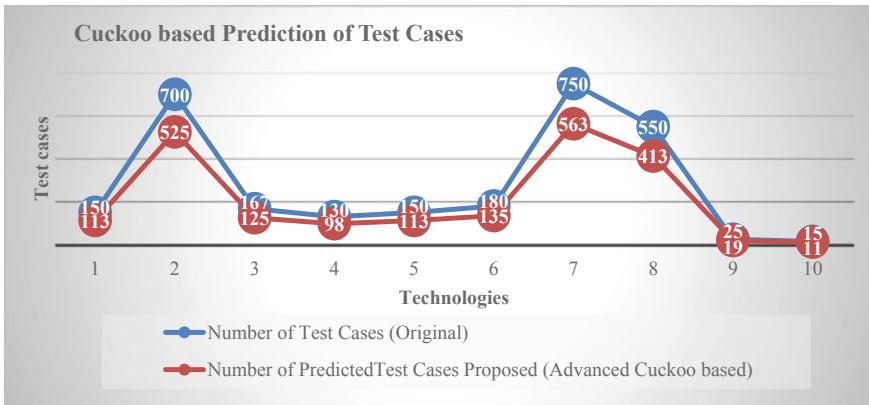
**Fig. 4** Number of defects and No. of defects

development. Compares with the algorithm of the future. A specific task's execution time is defined by how long the system performs that task, including runtime or system service execution time.

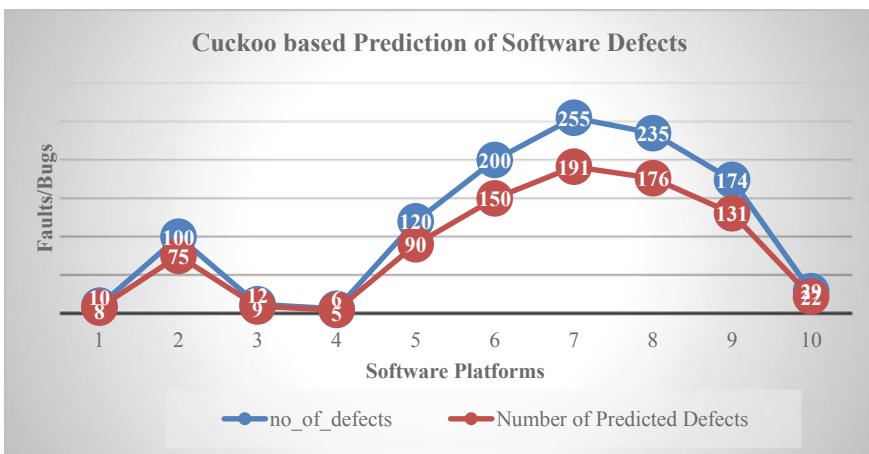
Prophecy is the announcement of an unexpected event. It is often, but not always, based on experience or knowledge. There is no agreement with the exact difference between the two terms; Different authors and categories explain different meanings (Figs. 3, 4, 5 and 6).

## 5 Conclusion

Software quality depends entirely on reliable and error-free software. Therefore, it is necessary to find a solution to fix software errors. Define a software error or



**Fig. 5** Cuckoo based prediction of test cases



**Fig. 6** Cuckoo based prediction of software defects

bug that could lead to an error, bug, error, and bug, error in a computer program or system or lead to a planned result. Some methods for assessing software damage are widely adopted, including Bayesian networks, SVMs, Innocent BIOSes, and other methods. Here, for clustering, we use the modified Cat-fish Algorithm (MFCM) in our recommended technology. In the future stage estimation process, the hybrid neural network and the MCS algorithm are used for the best estimation rate. In addition, we improve the efficiency of software error detection methods using MFCM and HHN with MCS. We analyzed the literature survey to find some software flaws in the software industry assessment process. Here, some of them have been greatly improved. Several strategies have been observed for improved efficiency; However, it has several limitations.

## References

1. Balogun, A.O., et al.: Performance analysis of feature selection methods in software defect prediction: a search method approach. *Appl. Sci.* **9**(13), 2764 (2019)
2. You, X., Ma, Y., Liu, Z.: An improved twin support vector machine based on multi-objective cuckoo search 291. An improved artificial bee colony algorithm for solving parameter identification problems. *Int. J. Comput. Sci. Math.* **8**(6), 570–579 (2017)
3. Sirisati, R.S.: Dimensionality reduction using machine learning and big data technologies. *Int. J. Innov. Technol. Explor. Eng. (IJITEE-2278-3075)* **9**(2), 1740–1745 (2019)
4. Sirisati, R.S.: Machine learning based diagnosis of diabetic retinopathy using digital fundus images with CLAHE along FPGA methodology. *Int. J. Adv. Sci. Technol. (IJASt-2005-4238)* **29**(3), 9497–9508 (2020)
5. Malhotra, R.: A systematic review of machine learning techniques for software fault prediction. *Appl. Soft Comput.* **27**, 504–518 (2015). <https://doi.org/10.1016/j.asoc.2014.11.023>
6. Rong, X., Cui, Z.: Hybrid algorithm for two-objective software defect prediction problem. *Int. J. Innov. Comput. Appl.* **8**(4), 207–212 (2017)
7. Yang, W.-H., Liu, J.-R., Zhang, Y.: A new local-enhanced cuckoo search algorithm. *Int. J. Comput. Sci. Math.* **8**(2), 175–182 (2017)
8. Yang, X.S., Deb, S.: Cuckoo search via levy flights. In: NaBIC 2009: World Congress on Nature and Biologically Inspired Computing, Coimbatore, India, pp. 210–214 (2010)
9. Yang, X.S., Deb, S.: Multiobjective cuckoo search for design optimization. *Comput. Oper. Res.* **40**(6), 1616–1624 (2013)
10. AL-Saati, N.A., Abd-AlKareem, M.: The Use of Cuckoo Search in Estimating the Parameters of Software Reliability Growth Models. arXiv preprint [arXiv:1307.6023](https://arxiv.org/abs/1307.6023) (2013)
11. Cai, X., et al.: An under-sampled software defect prediction method based on hybrid multi-objective cuckoo search. *Concurrency Comput. Pract. Exp.* **32**(5), e5478 (2020)
12. Cao, Y., et al.: An improved twin support vector machine based on multi-objective cuckoo search for software defect prediction. *Int. J. Bio-Inspired Comput.* **11**(4), 282–291 (2018)
13. Han, W., et al.: Cuckoo search and particle filter-based inversing approach to estimating defects via magnetic flux leakage signals. *IEEE Trans. Magnet.* **52**(4), 1–11 (2015)

# Human Facial Expression Recognition Using Fusion of DRLDP and DCT Features



M. Avanthi and P. Chandra Sekhar Reddy

**Abstract** Recognition of facial expressions is a major challenge in the field of computer vision. Using single-function models, the level of acknowledgment even in controlled capture conditions is considerably small. This paper proposed a method for facial emotion classification with the fusion of dimensionality reduced local directional pattern and discrete cosine transform features using SVM classifier. Local characteristics are extracted utilizing DRLDP, and global characteristics are extracted from facial expression images using DCT. SVM is used to classify the face images into six emotions (surprise, smile, sad, anger, fear and disgust). This method is experimented on JAFFE database and compared with existing approaches shows higher classification rate.

## 1 Introduction

Facial expression recognition plays an important role in computer vision-based applications like human–computer interaction, video interaction, cataloging, biometrics, including image recovery, with security, etc. Facial expression was its adjustments in the face in support of the inner emotional states as well as intentions of an individual person. Emotion is a familiar word used at a given moment for a person's feelings like surprise, smile, sad, anger, fear and disgust. Generally, emotions are identified with very little attempt of the human intelligence. Facial emotions machine identification and classification are cumbersome to realize people's feelings. An algorithmic methodology of classification is used for the labeling in one of the predefined sequences of provided input data. A classification algorithm is a template which executes the input data category.

One is the geometric methods of extraction based on features, while the other is the technique for extraction of features based on appearance. Geometric characteristic methods [1] derive the position but structure of facial components includes nose, eyes, mouth but eyebrows. The remaining part of the paper is organized as follows.

---

M. Avanthi (✉) · P. Chandra Sekhar Reddy  
CSE Department, GRIET, Hyderabad, Telangana, India

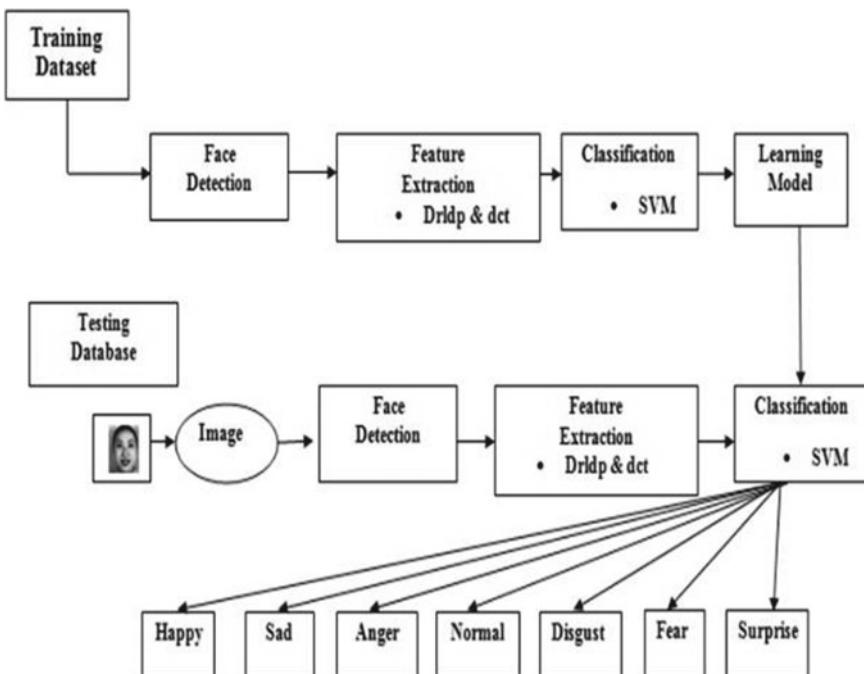
Section 2 proposed the methodology adopted; experimental results are given in Sect. 3 and conclusion in Sect. 4.

## 2 Methodology

After acquisition, the next sequence is to extract the information from input data, attributes such as eyes, nose, cheek, mouth, in case of geometric feature-based technique. Two main methods are used for the production of facial expressions (Fig. 1).

### 2.1 Local Directional Pattern

The local directional pattern is an 8-bit code representing edge responsiveness value. Kirsch masks ( $M_0, \dots, M_7$ ) shown in Fig. 2 are used to find edge response value in eight directions. LDP is computed considering only three prominent edge responses.



**Fig. 1** Architecture of facial expression

$$\begin{array}{cccc}
 \begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} & \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} & \begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} & \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \\
 \text{East } M_0 & \text{North East } M_1 & \text{North } M_2 & \text{North West } M_3 \\
 \begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} \\
 \text{West } M_4 & \text{South West } M_5 & \text{South } M_6 & \text{South East } M_7
 \end{array}$$

**Fig. 2** Edge response masks of Kirsch in eight directions

## 2.2 Facial Expression Recognition Using DRLDP

In this, human face reorganization method is shown in Fig. 3. The input images are preprocessed to decrease the noise, lighting recompense plus resizing. Then DCT is utilized to extract the feature vector. DRLDP [2] is utilized to diminish the measurements of extracted features. Features be particular as input to SVM classifier pro training the model. Then knowledge information base is updated. SVM classifies the test image into six different expressions such as shock, fear, sadness, joyfulness, vexation and disgust.

### 2.2.1 Dimensionality Reduced Neighborhood Directional Examples

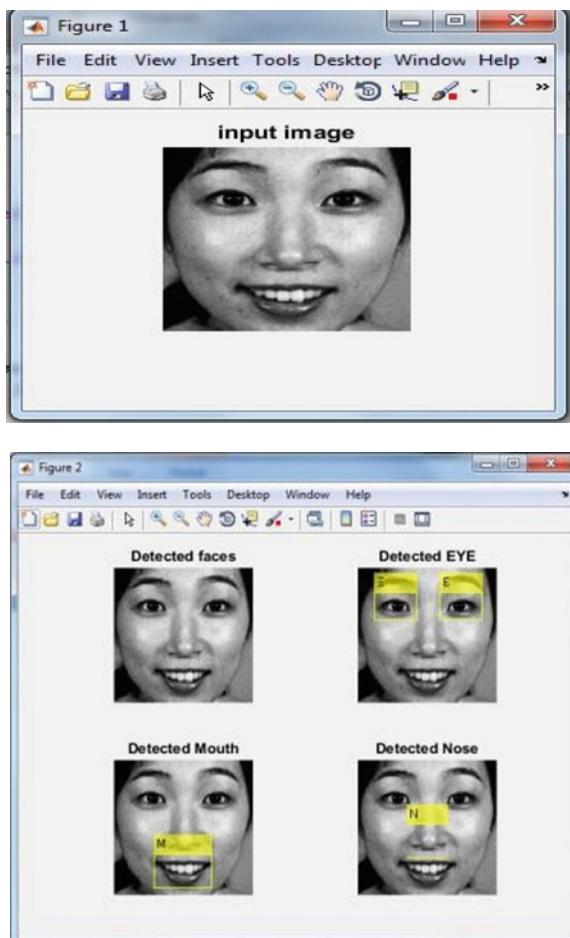
The suggested size reduced neighborhood positional example (DRLDP) is an eight-piece code assigned to each size three to three sub-districts. That software speaks to the square's textural instance. LDP is a single eight-piece code by each 18 a 3/3 square pixel. Both for square, the suggested DRLDP estimates a single eight-piece script. Models such as that of LDP are the suggested DRLDP statistics; however, differences arise in view of the fact which post handling of LDP design elements acquired for a square reduces the illustrations to a solitary eight-piece code.

For example, believe a picture I of size  $A \times B$ . Let  $p = A/t$ , and  $q = B/u$ .

For the picture I, the number of sub-images shaped is  $= A \times B t \times u \equiv p \times q$ .

The size of every sub-picture is  $ai = A/p \times B/q \equiv A \times B/a$  pixels. We describe the RR as the ratio of the no. of pixels in the input picture mapped onto the no. of pixels in the reduced image.

$$\text{Redundancy Ratio(RR)} = \frac{\text{No. of pixel in the input image}}{\text{No. of pixel in reduced image}}$$

**Fig. 3** Detected features

$$\text{For an image } I, \text{RR} = \frac{A \times B}{\frac{A \cdot B}{p \times q}} = p \times q = a$$

On the basis of two parameters  $t$  and  $u$ , the generalized DRLDP is supported. The general practice is to describe filters of size  $n \times n$ , i.e., a square mask. Therefore, it is unspecified that  $t$  plus  $u$  are equal.

### 2.3 Discrete Cosine Transformation

Two-dimensional DCT is used mainly to exclude the worldwide highlights from the exterior presence studies. The full face image is provided as a submission to DCT.

From the start, the image is divided into sub-image squares ( $8 \times 8$ ), and then subsequently discrete cosine transformation is used to extract the coefficients from each square. DCT produces one coefficient of DC as well as sixty-three coefficients were also dissimilar by each sub-square. Appropriately, it is registered again from above, and it left coefficients. Every sub-squares separated coefficients include ordinary vitality including recurrence information of under-square picture variety. Additionally, the upper as well as left sub-square districts speak to the information on the edge as well as directional substrate.

## **2.4 Support Vector Machine**

Support vector machine—similarly, it is a convincing AI technique or data characterization process; it introduces knowledge visualization into an elevated directional element space, as well as later finds a straight separation of the hyperplane some of the most extreme edges to differentiate data in the specified higher-dimensional space. SVM makes parallel choices, because of that multi-class grouping is capable, and this method teaches double classifiers to split one articulation as a whole, but also produces the largest yield of dual scheme class.

## **2.5 Classification**

For instance, upbeat, shock, outrage, tragic, dread, appall and unbiased will be used for intonation orders, and so on, and multi-class SVM is obtained. Seven categories are used for characterizing knowledge here. SVM is used for the conversion of Gabor highlights into vector structure. At the stage where the photo is provided as details again for test, Gabor is rendered on the direction of such an image, but instead transformed into a vector afterward. The information is isolated in two sections in SVM—training set and testing set, each involving the outline of the property. Each model is containing one objective method class name and a few traits.

## **2.6 Fusion**

By suggesting a detailed design clustering technique, the component vectors disentangled from either the input images are entangled. Straight, presently, permutation combination strategy is 20 used to intertwine the component vectors through using DCT as well as DRLDP, which have been removed again from data images. A mixture is used to enhance precision, increase efficiency and extend the power of the frame. Currently, combination schemes for summation as well as PCA are being used to

intertwine the neighborhood as well as the worldwide illustrates extracted from facial images.

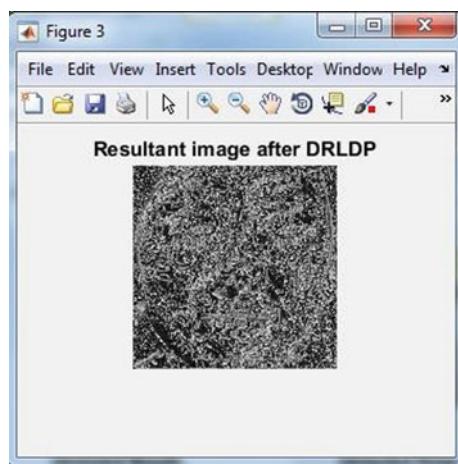
### 3 Experiments and Results

See Figs. 3 and 4.

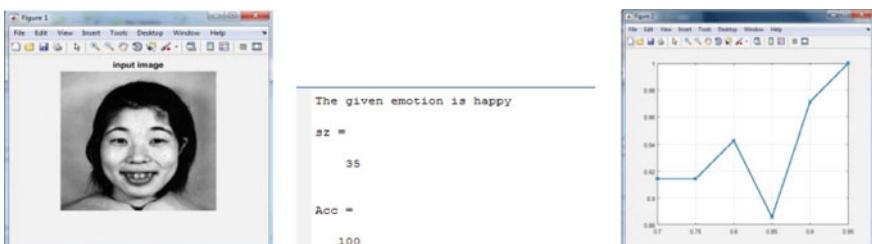
See Figs. 5 and 6.

See Fig. 7.

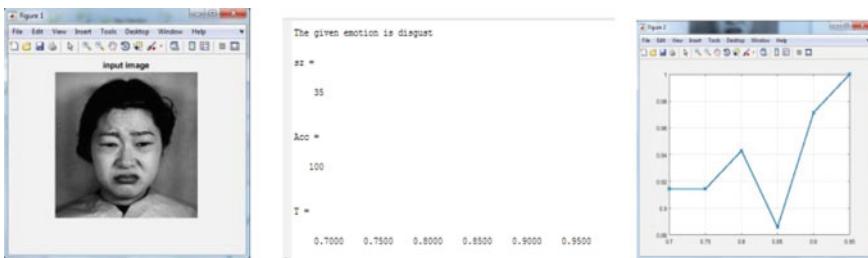
**Fig. 4** Result of the DRLDP



**Facial expression recognition results with LDP**

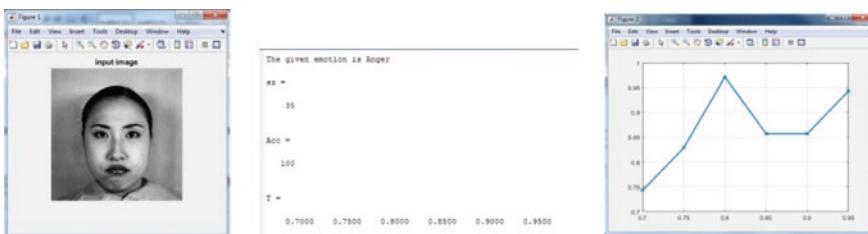


**Fig. 5** Accuracy of the input image



**Fig. 6** Accuracy of the Input Image

### Facial expression recognition results with DRLDP



**Fig. 7** Accuracy of the input image

## 4 Conclusion and Future Scope

The need for identification of facial expressions is rising rapidly. This approach is based on the combination of local and global significant features. This paper proposed a method for facial emotion classification with the fusion of dimensionality reduced local directional pattern and discrete cosine transform features using SVM classifier. Local characteristics are extracted utilizing DRLDP, and global characteristics are extracted from facial expression images using DCT. These features with SVM have classified considered database images for emotions with higher classification rate.

## Reference

1. H.-B. Deng, L.-W. Jin, L. Zhen and J. C. Huang, A New Facial Expression Recognition Method Based on Local Gabor Filter bank and PCA plus LDA International journal of Information Technology, vol. 11, (2005).
2. R. S.P., P.V.S.S.R. C.M., Dimensionality reduced local directional pattern (DRLDP) for face recognition, *Expert Systems with Applications*, (2016) 63 , pp. 66–73.

# Brain Tumor Classification and Segmentation Using Deep Learning



Manohar Madgi, Shantala Giraddi, Geeta Bharamagoudar,  
and M. S. Madhur

**Abstract** The brain is human body's most powerful organ and is responsible for regulating and maintaining all the body's essential life capabilities. Tumors are the outcome of anomalous and uninhibited cell division. A tumor is an aggregation of tissue that is formed by tremendous cell growth, which continues to grow. A brain tumor is produced in the brain itself or is grown and relocated in another place. No distinguishing cause for the growth of brain tumors has been recognized till date. Although tumors are not very common in the brain (worldwide brain tumors account for just 1.8% of total tumors recorded), the mortality rate of malignant brain tumors is very high due to criticality of the organ. Early detection of brain tumor is a difficult job for doctors. In this paper, authors introduced and implemented a method for classifying brain images with magnetic resonance as normal or abnormal. The abnormal images are further segmented to detect the brain tumor. Classification accuracy achieved is 100%. For segmentation, sensitivity achieved is 85%. Segmentation helps physicians to decide the course of treatment.

## 1 Introduction

The brain is a complex organ as it is made out of different cells. It includes 50–100 billion neurons. A Brain tumor can impact individuals of all ages. Tumors affected approximately 80,271 people in India in the year 2007. The National Cancer Institute (NCI) determined that in 2009 22,070 cases of brain tumor and the other disease of the central nervous system (CNS) were reported in the USA.

---

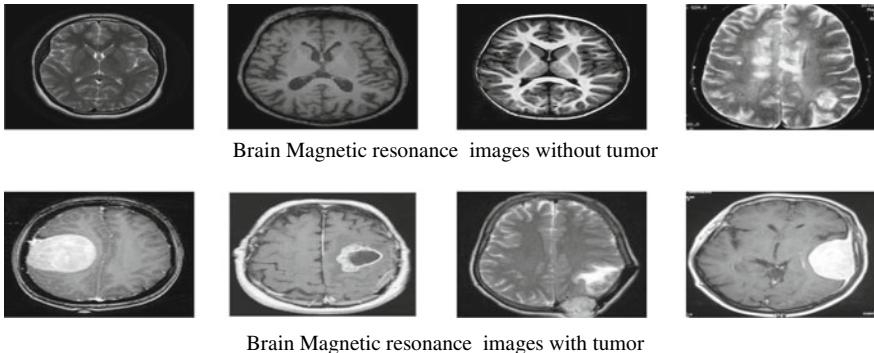
M. Madgi · G. Bharamagoudar · M. S. Madhur (✉)

Department of Computer Science and Engineering, KLE Institute of Technology, Hubballi, Karnataka 580027, India

S. Giraddi

Department of Computer Science and Engineering, KLE Technological University, Hubballi, Karnataka 580031, India

e-mail: [shantala@kletech.ac.in](mailto:shantala@kletech.ac.in)



**Fig. 1** Magnetic resonance images of brain without and with tumor

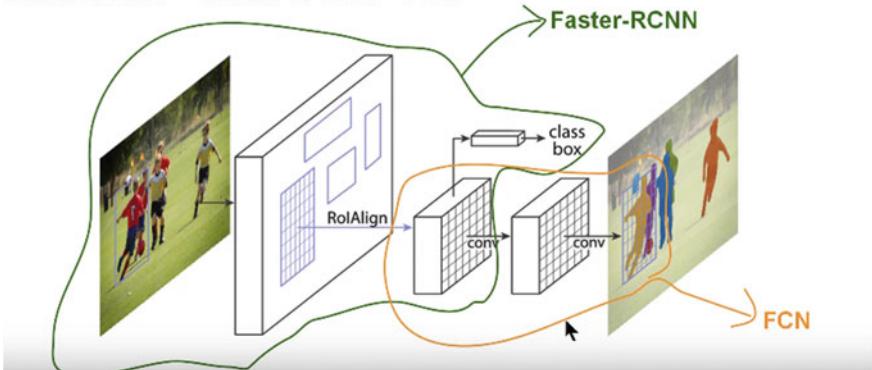
There are generally two forms of brain tumor, primary and secondary. A main brain tumor is a tumor that originates in the brain. Secondary tumors begin at other parts of the body and spread to the brain. The main tumor can be harmless or injurious. The healthy brain tumor is referred to as benign, while the dangerous tumor is referred to as malignant tumor. The benign kind of tumor grows bit by bit, and it rarely spreads to various regions of the body. It has distinct boundaries. Ultimately, this sort of tumor can be eliminated by operation and there are fewer chances to return. The malignant sort of tumor rapidly increases than the benign kind of tumor. Healthy brain cells also get affected by this sort of tumor. Such a tumor can appear back extensively after a clinical method. It can spread to various regions of the brain or spinal line. Detection of brain tumor is a challenging task. Diagnosis begins with MRI. Even though other modalities like computed tomography (CT) and ultrasound exist, MRI is the most appropriate for brain tumor detection. MRI plays an important role, as mode of treatment depends on the nature of the tumor. Type of tumor can be determined with a biopsy. Figure 1 shows brain MRI without and with tumor.

### 1.1 Mask R-CNN Architecture

Mask R-CNN uses the region proposal network (RPN) to spot items that divide the image into anchors where the image is traversed by sliding windows and areas of interest. The groundwork of this architecture is a ResNet101-CNN, in which initial layers detect low-level and higher-level characteristics recognized by next layers. Mask R-CNN uses function pyramid network (FPN) to raise the regular element extraction pyramid by unveiling an extra pyramid that takes suggestive-level highlights and transfers them down to the lower tiers, allowing all levels to address highlights at both upper and lower levels.

Mask R-CNN is a continuation of faster R-CNN, and it consists of a furthermore fully convolution network FCN. Faster R-CNN performs object detection by

### Mask R-CNN → Faster R-CNN + FCN



**Fig. 2** Faster R-CNN extended by FCN to form mask R-CNN

providing a bounding box over the object in a scene. Faster R-CNN comes with two outputs, class label for the object and bounding box over the object. Fast R-CNN is extended by FCN which performs the segmentation by classifying each pixel. The working of mask R-CNN is shown in Fig. 2.

**FCN:** Using different convolution blocks and max pool layers, this model first decompresses an image up to 1/32 of its original size. At this point of granularity, it then makes a class prediction. Finally, it resizes the image to its original dimensions using sampling and deconvolution layers.

## 2 Related Work

Malati et al. [1] have suggested completely automated brain tumor segmentation using neural convolution network. The suggested research performs brain tumor segmentation using tensor flow, in which high-level mathematical functions are implemented using the anaconda frameworks. Patient survival rates are improved through early diagnosis of brain tumor. Brain tumor segments are classified into four groups such as edema, non-enhancing tumor, tumor enhancement and necrotic tumor. The result shows that the approach applied only helps to identify tumor enhancement and assign tumor to the specific area of the tumor.

Talo et al. [2] have presented an approach to classify abnormal brain MR images using pre-trained deep transfer learning. ResNet34 model is used with data increase, optimum learning rate finder and fine-tuning. It achieved classification accuracy of 100%.

Wang et al. [3] have suggested the use of cascade of CNNs for brain tumor segmentation. They have proposed a 2.5D network which is a trade-off between memory requirement and model complexity. It has low memory requirements. Authors

have also employed run time augmentation which has yielded better segmentation accuracy.

He et al. [4] have presented an approach, called mask R-CNN for object segmentation. The approach detects the object as well as generates a segmentation mask. The design of mask R-CNN is a continuation of faster R-CNN. To train mask R-CNN is easy and straightforward.

Chinmayi et al. [5] using Bhattacharya coefficient explored a method for segmentation and classification of brain tumor MRI. They removed the unwanted parts of the skull using anisotropic diffusion filter. Further, a fast-bounding box of algorithms extracts the tumor region. They used CNN's deep learning to train the MRI brain images. Compare the effects of the proposed approach in terms of precision, similarity, index, PSNR and MSE. The findings will assist the radiologist in determining a tumor's size and location.

Dong et al. [6] have suggested noninvasive magnetic resonance techniques as a brain tumor screening method for detecting brain tumors without ionizing radiation. Manual segmentation of the volumes of the 3D MRI requires longer, and the output is focused mainly on the experience of the operator. So, the author suggested a deep convolution network focused on u-net. They perform this segmentation on BRATS 2015 datasets, which involves 220 high-grade brain tumor glioma and 54 low-grade cases of tumor. They compared the efficiency of our proposed approach to the deep neural network based on manual delineated ground truth. U-net provides the excellent results for key tumor areas.

Havaei et al. [7] have proposed a segmentation of brain tumor using deep neural networks to glioblastomas MRI image. This form of brain tumor occurs anywhere in the brain and has every shape, size and contrast as well. The article makes use of the convolution neural network as an algorithm for machine learning. It utilizes both local and global characteristics for segmentation of tumors. For research work, the author uses the BRATS dataset.

Cui et al. [8] developed a novel, automatic segmentation based on cascaded deep learning convolution neural networks. It has two subnetworks: tumor location network (TLN) and intra-tumor classification network (ITCN). The tumor region from the MRI brain slice is segregated by helping to mark the identified tumor region into multiple subregions using tumor localization, network and ITCN. The research was performed on a sample with 220 cases of high-grade glioma (HGG) and 54 cases of low-grade glioma (LGG), multimodal brain tumor segmentation (BRATS, 2015). The measurement can be achieved by coefficient of dice, positive predictive value (PPV) and sensitivity.

Shantala et al. [9] have carried out the analysis on exudate detection in retinopathy images using DBSCAN clustering and fuzzy classifiers. Accuracy of 90% is obtained with image-based evaluation.

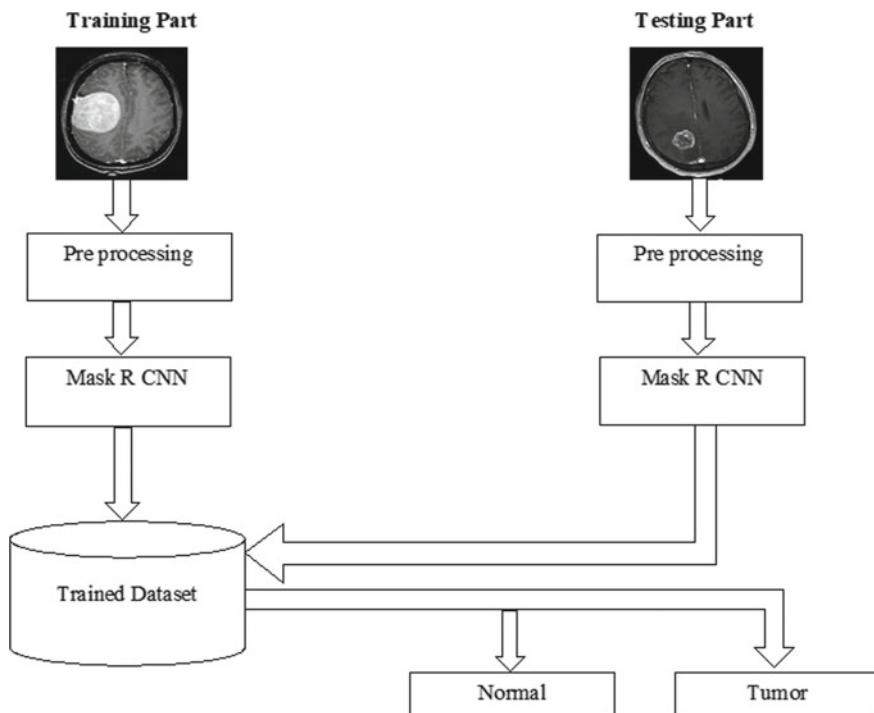
Shantala et al. [10] have conducted study to detect retinopathy images using DBSCAN clustering and neural networks that spread the back. The authors performed image-based as well as lesion-based result analysis. Accuracy of 100% is obtained with image-based performance analysis.

### 3 Methodology

**Dataset Description:** For conducting the brain tumor detection, Kaggle Brain MRI Images dataset is used. It has 253 images, 98 normal images and 155 annotated abnormal or deceased images. The proposed method was implemented using Python TensorFlow version 1.15.0 and Keras libraries version 2.2.5. The tests were performed on processor Intel(R) Core(TM) i3-5005U CPU @ 2.00 GHz, RAM 4.00 GB, operating system 64 bits, a processor based on ~64. The images are resized to  $256 \times 256$ . Eighty percentage images are used for preparation, evaluation 10% and checking 10%. The training parameters are learning rate is 0.0001 and epoch's number is 15.

**Transfer Learning:** Deep learning requires a large number of samples. Mask R-CNN requires high time for training where weights are initialized randomly. To overcome both these drawbacks, we can use transfer learning. In transfer learning, pre-trained models serve as initial points for training our model. We have used a pre-trained model mask\_rcnn\_coco.h5 which has been trained on MS COCO dataset. MS COCO dataset is object detection and segmentation dataset.

The approach suggested is shown in Fig. 3. The dataset consists of images of various sizes.



**Fig. 3** Proposed methodology of classification and segmentation

Dataset is cleaned by finding the extreme points of the image and crops them into rectangular shape. Then, images are annotated using the VIA tool, and annotated data is stored in a JSON file. After this, we copy the mask R-CNN archive which has the design for mask R-CNN model along with annotated data files. Mask R-CNN uses ResNet101 architecture to select the component outline from the images. These outlines of the components are then transferred through a region proposal network (RPN) that outputs the bounding boxes for the image. Over these image bounding boxes, we then put a RoI aligned layer to deliver all images to the same size. These regions are then transferred across a fully connected network to predict the class mark and bounding boxes. Next, we will measure the region of interest in order to reduce the calculation time. We measure the intersection over union (IoU) with the boxes of ground truth for all of the predicted regions. IoU is computed with Eq. 1:

$$\text{IoU} = \frac{\text{Area of intersection}}{\text{Area of union}} \quad (1)$$

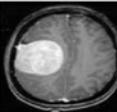
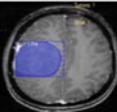
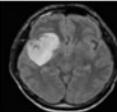
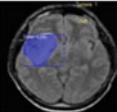
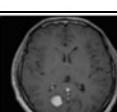
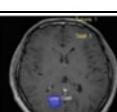
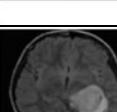
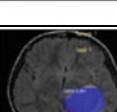
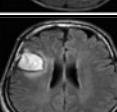
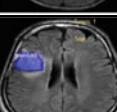
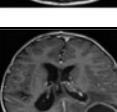
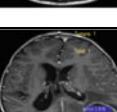
Now, we regard that as an area of interest whose IoU value is higher than or equivalent to 0.5. Otherwise, we are unconcerned about that specific region. We are doing this for all regions and then consider only one set of regions under which the IoU has been greater than 0.5. We can add a mask branch to the current architecture when we have the RoIs centered on the IoUs values. It returns the mask for segmentation for each area that contains an object. For each region, it returns a mask of size  $28 \times 28$  which is then scaled up for inference.

## 4 Results and Discussion

In this section, we illustrate the outcomes the proposed framework has achieved. The device provided is validated using 10% of images. Figure 4 shows the sample images of classification/segmentation results along with the accuracy. The first column shows the original images, second column displays the segmented images, third column indicates whether the images is normal or abnormal, and last column displays the confidence score. A classifier is assigned to each segmented region of a class label and a confidence score indicating the “confidence of the classifier” that the label is right. The method of segmentation and classification where the input of the classifier is used to achieve a final “stable” segmentation.

## 5 Conclusion

Brain tumor classification and segmentation using deep learning is achieved through an innovative technique. In this paper, we exhibit the mask R-CNN design which is mainly used for object detection, object localization and instance segmentation of

Original Image	Segmented Image	Classification	Confidence Score
		Abnormal	94%
		Abnormal	99%
		Abnormal	92%
		Abnormal	85%
		Abnormal	96%
		Abnormal	97%

**Fig. 4** Segmentation results of mask R-CNN

brain MRI images. Segmentation is a challenging task, and mask R-CNN is applied to attain extreme excellence results. The exhibited technique has invariant characteristics in terms of dimension, separation and magnitude of brain tumor. Experimental growth shows that the exhibited technique works fine in improving, separating and getting the brain tumor from MRI images. There are various very much alike jobs in medical image analysis for which probably mask R-CNN-based technique could be not difficult to enhance the efficiency without major alteration or customization. The procedure used for selecting right characteristics improves to categorize the fundamental work into normal and abnormal tissue which may decrease the complicatedness. In the end, it is accomplished that the performed analysis is of notable significance in brain tumor classification and segmentation using deep learning which is one of the challenging assignments in medical image treatment. This work will be beneficial for creating modernized collection of designs for brain tumor detection.

which may support higher effective results than current technique. Occurrences here include the practice of separating the brain's left ventricle, where correct division can be used to measure the ejection fraction of a heart patient to render better results or division of the liver to tumor. Future implementation will analyze the effectiveness and efficiency of mask R-CNN established designs for a collection of specific assignments.

**Acknowledgements** We would like to express our thanks to Dr. Basavaraja S. Anami, Principal, K. L. E. Institute of Technology, Hubballi, Karnataka, India, for his valuable suggestions.

## References

1. Malathi, M., Sinthia, P.: Brain Tumor segmentation using convolutional neural network with tensor flow. *Asian Pac. J. Cancer Prev.* **20**, 2095–2101 (2019)
2. Talo, M., Baloglu, U.B., Yildirim, O., Rajendra Acharya, U.: Application of deep transfer learning for automated brain abnormality classification using MR images. *Cogn. Syst. Res.* **54**, 176–188 (2019)
3. Wang, G., Li, W., Vercauteren, T., Ourselin, S.: Automatic brain tumor segmentation based on cascaded convolutional neural networks with uncertainty estimation. *Front. Comput. Neurosci.* **13**, 56 (2019)
4. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
5. Chinmayi, P., Agilandeswari, L., Prabu Kumar, M., Muralibabu, K.: An efficient deep learning neural network based brain tumor detection system. *Int. Pure Appl. Math.* **117**, 151–160 (2017)
6. Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y.: Automatic brain tumor detection and segmentation using U-net based fully convolutional networks. *MIUA* **3**, 1–12 (2017)
7. Havaei, M., Davy, A., Warde Farley, D.: Brain tumor segmentation with deep neural networks. *Med. Image Anal.* **35**, 18–31 (2017)
8. Cui, S., Mao, L., Jiang, J., Liu, C., Xiong, S.: Automatic semantic segmentation of brain gliomas from MRI images using a deep cascaded neural network. *J. Healthcare Eng.* 1–14 (2018)
9. Giraddi, S., Pujari, J.: Automated detection of exudates using DBSCAN clustering and fuzzy classifier. *Int. J. Adv. Res. Comput. Sci.* **3**(6), 129–133 (2012)
10. Giraddi, S., Pujari, J., Giraddi, S.: Exudates detection with DBSCAN clustering and back propagation neural network. *Int. J. Comput. Appl.* **86**(19), 16–20 (2014)

# A Hybrid Approach Using ACO-GA for Task Scheduling in Cloud



Simran Shrivastava, Sonika Shrivastava, and Lalit Purohit

**Abstract** Cloud computing is a fast-growing technology in today's world that offers a wide variety of services based on the need of its users. Cloud provides services like computing, storage, and networking which are accessible through the Internet. In the computing cloud, task scheduling plays a vital role in optimizing resource utilization and providing quality of service (QoS) to the customer. Efficient task scheduling is needed to fulfill the user requirement and system performance. Traditional task scheduling algorithms in a cloud platform like max–min, shortest job first, and round-robin are not so effective in reducing makespan and cost. The main motive behind this research is to apply a combination of the genetic algorithm for task scheduling in a cloud environment and analyzes its effect, as the genetic algorithm can efficiently solve NP-hard problems. This paper proposes a hybrid approach task that combines the advantages of the two most widely used evolutionary algorithms: genetic algorithm (GA) and ant colony optimization (ACO) for improving the scheduling in the cloud using hybrid approach to overcome the limitation of unnecessary diversity. The experimental result conclusively proved that there is an 18% to 20% reduction in cost and makespan by ACO-GA-based task scheduling in the cloud as compared to simple GA and ACO.

## 1 Introduction

In this rapidly developing world, the use of a cloud computing environment is inevitable. The demand for computation and storage is fulfilled with the help of a cloud platform. Cloud provides all types of resources, including servers, storage, networking, etc., over the Internet based on the pay-as-you-go pricing model [1]. Cloud services are divided into three categories: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS) model, and those service models provide access to the computing, storage, and networking resources at minimum cost and also provide a platform for programmers and developers to

---

S. Shrivastava (✉) · S. Shrivastava · L. Purohit

Shri Govindram Seksaria Institute Technology and Science, Indore, Madhya Pradesh, India

deploy, run, and develop the software. Many industries such as banking, health care, and education are using the cloud on daily basis. Cloud is a parallel and distributed computing platform that dynamically provides computing resources based on SLA established between the CSP and users. With the increasing usage of this business computing model, proper scheduling of tasks is an important issue [2]. The main aim of the task scheduling (TS) technique is to assign the incoming task to the possible resources efficiently [3]. Mainly, task scheduling processes are employed in infrastructure as a service model for offering a high QoS [4]. A proper task scheduling algorithm is needed for efficient system performance. In the past few years, several research activities had been carried out to optimize the task scheduling (TS) algorithms [5] and they proposed algorithms such as max–min [6], genetic algorithm (GA) [6], ant colony optimization (ACO) [7], and particle swarm optimization [8]. The main problem with previous applied evolutionary algorithms in the TS domain is that while reducing the total cost and computational time, they stuck in local optimum and result in slow convergence. So, the main aim of this paper is to address the above-mentioned problems and to reduce makespan and cost for task distribution in the cloud environment.

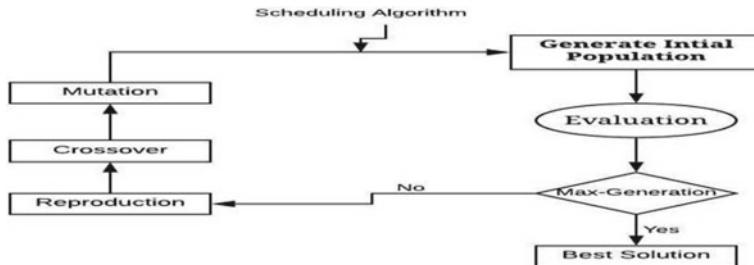
The rest of the paper is organized as follows. The overview of genetic-based TS in the cloud is covered in Sect. 2. Section 3 explains the proposed algorithm and followed by Sect. 4 which provides the detail of performance evaluation with the experimental results. Conclusion and future work are summarized in Sect. 5.

## 2 Evolutionary Algorithm for Task Scheduling (TS) in Cloud

In a dynamic environment such as the cloud, allocation of the task requires some good strategy to get better performance [9]. The detailed study of a different genetic algorithm and ant colony optimization algorithm applied in the cloud environment is covered in this section.

### 2.1 Genetic Algorithm (GA) for Task Scheduling (TS)

GA is an evolutionary algorithm that is inspired by biological evolution and uses genetic operators for better results [10]. In the simple GA, population is a subset of all the possible solutions in the given problem, the chromosome indicates task allocation information, and the length of the chromosome is the same as the number of input tasks. Individual is a set of genes, and that gene represents each task id of the chromosome [11]. Figure 1 shows the basic steps of simple GA in the cloud environment. In [12], Duan et al. proposed the adaptive incremental genetic algorithm (AIGA) that uses the probability of mutation and crossover operator to change



**Fig. 1** Genetic algorithm for task scheduling

the genetic operator's results according to the individuals. Standard genetic algorithm (SGA) [12] uses biological concepts but has some limitations like infeasible solutions.

In [13], the multi-agent GA (MAGA) uses two types of methods. For the first phase, it uses self-learning operators and in the second phase uses min–min and genetic algorithm. It gives superior results as compared to traditional genetic algorithms. This algorithm handles high-dimensional function. It provides an effective solution but is stuck in slow convergence.

## 2.2 Ant Colony Optimization (ACO) for Cloud TS

ACO is a heuristic algorithm to find an optimal solution for complicated problems. In this algorithm, the ant moves on the problem space, searches for better computing resources for a required task, and releases a pheromone to build a solution; the values may be modified during run time by the ant. Artificial ant has its memory to save the path [14]. In Kumar [15] proposed the best choice resource allocation for all tasks dynamically and reduce task completion time, but there is no precedence constraints between task and required more computation time. In [16], slave ant used to attain global optimization by ignoring long paths and also use minimal pre-processing overhead for slave ants and uses machine learning technology to optimize the performance of the system. Table 1 shows the comparative analysis of different evolutionary algorithms used for TS.

## 3 Proposed Hybrid ACO-GA Approach for Task Scheduling (TS) in Cloud

The problem of mapping tasks on cloud computing resources comes in the category of the NP-hard problem [17]. To get a better result in terms of time and space, there are various meta-heuristic algorithms used in TS, but they have some drawbacks

**Table 1** Comparative study of evolutionary algorithm for TS in cloud

Author	Methods	Objective	Limitation
Pande [4]	Task partitioning algorithm	Makespan and cloud resources	Work on communication time
Yiqiu [3]	Improved genetic algorithm	Load balancing, set of resources	Only work in static scheduling
Yin [9]	Task allocation model	Local balancing, execution cost	Accuracy is required
Zhu [13]	Multiple agent algorithm	Multiple QoS constraints	Premature convergence
Gang [15]	Hybrid heuristic GA	Cost	Not independent

like premature convergence, infeasible solution, and trap in a local optimum [18]. In this work, the ACO-GA hybrid approach is used with the help of ACO, it finds out superior solutions for TS, and GA will control the problem of ACO, i.e., gradually increase in running time with every generation, which reduces the overall makespan and cost of TS in the cloud. The flowchart of the proposed approach is explained in Fig. 2.

### Phase I: Initialization Population

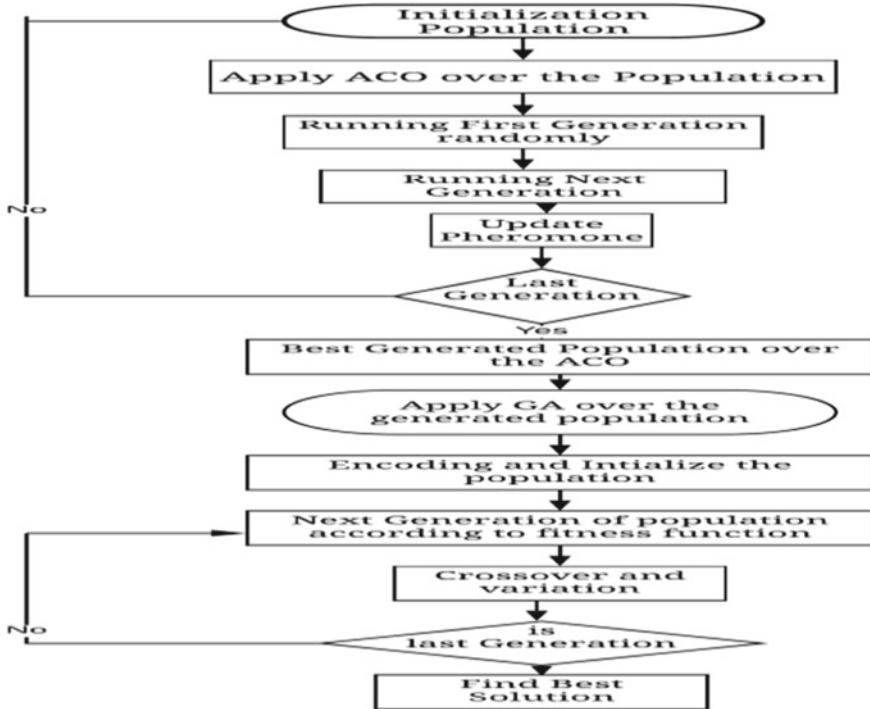
In this hybrid ACO-GA algorithm, first iteration is initialized by the random population method.

### Apply ACO over the population

ACO works on the whole population. It is applied such that every ant finds the path in each iteration when their previous iteration is complete. The pheromone matrix of every ant updates the global pheromone using the matrix model as shown in Eq. (1). In this phase, population is generated based on the total number of iteration of an ant and it is denoted in the equation by MAX\_GEN; as the previous iteration is complete, every ant releases pheromone which is denoted by deltaT as given in Eq. (2).

### Phase II: Apply GA over the population

GA uses genetic operators over the generated population given by phase I. In the selection step, tournament selection method is used to select the best chromosome from a group of the chromosome [15]. Cross over operator, select a random number, and return the offspring chromosome that contain both gene and VM [19]. In the mutation operator occurrence is based on the mutation rate variable. After some iteration, the fitness function gradually improved. The idea of the ACO-GA algorithm is that first run the ACO and then GA for intermediate result and then calculate the fitness value according to the specified objective. In the existing algorithm, if same



**Fig. 2** Flowchart of ACO-GA algorithm

individual is used for finding an optimal solution, they affect unnecessary diversity [20], but using a combination of the proposed approach improves the existing drawback. The implementation of the proposed algorithm is shown in algorithm 1.

$$P = \text{Phermone} \left[ \begin{pmatrix} p_0 & \dots & p_n \\ \vdots & \ddots & \vdots \\ p_{k0} & \dots & p_{km} \end{pmatrix} * (1 - rho) \right] \quad (1)$$

$$\Delta t = \frac{Q}{\text{Max}(CVm)} \quad (2)$$

$$P_{ij}^k(t) = \left\{ [T_{ij}(t)]^\alpha \cdot [r_{ij}(t)]^\beta / \sum_{l \in \text{AllowTask}(t), h \in v_m} [T_{lj}(t)]^\alpha \cdot [r_{lj}(t)]^\beta \right\} \quad (3)$$

$$\text{Fitness} = \frac{Q'}{r * \text{max}(CVm) + (1 - r) * D} \quad (4)$$

$$D = pe * cPe * \sum CVm(t) + Cpe * mem + CpeS * size + CpeB * bw \quad (5)$$

---

**Algorithm 1** For ACO-GA-based TS in cloud
 

---

**Input:** Cloudlets (Task: T1, T2,...) and set of resources [VM1, VM2, VM3...]

**Output:** Best Solution. // allocate over the VM

**For**  $i = 0$  to  $p$

$\text{Popu}(p) \leftarrow \text{randomize}()$  in Eq. 1 // initialize population

**End For**

**While** not reach half of  $n$  **do** // $n$  is number of iteration

**While** not reach MAX\_GEN **do**

Calculate DeltaT in Eq. 2

At each iteration every ant find path and calculate Pheromone  $P_{ij}^k(t)$

**Repeat**

**Repeat**

Set  $\text{Generatedpopulation}_j$  as chromosome //  $j$  is the index

Initialize generated chromosome and apply GA operators

$\text{Chromosome}_i \leftarrow \text{tournament}(\text{popu})$  // selection operator

$\text{Chromosome}_j \leftarrow \text{tournament}(\text{popu})$

$\text{offspring\_chromosome}_j \leftarrow \text{crossover}(\text{chromosome}_j, \text{chromosome}_i)$

$\text{Generatedpopulation}_j \leftarrow \text{mutation}(\text{offspringchromosome}_j)$

**Repeat**

---

Various parameters computed are used for ACO-GA discussed in Eqs. (1)–(5), where  $Q'$  is common,  $r$  is control weight, and Pe is the number of processing elements in a single virtual machine. CPe is the cost of a single element. CpeS is storage cost, CpeB is the cost of bandwidth, bw is the amount of bandwidth in Eq. (5) [19], and Rho is denoted as a weight parameter.

## 4 Implementation and Result Analysis

The proposed algorithm is implemented using a CloudSim simulator [21, 22]. For this experiment, three scenarios are considered with varying number of tasks. The algorithm starts with the 20 random tasks called ants. The algorithm uses a trail step to update the pheromone values and calculate the fitness function; beta is the influence of the virtual machine instance. The count of ants is twice as the allocation of the VM. So that, generation controlled in the first stage is double. In the GA algorithm phase use, ACO generated a solution called population. The crossover method and the mutation rate operator in the GA phase are defined in Table 2.

Furthermore, the workload data used a different number of tasks (20–1000) for evaluating the algorithm with various scenarios. In the ACO-GA algorithm, number of iterations is used to reach the optimal solutions. The experiments are repeated more than 30 times given the average values of results compared with other similar approaches. The size of data is changed to check the ability of the ACO-GA to reduce makespan and cost.

**Table 2** ACO-GA-based TS parameters

Parameter	Value
Population size	50
Alpha	0.3
Mutation rate	0.05
Crossover	Single point
Iteration	10–100

For each scenario, the number of tasks and the number of iterations used are shown in Table 3. In the first scenario, it has a small search space and gives a fast and optimal result as compared to other existing approaches. ACO-GA algorithm performs better in the second and third scenario as compared to other approaches [21]. The third scenario represents the worst-case scenario. Experiment was done in all the three scenarios to evaluate the effect on cost and time as compared with individual approach based on ACO and GA as shown in Table 4. These experiments are carried out with gradually increasing the number of iteration and number of tasks with the allocation on different virtual machines.

At last, the comparative graph for makespan and cost is shown in Figs. 3 and 4. It is observed that with the increase in the iteration time, more is the cost-saving. The experimental result conclusively has proven that there is the reduction in cost and makespan by ACO-GA-based TS in the cloud as compared to simple GA and ACO and also this approach avoids unnecessary diversity and randomness.

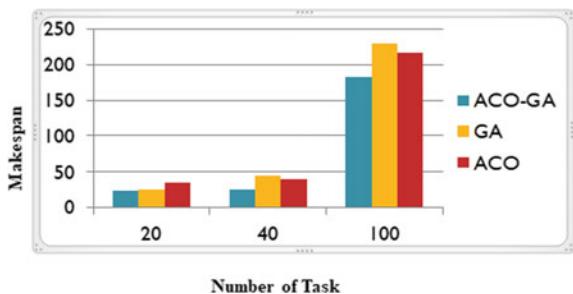
**Table 3** ACO-GA algorithm parameters

Scenario	Task	Iteration
First	20	10
Second	40	30
Third	100	50

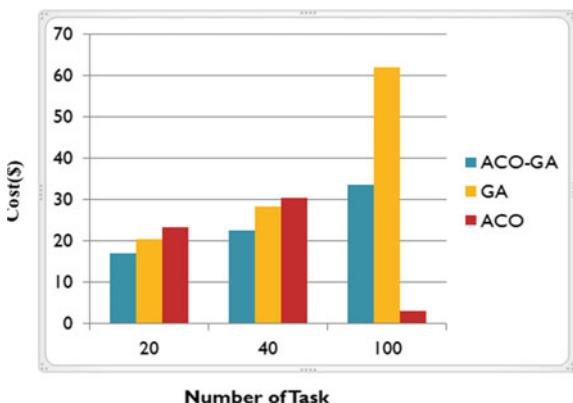
**Table 4** Result of ACO-GA-based task scheduling in different scenarios

Approach based on	Scenario first		Scenario second		Scenario third	
	Makespan (s)	Cost (\$)	Makespan (s)	Cost (\$)	Makespan (s)	Cost (\$)
ACO-GA	22.9	17.06	24.6	24.3	183.3	35.05
GA	30.4	20.3	44.6	28.6	230.2	61.3
ACO	35.2	23.4	39.3	30.4	216.3	52.3

**Fig. 3** ACO-GA makespan compared with simple GA and ACO over a different number of tasks



**Fig. 4** ACO-GA cost compared with simple GA and ACO over a different number of tasks



## 5 Conclusions

The GA-based algorithms used for task scheduling in the cloud environment were only considering single objective function, and they were affected with poor resource utilization and local optimal solution. The proposed ACO-GA-based task scheduling algorithm in the cloud takes into consideration both cost and makespan and implements the ACO-GA with different fitness functions. The experimental results state that proposed ACO-GA algorithm obtains 18–20% better results in terms of cost and makespan as compared to simple GA and ACO.

## References

- Buyya, R., Yeo, C., Broberg, J.: Cloud computing and emerging IT platforms: vision, and reality for delivering computing. In: 5th utility, Future Generation Computer Systems, vol. 25, pp. 599–616 (2009)
- Purohit, L., Kumar, S., Kshirsagar, D.: Analyzing genetic algorithm for web service selection. In: 1st International Conference on Next Generation Computing Technologies (NGCT), Dehradun, pp. 999–1001 (2015)

3. Arunarani, A., Manjula, D.: Task scheduling techniques in cloud computing a literature survey. In: Future Generation Computer Elsevier, vol. 91, pp. 407–415 (2019)
4. Pande, S.K., Pande, S.K., Das, S.: Task partitioning scheduling algorithms for heterogeneous multi-cloud environment. *Arab J. Sci. Eng.* **43**, 913–933 (2018)
5. Padhy, R.P., Patra, M.R., Satapathy, S.C.: Cloud computing: security issues and research challenges. *IRACST Int. J. Comput. Sci. Inf. Technol. Secur.* **1**(2), 136–145 (2011)
6. Yiqiu, F., Xia, X., Junwei, G.: Cloud computing task scheduling algorithm based on improved genetic algorithm. In: IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, pp. 852–856 (2019)
7. Gupta, A., Garg, R.: Load balancing based task scheduling with ACO in cloud computing. In: International Conference on Computer and Applications (ICCA), Doha, pp. 174–179 (2017). <https://doi.org/10.1109/COMAPP.2017.8079781>
8. Gawali, M., Shinde, S.: Task scheduling and resource allocation in cloud is using a heuristic approach. *J. Cloud Comput. Adv. Syst. Appl.* **7**(4), 1–16 (2018)
9. Yin, S., Ke, P., Tao, L.: An improved genetic algorithm for task scheduling in cloud computing. In: 13th IEEE Conference on Industrial Electronics and Applications, Wuhan, pp. 526–530 (2018)
10. Purohit, L., Kumar, S.: A study on evolutionary computing based web service selection techniques. *Artif. Intell. Rev.* (2020). <https://doi.org/10.1007/s10462-020-09872-z>
11. Kairong, D., Fong, S., Weng, S., Wei, S., Sheng-Uei, G.: Adaptive incremental genetic algorithm for task scheduling in cloud environments. *Symmetry* **10**, 168 (2018). <https://doi.org/10.3390/sym10050168>
12. Jang, S., Young, T.: The study of Genetic algorithm based task scheduling for cloud computing. *Int. J. Control Autom.* **5**(4) (2012)
13. Zhu, K., Song, H., Liu, L., Gao, J., Cheng, G.: Hybrid genetic algorithm for cloud computing applications. In: IEEE Asia-Pacific Services Computing Conference, Jeju Island, pp. 182–187 (2011)
14. Verma, P., Shrivastava, S., Pateriya, R.K.: Enhancing load balancing in cloud computing by ant colony optimization method. *Int. J. Comput. Eng. Res.* **6**, 277–284 (2017)
15. Purohit, L., Kumar, S.: Replaceability based web service selection approach. In: IEEE 26th International Conference on High Performance Computing, Data, and Analytics (HiPC), Hyderabad, India, pp. 113–122 (2019). <https://doi.org/10.1109/HiPC.2019.00024>
16. Gang, L., Zhipun, W.: Ant colony optimization task scheduling algorithm for SWIM based on load balancing. In: Future Internet (2019)
17. Moon, Y., Yu, H., Gil, J.: A slave ants based ant colony optimization algorithm for task scheduling in cloud computing environments. *Hum. Cent. Computer. Inf. Sci.* **7**, 28 (2017)
18. Abualigah, L., Diabat, A.: A novel hybrid antlion optimization algorithm for multi-objective task scheduling problems. In: "Cloud Computing Environments". Cluster Computing (2020). <https://doi.org/10.1007/s10586-020-03075-5>
19. Ahmad, M.M., Hanan, B.A.: Workflow scheduling using hybrid GA-PSO algorithm in cloud computing. In: Hindawi Wireless Communications and Mobile Computing, pp. 1–17 (2018)
20. Tawfeek, M.A., El-Sisi, A., Keshk, A.E., Torkey, F.A.: Cloud task scheduling based on ant colony optimization. In: 8th International Conference on Computer Engineering and Systems, pp. 64–69 (2013)
21. Kumar, P., Verma, A.: Scheduling using improved genetic algorithm in cloud computing for independent tasks. In: International Conference on Advances in Computing, Communications and Informatics, ICACCI, India, pp. 137–142 (2012)
22. Shrivastava, S., Pateriya, R.K.: Efficient storage management framework for software defined cloud. *J. Internet Technol. Secur. Trans.* **7**(4), 317–3291 (2017)

# K-Means Algorithm-Based Text Extraction from Complex Video Images Using 2D Wavelet



Divya Saxena and Anubhav Kumar

**Abstract** In this paper, the K-means algorithm for information extraction in the form of text from complex video images using 2D wavelet is presented. Haar wavelet and K-means algorithm-based simple hybrid approach are proposed in this paper. Haar wavelet is used to efficiently convert grayscale image to edge image. K-means algorithm is used for the information localization and segmentation process to split the background pixels from the text images. Large non-text background interrupts to extract the text information from any video images, so that non-text background can be identified and eliminated from the text images with the help of the hybrid approach. This algorithm is evaluated on the complex background video frames. The recall rate and the precision rate are obtained 99.01 and 95.75% of the proposed algorithm in video images.

## 1 Introduction

Digital technologies are gaining new dimensions every day, in which image processing has made valuable contributions to digital technologies. Extracting text and information from digital images is a critical system, and it is a very important application of multimedia communication in the digital era. In present years, researchers and companies have been active on text and information extraction. The contrast area of images, including text, can be extracted from the edge of the images. To extract text and information from video frames or images, the text area has to be segmented for which the segmentation process is used, but eliminating the non-text background in the text and information extraction is a complex process, with which the size of the small non-text element is equal to the text area, making the text extraction complex. In-text and information extraction from images and video

---

D. Saxena (✉)

Applied Science and Humanities, ABES Engineering College, Ghaziabad, Uttar Pradesh, India

A. Kumar

Department of Electronics and Communication Engineering, Raj Kumar Goel Institute of Technology and Management, Ghaziabad, Uttar Pradesh, India

images are used by different text segmentation, localization and edge-based methods such as differential evolution, machine learning, rotation proposals, block projections, character segmentation, edge-based, Gabor filter, connected component, and morphological approach [1–9].

Research in machine learning is coming forward very fast in this digital era. Although the K-means algorithm was known since the mid-1960s, the K-means algorithm has contributed a lot in fields such as clustering [6], medical imaging [7], and text extraction from images [8]. K-means machine learning algorithm is a popular and efficient unsupervised algorithm in machine learning. K-means clustering can work quickly on unlabeled data with shortcomings such as the number of repetitions. In this paper, the 2D Haar wavelet is used to extract the contrast edge of the text and images, while the text is extracted from the morphological process after using the K-means algorithm to segment the images.

After the introduction, text extraction is illustrated as follows. The proposed approach and outcome analysis are briefly discussed in Sect. 2 and Sect. 3. The conclusion is elaborated in the last section.

## 2 Proposed Algorithm

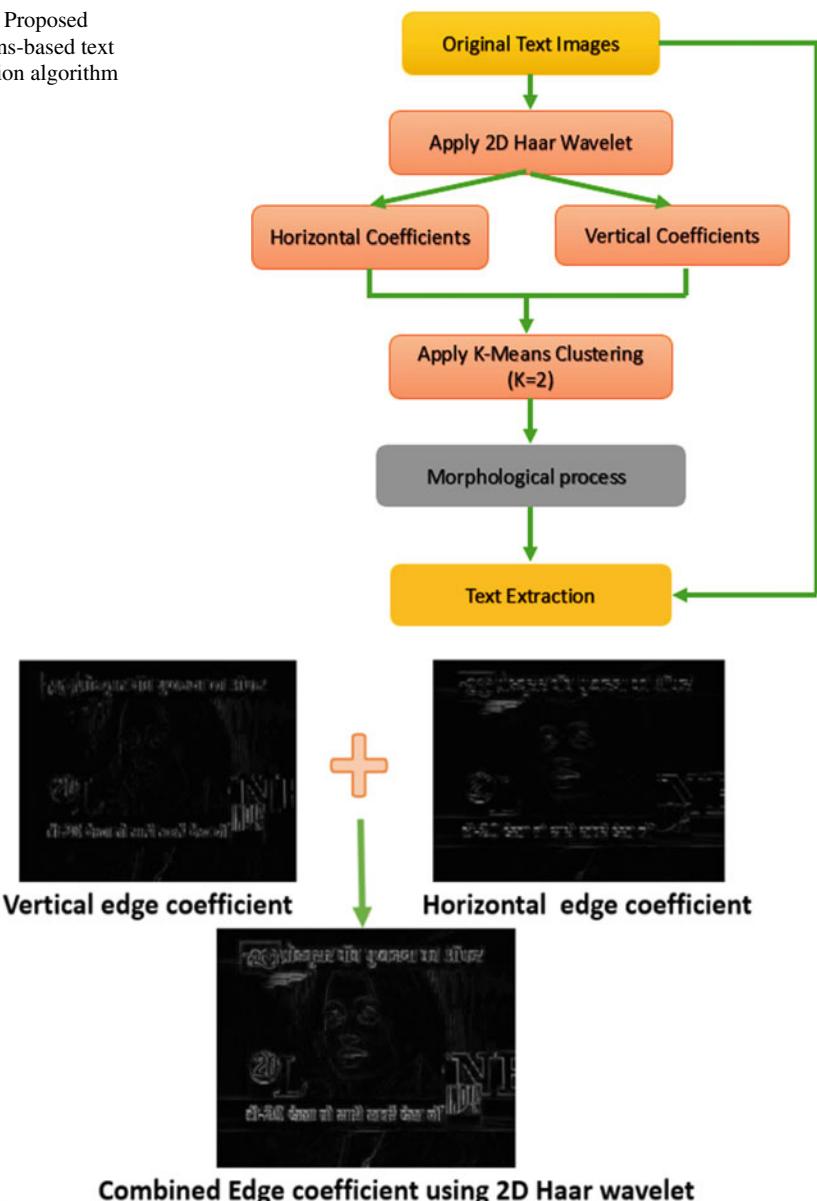
The proposed algorithm is divided into 4 parts, which include preprocessing, Haar wavelet-based edge conversion, text segmentation from K-means algorithm, and text extraction with morphological approach in the last. The proposed algorithm flowchart with all steps is shown in Fig. 1.

In preprocessing, the color image has been changed to the grayscale image, and the grayscale image is shown in Fig. 2. After that, Haar wavelet is used to determine the text edge from the images. Large coefficients of wavelet are closely related to edges in the images with irregular texture in the images [9]. Concentrations of any edge pixels in any text are around the text itself, so that the edges can be correctly extracted with wavelet. The Haar wavelet can be represented as

$$\psi(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2 \\ -1 & \text{if } 1/2 \leq t < 1 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

2D Haar wavelet decomposed the original images into sub-band, which include the LL, LH, HL, and HH coefficients. HL and LH sub-bands are representing the vertical edge coefficient and the horizontal coefficient. The concentration of the edge of the text is determined from its horizontal and vertical coefficients. Vertical and horizontal components have been combined to found the edges of the text efficiently, which gives a highly focused edge image, and the text or non-text element can be separated in the next parts of the algorithm.

**Fig. 1** Proposed K-means-based text extraction algorithm

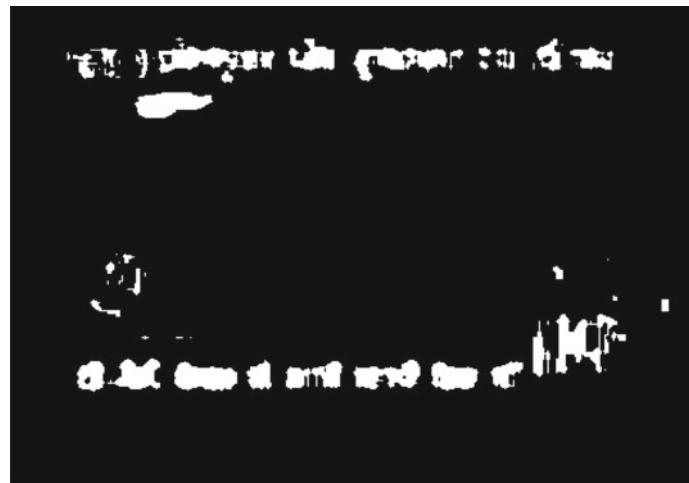


**Fig. 2** Edge image after 2D Haar wavelet

It is necessary to remove the non-text edge of the image, so that text information can be extracted successfully. K-means [10] algorithm is used to eliminate the non-text element and segment of the actual text area. So that, apply the K-means algorithm on the edge image. In this,  $k = 2$  value is taken for initial centroid. K-means algorithm is based on the initial centroid value of  $k$ . A cluster-based edge component is found by this algorithm. The edge image after the K-means algorithm is shown in Fig. 3.

The intensity of the edge pixels is also regulated by the centroid and by adjusting the value of  $k$ . The large non-text background and non-text elements are eliminated from the K-means algorithm. 8 by 8 median filtering applied in to edge image, so that small noise, non-text, and non-cluster element are removed. Figure 4 displays the edge picture after 8 by 8 median filter.

**Fig. 3** Edge image after K-means algorithm

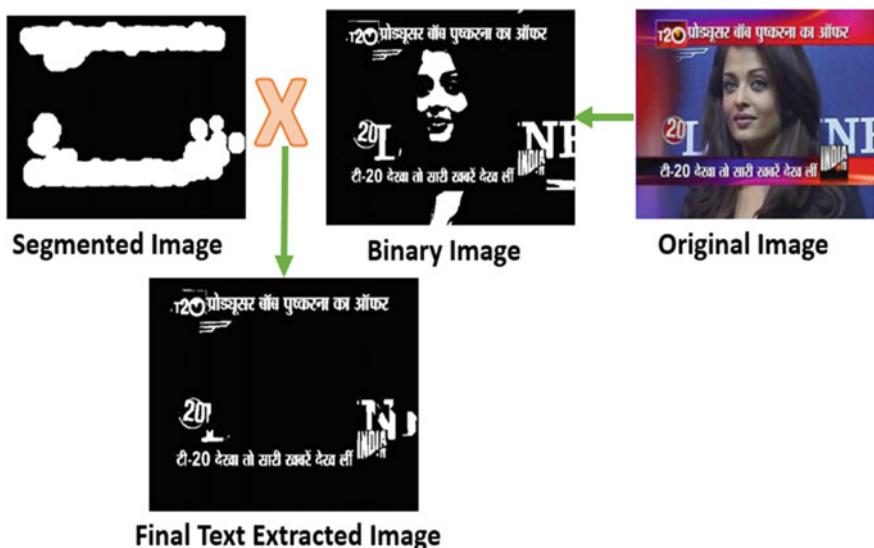


**Fig. 4** Edge image after 8 by 8 median filter

After small edge removal, the morphological approach is applied to the edge element so that edge text words are connected. The area-based segmented image after morphological approach is shown in Fig. 5. Multiplication property [3–5] is applied between the segmented image and the binary component of the original text image for text extraction. Final text extracted process from the image is shown in Fig. 6.



**Fig. 5** Image after morphological approach



**Fig. 6** Final text extraction process

### 3 Simulated Result Analysis

In simulation and result analysis, 20 text images with complex background and different types of videos such as news, live shows, and movies have been taken from the Internet randomly. MATLAB software is used to analyze the proposed algorithm. A total of 204 words are extracted from 20 text images. According to Fig. 7, proposed algorithm extracted correctly text words from news and movies text images. The recall rate and the precision rate are used to analyze as per Eqs. (2) and (3).

$$\text{Recall Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words} + \text{False Negatives}} * 100 \quad (2)$$

$$\text{Precision Rate} = \frac{\text{Correctly detected words}}{\text{Correctly detected words} + \text{False Positives}} * 100 \quad (3)$$

In Table 1, comparative analysis with the proposed algorithm is evaluated with existing algorithms. In this analysis, the proposed text extraction algorithm has a 99.01% recall rate with a 95.75% precision rate. Recall rate is almost equal to other similar approaches, but the precision rate is higher due to non-text pixel removal using the K-means segmented approach. This algorithm can be applicable as a text and information retrieval in digital and multimedia technology.



**Fig. 7** Original image and text extraction image

**Table 1** Analysis of proposed K-means text extraction algorithm with existing algorithm

Algorithm	Recall rate (%)	Precision rate (%)
Proposed	99.01	95.75
Karpagam [1]	77	78
Kumar [2]	99.61	96.20
Kumar et al. [3]	95.3	–
Kumar [5]	99.11	94.67
Kumar [11]	96.20	87.54

## 4 Conclusion

In this paper, text extraction from complex video images is done using a hybrid approach of the Haar and K-means algorithm. Effectively, non-text is removed from complex video images so that the false alarm of the algorithm is decreased. The recall rate and precision rate of the proposed algorithm are 99.01 and 95.75%. The recall rate and precision rate evaluated the effectiveness of the presented algorithm, and it can be used in digital communication for text and information retrieval algorithms.

## References

1. Karpagam, A.V., Manikandan, M.: Text extraction from natural scene images using Renyi entropy. *J. Eng.* **8**, 5397–5406 (2019)
2. Kumar, A.: An efficient text extraction algorithm in complex images. In: 2013 Sixth International Conference on Contemporary Computing (IC3), pp. 6–12. IEEE (2013)
3. Kumar, A., Kaushik, A.K., Yadav, R.L.: A robust and fast text extraction in images and video frames. In: International Conference on Advances in Computing, Communication and Control, pp. 342–348. Springer, Berlin (2011)
4. Kumar, A., Awasthi, N.: An efficient algorithm for text localization and extraction in complex video text images. In: 2013 2nd International Conference on Information Management in the Knowledge Economy, pp. 14–19. IEEE (2013)
5. Kumar, A.: An efficient approach for text extraction in images and video frames using gabor filter. *Int. J. Comput. Electr. Eng.* **6**(4), 316 (2014)
6. Mustafi, D., Sahoo, G.: A hybrid approach using genetic algorithm and the differential evolution heuristic for enhanced initialization of the k-means algorithm with applications in text clustering. *Soft Comput.* **23**(15), 6361–6378 (2019)
7. Jaroš, M., et al.: Implementation of K-means segmentation algorithm on Intel Xeon Phi and GPU: application in medical imaging. *Adv. Eng. Softw.* **103**, 21–28 (2017)
8. Akter, S., et al.: An extractive text summarization technique for Bengali document(s) using K-means clustering algorithm. In: 2017 IEEE International Conference on Imaging, Vision and Pattern Recognition (icIVPR). IEEE (2017)
9. Mallat, S.: A Wavelet Tour of Signal Processing, Elsevier (1999)
10. Arthur, D., Vassilvitskii, S.: K-means++: the advantages of careful seeding. In: SODA ‘07: Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 1027–1035 (2007)
11. Kumar, A., Kaushik, A.K., Yadava, R.L., Saxena, D.: An edge-based algorithm for text extraction in images and video frame. In: Advanced Materials Research, vol. 403, pp. 900–907. Trans Tech Publications Ltd. (2012)

# Opinion Mining-Based Conjoint Analysis of Consumer Brands



Kumar Ravi, Aishwarya Priyadarshini, and Vadlamani Ravi

**Abstract** Nowadays, online customer reviews in the form of feedback and ratings are available for a large number of product or service-based categories. These product reviews signify individually perceived strengths and shortcomings of a given product category, which can essentially provide design engineers and market researchers with invaluable insights into product design. In order to analyze the available individual customer opinions and turn them into aggregate consumer preferences that can be used in the product development and improvement process, we proposed a hybrid of sentiment analysis and conjoint analysis to determine the best brand among a list of brands for a specified product category. The outcome of this hybrid model is compared with the hybrid of sentiment analysis and multi-criteria decision making. The proposed framework estimates the relative effect of product attributes and brand names on the overall preference of consumers. We demonstrated the effectiveness of our approach using product review data taken from the literature. Our proposed approach turned out to be useful in determining the best of market availability.

## 1 Introduction

Conjoint analysis is a statistical technique used in market research to determine how the potential customers assess different attributes (feature, function, benefits) that make up an individual product or service-based category. The objective of conjoint analysis is to analyze the e-customer reviews and determine what combination of

---

K. Ravi

IoTWoRKS, HCL Technologies Ltd, Plot 3A, SEZ, Noida 201301, Uttar Pradesh, India

A. Priyadarshini

Department of Computer Science and Engineering, IIIT Bhubaneswar, Bhubaneswar 751003, India

V. Ravi (✉)

Center of Excellence in Analytics, Institute for Development and Research in Banking Technology, Castle Hills Road No. 1, Masab Tank, Hyderabad 500057, India  
e-mail: [vravi@idrbt.ac.in](mailto:vravi@idrbt.ac.in)

a limited number of attributes is most significant and essentially influences respondent's choice of product category. Conjoint analysis is evolved in the form of linear programming for rank-ordered data, self-explicated research, adaptive conjoint analysis (ACA), choice-based approaches, best-worst scaling, etc. This mathematical model is used by design engineers and market researchers to visualize, extract the underlying relationships among the attributes and the utilities of the attributes that a potential consumer associates with it. A great deal of market research commissioned is largely descriptive, which is beneficial to illustrate demographics, usage patterns, feelings, and preferences of the individuals or the e-customers. So, in the competitive market, the design engineers and researchers require tools, which can be used to predict what customers will purchase when faced with a variety of available brands and myriad product characteristics.

Individuals apply an assortment of heuristics while assessing and selecting product alternatives in the marketplace. Many product concepts consist of a huge array of features largely differentiated by brand, packaging, and prices. To decide which products to trade, executives may use their intuition or the recommendation of design engineers, or they may also analyze the present-day market trends. However, these strategies are short-termed and reactive.

In consumer-oriented firms, the potential products are often evaluated majorly via market assessments. Voluminous online customer reviews offer tremendous prospects for generating surveys or questionnaires without spending a lot of time in formulating product specific surveys. Instead of applying combinatorics to prepare questionnaires, anyone can obtain the real requirements of live customers/consumers. Consumers are shown a few product concepts and asked relevant questions concerning their purchase interests. The process provides essential information about the consumers' perception about certain features of brand profiles and their expectations out of the product category expressed as low rated reviews.

In this study, we proposed a hybrid of sentiment analysis and conjoint analysis to determine preference of customers regarding different attributes of digital camera and mobile phone brands. In addition, the obtained preference rankings are compared with those of Ravi and Ravi [1]. To apply conjoint analysis, we considered the economic preference model proposed in Decker and Trusov [2]. They considered the distribution of words taken from cons and pros of review, whereas we computed sentiment score for a set of identified aspects using dictionary approach in a systematic manner. Further, they did not compare their results with any method that yields ranks, whereas we compared our results with that of [1], where sentiment analysis and MCDM hybrids were developed. These are our significant contributions.

The remainder of this paper is organized as follows. We presented literature survey related to conjoint analysis and aspect-based sentiment analysis in Sect. 2. The proposed approach is presented in Sect. 3. Experiment setup is presented in Sect. 4. The results and discussion are presented in Sect. 5. Finally, conclusion and future directions of research are presented in Sect. 6.

## 2 Literature Review

Aspect-level sentiment analysis can be performed mainly using three approaches involving machine learning [3], lexicon-based approaches [4], and a hybrid of both [5]. Murthy et al. [6] reviewed literature available for text classification. Al-Smadi et al. [7] considered morphological, syntactic, and semantic features to perform aspect-based sentiment analysis in the domain of hotel industry. They worked on three tasks, namely aspect category identification, opinion target expression extraction, and sentiment polarity identification. Ma et al. [8] proposed a position attention mechanism to determine sentiment expressed on multiple aspects in the same sentence. The position attention is able to model the explicit position context between the aspect and its context words. Along the same lines, Zhao et al. [9] proposed a hybrid model based on graph convolutional networks (GCN). First, they employed bidirectional encoder representations from transformers (BERT) with position encoding to capture embedding based on aspect and its context. Second, they employed GCN to model the sentiment dependencies between different aspects in one sentence. Nguyen and Nguyen [10] considered intra-attention and interactive-attention along with sentiment lexicon to perform aspect-level sentiment analysis. They considered lexicon-aware attention to be fused with deep neural networks, which can capture sentiment context conditioned on the informative aspect words. Chen et al. [11] proposed multi-source data fusion considering aspect-level corpora, sentence-level corpora, word-level sentiment lexicons, and BERT. BERT was considered to generate aspect-specific sentence representations for sentiment classification. Murty et al. [12] proposed a similarity measure for clustering problems.

Fan et al. [13] reviewed available literature on information fusion for ranking products using online reviews. Mokonyama and Venter [14] determined the preference of customers for public transport contracts in South Africa. They observed that reliability, security, staff respect, and service frequency are major attributes of contracts. Wu et al. [15] computed part-worth for six attributes, viz. power, appearance, safety, fuel efficiency, price, and gadgets for eight sub-compact cars in Thailand. The order of preference for attributes obtained as appearance, fuel efficiency, price, safety, power, and gadgets. Dauda and Lee [16] evaluated the perceptions of Nigerian banking customers' on bank service quality. They applied a discrete choice-based method on 1245 survey-based data samples. They recommended to reduce transaction errors, transaction costs, waiting time, and initial online learning time. Yang et al. [17] fused fuzzy sets and graph models to determine review helpfulness. Using 1158 online reviews, they presented different applications of fusion, viz. product recommendation, market analysis, customer satisfaction analysis, product defect identification, and consumer preference analysis. Anand et al. [18] employed conjoint analysis to determine the preference of attributes of cell phones as brand, price, features, and colors. They forecasted probable sales for the next ten periods using the choice probabilities-based forecasting diffusion model.

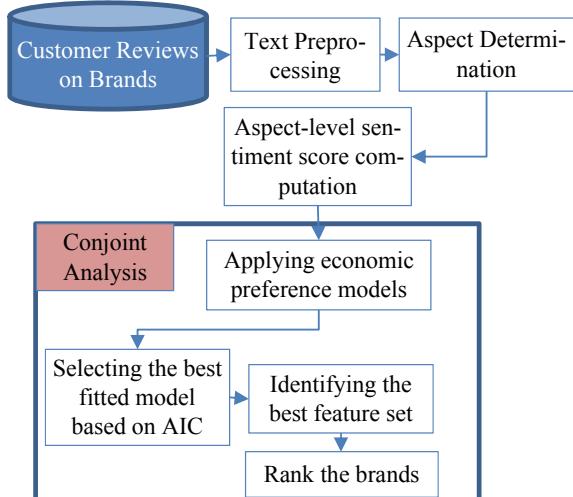
Kodapanakkal et al. [19] employed conjoint analysis approach to understand people's preference in accepting big data technologies in six sectors, namely criminal

investigations, crime prevention, citizen scores, health care, banking, and employment. They found data protection and outcome favorability as prime factors in technology adoption. Wang et al. [20] proposed inconsistent ordered choice model (IOCM) based on online review and review rating. They categorized a list of features of auto cell into four categories, viz. indifferent features, must-have features, exciting features, and performance features. Most recently, Mitra and Jenamani [21] proposed Online Brand IMage (OBIM) score by computing sentiment and co-word network analysis of product features like favorability, strength, and uniqueness. They applied OBIM for SWOT analysis and sentiment concept mapper.

### 3 Proposed Approach

In this study, we developed a hybrid of preference economic-based conjoint methods and sentiment analysis to rank a set of brands based on electronic word of mouth. The obtained rank is compared with the rank obtained using the hybrid of sentiment analysis and multi-criteria decision-making (MCDM) methods. The proposed approach, whose schematic is depicted in Fig. 1, has six stages, viz. data collection, text preprocessing, aspect determination, aspect-level sentiment score computation, applying conjoint analysis or economic preference models and finally ranking the brands. The detailed steps of applying MCDM methods are available in [1]. In this study, we presented the steps involved in the hybrid approach of sentiment analysis and conjoint analysis. The proposed approach yields a ranking for a set of brands which are compared with those obtained for the same set of brands by Ravi and Ravi [1].

**Fig. 1** Schematic of the proposed approach



The details of data collection and text preprocessing are presented in Sects. 4.1 and 4.2, respectively. Based on the frequency of occurrence, we considered a set of aspects to compare a set of brands. Based on the corpus prepared for one brand, we computed the sentiment score for each combination of aspects and product brand. For each aspect, we obtained aggregate sentiment score with respect to negative and positive polarity.

For conjoint analysis, we considered different econometric preference models proposed in [2]. We applied Poisson regression (PR) and negative binomial regression (NBR) as represented using Eqs. (1) and (3), respectively. For modeling the homogeneous preferences, we applied PR with conditional mean,  $\lambda > 0$ , which is presented in Eq. (1). Here, let  $k = 1 \dots K$  in the subscript denote the individual product reviews,  $l = 1 \dots L$  denotes the subscript of the functional product attributes, and  $h \in \{1, 2\}$  are the respective levels or attributes of  $l$ . Using these notations, the opinion data can be converted into binary data using

$$x_{klh} = \begin{cases} 1 & \text{if the functional attribute } l \text{ takes level } h \text{ in review } k \\ 0 & \text{otherwise} \end{cases}$$

Therefore,  $x_{kl1} + x_{kl2} \leq 1, \forall k, l$ . For denoting the brand name effects, we introduce another variable

$$\tilde{x}_{km} = \begin{cases} 1 & \text{if product discussed in review belongs to brand } m \\ 0 & \text{otherwise} \end{cases}$$

where  $m = 1 \dots M$ , and  $M$  represents the number of brands considered for computation which in our case is 4, irrespective of the dataset (digital camera or mobile phones) and  $\sum_{m=1}^M \tilde{x}_{km} = 1, \forall k$ . Hence, the observed rating  $y_k$  of the product in a certain review  $k$  has been defined in the function below by Poisson regression with conditional mean,  $\lambda > 0$ .

$$\text{Prob}^{PR}(Y = y) = \frac{\lambda^y}{y!} e^{(-\lambda)} \quad \text{with } \lambda = \exp\left(\alpha + \sum_{l=1}^L (\beta_{l1}x_{l1} + \beta_{l2}x_{l2}) + \sum_{m=1}^M \delta_m \tilde{x}_m\right) \quad (1)$$

The above equation determines the probability of observing product rating  $y$  given a certain pos/neg summary coded by means of explanatory variables  $x$  and  $\tilde{x}$ . Similarly, for modeling the heterogeneous preferences, we applied the NBR approach, because a certain amount of variation in the observed data can be credited to unfamiliar sources (apart from variables  $x$  and  $\tilde{x}$ ), the use of an appropriate statistical distribution for demonstrating this “unobserved” heterogeneity seems promising. Let  $\varepsilon$  be a random variable that denotes the “unobserved” heterogeneity in the dataset to be studied and that is not correlated with  $x$  and  $\tilde{x}$  but satisfies  $E(\varepsilon) = 0$ . The conditional mean  $\lambda$  is replaced with a random variable  $\Lambda$  as presented in Eq. (2)

$$\Lambda = \lambda \cdot e^\varepsilon = \lambda \cdot \tilde{\varepsilon} \text{ with } E(\Lambda) = \lambda \quad (2)$$

The probability of observing a certain value  $y$  can then be computed using the following mixture distribution with  $\gamma > 0$ .

$$\text{Prob}^{\text{NBR}}(Y = y) = \binom{y + \gamma - 1}{y} \cdot \frac{\lambda^y \gamma^\gamma}{(\lambda + \gamma)^{y+\gamma}} \text{ with } \gamma > 0 \quad (3)$$

Here,  $\gamma$  represents the scale parameter and as well as the shape parameter for convenience of computation [2].

Ducker and Trusov [2] demonstrated that these models can yield competitive results as conjoint analysis techniques. Employed economic preference model estimates parameters associated with a set of criteria selected for a set of brands. The parameters reflect the consumer preference of one brand over another, which can be converted in the form of brand ranking. As user reviews are considered to develop the model, it can help us understand the heterogeneity of users' preferences for evaluation of the product brands. We considered the Akaike Information Criterion (AIC) to compare the performance of different economic preference models. AIC tests for the goodness of fit for all the statistical models, thereby indicating their quality. These statistical models can be employed to estimate the relative preference of brands and functional attributes using Eqs. (4) and (5) [2]:

$$\phi^{\text{Attribute}}(x_l) = (\exp(\hat{\beta}_{l1}) + \exp(\hat{\beta}_{l1}) - 2) \times 100\% \quad (4)$$

$\phi^{\text{Attribute}}$  is called as the “backlog of impact”, which measures the average sensitivity of the product evaluation to variations of the quality of functional attributes. The relative strength of effect between two brands  $m_1$  and  $m_2$  allows for tentative inferences

$$\phi^{\text{Brand}}(\tilde{x}_{m_1}, \tilde{x}_{m_2}) = (\exp(\hat{\delta}_{m_1}) + \exp(\hat{\delta}_{m_2})) \times 100\% \quad (5)$$

about the relative brand value or power from the consumer's perspective.

## 4 Experimental Setup

### 4.1 Data Details

For experimental purposes, we considered reviews on the digital camera and mobile phone brand collected from Twitter and [www.amazon.com](http://www.amazon.com). The distribution of reviews of two different datasets considered for experiments is presented in Table 1.

**Table 1** Dataset details

Digital camera brand	No. of reviews	Cell phone brand	No. of reviews
Canon (D1)	676	Apple iPhone (M1)	722
Nikon (D2)	676	HTC One (M2)	870
Panasonic (D3)	676	Nokia Lumia (M3)	699
Sony (D4)	676	Samsung Galaxy (M4)	873

We considered reviews or tweets containing at least one aspect. Further details about the dataset are presented in [1].

## 4.2 *Text Preprocessing and Sentiment Score Computation*

We performed text preprocessing like tokenization, length-based token filtering, stop word removal, POS tagging, lemmatization, and parsing. For criteria determination, we obtained the frequency of all tokens. Based on the highest frequency and context of the domain, we selected nine product features as criteria for both datasets. We also considered the synonyms of each aspect to better capture sentiment concerning an aspect. For sentiment score computation, we considered all opinion words, viz. nouns, verbs, adjectives, and adverbs associated with an aspect. To determine the association, we considered the parse graph approach as explained in [1]. A set of opinion words that appeared in the neighborhood of an aspect were considered as associated opinion words. The neighborhood of 1 was decided based on the length of the shortest path. The sentiment score was obtained using three sentiment dictionaries, viz. SentiWordNet [20], SenticNet 3.0 [21], and SO-CAL [22]. We considered positive and negative sentiment scores separately for each combination of an aspect and a review for our conjoint analysis experiment. To determine the target variable for regression purpose, we obtained summation of sentiment of all aspects appeared in one review together. The sentiment sum was bucketed into five categories based on quintile. In [2], the positive and negative sentiment were added together to obtain final sentiment.

## 5 Results and Discussion

The parameters obtained for different economic preference models on the digital camera and the cell phone dataset are presented in Tables 2 and 3, respectively. We can observe that a relatively large number of attributes contribute significantly toward

**Table 2** Model goodness of fit measure for digital camera dataset

Fitting parameter	PR	NBR (Link = log)	NBR (link = sqrt)
AIC	6820.0	6822.0	6784.0
Residual deviance and degree of freedom (DF)	876.79 and 2035	876.77 and 2035	838.73 and 2035

**Table 3** Model goodness of fit parameters for mobile phone dataset

Fitting parameter	PR	NBR (Link = log)	NBR (link = sqrt)
AIC	5179.5	5181.5	5133.0
Residual deviance and DF	600.8 and 1574	600.79 and 1574	552.35 and 1574

the customer preference of a given product. Akaike Information Criterion (AIC) is an estimator of the relative quality of the statistical models for a given set of data. PR is employed to model homogenous preferences, whereas NBR is employed to model heterogeneous preferences.

We can observe that the different parameter vectors reveal significant information about the aspects. For both datasets, the intercepts that are significant in each case are of approximately the same magnitude. For the digital camera dataset, PR yielded the best AIC value of 6820 and residual deviance of 876.79 as presented in Table 2. All aspects are having a significant coefficient at least at the 5% level (two-tailed  $p$  values) except *batterypos*, *batteryneg*, *dslrneg*, *priceneg*, and *pricepos*. We can observe that the signs of coefficients are associated with the polarities of the sentiment. Based on positive sentiment, the order of importance of different aspects turned out to be in the decreasing order as *picture*, *brand*, *lens*, *processor*, *appearance*, *flash*, and *dslr*. It implies that increasing sentiment of these aspects would increase the favorability of the digital camera brand. Based on negative sentiment, the order of importance for features is *picture*, *processor*, *brand*, *lens*, *flash*, *appearance*, and *battery*. It indicates that increasing these parameters would lead to a decrease in the reputation of the brand. To model heterogeneous preferences, we employed NBR with different link function as presented in Tables 2 and 4. NBR yielded the best residual deviance. Hence, we considered NBR to rank different digital camera brands. As per NBR parameters, all aspects are having significant coefficients except *batterypos*, *dslrneg*, *priceneg*, and *pricepos*.

According to Eqs. (3) and (4), we estimated parameters  $\hat{\beta}_{11} \dots \hat{\beta}_{19}$  and  $\hat{\delta}_{m_1} \dots \hat{\delta}_{m_4}$  for both datasets. Contribution of D1 is  $\exp(\hat{\delta}_{m_1}) = \exp(0.055) = 1.06$ , which is strongest in the brand name. The D4 has a relatively negative impact with  $\exp(\hat{\delta}_{m_4}) = 0.96$ . In terms of features, the highest contributing feature is *picture* with effect of  $\exp(\hat{\beta}_{17}) = \exp(0.33) = 1.39$ . The least contributing feature is *processor* with effect of  $\exp(\hat{\beta}_{17}) = \exp(-0.246) = 0.78$ . Based on this calculation, the decreasing order of importance for positive sentiment is obtained as *picture*, *brand*, *lens*, *processor*, *flash*, *appearance*, and *dslr*. It implies that the picture quality is the prime feature for a camera as per user opinion. For the negative sentiment, the decreasing order

**Table 4** Parameter estimates for digital camera dataset

	PR			NBR (link = sqrt)		
	Estimate	Std. error	p-value	Estimate	Std. error	p-value
(Intercept)	0.757317	0.035833	***	1.400368	0.030714	***
D4	-0.059292	0.040058	0.14	-0.04382	0.033549	0.19
D1	0.066757	0.038735	.	0.055238	0.033587	0.1
D3	0.003552	0.036236	0.92	0.003378	0.0319	0.91
D2	-	-	-	-	-	-
appearanceneg	-0.191914	0.06717	**	-0.1728	0.053007	**
appearancepos	0.181764	0.033988	***	0.192088	0.0318	***
batteryneg	-0.116208	0.074134	0.12	-0.10389	0.062966	.
batterypos	0.076922	0.051426	0.13	0.077162	0.047475	0.1
brandneg	-0.276698	0.084856	**	-0.23066	0.062921	***
brandpos	0.245555	0.032339	***	0.253249	0.029507	***
dslrneg	-0.124558	0.114923	0.28	-0.12252	0.092608	0.19
dslrpos	0.113197	0.038341	**	0.127156	0.034837	***
flashneg	-0.22955	0.10967	*	-0.19649	0.085358	*
flashpos	0.183707	0.05924	**	0.193795	0.057589	***
lensneg	-0.26253	0.073393	***	-0.20528	0.053552	***
lenspos	0.20377	0.030294	***	0.204311	0.027234	***
pictureneg	-0.368778	0.074283	***	-0.23874	0.04914	***
picturepos	0.353881	0.029309	***	0.332494	0.025313	***
priceneg	-0.431763	0.380227	0.26	-0.4307	0.290668	0.14
pricepos	0.128494	0.185898	0.49	0.093189	0.179303	0.6
processorneg	-0.327797	0.07395	***	-0.24605	0.051862	***
processorpos	0.192479	0.030343	***	0.196781	0.027615	***

Significance levels (two-tailed p values): \*\*\* = 0.001, \*\* = 0.01, \* = 0.05, . = 0.1

of importance of attributes is *battery*, *appearance*, *brand*, *flash*, *lens*, *picture*, and *processor*. So, based on the above information, the decreasing order of the brand for the digital camera is **Canon → Nikon → Panasonic → Sony**.

For the mobile phone dataset, NBR with link *sqrt* yielded the best AIC of 5133.0 and residual deviance of 552.35 as presented in Tables 3 and 5, respectively. All aspects are significantly contributing to the model except *appearanceneg*, *memorystos*, and *processorneg*. Based on the coefficients obtained in Table 5, HTC One is the strongest mobile brand as per user opinions. The decreasing order of importance for positive sentiment is obtained as *camera*, *screen*, *brand*, *processor*, *price*, *appearance*, *sound*, *battery*, and *memory*. It indicates that the camera is one of the major aspects while selecting a mobile phone. These coefficients yield the decreasing order of the brand as **HTC One → Nokia Lumia → Apple iPhone → Samsung Galaxy**.

**Table 5** Parameter estimates for mobile phone dataset

	PR			NBR (link = sqrt)		
	Estimate	Std. error	p-value	Estimate	Std. Error	p-value
(Intercept)	0.86674	0.03892	***	1.47364	0.03524	***
M4	-0.08026	0.04253	.	-0.06187	0.03657	.
M1	-0.07585	0.04422	.	-0.05997	0.03835	0.11
M3	-0.04025	0.04361	0.36	-0.03035	0.03811	0.42
M2	-	-	-	-	-	-
batteryneg	-0.37091	0.0791	***	-0.28312	0.05621	***
batterypos	0.17309	0.04917	***	0.19129	0.04589	***
brandneg	-0.4799	0.10929	***	-0.3282	0.07112	***
brandpos	0.29148	0.03914	***	0.30607	0.03588	***
cameraneg	-0.41724	0.0937	***	-0.30319	0.06379	***
camerapos	0.30339	0.0343	***	0.31193	0.03203	***
appearanceneg	-0.15581	0.12057	0.2	-0.1103	0.09475	0.244
appearancepos	0.233	0.04646	***	0.25578	0.04417	***
memoryneg	-0.53205	0.10449	***	-0.33998	0.06652	***
memorypos	-0.03825	0.0615	0.53	0.00904	0.05239	0.86
priceneg	-0.3292	0.09916	***	-0.24204	0.06817	***
pricepos	0.24336	0.04007	***	0.2591	0.0364	***
processorneg	0.06132	0.1883	0.74	0.06571	0.14106	0.64
processorpos	0.25638	0.0552	***	0.27574	0.05372	***
screenneg	-0.35304	0.091	***	-0.24957	0.06044	***
screenpos	0.30285	0.03346	***	0.31143	0.03071	***
soundneg	-0.38759	0.11267	***	-0.30421	0.07924	***
soundpos	0.23772	0.04583	***	0.24481	0.04364	***

Significance levels (two-tailed  $p$  values): \*\*\* = 0.001, \*\* = 0.01, \* = 0.05, . = 0.1

As per Eq. (4), the percentage backlog impact of digital camera brand and their aspects are presented in Tables 6 and 7, respectively. Similarly, the percentage backlog impact of digital camera brands and their aspects are presented in Tables 8 and 9, respectively. The first row of Table 6 indicates that D1 has the highest relative

**Table 6** Percentage backlog of impact (digital camera brands)

	D1	D2	D3	D4
D1	0	5.65	5.31	9.96
D2	-5.65	0	-0.34	4.3
D3	-5.31	0.34	0	4.65
D4	-9.96	-4.3	-4.65	0

**Table 7** Percentage backlog of impact (digital camera aspects)

Aspect	Relative strength	Aspect	Relative strength
Price	-25.2	Lens	4.11
Battery	-1.85	Appearance	5.31
Processor	-0.06	Brand	8.22
dslr	2.03	Picture	18.2
Flash	3.55		

**Table 8** Percentage backlog of impact (mobile phone brands)

	M1	M2	M3	M4
M1	0	-5.82	-2.83	0.18
M2	5.82	0	2.99	6
M3	2.83	-2.99	0	3.01
M4	-0.2	-6	-3.01	0

**Table 9** Percentage backlog of impact (mobile phone aspects)

Aspect	Relative strength	Aspect	Relative strength
Memory	-27.91	Camera	10.45
Battery	-3.58	Screen	14.45
Sound	1.5	Appearance	18.7
Brand	7.83	Processor	38.5
Price	8.08		

strength among all brands. It implies that brand D1 highly influences the perception of consumers toward digital camera brand. Hence, the brand D1 is suitable for well-directed image campaigns and brand image transfers in new product introduction processes. Among the backlog of digital camera aspects, a picture has 18.2% more effect on positive sentiment compared to a negative one. On the other hand, the price has a relative strength of -25.2%, which indicates that an impairment with respect to the attribute, *price*, has a higher influence on product evaluation than an improvement of the same magnitude. The first row of Table 8 indicates that M2 has the highest relative strength compared to the rest. From Table 9, we can infer that the *processor* is highly sensitive to positive sentiment whereas *memory* is highly sensitive to the negative sentiment.

By applying MCDM techniques, the order of brands for the digital camera is **Panasonic → Nikon → Canon → Sony**, while the order for the mobile phone brands is **HTC One → Nokia Lumia → Apple iPhone → Samsung Galaxy** [1]. As per the preference model, the order of the digital camera brand is **Canon → Panasonic → Nikon → Sony**. The order of the mobile phone brand is **HTC One → Nokia Lumia → Apple iPhone → Samsung Galaxy**. Here, both approaches yielded

the same ranking for mobile phone brand but not for digital camera brand. In some techniques of MCDM, we could not obtain same rankings for a set of alternatives due to compensation features of multi-criteria decision-making methods. That is, if one brand is ranked lower with respect to a criterion, then it may be compensated by another criterion. Thus, the trade-off occurs, and the brands get ranked with respect to an unknown utility function comprising the criteria used by the users while selecting the brands. Further, the philosophy and underlying theory of conjoint analysis are different from that of MCDM methods. Therefore, the different ranks are expected.

## 6 Conclusions and Future Directions

In this study, we applied a hybrid of sentiment analysis and conjoint analysis to determine preferences of consumers regarding different consumer brands. By employing different preference models, we computed the relative strength of different product aspects which contribute to determine the preference. The ranks obtained using conjoint analysis models are compared with those obtained by using a hybrid of sentiment analysis and MCDM methods. We observed that the obtained ranks are identical for one dataset and partially similar for the other dataset. In future, we apply such hybrids on a huge corpus of reviews to understand the aggregate opinions of consumers. Further, we can combine different quantitative and qualitative features to apply conjoint analysis to rank the products and services of commercial banks and insurance companies. Conjoint analysis needs to be redesigned before applying to some related fields like political research communication [23].

## References

1. Ravi, K., Ravi, V.: Ranking of branded products using aspect-oriented sentiment analysis and ensembled multiple criteria decision-making. *Int. J. Knowl. Manag. Tour. Hosp.* **1**, 317–359 (2018)
2. Decker, R., Trusov, M.: Estimating aggregate consumer preferences from online product reviews. *Int. J. Res. Mark.* **27**, 293–307 (2010). <https://doi.org/10.1016/j.ijresmar.2010.09.001>
3. Do, H.H., Prasad, P.W.C., Maag, A., Alsadoon, A.: Deep learning for aspect-based sentiment analysis: a comparative review. *Expert Syst. Appl.* **118**, 272–299 (2019)
4. Mowlaei, M.E., Abadeh, M.S., Keshavarz, H.: Aspect-based sentiment analysis using adaptive aspect-based lexicons. *Expert Syst. Appl.* **148**, 113234 (2020)
5. Ravi, K., Ravi, V.: A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowl. Based Syst.* **89**, 14–46 (2015)
6. Murty, M.R., Murthy, J.V.R., Reddy, P.P., Satapathy, S.C.: A survey of cross-domain text categorization techniques. In: 2012 1st International Conference on Recent Advances in Information Technology (RAIT), pp. 499–504 (2012). <https://doi.org/10.1109/RAIT.2012.6194629>
7. Al-Smadi, M., Al-Ayyoub, M., Jararweh, Y., Qawasmeh, O.: Enhancing Aspect-Based Sentiment Analysis of Arabic Hotels' reviews using morphological, syntactic and semantic features. *Inf. Process. Manag.* **56**, 308–319 (2019). <https://doi.org/10.1016/j.ipm.2018.01.006>

8. Ma, X., Zeng, J., Peng, L., Fortino, G., Zhang, Y.: Modeling multi-aspects within one opinionated sentence simultaneously for aspect-level sentiment analysis. *Futur. Gener. Comput. Syst.* **93**, 304–311 (2019). <https://doi.org/10.1016/j.future.2018.10.041>
9. Zhao, P., Hou, L., Wu, O.: Modeling sentiment dependencies with graph convolutional networks for aspect-level sentiment classification. *Knowl.-Based Syst.* **193**, 105443 (2020). <https://doi.org/10.1016/j.knosys.2019.105443>
10. Nguyen, H.-T., Nguyen, L.-M.: ILWAANet: an Interactive Lexicon-Aware Word-Aspect Attention Network for aspect-level sentiment classification on social networking. *Expert Syst. Appl.* **146**, 113065 (2020). <https://doi.org/10.1016/j.eswa.2019.113065>
11. Chen, F., Yuan, Z., Huang, Y.: Multi-source data fusion for aspect-level sentiment classification. *Knowl.-Based Syst.* **187**, 104831 (2020). <https://doi.org/10.1016/j.knosys.2019.07.002>
12. Murty, M.R., Murthy, J.V.R., Reddy, P.P., Naik, A., Satapathy, S.C.: Homogeneity separateness: a new validity measure for clustering problems. In: *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India*, vol I, pp. 1–10. Springer (2014)
13. Fan, Z.-P., Li, G.-M., Liu, Y.: Processes and methods of information fusion for ranking products based on online reviews: an overview. *Inf. Fusion.* **60**, 87–97 (2020). <https://doi.org/10.1016/j.inffus.2020.02.007>
14. Mokonyama, M., Venter, C.: Incorporation of customer satisfaction in public transport contracts—a preliminary analysis. *Res. Transp. Econ.* **39**, 58–66 (2013). <https://doi.org/10.1016/j.retrec.2012.05.024>
15. Wu, W.Y., Liao, Y.K., Chatwuthikrai, A.: Applying conjoint analysis to evaluate consumer preferences toward subcompact cars. *Expert Syst. Appl.* **41**, 2782–2792 (2014). <https://doi.org/10.1016/j.eswa.2013.10.011>
16. Dauda, S.Y., Lee, J.: Quality of service and customer satisfaction: a conjoint analysis for the Nigerian bank customers. *Int. J. Bank Mark.* **34**, 841–867 (2016)
17. Yang, S.-B., Shin, S.-H., Joun, Y., Koo, C.: Exploring the comparative importance of online hotel reviews' heuristic attributes in review helpfulness: a conjoint analysis approach. *J. Travel Tour. Mark.* **34**, 963–985 (2017). <https://doi.org/10.1080/10548408.2016.1251872>
18. Anand, A., Bansal, G., Aggarwal, D.: Choice based diffusion model for predicting sales of mobile phones using conjoint analysis. *J. High Technol. Manag. Res.* **29**, 216–226 (2018). <https://doi.org/10.1016/j.hitech.2018.09.008>
19. Kodapanakkal, R.I., Brandt, M.J., Kogler, C., van Beest, I.: Self-interest and data protection drive the adoption and moral acceptability of big data technologies: a conjoint analysis approach. *Comput. Human Behav.* **108**, 106303 (2020). <https://doi.org/10.1016/j.chb.2020.106303>
20. Wang, A., Zhang, Q., Zhao, S., Lu, X., Peng, Z.: A review-driven customer preference measurement model for product improvement: sentiment-based importance–performance analysis. *Inf. Syst. E-bus. Manag.* **18**, 61–88 (2020). <https://doi.org/10.1007/s10257-020-00463-7>
21. Mitra, S., Jenamani, M.: OBIM: a computational model to estimate brand image from online consumer review. *J. Bus. Res.* **114**, 213–226 (2020). <https://doi.org/10.1016/j.jbusres.2020.04.003>
22. Baccianella, S., Esuli, A., Sebastiani, F.: SentiWordNet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In: *LREC*, pp. 2200–2204 (2010)
23. Cambria, E., Olsher, D., Rajagopal, D.: SenticNet 3: a common and common-sense knowledge base for cognition-driven sentiment analysis. In: *Proceedings of the twenty-eighth AAAI conference on artificial intelligence*, pp. 1515–1521. AAAI Press (2014)
24. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede., M.: Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **37**, 267–307 (2011)
25. Knudsen, E., Johannesson, M.P.: Beyond the limits of survey experiments: how conjoint designs advance causal inference in political communication research. *Polit. Commun.* **36**, 259–271 (2019). <https://doi.org/10.1080/10584609.2018.1493009>

# Task Scheduling in Cloud Using Improved Genetic Algorithm



Shyam Sunder Pabboju and T Adilakshmi

**Abstract** Cloud computing usually needs to process a large number of computing tasks, and task scheduling strategies play a key role in determining the efficiency of cloud computing. How to allocate computing resources reasonably and schedule task operations effectively so that the time and cost required to complete all tasks are shorter is an important issue. This paper proposes an Improved Genetic Algorithm (I-GA) that considers time and cost constraints. The result of scheduling by this algorithm can not only make the task completion time shorter, but also cost less. Through experiments, I-GA is compared with Genetic Algorithm (T-GA) considering time constraints and Genetic Algorithm (C-GA) considering cost constraints. The experimental results show that this algorithm is an effective task scheduling algorithm in cloud computing.

## 1 Introduction

Since the concept of cloud computing was put forward, it has become a hot research direction. Paper [1] summarizes the current technology used in cloud computing, analyzes the technical meaning behind it and the current cloud computing implementation solutions adopted by participating companies in cloud computing, and provides three specific cloud computing examples in the industry, including Google's GCP that Cloud computing platform, Amazon's EC2 and Microsoft's Azure. Paper [2] defines the concept of cloud computing, and analyzes and describes the problems and opportunities faced by the development of cloud computing.

Cloud computing provides infrastructure, platform, and software services. Its basic principle is to divide the computing processing work into multiple smaller

---

S. S. Pabboju (✉)

Department of CSE, MGIT Hyderabad, Telangana, India

e-mail: [pshyamsunder\\_cse@mgit.ac.in](mailto:pshyamsunder_cse@mgit.ac.in)

T. Adilakshmi

Department of CSE, Vasavi College of Engineering, Hyderabad, Telangana, India

e-mail: [t\\_adilakshmi@staff.vce.ac.in](mailto:t_adilakshmi@staff.vce.ac.in)

subtasks through the internet, and then a huge system consisting of various servers will be searched for, calculate and analyze the results, finally return to the user [3]. Cloud computing popularly known for the development of distributed, grid, and parallel computing. Which can be used in the business model as pay for use. Authors in [4] comprehensively compare grid computing and cloud computing, and points out the business model of cloud computing as “on demand, pay for use”.

Due to the higher number of computing tasks faced by cloud computing, task scheduling and resource allocation are the key and difficult points that determine the efficiency of cloud computing. At present, there is not much research on task scheduling and resource allocation in cloud computing, while grid computing has conducted extensive research on related issues [5–8]. To a certain extent, the two are similar. A task scheduling algorithm in grid computing, the main goal is to minimize the time required to complete all tasks. Most task scheduling algorithms optimize task scheduling with this goal. However, in the cloud computing model, the cost required for task execution is also a factor that cannot be ignored. Resources with different computing capabilities have different usage costs. For time-sensitive applications, provide resources with strong processing capabilities, so that the task is completed in a shorter time. For cost-sensitive user applications, provide resources with lower processing costs, so that the cost of task completion is lower. In this research work, by studying how to schedule task so that it allocates resources in cloud computing reasonably and efficiently, this work came up with GA-based task scheduling algorithm named Improved Genetic Algorithm (I-GA) that considers time and cost constraints Genetic Algorithm. Further, through simulation experiments, the efficiency of this work is verified.

## 2 Task Scheduling Problem in Cloud

In the present situation, most of the cloud computing platforms adopt the Map/Reduce programming model proposed by Google for parallel processing of large-scale data sets. Through the two stages of Map and Reduce, the larger task is divided into multiple smaller subtasks, and then allocated to multiple computing resources for parallel execution, and the final running result is obtained. Under the Map/Reduce programming model, how to divide more number of subtasks is the major problem to be considered. Cloud service providers need to extend services to various users at given instance of time, taking into account the response time of each user, and at the same time, considering the cost of the service. The available task scheduling algorithms focus primarily on optimization as one of the major concerns, which tends to result in a shorter task completion time and a higher cost. Therefore, this paper proposes IGA to improve scheduling of tasks and allocation of resources in cloud computing in order to enhance efficiency of cloud computing.

Resources in cloud computing include processors, storage, networks, etc. The use of resources is used on-demand and paid for by usage. This paper treats the resources in cloud computing as computing resources and makes the following assumptions:

1. The input of a task is a batch of subtasks that are decomposed into multiple larger computing tasks, and the granularity of subtasks is uniform, that is the required running time is not much different.
2. The number of subtasks is far greater than the number of resources.
3. The time needed by the subtask to run on each cloud computing resource is known.
4. The cost of task running unit time on each computing resource node is known.

Here we will be using  $N$  to represent the number of subtasks,  $M$  to indicate required no. of cloud computing resources, and it makes use of  $M \times N$ 's ETC (Expect Time to Complete) matrix [9] ( $\text{ETC}(i,j)$ ) which indicates the time required by the  $j$ -th subtask is completed on the  $i$ -th computing resource to calculate the time required for the task queue to run on each computing resource. Use  $\text{RCU}(i)$  (Resource Cost per Unit) to represent the cost of running tasks in each computing resource unit. The subtask is represented by  $T_j$ , and the computing resource is represented by  $P_i$ . Then the subtask instance can be described as  $(T_j, P_i, \text{ETC}(i, j), \text{RCU}(j))$ ,  $i \in [1, M], j \in [1, N]$ .

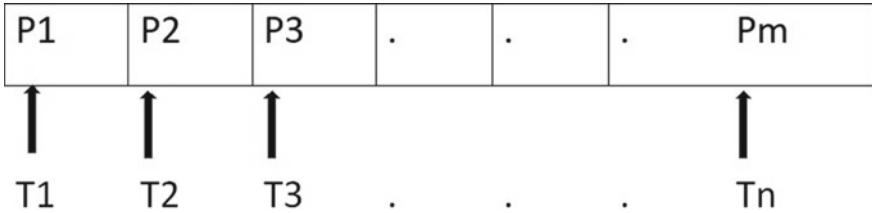
Under the above assumptions, the resource scheduling issue in cloud computing may be defined as the best way to allocate subtasks to different cloud computing resources effectively such that the time taken to complete the task is less and the cost is lower.

### 3 Improved Genetic Algorithm (I-GA) for Scheduling

Genetic Algorithm (GA) is an adaptive probability optimization technology that is suitable for complex system optimization based on biological genetics and evolution mechanisms inspired by biological simulation technology by Professor Holland. In short one can rate this as an efficient, parallel, and global search method, which implicitly inherits and accumulate knowledge regarding problem search space when using search mechanism, and is adaptive control search process to find the best possible optimal solution [10].

#### 3.1 Encoding and Decoding of Chromosomes

There are many ways to encode chromosomes, either direct encoding (encoding the execution status of the task) or indirect encoding. This work makes use of the resource-task indirect coding method proposed in [8] to code the resources occupied by subtasks. The chromosome length is determined by the number of subtasks, and expected value of individual gene in the chromosome is the resource number occupied by the subtask corresponding to that position, as shown in Fig. 1. Among them,  $T_i$  represents the task number, and  $P_i$  represents the resource number occupied when the task  $T_i$  is executed. When generating the initial population, the resource number  $P_i$



**Fig. 1** Task code

in each chromosome is randomly generated. After crossover and mutation operators, the subtask  $T_i$  may occupy any available resource, so the optimal solution must correspond to a certain chromosome code. Assuming there are 10 subtasks and 3 available resources, chromosome length is 10 and individual gene value is a random number between 1 and 3. For example, the following chromosome code is randomly generated: {2, 1, 2, 1, 3, 1, 3, 3, 2, 3}.

First subtask will be represented by this chromosome that runs on second cloud resource, the second subtask runs on the first cloud resource, and so on, the tenth subtask runs on the third resource.

After the chromosome is generated, it must be decoded to get the distribution of subtasks which runs on different resources. The subtasks are classified according to the occupied resources, and multiple sets of subtask sequences classified according to resource numbers are generated. Decode the above chromosome as:

$$P_1 : \{T_2, T_4, T_6\}, P_2 : \{T_1, T_3, T_9\}, P_3 : \{T_5, T_7, T_8, T_{10}\}$$

The sequence of subtasks allocated to each computing resource is obtained by decoding, and the time taken for each computing resource to finish the sequence of subtasks can be calculated using the ETC matrix. Due to the characteristics of concurrent processing among various computing resources in cloud computing, the threshold value of the above calculation results is considered as the time when all subtasks are completed:

$$\text{sumTime}(i) = \sum_{j=1}^n \text{Time}(i, j), i \in [1, M] \quad (1)$$

$$\text{completeTime}(I) = \max(\text{sumTime}(i)) \quad (2)$$

In the above equation,  $n$  represents the number of individual subtasks allocated to each computing resource, and  $\text{Time}(i, j)$  indicates the time required to complete the execution of the  $j$ -th subtask allocated to the cloud computing resource  $P_i$ . At this point, using the  $\text{sumTime}(i)$  obtained above, combined with  $\text{RCU}(i)$ , the cost required to complete all subtasks can be obtained:

$$\text{complete cost}(I) = \sum_{i=1}^M \text{sum Time}(i) \times \text{RCU}(i) \quad (3)$$

The task scheduling algorithm proposed in this paper needs to consider the time and cost of all subtasks at the same time. Here, using the greedy algorithm, the maximum time and maximum cost constraints required for the completion of all subtasks are:

$$\begin{aligned} \text{MaxTime} &= \text{completeTime}_{\text{MIN}} + \\ &t \times (\text{completeTime}_{\text{MAX}} - \text{completeTime}_{\text{MIN}}) \end{aligned} \quad (4)$$

$$\begin{aligned} \text{MaxCost} &= \text{completeCost}_{\text{MIN}} + \\ &c \times (\text{completeCost}_{\text{MAX}} - \text{completeCost}_{\text{MIN}}) \end{aligned} \quad (5)$$

Among them,  $\text{completeTime}_{\text{MIN}}$  and  $\text{completeCost}_{\text{MIN}}$  are the results of greedy selection based on the possible minimum time and minimum cost needed for the subtasks to be get executed on each cloud computing resource.  $\text{completeTime}_{\text{MAX}}$  and  $\text{completeCost}_{\text{MAX}}$  are just the opposite, which are based on the subtasks running on each computing resource. The maximum time and maximum cost required to complete the results obtained by greedy selection. The  $t \in [0, 1]$  in the formula (4) is the time factor, and the  $c \in [0, 1]$  in the formula (5) is the cost factor. Their values are specified by the user according to their needs, and  $t$  and  $c$  are inversely proportional. For time-sensitive applications,  $t$  should be specified as the smaller value between [0.2, 0.5], and  $c$  should be specified as the larger value between [0.5, 0.8]. For cost-sensitive applications,  $c$  should be specified as the smaller value between [0.2, 0.5], and  $t$  should be specified as the larger value between [0.5, 0.8].

### 3.2 Generation of Initial Population

If the population size is considered to be  $S$ , where  $M$  indicates the number of resources and  $N$  indicates number of subtasks, then the initialization description is:  $S$  chromosomes are generated by the system randomly,  $N$  is the length of chromosome, and the gene value lies in  $[1, M]$  and it is a random number.

### 3.3 Fitness Function

The selection of the fitness function of the genetic algorithm is very important, which directly affects the convergence speed of the genetic algorithm and the search for the optimal solution. Individuals with greater fitness are more likely to be inherited

to the next generation. Individuals with less fitness are less likely to be inherited to the next generation. Task scheduling needs to consider the time and cost required to complete all subtasks.

The fitness function of the defined time is:

$$F_{\text{time}}(I) = \frac{1}{\text{complete Time}(I)} \times u_{\text{LB}} \quad (6)$$

$$u_{\text{LB}} = \frac{\sum_{i=1}^M \text{sumTime}(i)}{M \times \text{completeTime}(I)} \quad (7)$$

The  $u_{\text{LB}}$  in formula (6) is defined as the load factor of balancing tasks, and its value can be obtained by formula (7), which represents the utilization of each computing resource. The larger the value of  $u_{\text{LB}}$ , the greater the utilization of cloud computing resources, and the smaller the value of  $\text{completeTime}(I)$  will be.

The fitness function that defines the cost is:

$$F_{\text{cost}}(I) = \frac{1}{\text{complete cost}(I)} \quad (8)$$

In the fitness function that only considers time constraints, the greater the utilization of cloud computing resources and the time will be proportionately less to complete all subtasks, further fitness value will be high. In the fitness function that only considers cost constraints, as the cost of completing all subtasks is less, the greater is the fitness value. Accordingly, the fitness function considering the time-cost constraint can be defined as:

$$\text{Fitness}(I) = \alpha \times F_{\text{time}}(I) + \beta \times F_{\text{cost}}(I) \quad (9)$$

In above equation  $\alpha \in [0, 1]$ ,  $\beta \in [0, 1]$ ,  $\alpha + \beta = 1$ . When  $\alpha = 1$ ,  $\beta = 0$ , the result of algorithm scheduling is the shortest time scheduling to complete all subtasks. When  $\alpha = 0$ ,  $\beta = 1$ , the result of algorithm scheduling is the minimum cost scheduling of all subtasks.

### **3.4 Genetic Manipulation**

#### **3.4.1 Selection Operation**

Selection operation is the process of selecting individuals with strong adaptability in a population to generate a new population. As per the principle of “survival of the fittest”, if the fitness of an individual is high, then greater will be the probability

of being selected and inherited to the next generation, so that the fitness value of the individual in the population is constantly reaching the optimal solution. I-GA makes use of roulette selection as the selection operation operator and determines the probability of an individual being selected by the fitness calculation formula (9).

$$P(i) = \frac{\text{Fitness}(i)}{\sum_{j=1}^S \text{Fitness}(j)} \quad (10)$$

The fitness function takes into account the time and cost constraints of task scheduling. Through the above selection operations, there are both individuals with shorter task finish time and individuals with less task finish cost in the population.

### 3.4.2 Crossover and Mutation

Crossover is termed as operation in the main method to produce new generation individuals, which can be used as a global search item of a given genetic algorithm. Mutation operation will be used to enhance local search the ability of the given genetic algorithm, which preserves the diversity of the population, and premature phenomena will be prevented. I-GA uses the adaptive genetic algorithm (AGA) proposed in [11], and improves the calculation formulas of crossover probability and mutation probability, and adaptively adjusts the probability of crossover and mutation operations.

$$P_c = \begin{cases} P_{cl} - (P_{cl} - P_{c2}) \times (f_{\max} - f') / (f_{\max} - f_{\text{avg}}), & f' \geq f_{\text{avg}} \\ P_{c1}, & f' < f_{\text{avg}} \end{cases} \quad (11)$$

$$P_m = \begin{cases} P_{m1} - (P_{m1} - P_{m2}) \times (f_{\max} - f) / (f_{\max} - f_{\text{avg}}), & f \geq f_{\text{avg}} \\ P_{m1}, & f < f_{\text{avg}} \end{cases} \quad (12)$$

In the formula,  $f_{\max}$  indicates the maximum fitness value in the group.  $f_{\text{avg}}$  indicates the average fitness value of each generation of the group.  $f'$  indicates the larger fitness value of the two individuals to be crossed.  $f$  indicates the fitness value of the muted individual [11].

## 3.5 I-GA Reschedule

After each generation of genetic operations ends, I-GA uses the MaxTime and Max-Cost constraints obtained at the beginning of the algorithm to compare with the chromosomes of the optimal subtask scheduling results. If the optimal subtask scheduling result generated by the running of the I-GA algorithm does not satisfy the constraint

condition that  $\text{completeTime}(I)$  is less than or equal to  $\text{MaxTime}$ , or  $\text{completeCost}(I)$  is less than or equal to  $\text{MaxCost}$ , the I-GA algorithm needs to be rerun to generate a new optimal subtask scheduling result.

## 4 Simulation Results

Work implemented in this paper is based on CloudSim [12–15] platform. The same is used to simulate the cloud environment. Under the same conditions, I-GA, T-GA (Time constraints Genetic Algorithm) and C-GA (Cost constraints Genetic Algorithm) are used for comparative experiments. The values of various parameters used for this work are shown in Table 1.

The first and foremost condition of the algorithm:  $S$  is 60,  $M$  is 5,  $N$  is 1000, further the system generates ETC matrix and RCU array.

Algorithm Termination condition: Taking into account the overhead of using genetic algorithm to schedule tasks, set the maximum evolution algebra (MaxGn) to 100, and the algorithm terminates without waiting for the algorithm to converge.

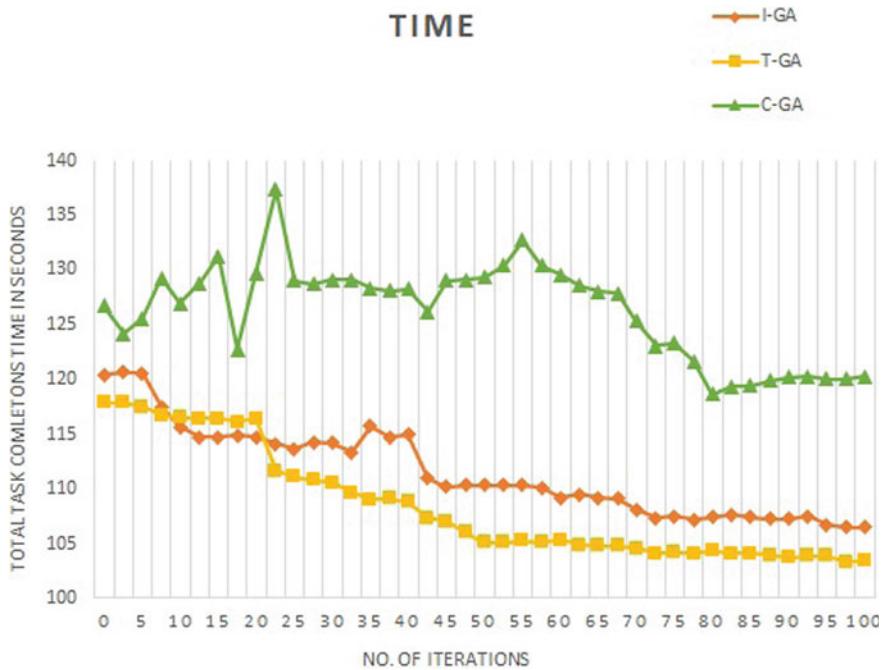
The experimental results are shown in Figs. 2 and 3.

It can be seen from Fig. 2 that in the early phase of genetic algorithm evolution, the total task finish time required for the optimal subtask scheduling results generated by I-GA and T-GA and C-GA operations are not much different. With the growth of evolutionary algebra, I-GA, and T-GA scheduling results have become more obvious in the optimization of the total task completion time, while C-GA scheduling results have no obvious optimization effect on task completion time in total.

One can observe from Fig. 3 that in the early phase of genetic algorithm evolution, the total task completion cost required for the optimal subtask scheduling results produced by I-GA and T-GA and C-GA operations is not much different. With the growth of evolutionary algebra, I-GA and C-GA scheduling results have become more and more obvious in the optimization of the total task completion time, while T-GA scheduling results have no obvious optimization effect on the total task finish time.

**Table 1** Simulation parameters

Major Parameters	I-GA	T-GA	C-GA
$t$	0.5	—	—
$c$	0.5	—	—
$\alpha$	0.3	1	0
$\beta$	0.7	0	1
$P_{c1}$	—	0.9	—
$P_{c2}$	—	0.61	—
$P_{m1}$	—	0.1	—
$P_{m2}$	—	0.01	—



**Fig. 2** Comparison of total task completion time

In summary, the subtask scheduling results obtained by T-GA can achieve the shortest total task completion time, but the effect of optimizing the task completion cost is not obvious. The subtask scheduling results obtained by C-GA can achieve the smallest total task completion cost, but The optimization effect of task completion time is not obvious. I-GA also considers the constraints of time-cost on the algorithm, so the result of subtask scheduling obtained makes the total task finish time shorter and cost involved will be less.

## 5 Conclusion

Task scheduling algorithms usually use the total time taken for the process to complete the task as the standard to schedule tasks. This paper combines the characteristics of cloud and proposes an improved scheduling algorithm for cloud based on GA considering time-cost constraints. The algorithm combines the total time taken for task to complete its work with the total cost of task completion is also used as a standard to schedule tasks. Experimental results show that this algorithm can achieve a more reasonable task scheduling in the cloud computing platform and produce ideal task scheduling results. In future work, we will focus on resource load balancing



**Fig. 3** Comparison of total task completion cost

for dynamic task scheduling in cloud computing, and consider the impact of other resources in cloud computing, such as network service quality, data distribution, and other factors on task scheduling results.

## References

1. Jakobik et al.: Non-deterministic security driven meta scheduler for distributed cloud organization. *Simul. Model Pract. Theory* 67–81 (2017)
2. Douglas et al.: Experimental assessment of routing for grid and cloud. In: 10th International Conference on Networks pp. 341–346 (2011)
3. Alhakami et al.: Comparison between cloud and grid computing: review paper. *Int. J. Cloud Comput.* **2**(4) 1–21 (2012)
4. Hao, Y. et al.: An adaptive algorithm for scheduling parallel jobs in meteorological Cloud. *Knowl. Based Syst.* (2016) 226–240
5. Khorandi et al.: Scheduling of online compute-intensive synchronized jobs on high performance virtual clusters. *J. Comput. Syst. Sci.* 1–17 (2017)
6. Chongdarakul et al.: Efficient task scheduling based on theoretical scheduling pattern constrained on single I/O port collision avoidance. *Simul. Model. Pract.* 171–190 (2016)

7. Cao, Q., et al.: An optimized algorithm for task scheduling based on activity based costing in cloud computing. In: 3rd International Conference on Bioinformatics and Biomedical Engineering, pp. 34–37 (2009)
8. Guo, L., et al.: Task scheduling optimization in cloud computing based on heuristic algorithm. *J. Netw.* **547**–553 (2012)
9. Buyya, R., et al.: GridSim: a toolkit for the modeling and simulation of distributed resource management and scheduling for grid computing. *J. Concurr. Comput.* **13**–15 (2002)
10. Calheiros, R.N., et al.: CloudSim: a novel framework for modeling and simulation of cloud computing infrastructures and services. Technical Report, GRIDS-TR-2009-1, Grid Computing and Distributed Systems Laboratory (2009)
11. Buyya, R. et al.: Calheiros, modeling and simulation of scalable cloud computing environments and the CloudSim toolkit: challenges and opportunities. *High Perform. Comput. Simul.* **1**–11 (2009)
12. Zhong-wen, G., et al.: The Research on cloud computing resource scheduling method based on Time-Cost-Trust model. In: 2nd International Conference on Computer Science and Network Technology (ICCSNT), p. 10 (2009)
13. Wu, H., et al.: A priority constrained scheduling strategy of multiple workflows for cloud computing. In: 14th International Conference on Advanced Communication Technology (2012)
14. Zhang, X., et al.: Locality-aware allocation of multi-dimensional correlated files on the cloud platform. *J. Distrib. Parallel Databases* **33**(3), 353–380 (2015)
15. Mukundan, et al.: Efficient integrity verification of replicated data in cloud using homomorphic encryption. In: *J. Distrib. Parallel Databases* **32**(3), 507–534 (2014)

# Sentiment Analysis for Telugu Text Using Cuckoo Search Algorithm



G. Janardana Naidu and M. Seshashayee

**Abstract** In recent times, most of the microblogging applications have started allowing people to express their opinions and feeling toward entities in regional languages. This resulted in good demand for sentiment analysis in regional languages. Telugu is a morphologically rich agglutinative south Indian language with nearly 100 million native speakers. In this paper, various unsupervised machine learning algorithms are explored for the classification of Telugu text into negative or positive classes.

## 1 Introduction

In recent times, most of the microblogging applications have started allowing individuals to give their feelings and opinions toward products or issues of public interest in regional languages. As a result, huge volumes of text are made available in regional languages. This resulted in good demand for sentiment analysis in regional languages along with other research areas like text summarization, and machine translation. Telugu is a morphologically rich agglutinative south Indian language with nearly 100 million native speakers. Recently many researchers have started showing interest in the research area of sentiment analysis for Telugu language. In this paper, we explored various unsupervised learning algorithms for the classification of Telugu text into negative or positive classes. The dataset used for the current study is the Annotated Corpus for Telugu Sentiment Analysis (ACTSA) [1] which is a collection of Telugu text taken from various e-newspapers. A few changes to the dataset have been made so has to be appropriate for the study. The modified dataset undergoes a few preliminary steps before the unsupervised algorithms are applied to it. The preliminary steps the dataset goes through are pre-processing and feature extraction. Two unsupervised algorithms are used, namely K-means and Cuckoo Search algorithm for the classification of Telugu text into negative or positive classes.

---

G. Janardana Naidu ( · M. Seshashayee

Department of Computer Science, GITAM (Deemed to be University), Visakhapatnam, India

## 2 Related Work

Support Vector Machines (SVM) was used by Mullen and Collier [2] to bring together information from diverse sources and assign values to selected words and phrases so that texts can be classified as positive or negative. Pandey et al. [3] proposed a method to find the appropriate cluster heads from Twitter dataset's sentimental content based on K-means and cuckoo search. Lee and Cheoah [4] proposed a method to predict the usefulness of online reviews based on neural networks.

Das and Bandyopadhyay [5] generated SentiWordNets for Indian languages which include Telugu. Computational techniques for generation of sentiment lexicons for Indian Languages were proposed by Das and Bandyopadhyay [5]. Dr. Sentiment a tool which automatically creates the PsycoSentiWordNet an extension of Senti-WordNet was proposed by Das and Bandyopadhyay 2011 [6]. PsycoSentiWordNet is created involving the Internet population, and presently sentiment knowledge and human psychological knowledge on a few aspects.

Mukku et al. [7] explored various supervised learning algorithms for the classification of Telugu text into negative or positive classes.

## 3 Dataset

Annotated Corpus for Telugu Sentiment Analysis (ACTSA) [1] has a collection of Telugu sentences taken from various e-newspapers and preprocessed and manually annotated by Telugu native speakers using annotated guidelines. Dataset contains 1489 positive sentences, 1441 negative sentences, 2475 neutral sentences and total 5410 sentences.

## 4 Methodology

This approach involves three steps, pre-processing, feature extraction and clustering.

### 4.1 Pre-processing

Each sentence is converted into tokens using python regular expression which uses spaces as a delimiter. As a next step, word embedding is created by using Word2Vec tool provide by Gensim. Word2Vec model created using 5 windows and 200 feature dimension vector and vocabulary table was built. As a last step model was trained and saved for further use.

## 4.2 Clustering

### K-means

K-means data clustering method groups n points into K groups by iteratively minimizing the distance of the K-cluster heads from the data points. Euclidean distance or cosine means are used to calculate the distance. Centroids are iteratively calculated by minimizing the squared sum of distance between data point and centroid. In this paper, scikit-learn package in python is used to implement K-means algorithm.

Algorithm 1

1. Randomly K-cluster heads are specified
2. Data points are assigned to the cluster to which the distance is minimum
3. Eq. (1) is used to calculate the new cluster head.

$$c_i = \frac{1}{n_i} \sum_{d_i \in s_i} d_i, \quad i = 1, 2, 3, 4 \dots \quad (1)$$

4. Continue Step 2 and 3 until it converges

### Cuckoo Search Algorithm

The algorithm is based upon the breeding behavior of the cuckoo bird. A few species of cuckoo engage the obligate brood parasitism by using other species nests to lay their eggs. There are three idealized rules for the cuckoo search algorithm [8]. The rules are as follows:

- (1) At a given time, a randomly chosen nest is taken and the cuckoo lays one egg in it.
- (2) The nests with best quality eggs are sent to the next generation.
- (3) The host nests used in the algorithm are fixed, and the eggs laid by cuckoo are discovered with a probability  $p_a$ .

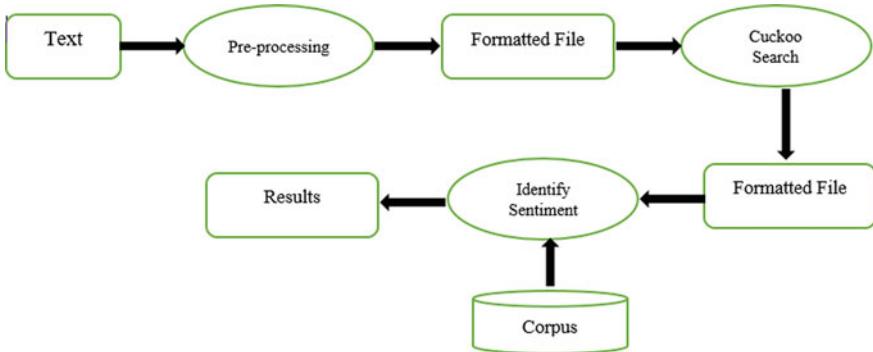
Then generate an initial population of n host nests  $x_i$  ( $i = 1, 2, \dots, n$ ). A cuckoo is selected randomly using levy flight Eq. (2) at each iteration. This process of selection runs until the convergence is reached.

$$X_i^{(t+1)} = X_i^{(t)} + \alpha * \text{LF} \quad (2)$$

where LF is a value from the Levy distribution,  $\alpha > 0$  is the step-size,  $i = 1, 2, \dots, n$ ,  $n$  is the number of nests considered.

It is assumed that each nest is a solution and the cuckoo egg represents a new solution. Then comparison of the fitness of the existing solution ( $F_n$ ) with the new solution ( $F_c$ ) is done. If  $F_c > F_n$ , then replace the solution.

The fitness of the cuckoo is calculated using



**Fig. 1** Architecture for cuckoo search (CS) for sentiment analysis

$$f_i = f(X_i) \quad i = 1, 2, \dots, n \quad (3)$$

And

$$f(k) = \sum_{k=1}^K \sum_{i=1}^{n_k} (x_i - c_k)^2 \quad (4)$$

## 5 Result

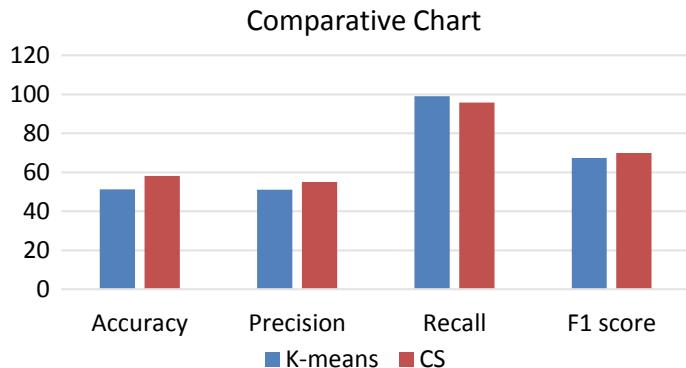
Two experiments are done. K-mean algorithm is used to verify the sentiment of Text as first experiment. Second experiment is conducted by using cuckoo search algorithm to find initial centroid. Cuckoo Search (CS) algorithm provides the best result in clustering compared to K-means. All the programs are written in python using scikit-learn,<sup>1</sup> gensim<sup>2</sup> packages (Fig. 1).

## 6 Conclusion

The results show that the use of Cuckoo Search is better in comparison with the use of K-means. The reason being the capacity of cuckoo search to identify the centroid of the cluster is better than K-means. As mentioned, the reason for use of unsupervised learning algorithms is the lack of annotated datasets for Telugu texts in multiple domains. Future work can be done to explore more unsupervised algorithms and

<sup>1</sup><https://scikit-learn.org/stable/install.html>.

<sup>2</sup><https://pypi.org/project/gensim/>.



**Fig. 2** Comparative chart

**Table 1** Comparative table

	Accuracy	Precision	Recall	F1-Score
K-means	51.2	51.1	99.1	67.4
Cuckoo search	58.1	55.1	95.7	69.9

to identify a suitable algorithm for sentiment analysis of Telugu language (Fig. 2; Table 1).

**Acknowledgements** I would like to thank Dr. D. Sasi Raja Sekhar, Assoc. Professor in St. Mary's Group of Institutions, Hyderabad, who motivated me toward research in NLP.

## References

1. Mukku, S.S., Radhika, M.: ACTSA: annotated corpus for telugu sentiment analysis. In: Proceedings of the First Workshop on Building Linguistically Generalizable NLP Systems, pp 54–58 (2017)
2. Mullen, T., Collier, N.: Sentiment analysis using Support Vector Machines with diverse information sources. EMNLP **4**, 412–418 (2004)
3. Pandey, A.C., Rajpoot, D.S., Saraswat, M.: Twitter sentiment analysis using hybrid cuckoo search method. Inf. Process Manage. **53**, 764–779 (2017)
4. Lee, S., Choeh, J.Y.: Predicting the helpfulness of online reviews using multilayer perceptron neural networks. Expert Syst. Appl. **41**(6), 3041–3046 (2014)
5. Das, A., Bandyopadhyay, S.: SentiWordNet for Indian Languages. In: The 8th Workshop on ALR, COLING 2010, pp 56–63 (2010)
6. Das, A., Bandyopadhyay, S.: Dr sentiment knows everything!. In: ACL/HLT 2011 Demo Session, pp. 50–55 (2011)
7. Mukku, S.S., Choudhary, N., Mamidi, R.: Enhanced sentiment classification of Telugu text using ML techniques. In: 25th International Joint Conference on Artificial Intelligence (2016)
8. Yang, X.-S., Suash, D.: Cuckoo search via levy flight. World congress on nature and biologically inspired algorithms. EEE Publication, pp 210–214 (2009)

# Automation of Change Impact Analysis for Python Applications



T. Jalaja, T. Adilakshmi, and P. S. R. Abhishek

**Abstract** Software applications have to be updated on an ongoing basis as per the requirements of the user or the client or to fix any bugs. The cost of this maintenance phase of a software application typically costs more than 50% of all other software development life cycle (SDLC) phases. When a change request (CR) is received from the client, developers have to work upon the request. It may be possible that as a result of addressing this CR, it may impact other areas of the existing application functionality. This mandates that developer working on this CR should have complete knowledge of the application codebase and working scenarios. Quite often, developer may be lacking this extensive expertise and also may not have the time required to spend in doing the detailed impact analysis. There are some change impact analysis (CIA) tools which can aid developer in finding out almost all impacted elements for a given CR. But no such tool or prototype exists for Python applications which can aid developer in finding out impacted elements for a CR. This paper aims at developing an impact analysis tool (HETeye) which will help the developer to know what all elements of the Python source code will be affected and the dependencies of affected elements upon giving a CR and blacklist words which help in filtering the results even further.

## 1 Introduction

Software applications will keep evolving for years and decades. For example, Facebook started as a simple chat application in 2004, and now, it evolved into a widely adopted social media platform and a tech giant [1] by adding new features and building services over the course. During the evolution process, there would have been some code changes and integrations. However, majority of the codebase would

---

T. Jalaja (✉) · T. Adilakshmi · P. S. R. Abhishek  
Vasavi College of Engineering, Hyderabad, Telangana, India  
e-mail: [jalaja.t@staff.vce.ac.in](mailto:jalaja.t@staff.vce.ac.in)

T. Adilakshmi  
e-mail: [t\\_adilakshmi@staff.vce.ac.in](mailto:t_adilakshmi@staff.vce.ac.in)

have remained the same. Over the duration, the code becomes legacy, which makes it more tedious to maintain as it requires more resources and time to perform any task on it.

Standards organizations like ISO [2] and IEC [3] mandate to perform change impact analysis (CIA) prior to making any software changes to the application for some perilous sectors like railway, aviation, medical and few other sectors. These standards do not state on how to perform the change impact analysis. In most applications, it is done manually with help of diagrams and charts like system dependency graph (SDG). Sometimes, developers will discuss to find out the possible impacted elements based on their prior knowledge of the codebase. But these methods are labor-intensive, time consuming and less accurate. In such scenarios, CIA tools aid the developers in finding out impacted elements with less efforts, thereby improving the productivity.

There are few existing CIA tools available for languages like JAVA, C and C++. Some of these tools like *ImpRec* [4] use information retrieval (IR) and recommendation system for software engineering (RSSE), where the system is trained on information of what files were impacted for a change and start giving recommendations after training. These methods are inefficient as they take time to train and build a knowledge base for each application.

Natural language processing (NLP) is a technology that gives the machine ability to understand and analyze the human languages.

The existing NLP tools can process human language and derive the sentence meaning and context, in a considerable amount of time. They can also be used to check the similarities between word and sentences or the relationship of words with a sentence. Using the parts-of-speech (PoS) tagger, we can get the significance of the words in a sentence.

Python is one of the languages increasing in popularity over the last decade [5] for various reasons, and there is no proven CIA tool available for Python applications.

## 2 Literature Survey

### 2.1 Change Impact Analysis

A key factor for success of software is that they can be modified with less efforts and resources compared to other engineering disciplines. But making changes to a part of code can have impact on other parts of the code related to it, which increases the failure rate of the software when a change is made. This is referred to as *ripple effect* or *side effect*.

To overcome this, the impacted parts of code should also be modified according to the changes being made. But understanding the impact of a change on the code is a tedious and time-consuming task. To reduce the failure rate, the change impact

analysis is mandated by standards organizations like ISO and IEC in safety-critical sectors like aviation [2, 3].

Bohner et al. defined CIA as '*the process of identifying the potential consequences of a change, or estimate what needs to be modified to accomplish a change*' in the book Software Change Impact Analysis [6].

## 2.2 Existing Tools

Software engineering projects constitute of thousands of heterogeneous artifacts, and the relation between them is even more complex. Analyzing the impact manually between them is time taking and also requires prior experience with that project.

There is a lot of research being conducted on CIA, to ease the process by building tools which can aid developer in understanding the impact with less effort.

There are some tools built to perform CIA, and most of them are built using static code analyzers, information retrieval (IR) systems, recommendation system for software engineering (RSSE) or with a combination of those technologies.

Though they helped developers in perform CIA, their efficiency can still be improved.

We can mimic the human process of understanding the purpose of the code element through its name by using natural language processing (NLP).

With NLP libraries and models like NLTK [7] and Open AI's [8] offerings, we can build a tool which can analyze the change request and relevance of a code element with that change request

## 3 Methodology

### 3.1 Proposed Work

In this paper, we aim at developing a CIA tool (*HETeye*, pronounced as *HET-I*) for Python applications which is efficient and faster. The technologies like natural language processing (NLP) and an enhanced algorithm are used to process the change request and output the affected elements.

In this section, the *HETeye* tool is discussed, which takes the change request, path of the source code and blacklisted words as input. *HETeye* then processes the change request based on parts of speech and stop words to identify the word with emphasis. Then, synonyms of these emphasized words are generated based on the context of change request.

*HETeye* computes rank for each class, method, method argument, attribute, decorators and imported libraries, where the rank represents the number of words associated with user input present in the name of the method, class or other code elements. The higher the rank, the more it is related to the change request.

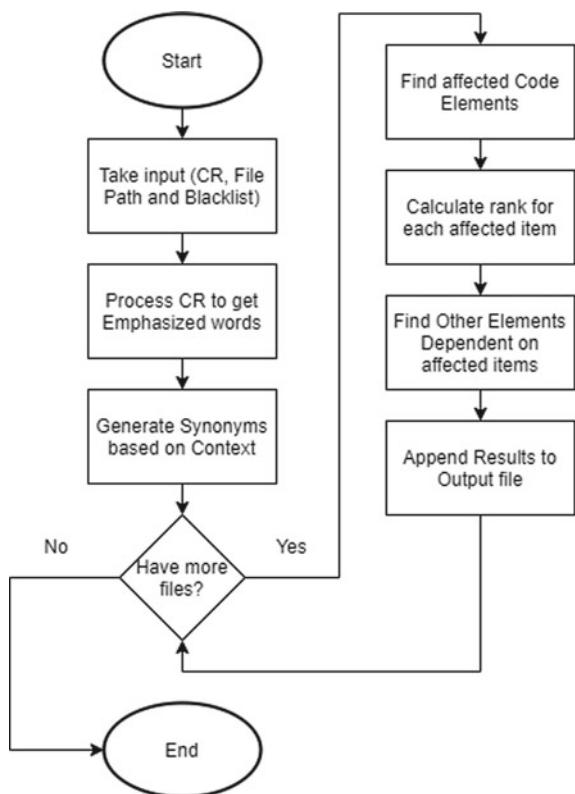
Blacklist words is a list of words given by user as input to further filter the results as per his needs. The classes, methods and attributes with rank 1 will not be considered if the only word its relative with is in blacklist. This comes in handy to filter out the results or to filter out other common programming constructs.

*HETeye* will output the affected elements, file wise with the elements sorted according to their ranks and their dependencies if any are present.

### 3.2 Proposed Algorithm

Figure 1 depicts the flowchart of the proposed algorithm.

**Fig. 1** Flowchart of proposed algorithm



### 3.3 Implementation Steps

**Classification of Elements** This phase will tokenize the change request and tag the words using their respective parts of speech which are used to find out the emphasized words. Then, the synonyms of those emphasized words are generated based on the context of change request.

**Getting Impact Set** This phase uses the output of the previous phase to search for classes, methods and attributes whose names match with any of the words generated in previous set. It also makes a list of dependencies (abstract tree) of the classes and methods in the result.

### 3.4 Performance Metrics

Accuracy and precision are used to evaluate the performance of the *HETeye* tool.

Confusion matrix depicts how the actual values differ from the predicted values with four attributes. Figure 2 gives an idea on how the data is categorized under those four attributes.

Accuracy is the proportion of correct classifications from overall number of cases.

$$\text{Accuracy} = \frac{\text{True Negative} + \text{True Positive}}{\text{True Negative} + \text{False Negative} + \text{True Positive} + \text{False Positive}} \quad (1)$$

Precision is the proportion of correct positive classifications from overall positive classifications.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} = \frac{\text{True Positive}}{\text{Total Positive}} \quad (2)$$

**Fig. 2** Confusion matrix

		Predicted Class	
		Normal	Attack
Actual Class	Normal	True Negative (TN)	False Positive (FP)
	Attack	False Negative (FN)	True Positive (TP)

**Table 1** Sample CRs showing the corresponding TP, TN, FP, FN and accuracy for gensim

Change request	TN	TP	FN	FP	Accuracy	Precision
Improve performance of author model	14012	6	2	5	0.99950	0.545455
Change corpus indexing method	14016	7	1	1	0.99985	0.875
Increase window size while summarization	14017	6	2	0	0.99985	1
Use tidygraph instead of igraph	14024	1	0	0	1	1
Decrease minimum words of article in wiki corpus	14006	10	0	9	0.99935	0.526316
Add functionality to change base dir	14013	7	5	0	0.99964	1

## 4 Results

The accuracy and precision of the *HETeye* tool were tested upon three applications taken from GitHub. Each application was built to serve different purpose. The average accuracy and precision over these three applications will be considered as the average accuracy and precision of the *HETeye* tool.

### 4.1 Gensim

The first application is Gensim, which is a Python library for *topic modeling*, *document indexing* and *similarity retrieval* with large corpora [9].

Table 1 shows some of the change requests for this application and their corresponding TP, TN, FP, FN, accuracy and precision values.

The average for Gensim accuracy is 0.999702911.

The average for Gensim precision is 0.824461722.

### 4.2 Zulip

The second application is Zulip, an open-source group chat application that combines the immediacy of real-time chat with the productivity benefits of threaded conversations. It is developed using Django Framework [10].

Table 2 shows some of the change requests for this application and their corresponding TP, TN, FP, FN, accuracy and precision values. Figure 3 shows accuracy and precision values of each CR along with average lines.

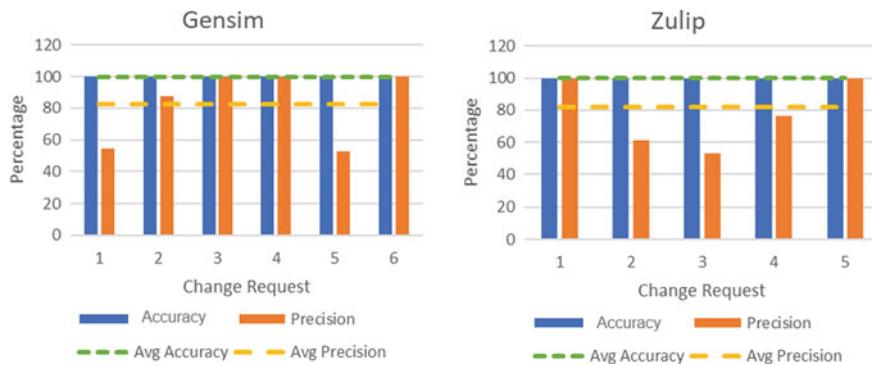
The average accuracy for Zulip is 0.999187992.

The average precision for Zulip is 0.781900452.

In some change requests, the false positive percentage is a bit higher. Though those cases are not affected, those elements are still related to the change request, this trend was not seen in other applications being tested.

**Table 2** Sample CRs showing the corresponding TP, TN, FP, FN and accuracy for Zulip

Change request	TN	TP	FN	FP	Accuracy	Precision
Change date time format	40546	79	15	0	0.99963	1
Remove statsd integration	40555	65	0	20	0.99950	0.764705
Implement ip rate based limiting	40555	45	0	40	0.99901	0.529412
Replace mock path	40406	144		90	0.99778	0.615385
Remove the feature to mute topics	40532	108	0	0	1	1

**Fig. 3** Bar chart of Gensim CR's (right) and bar chart of Zulip CR's (right)

### 4.3 Bokeh

The third application is Bokeh, an interactive visualization library for modern Web browsers built using Python. The majority of the codebase is in Python and typescript [11]. Figure 4 shows the accuracy and precision of few Bokeh change requests.

Table 3 shows some of the change requests for this application and their corresponding TP, TN, FP, FN, accuracy and precision values.

The average accuracy for Bokeh is 0.999929766.

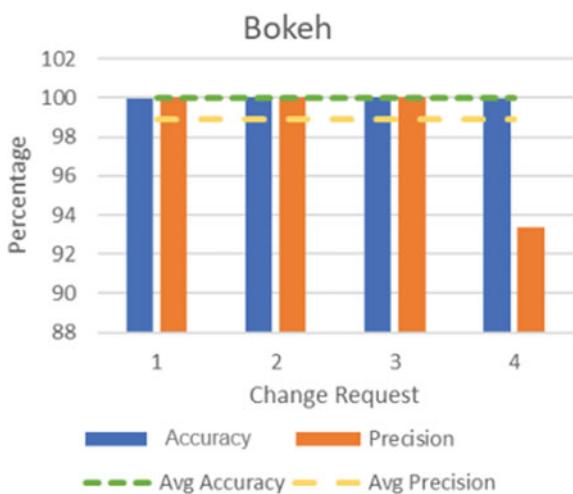
The average precision for Bokeh is 0.986666667.

### 4.4 Result

The average accuracy and precision of the tool are 0.99960 and 0.86434, respectively. Figure 5 shows the accuracy and precision value of all CR's of all applications.

Figure 6 shows how the output of the impact analysis tool looks like, file-wise classification of elements along with their ancestors and sorted according to their rank. The number of lines affected depicts the number of occurrences of affected items in that file.

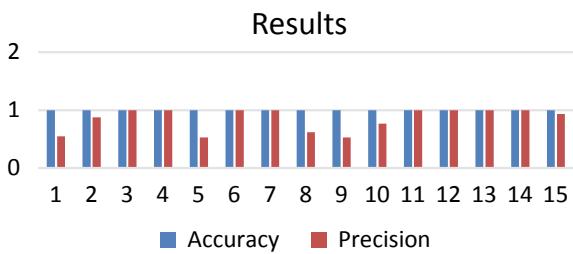
**Fig. 4** Bar chart of Bokeh CR's



**Table 3** Sample CRs showing the corresponding TP, TN, FP, FN and accuracy for Bokeh

Change request	TN	TP	FN	FP	Accuracy	Precision
Replace setter_id with better mechanism	48221	195	15	0	0.99969	1
Replace yaxis, ygrid with y_axis, y_grid	48221	391	0	0	1	1
Replace xaxis, xgrid with x_axis, x_grid	48221	512	0	0	1	1
Redesign set_dataset_visual	48221	28	0	2	0.999959	0.933333

**Fig. 5** Bar chart of CR's results



## 5 Conclusion and Future Work

The developed software tool displays the impacted set of classes, methods, attributes and other elements like imports and arguments with an average accuracy of 99.96% and average precision of 86.43% which was calculated by comparing the manual result to output given by the tool. This tool was developed using Python and takes only Python applications as input.

It can be further improved by generating some blacklist words automatically based on the context of the change request and by listing the elements associated

```
--C:/impact analysis/gensim\gensim\sklearn_api\atmodel.py---
The number lines effected are 4
Items that may be Impacted are
1 Attributes:
Rank
-----
2 AuthorTopicModel

0 Functions:
0 Classes:
1 Other Calls:
Rank Func           Ancestors
----- -----
2 AuthorTopicModel      -
```

**Fig. 6** Example output (snippet)

with library methods whose implementation cannot be altered. The Open AI's GPT [8] is developed for conversation generation but, it can be used to get the in-depth meaning and context of a sentence. This API can be used to understand CR better, thus improving the accuracy.

## References

1. Britannica–Facebook: <https://www.britannica.com/topic/Facebook>
2. International Organization for Standardization: ISO 26262-1:2011Road Vehicles—Functional Safety (2011)
3. IEC 61511-1 ed 1.0, Safety Instrumented Systems for the Process Industry Sector (2003)
4. Borg, M., Wnuk, K., Regnell, B., Runeson, P.: Supporting change impact analysis using a recommendation system: an industrial case study in a safety-critical context. In: IEEE Transactions on Software Engineering (2016)
5. <https://stackoverflow.blog/2017/09/06/incredible-growth-python/>
6. Bohner S., Arnold R.: Software change impact analysis. (IEEE Computer Society Press, Los Alamitos, 1996)
7. Nltk: <https://www.nltk.org/>
8. Language Models are Few-Shot Learners: <https://arxiv.org/abs/2005.14165>
9. Gensim: <https://github.com/RaRe-Technologies/gensim>
10. Zulip: <https://github.com/zulip/zulip>
11. Maps Model Importer: <https://github.com/eliemichel/MapsModelsImporter>

# Enhancing Item-Based Collaborative Filtering for Music Recommendation System



M. Sunitha, T. Adilakshmi, and Mir Zahed Ali

**Abstract** Music recommendation system is a tool designed to help users to find interesting music from huge volumes of digital collections. Content-based filtering method and collaborative filtering method are the most used by researchers in the field of music recommendation. The objective of the paper is to enhance generic item-based CF. In item-based CF, items are grouped into clusters based on their similarity. The item clusters thus formed are used in recommendation process. As the numbers of items in clusters are very large, framing recommendation vector based on only item-based CF might not result in a very successful and accurate recommendation system. This paper proposes a research work to enhance generic item-based CF, by combining with K-nearest neighbor method. Music taste of a user varies based on the time of a day. To include this parameter in generic item-based CF, context information is defined for each item based on the time of the day. Context information is combined with item-based CF to further enhance the proposed system. Proposed methods to enhance the item-based collaborative filtering are experimentally verified by using a standard benchmark dataset Last.fm which is obtained from million song dataset. Results show an improvement over generic item-based CF model.

## 1 Introduction

Due to the availability of open sources on the Internet, a greater number of users started using the Web and contributing content to the Web. This led to the exponential growth of information on the Web. However, this prompted two important issues as follows:

Data collection: Availability of different types of data allowed users to use different tools to work on unstructured and semi-structured data.

---

M. Sunitha (✉) · T. Adilakshmi · M. Z. Ali  
CSE Department, Vasavi College of Engineering, Hyderabad, India  
e-mail: [m.sunithareddy@staff.vce.ac.in](mailto:m.sunithareddy@staff.vce.ac.in)

Searching time: Because of the huge amount of data, users find it difficult to search the items useful and interesting. This is known as information overloading.

Recommendation systems are designed to solve the problem of users with respect to searching time. They work as information filtering tools and provide suggestions to the users with less/no user intervention. There are many fields such as books, movies, and news where recommendation systems are playing a vital role. This research paper addresses a recommendation system in music field. Music is such a vibrant item, and with the increased usage of Internet, music service providers such as Amazon music, Gaana, and Wink music are using music recommendation systems to understand the user behavior, and thereby, increase their business. Music recommendation system is an information filtering tool which predicts the songs interesting to a user based on the past history of the users. Because of the recommendation systems, music industry is able to handle huge volumes of data available via digital platforms and satisfy the needs of different users.

Most of the music recommendation systems are based on popularity of an item, i.e., recommend popular songs to users. Even though it is very simple and basic recommendation method, it does not add any personalization to a specific user. The proposed music recommendation system in this paper understands user behavior and recommends songs interesting to the user.

The other commonly used approaches for music recommendation are content-based and collaborative filtering. The limitation with content-based approach is availability of content information for the items. Collaborative filtering method faces challenges like cold start, sparsity, and long tail [1].

The rest of the paper is organized as follows: Sect. 2 describes related work; Sect. 3 showcases various methods to enhance item-based collaborative method, and results are given in Sect. 4. Section 5 explains about conclusion and future scope. References are given in the next section.

## 2 Related Work

The objective of a music recommendation system is to suggest music interesting to the users and help users in discovering new artists, songs based on their interest. Music service providers such as Allmusic, Pandora, Audiobaba, Mog, Musicover, Spotify, and Apple Genius collected millions of users' behavior and suggested music based on their interest. Music is different from other items such as books, movies, and restaurants. Recommendation systems need data about the users to perform recommendations. The framework for recommendation system consists of a set of users represented as  $\{U_1, U_2, \dots, U_n\}$  and a set of items represented as  $\{I_1, I_2, \dots, I_m\}$ . Based on the users and items, the framework constructs the basic data structure used in the process of recommendation system known as user-item rating matrix. Users are on rows, and items are on columns used to form the user-item matrix. In music recommendation system, users are the listeners, and songs are the items as shown in Fig. 1.

Users/Songs	S1	S2	S3	S4
U1	2	0	0	3
U2	0	0	6	4
U3	3	4	0	3
U4	0	5	6	1
U5	4	8	0	0
U6	0	0	4	3

**Fig. 1** Sample user-song matrix in music recommendation system

Most of the research in this field focused on suggesting a list of artists and a sequence of songs (playlist addressing personal interest of user). Demographic-based model, collaborative filtering model, and content-based model are used by the researchers in the field of music recommendation.

## 2.1 *Demographic-Based Model*

It is one of the basic models to include demographic information related to user or item for recommendation. Most used information about items is title of the song, artist's name, and lyrics to find the recommendation list for the target users. This method looks very easy and fast, but this model requires user to know about the information of the items. The major drawback of this method is that users will never get a chance to listen to novel songs.

## 2.2 *Collaborative Filtering (CF) Model*

CF model is one of the most popular models used for music recommendation. The basic working principle of a CF-based model is when two users agree in the past, then they will agree in the future also, i.e., if two users like similar kind of songs in the past, then we might use the list of songs listened by one of the users to prepare recommendation list for other user. Nearest neighbors are used as one of the most common approach in CF-based methods to provide recommendations. Memory-based, model-based, and hybrid collaborative filtering are the three different varieties into which most of the research work done in music recommendation system based on CF can be categorized [1, 2]. Memory-based collaborative filtering predicts the items based on the entire collection of previous ratings. Nearest neighbors are one of the commonly used memory-based method. As the name suggests, it finds the nearest neighbors of

a target user or item for recommendations. Model-based collaborative filtering, in contrast to memory-based CF, uses machine learning and data mining algorithms to build a model which captures the underlying relation between users and items. Based on the known model, the recommendation system makes recommendations for target users. Features of both model-based and memory-based methods are combined into a hybrid collaborative filtering model. The literature proves that hybrid CF model performs better compared to individual methods [1].

### 2.3 *Content-Based Model*

These methods are based on the metadata related to songs such as timbre and rhythm [3, 4]. For any target user, finding and recommending list of songs similar to the songs listened by her/him is done. Similarities between the songs are identified based on the features or content of it. Thus, the name is given as content-based recommendation system.

## 3 Enhancing Item-Based Collaborative Model

Item-based CF is a model-based CF. It uses clustering algorithm to group items into clusters and build a model represented in terms of item clusters. The clusters thus formed are used in framing recommendation vector as shown in [5]. But due to a large number of items in the dataset, recommendations based on only item clusters were not that accurate. So, to improve the performance of generic item-based CF, it is combined with the nearest neighbors and context information. This section explains different methods to enhance item-based model.

### 3.1 *Combining KNN with Item-Based-CF Model*

To enhance the accuracy of the music recommendation system, item-based CF is combined with K-nearest neighbor method. Item-based clusters address the scalability issue faced by most of the collaborative filtering methods. It builds a model in the form of finite and manageable item clusters from huge user-item rating matrix. But most of the research work in the music recommendation system is based on KNN with implicit or explicit rating matrix obtained from user experience. Let us consider the following user-item rating matrix of four users about bollywood top songs in the year 2019 as shown in Table 1 to illustrate the working of a simple KNN-based recommendation system.

In order to provide recommendations for the user Alice, let us consider K-nearest neighbor algorithm with  $K = 1$ . Alice already liked S1, so with KNN, we need to

**Table 1** Top-4 bollywood songs of 2019

User/song	O Saki Saki (S1)	Dheeme Dheeme (S2)	Bala Bala (S3)	Gunghroo (S4)
Christ	2	0	2	1
Bob	3	1	0	3
Suji	0	2	1	0
Alice	3	?	?	?

**Table 2** Euclidean distance from S1 to other three songs

Song/distance	S2	S3	S4
S1	$\sqrt{12}$	$\sqrt{5}$	1

find the nearest neighbors of S1. As per the data given in Table 1, the song vectors are formed for S1 as [2,3,0], S2 as [0,1,2], S3 as [2,0,1] and S4 as [1,3,0]. Euclidean distance is used to find the nearest neighbors for S1.

From Table 2, it can be seen that the first nearest neighbor for S1 is S4. As K is given as 1, so S4 is recommended to Alice. The major issue with simple KNN algorithm is the run time complexity of the algorithm to provide recommendations. There is no model building in KNN. Hence, it is called lazy learner. For each test user, the algorithm needs to find the nearest neighbors during the recommendation process in order to frame recommendation vector.

In the process of music recommendation with the user-item matrix of size  $200 \times 14,458$ , simple KNN algorithm takes a lot of time to provide recommendations for the test users as there 14,458 unique items. To handle this issue with simple KNN, in generic item-based CF, a model is built from the user-item matrix of size  $200 \times 14,458$  [5, 6]. Item clusters serve as the basic model to provide recommendations. Even with item-based CF, where the numbers of item clusters are finite and small in number, it sometimes fails to identify the items interesting for a test user. So, to capture the interest of a test user accurately and recommend the items interesting to test users are proposed in this paper. To enhance the performance of item-based CF recommendation system, other parameters and additional information are added in this research work.

The first approach to enhance generic item-based CF is to take advantage of KNN in case small-size data. As in item-based CF, items are grouped into item clusters. Each item cluster contains small number of items compared to overall items in the dataset. So, KNN is brought into work after a test user is mapped with the nearest item cluster. Instead of recommending all items from the mapped item cluster as shown in item-based CF [5, 6], KNN is used to find the items that really satisfy the needs of a test user. For a test user, 5% of the items already present in the test user-item vector are considered as seed to find nearest neighbors. For the items present in 5% of test users, mean vector is computed to aggregate the past history of the test user. Mean item vector is used to find nearest neighbors from mapped item cluster for any test user as shown by algorithm in Fig. 2.

Algorithm ITEM\_CF\_KNN\_Mean()

Input: Item clusters, test users

Output: Recommendation of songs to test users

Method:

Begin

1. Let m is the number of item clusters given as SC1, SC2,.....SCm
2. Let Ut  $\in \{U_1, U_2, \dots, U_{200}\}$  be a test user
3. Compute the mean\_Vector(Ut) for each test user from 5% of items present in test user item vector
4. Find the distance (Ut, SCv) where SCv  $\in \{SC_1, SC_2, \dots, SC_m\}$
5. Let SCv be the nearest item cluster for the test user Ut
6. Let SCv = {S1, S2, ..., Sl}
7. Find Euclidean distance between mean\_vector(Ut) and each S  $\in \{S_1, S_2, \dots, S_l\}$
8. Arrange the songs {S1, S2, ..., Sl} in the ascending order of distance from test user Ut
9. Add first K items from the list obtained in step 8 to the recommendation vector for test user Ut
10. Goto step 2 and select a new test user. Repeat steps 3 to 9 for all the test users

End

**Fig. 2** Combining KNN with item-based CF

In the process of enhancing music recommendation system, item-based CF is combined with KNN by considering top 5% of items from the test user-item vector as the seed to find the nearest neighbors from a mapped item cluster as shown by the algorithm in Fig. 3. The nearest neighbor for each item present in the top 5% of items of a test user is added to the recommendation vector.

### 3.2 Combining Item-Based CF with Context

The context in recommendation systems refers to the time of a day, location, age, etc. In the process of music recommendation system, context information is obtained from the time of a day. The rational for considering context from the time of the day is that users will have different music tastes at different times of the day. So, by considering context based on the time will enhance the performance of the recommendation system. The dataset obtained from Last.fm [7] for conducting this research work consists of the attributes given as userID, timestamp, AlbumID, AlbumName, TrackID, and Trackname as shown in Fig. 4. Timestamp which is mentioned in a 24-hour clock is used to obtain context information. Timestamp is divided into three different slots to define context information as given in Table 3.

Consider the data from Table 4 as an example to obtain the context of a song. Context of S1 is evening, S2 is morning, S3 is evening, and S4 is afternoon as they

```

Algorithm ITEM_CF_KNN_Top_items()
Input: Item clusters, test users
Output: Recommendation of songs to test users
Method:
Begin
  1. Let m is the number of item clusters given as SC1, SC2,.....SCm
  2. Let Ut ∈ { U1,U2,.....U200} be a test user
  3. Compute the mean_Vector(Ut) for each test user
  4. Find the distance ((Ut, SCv) where SCv ∈ { SC1, SC2,.....SCm })
  5. Let SCv be the nearest item cluster for the test user Ut
  6. Let SCv= {S1,S2,.....SI}
  7. Find top 5% of items listened by test user Ut
  8. Find the distance between ((each song ∈ top 5% of Ut), (each item ∈ SCv)
    where SCv ∈ { SC1, SC2,.....SCm })
  9. Arrange the songs {S1,S2,.....SI} in the ascending order of distance
    from the top 5% of test user Ut listened items
  10. Add first K songs from the list obtained in step 9 to the recommendation vec-
      tor for test user Ut. (remove the duplicates while adding K items from each
      test user song)
  11. Repeat the steps from 8 to 10 for all the items in top 5% for all the test users
End

```

**Fig. 3** Combining KNN for top items with item-based CF

have been heard more number of times during that time of the day. This context information is included in the user-item matrix as shown in Table 5.

In the process of including context information into music recommendation system, user-item matrix is modified to include the context information as shown in Table 5. Algorithm defined in Fig. 5 is used to combine context information with item-based CF.

Context information is used to filter items for a test user from a mapped item clusters as shown in Fig. 5. For any test user, the list of items from the mapped item cluster should be found. This should match the context of the test user seed items to form recommendation vector.

Context-based recommendation system is further enhanced by combining KNN. The steps are shown in the algorithm as given in Fig. 6.

## 4 Results

Proposed enhancements for music recommendation system are evaluated using the evaluation measures from Information retrieval systems known as precision and recall. Precision shows the accuracy of the proposed recommendation system as shown in Eq. (1), and recall shows the exhaustiveness of the music recommendation

UserID	time-Stamp	albumID	album-Name	trackID	trackName
user_000001	2008-01-14T14:38:16Z	b2bbf316-ebae-4275-a0b2-58e3758a0d47	Kuniyuki Takahashi	b3df7881-06ed-45f7-bbdf-93c67b6a9fa6	Think Of You (Piano Re-Edit)
user_000002	2008-02-24T12:17:53Z	315fa94f-3bb5-43db-a28e-406a83e9805c	Kett-car	a7dd46a7-9d3e-4679-88d2-a69e119b067e	Money Left To Bum

**Fig. 4** Sample user logs**Table 3** Defining context based on the timestamp

Time duration	Name of the context
4.00 am to 10.00 am	Morning
10.00 am to 6.00 pm	Afternoon
6.00 pm to 4.00 am	Evening

**Table 4** Obtaining the context of a song

Song	Morning	Afternoon	Evening
S1	3	0	12
S2	5	2	3
S3	2	1	6
S4	0	6	2

system as given in Eq. (2). Precision and recall are obtained from true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN). F-measure as shown in Eq. (3) represents harmonic mean of precision and recall.

**Table 5** User-item matrix with context information

User/item	S1	Context	S2	Context	....	S14458	Context
U1	12	Evening	5	Morning	....	6	Afternoon
U2							
....							
U200							

## Algorithm ITEM\_CF\_Context()

Input: User-item matrix with context, Item clusters

Output: Recommendation vector for test users

Method:

Begin

1. Let m is the number of item clusters given as SC1, SC2,.....SCm
2. Let Ut  $\in \{U1, U2, \dots, U200\}$  be a test user
3. Compute the mean\_Vector(Ut) for each test user
4. Find the distance ( $Ut, SCv$ ) where  $SCv \in \{SC1, SC2, \dots, SCm\}$
5. Let  $SCv$  be the nearest item cluster for the test user Ut
6. Let  $SCv = \{S1, S2, \dots, Sl\}$
7. Find top 5% of songs listened by test user Ut
8. Obtain the context of each song in top 5% from user-item matrix with context
9. Find the context of the songs in the nearest cluster
10. Add the songs with the same context as the context of top 5% songs of test users to the recommendation vector

End

**Fig. 5** Combining context with item-based CF

$$P(U_i) = \frac{\text{No. of songs listened by a test user from the recommended list}}{\text{Total no. of recommendations}} \\ = \frac{TP}{TP + FP} \quad (1)$$

$$R(U_i) = \frac{\text{No. of songs listened by a test user from the recommended list}}{\text{Total no. of songs listened by a test user}} \\ = \frac{TP}{TP + FN} \quad (2)$$

$$\text{F-measure}(U_i) = \frac{2 \times (P(U_i) + R(U_i))}{(P(U_i) \times R(U_i))} \quad (3)$$

The research work proposed in this paper has been experimentally verified by examining with the benchmark dataset obtained from Last.fm. The dataset is a collects behavior of 999 users for about three years. Information about users listening

Algorithm ITEM\_CF\_Context\_KNN()

Input: User-item matrix with context, Item clusters

Output: Recommendation vector for test users

Method:

Begin

1. Let m is the number of item clusters given as SC1, SC2,.....SCm
2. Let Ut  $\in \{U_1, U_2, \dots, U_{200}\}$  be a test user
3. Compute the mean\_Vector(Ut) for each test user
4. Find the distance (Ut, SCv) where SCv  $\in \{SC_1, SC_2, \dots, SC_m\}$
5. Let SCv be the nearest item cluster for the test user Ut
6. Let SCv = {S1, S2, ..., S1}
7. Find top 5% of songs listened by test user Ut
8. Obtain the context of each song in top 5% from user-item matrix with context
9. Find the context of the songs in the nearest cluster
10. Find the K nearest neighbour songs of top 5% songs of test users
11. Add the songs from K nearest neighbours with the same context as the songs in top 5% of test user to the recommendation vector

End

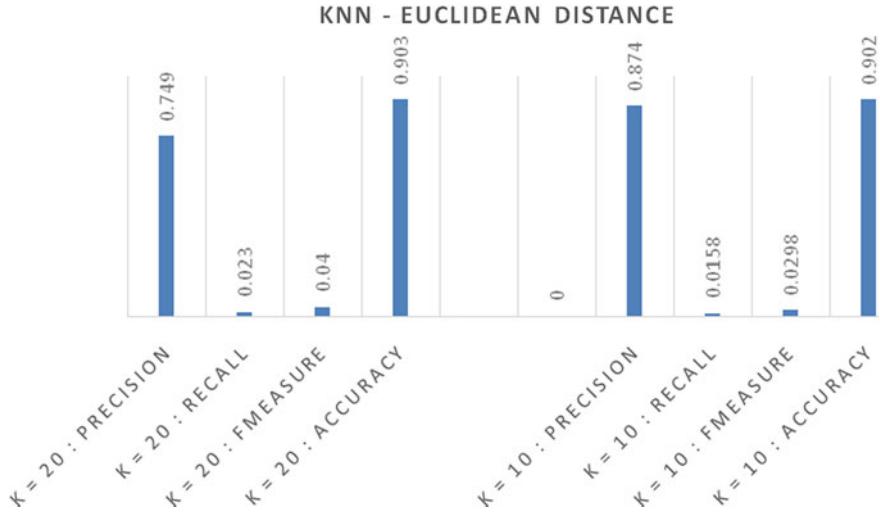
**Fig. 6** Combine context with KNN for item-based CF

history is maintained in dataset with attributes given as UserID, Timestamp, ArtistID, ArtistName, TrackID, and TrackName. Another dataset is used to give information about user demographic information with attributes given as UserID, age, gender, location, and registration date. Sample datasets screenshots are shown in Figs. 1 and 7, respectively.

In the process of enhancing item-based CF, KNN is combined to find items more similar to the items preferred by test user. KNN is implemented in different variations.

User ID	Gender	Age	Country	Registration date
user_000001	M	21	Japan	Aug 13, 2006
user_000002	F	24	Peru	Feb 24, 2006
user_000003	M	22	United States	Oct 30, 2005
user_000007	F	19	United States	Jan 22, 2006
user_000008	M	23	Slovakia	Sep 28, 2006

**Fig. 7** Sample user profile data



**Fig. 8** Performance evaluation of K-nearest neighbor item mean with item-based CF

First method of KNN implementation is to find cumulative mean of 5% items listened by a test user. The mean vector is used to find the nearest neighbors from the mapped item cluster. The second method of KNN implementation is to consider individual items of 5% of a test user and find the neighbors from the mapped item cluster. The last and third method of KNN implementation is to find 5% of the top items preferred by the test user and find the nearest neighbors from the mapped item cluster.

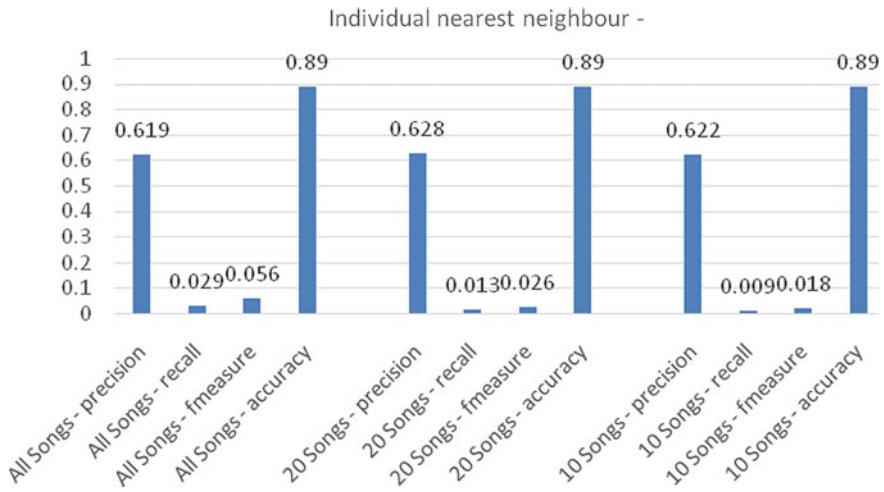
Results obtained for cumulative KNN are shown in Fig. 8. The results show a very good improvement over generic item-based CF in all cases. The recommendation system of the top-10 items performs well compared to other methods.

KNN with individual items of test user with only one nearest neighbor is shown in Fig. 9. This variation of KNN method is performed better compared to cumulative KNN.

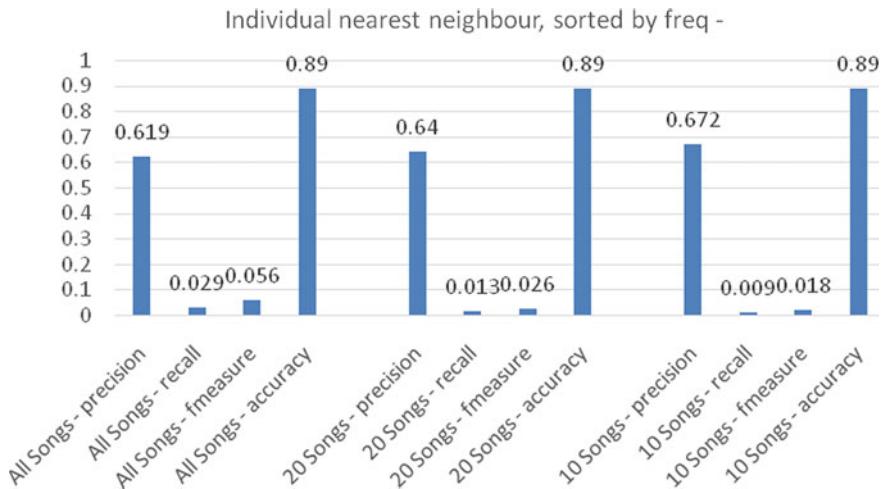
KNN with individual items of test user sorted by frequency with only one nearest neighbor is shown in Fig. 10. This variation of KNN method is performed better compared to individual item-based KNN as shown in Fig. 9.

Context information combined with generic item-based CF and results obtained are shown in Fig. 11. Precision of the proposed system is better compared to item-based CF but not as good as KNN-based enhancement. Context-based system is further improved by including KNN, and the results obtained are shown in Fig. 12. These results are good compared to only context combined with item-based CF.

Results obtained with generic item-based CF along with proposed enhancements are compared and shown in Fig. 13, and it proved that the proposed enhancements increased the precision of resulting recommendation systems.



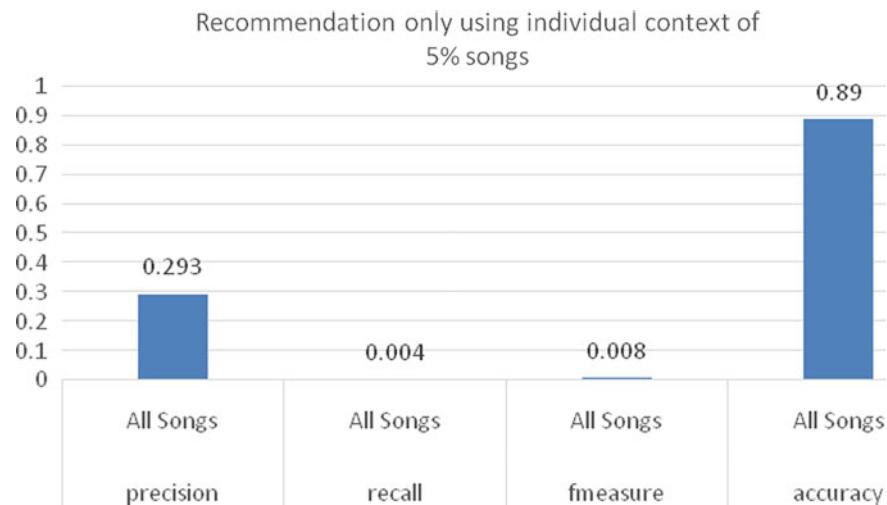
**Fig. 9** Performance evaluation individual K-nearest neighbor with item-based CF



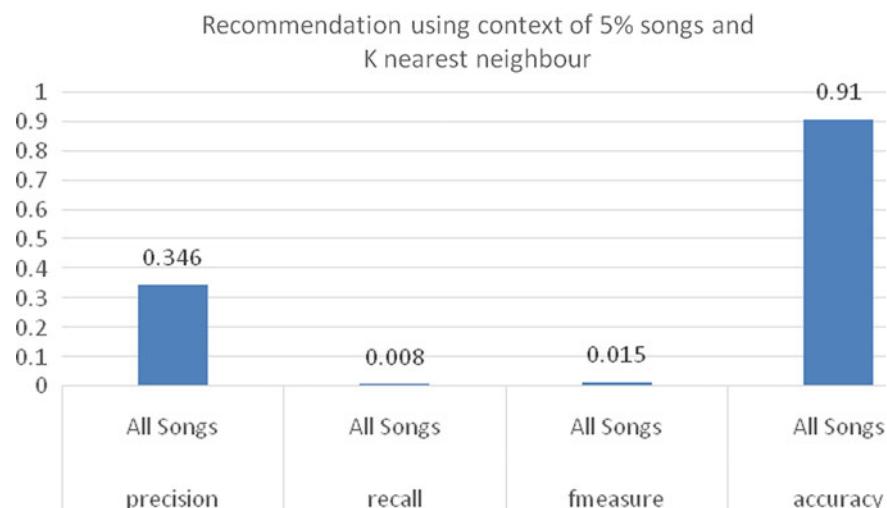
**Fig. 10** Performance evaluation individual K-nearest neighbor for top items with item-based CF

## 5 Conclusion and Future Scope

This research work enhances generic item-based CF. In item-based CF, item clusters are formed to build a model. Due to the size of dataset, item-based CF alone has not given good precision value. So, it is enhanced by combining KNN with item clusters. Once a test user is mapped to nearest item cluster, KNN is used to filter items from the mapped cluster and recommend only the items that satisfy the taste

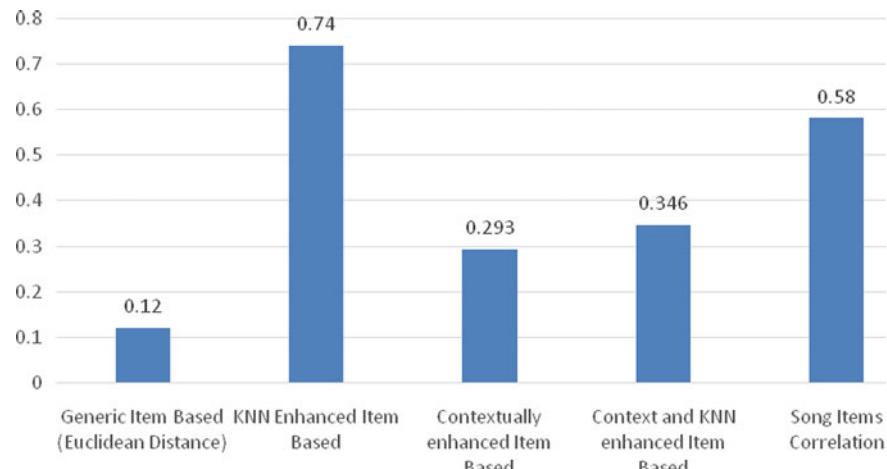


**Fig. 11** Performance evaluation of context-based recommendation with item-based CF



**Fig. 12** Performance evaluation of context-based recommendations using K-nearest neighbor with item-based CF

of the test user. KNN is implemented by considering 5% of the items present in test user-item vector. Randomly, 5% items are considered, and mean vector is computed to find the nearest neighbors. Another variation is considered randomly 5% of test user items as seed and find the individually nearest neighbor of each seed. Finally, KNN is implemented with top 5% of items as to form seed for test user and find the recommendations. The results thus obtained show that all methods performed



**Fig. 13** Comparison of precision for item-based CF with different enhancements proposed

with KNN enhanced the performance of generic item-based CF. Later, the research work proposed in this paper also included context as an additional parameter to recommend songs based on the time of a day. Context is obtained for an item based on the maximum number of times, and an item has been heard during a particular time of the day. Results show the improvement over generic item-based CF

#### Future scope:

The clustering algorithms used in item-based CF can be manipulated to the specific needs of the user and result in appropriate groupings of items. Regression analysis can be performed on the dataset to extract the user's mood and serve as a valuable parameter in music recommendation process.

## References

1. Elahi, M., Ricci, F., Rubens, N.: A survey of active learning in collaborative filtering recommender systems. *Comput. Sci. Rev.* **20**, 29–50 (2016)
2. Elahi, M., Ricci, F., Rubens, N.: Active learning in collaborative filtering recommender systems. In: Hepp, M., Hoffner, Y. (eds.) *E-Commerce and Web Technologies, Lecture Notes in Business Information Processing*, vol 188. Springer, pp 113–124 (2014). [https://doi.org/10.1007/978-3-319-10491-1\\_12](https://doi.org/10.1007/978-3-319-10491-1_12)
3. Ebrahimi Kahou, S., Michalski, V., Konda, K., Memisevic, R., Pal, C.: Recurrent neural networks for emotion recognition in video. In: *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ICMI'15*. ACM, New York, NY, USA, pp 467–474 (2015). <https://doi.org/10.1145/2818346.2830596>
4. Navya Sri, M., Ramakrishna Murty, M., et al.: Robust features for emotion recognition from speech by using gaussian mixture model classification. In: *International Conference and Published Proceeding in SIST Series*, Springer, vol. 2, pp. 437–444 (2017)

5. Sunitha, M., Adilakshmi, T., Ali, M.Z.: Enhancing item based collaborative filtering with item correlations for music recommendation system. *Int. J. Recent Technol. Eng. (IJRTE)* **9**(1) (2020). ISSN-2277-3878
6. Sunitha, M., Adilakshmi, T.: Comparison of user-based collaborative filtering model for music recommendation system with various proximity measures. *Int. J. Innov. Technol. Explor. Eng.* **8**(6S2), 15–21 (2019). ISSN: 2278–3075
7. Last. FM: A Popular Music Web Portal. <http://www.last.fm>

# Deep Learning-Based Enhanced Classification Model for Pneumonia Disease



S. Jeba Priya, S. Joshua Jaistein, G. Naveen Sundar,  
and T. Raja Sundrapandiyanleebanon

**Abstract** Pneumonia is an infectious disease in the lung which results in inflammation of the lung tissues and eventually leads to death of several millions worldwide. This paper aims to detect and classify the pneumonia using deep learning-based algorithm by examining the chest X-ray or computed tomography (CT) or combination of both. The field experts or clinical experts utilized the chest X-ray for diagnosing of pneumonia. This paper investigates the convolutional neural network (CNN) model for feature extraction and optimized the algorithm by increasing the number and layers. The proposed model performed in this study improved the average accuracy of 91% and with some other metrics that surpass the existing models in medical imaging.

## 1 Introduction

Image processing is used for object detection in the field of biomedical imaging to localize, identify and verify the target entity. In contrast to machine learning algorithms, convolutional neural networks (CNN) are inclined due to their multi-layered architecture. Deep learning models have accomplished robust results in many sectors [1–4]. Annually, about a tenth of the world's population is infected by pneumonia, out of which millions of patients face fatal risks [5]. Pneumonia has become one of the top causes of mortality in children and the elderly worldwide [6]. In fact, pneumonia is indeed a main cause of mortality from stroke, especially in patients with acute

---

S. Jeba Priya (✉) · S. Joshua Jaistein · G. Naveen Sundar · T. Raja Sundrapandiyanleebanon  
CSE Department, Karunya Institute of Technology and Sciences, Coimbatore 641114, Tamil Nadu, India

e-mail: [jebapriya@karunya.edu](mailto:jebapriya@karunya.edu)

S. Joshua Jaistein  
e-mail: [joshuajaistein@karunya.edu.in](mailto:joshuajaistein@karunya.edu.in)

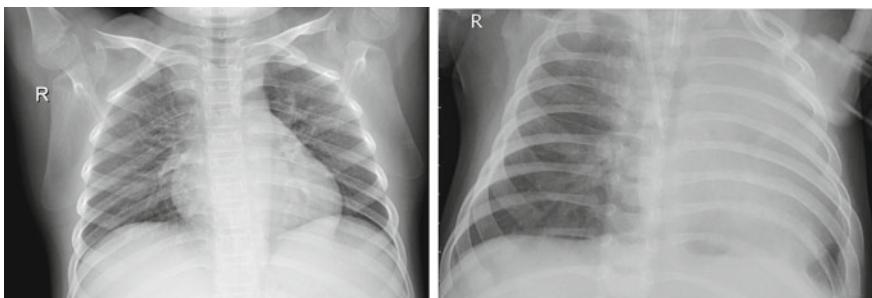
G. Naveen Sundar  
e-mail: [naveensundar@karunya.edu](mailto:naveensundar@karunya.edu)

ischemic stroke (AIS) [7]. The main objective of this paper is to automatically classify pneumonia-infected individuals from their chest X-ray images. The commonly used methodology to detect pneumonia bacterial infection worldwide is X-rays and computed tomography. Chest X-ray is one of the most popular, painless and non-invasive form of diagnosis. However, detection of pneumonia from X-rays is challenging even for experienced radiologists and is a time-consuming process [5]. This research paper depends on X-ray images rather than CT scans. Pneumonia typically occurs as a zone or field of increased opacity [8]. Therefore, an AI-driven program would be promising to enable doctors detect pneumonia correctly and in a timely manner and save more lives. Performance of the model is measured with accuracy, precision, recall and F1-score. Computer-aided diagnosis (CAD) for medical imaging has become more accurate and robust due to the advances in computational power and deep learning algorithms. Stunning level performance which surpass human intelligence is achieved by these algorithms, especially the convolutional neural networks (CNN) on encountering computer vision problems in medical imaging domain. In this research paper, an efficient method is suggested using appropriate dataset to train an efficient neural network such that the learned parameters can be used to detect pneumonia cases.

## 2 Literature Review

The development of deep learning and broad datasets allows algorithms to accomplish a broad variety of applications and spectacular results in the practice of radiology and medical imaging. Deep learning models are now commonly adopted in medical imaging systems because they can automatically extract features or by using certain pre-trained networks. Liang and Zheng [9], to anticipate one or more of fourteen potential diseases, a deeply related convolutional neural network was deployed (Fig. 1).

Anthimopoulos et al. [10] used a transfer learning with deep residual network for pediatric pneumonia diagnosis. Elhoseny and Shankar [11] used deep neural



**Fig. 1** Illustrates the differences in lungs of a normal versus pneumonia infected individual

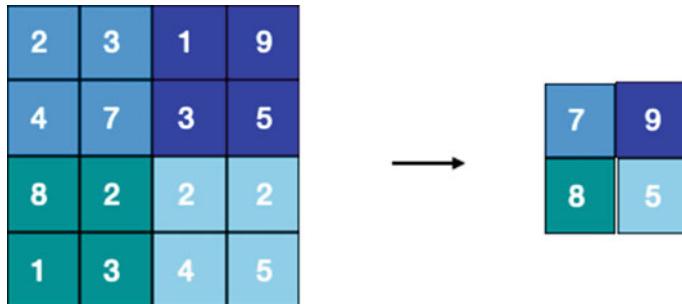
model cluster for localizing pneumonia for a large-scale chest X-ray dataset. Sirazit-dinoz and Kholiavchenko [12] presented a double CNN that automatically identifies huge front and transverse chest X-ray images mostly on MIMIC-CXR dataset and is currently the biggest dataset for chest X-ray scans. Varshini et al. [13], deep CNN was qualified to automatically identify respiratory infections via the ultrasound images. For class prediction, AlexNet and GoogLeNet, dual CNNs were used. Ref. [14], Mask RCNN-based recognition algorithm was adopted for pixel-wise segmenting and also integrated global and local characteristics.

Jain [15], to recognize intracellular respiratory disease trends, a CNN model was developed. Whose model incorporated five convolutionary layers, using a leaky activation feature of ReLU, average max pooling and triple thick layers and made crucial improvements throughout the training period by combining bounding box coordinates from different models. Rubin and Sanghavi [16], using chest X-ray pictures, the AG-CNN model technique was developed to diagnose thorax disorder. This study was performed on the dataset for Chest X-ray. He et al. [17], the topic of the highest recognition timescale to ANNs was concerned about. Two versions, MCPN but also MKNN, were introduced, classifying iteration-free MR high-resolution image precision. Tolerance and precision were used as success metrics for their models. It is evident that deep learning model can attain significant outcomes for pneumonia classification using chest X-ray images. Additional variation and tuning of hyperparameters can further improve the efficiency of the deep learning model. Therefore, the development of a more accurate deep learning-based pneumonia-affected patient classification model is the motivation for this work. Successive session elaborates the dataset and model succeeded by the experiment results and discussions. Finally, the section concludes the study.

### 3 Proposed Model

The image data is translated into the form of a matrix. The feature map is formed as a result of the operation between [18]. This method decreases the aspect ratios, making it simpler to process the frame. The feature detector preserves the essential portions of the image. The model contains of five convolutional layers and a pooling layer in-between each of them. These convolutional layers learn the characteristic representations of their images which have been fed into the neural net. Each neuron in a feature map has a field of reception. New feature maps are formed when the inputs are convoluted. At each position, maximal amount of features should be extracted for efficiency, so various feature maps within the same convolutional layer have different weights (Le Cun et al. 2015). Figure 2 describes feature detector pictorially.

Recently, the rectified linear units (ReLU) have become more popular. The rectified linear activation function removes the dilemma of the vanishing gradient, helping models to understand more easily and perform better. Xavier Glorot et al. in their milestone 2012 paper highlighted some of the benefits of the ReLU titled “Deep Sparse Rectifier Neural Network”. Mathematically, ReLU is defined as the following



**Fig. 2** Feature detector

formula.

$$f(x) = \max(0, x) \quad (1)$$

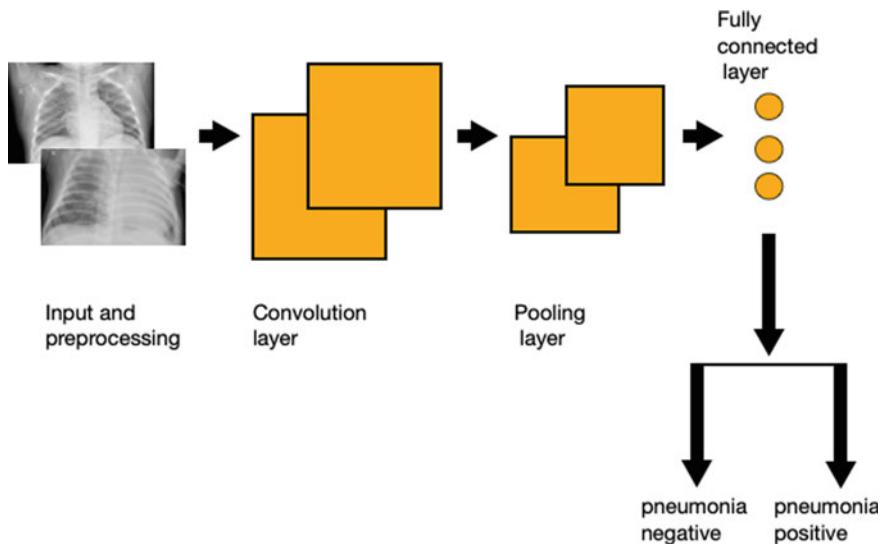
In order to decrease the spatial resolution of the feature maps, the pooling layers are introduced. It significantly reduces the number of parameters to learn and the amount of computation performed. The size of the pooling operation or filter is smaller than the feature map. In this paper, max pooling is performed, where the largest of the pixel values of an image segment is taken to be processed. The feature detector analyzes any pixel whether this pixel has an attribute.

Each neuron is bound to each and every neuron in another layer using the fully connected layer. It goes the final probabilities for each label.

Figure 3 represents the proposed architecture. The input image undergoes preprocessing, and the model contains a series of convolutional and pooling layer and finally to the fully connected layer which leads to the output of the neural network.

## 4 Result and Discussion

The dataset used is accessible on Kaggle under the name “Chest X-Ray Images Pneumonia Challenge” which contains images in grayscale. The 80:20 train-test scheme is used to train the chest X-ray pneumonia image classification model. We train our model on training data, gather the output and fine-tune the hyperparameters and use the testing data to measure the overall efficiency of the model proposed. For this paper, through measuring first-order moment projections and second-order moment projections of the gradient moment, Adam optimizer constructs separate adaptive learning thresholds for varying factors. When the accuracy of the model tends to show property of being overfit, the learning rate is decreased by three times. The training is finished until 150 epochs are constantly trained by the network. All the deep learning algorithms including the proposed model are deployed using the



**Fig. 3** Proposed architecture

Keras framework with Tensorflow-GPU backend on a Linux machine with NVIDIA GeForce GTX 1060 with 6 GB of dedicated graphics memory running CUDA 10.

To test the model response in a two-class challenge, evaluation metrics such as accuracy, precision, recall and F1-score are measured. In order to further assess the classification impact of the proposed model in this paper, we contrasted the five most frequently used CNN medical imaging architectures—VGG16, DenseNet121, XceptionNet and InceptionV3.

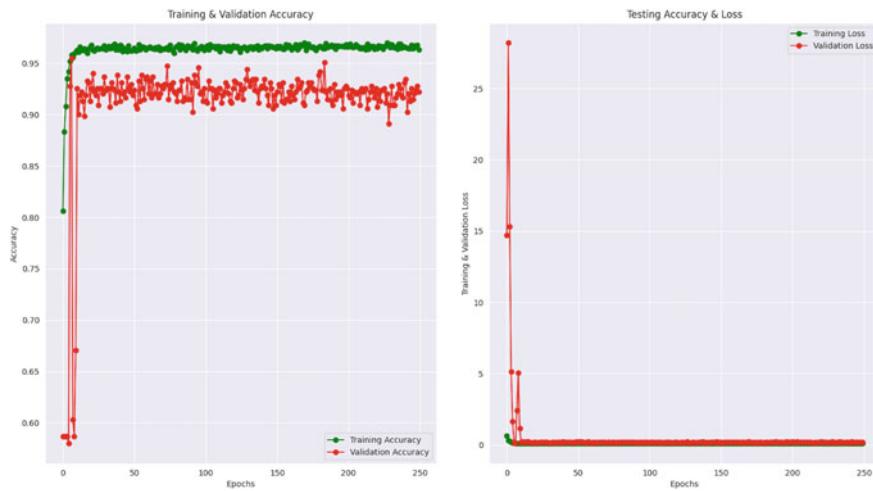
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (2)$$

Figure 4 plots the different accuracy and loss of the model with respect to epochs. The model showed a sudden increase in the accuracy and remained almost constant after 20 epochs. The loss tended to be very low.

A total of 6398 images of patients with and without pneumonia were considered for the prediction of pneumonia using chest X-ray scans. The various metrics are contrasted in Fig. 5. The medical prediction results achieved a respectable accuracy of 0.9102 (or) 91%.

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

Precision can be defined as the ratio of positive observations correctly determined to the total positive observations predicted. The highest precision of 0.9247 was obtained in this experiment



**Fig. 4** Model training accuracy and loss

	Accuracy	Precision	recall	F1-score
<b>Proposed model</b>	0.920	0.924	0.978	0.932
<b>DenseNet121</b>	0.8190	0.792	0.9270	0.869
<b>InceptionV3</b>	0.853	0.916	0.841	0.877
<b>Xceptionnet</b>	0.878	0.857	0.967	0.908
<b>VGG16</b>	0.742	0.723	0.951	0.840

**Fig. 5** Evaluation metrics—comparison table

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

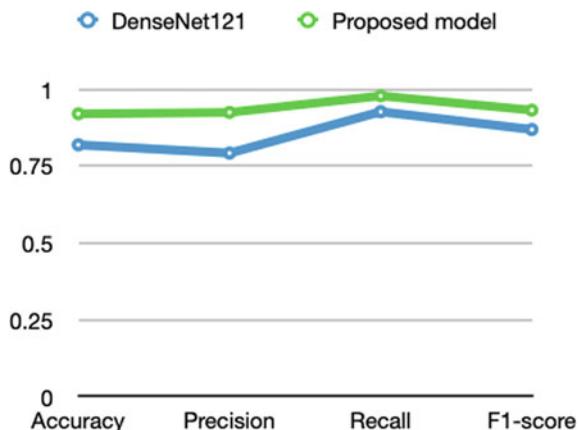
Recall is the number of accurately predicted positives of all results, the highest recall 0.9553.

$$\text{F1 Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision}) \quad (5)$$

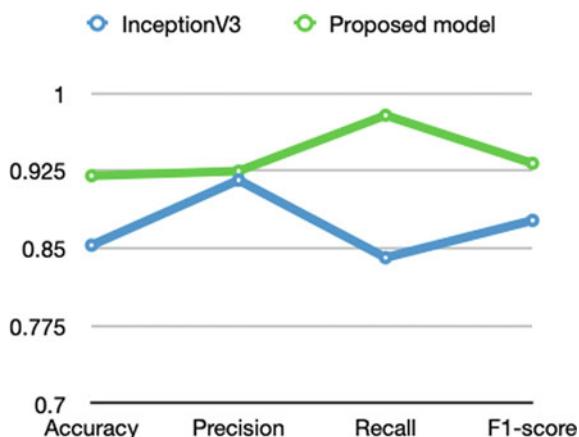
F1-score is a measure of a test's accuracy. A high F1-ranking means you have low false positives and low false negatives, and a respectable F1-score of 0.9324 was obtained in this experiment.

Figures 6, 7, 8 and 9 contrast the performance of the proposed model versus other popular neural networks with respect to four evaluation metrics, namely accuracy, precision, recall and F1-score.

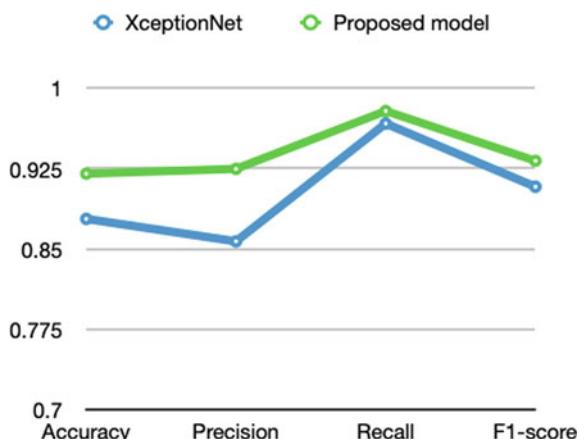
**Fig. 6** Model performance comparison with DenseNet121



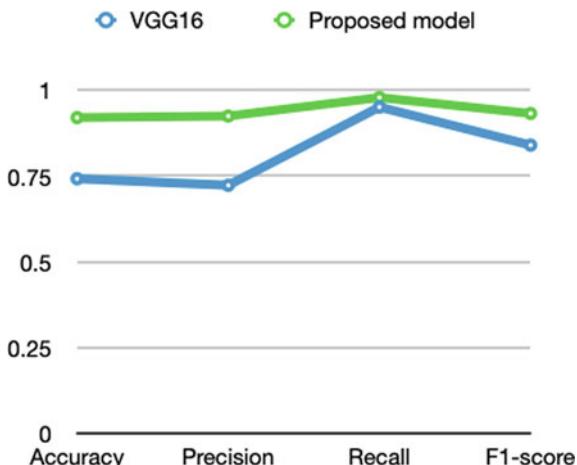
**Fig. 7** Model performance comparison with InceptionV3



**Fig. 8** Model performance comparison with XceptionNet



**Fig. 9** Model performance comparison with VGG16



## 5 Conclusion and Future Work

In this paper, we present an automatic diagnostic method that classifies X-ray images of pneumonia-affected individuals vs healthy individual. In the future research, we will continue to optimize the work by making the neural net perform it tasks on MRI and CT scans. The best way to expose the health effects of an algorithm in real time is by medical trials including AI and DL treatments with detailed systematic monitoring protocols. Figure 10 shows the confusion matrix of the proposed classifier model.

On recognition of the excellent results against certain performance metrics, the suggested model could be successfully included in the medical testing by health professionals for advanced diagnosis of pneumonia in patients.

Increase in the number of neural layers and convolutional layers certainly increases the computational complexity of the algorithm. The authors of this research paper plan to enhance the model by following the methods proposed by Maji and Mullins [19] to decrease the computational complexity of deep convolutional neural networks. We have taken permission from competent authorities to use the images/data as given in the paper. In case of any dispute in the future, we shall be wholly responsible.

**Fig. 10** Confusion matrix

	<b>Positive</b>	<b>Negative</b>
<b>Positive</b>	371	19
<b>Negative</b>	42	192

## References

1. Del Fiol, G., Michelson, M., Iorio, A.: A Deep Learning method to Automatically Identify Reports of Scientifically Rigorous Clinical Research from the Biomedical Literature: Comparative Analytic Study. <https://doi.org/10.2196/10281>
2. Comert, Z., Fetal Kocamaz, A.F.: Hypoxia Detection Based on Deep Convolutional Neural Network with Transfer Learning Approach. Springer International Publishing, Berlin (2019), pp. 239–248
3. Budak, U., Comert, Z., Cibuk, M., Sengur, A.: DCCMED-net: Densely Connected and Concatenated Multi Encoder-Decoder CNNs for Retinal Vessel Extraction from Fundus Images (2019). <https://doi.org/10.1016/j.mehy.2019.109426>
4. Budak, U., Comert, Z., Rashid, Z.N., Sengur, A., Cibuk, M.: Computer-aided diagnosis system combining FCN and Bi-LSTM model for efficient breast cancer detection from histopathological images. *Appl. Soft. Comput.* <https://doi.org/10.1016/j.jhmas/jrr047>
5. Levine, M.: The early clinical X-ray in the united states. Patient experiences and public perceptions. *J. Hist. Med. Allied Sci.* (2011). <https://doi.org/10.1093/jhmas/jrr047>
6. Shen, Y., Tian, Z., Lu, D., Huang, J., Zhang, Z., Li, X.: Impact of Pneumonia and Lung Cancer on Mortality of Women with Hypertension (2016). <https://doi.org/10.1038/s41598-0160023-2>
7. Ge, Y., Wang, Q.: Predicting post-stroke pneumonia using deep neural network approaches. *Int. J. Med. Inform.* <https://doi.org/10.1016/j.ijmedinf.2019.103986>
8. Rubin, J., Sanghavi, D.: Large Scale Automated Reading of Frontal and Lateral Chest X-rays Using Dual Convolutional Neural Networks (2018). arXiv preprint [arXiv:1804.07839](https://arxiv.org/abs/1804.07839)
9. Gaobo, L., Zheng, L.: A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput. Methods Progr. Biomed.* **187** (2019)
10. Anthimopoulos, M., Christodoulidis, S., Ebner, L.: Lung pattern classification for interstitial lung diseases using a deep learning neural network. *IEEE Trans. Med. Imag.* **35**(5), 1207–1216 (2016)
11. Elhoseny, M., Shankar, K.: Optimal bilateral filter and convolutional neural network based de-noising method of medical image measurements. *Measurements* **143**(2019), 125–135 (2019)
12. Siraztinoz, I., Kholiavchenko, M.: Deep Neural Network Ensemble for Pneumonia Localization from a Large-Scale Chest X-Ray Database
13. Varshini, D., Thakral, K., Agarwal, L.: Pneumonia detection using CNN based feature extraction. In: IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–7 (2019)
14. RSNA Pneumonia Detection Challenge. Kaggle. <https://www.kaggle.com/c/rsna-pneumonia-detection-challenge>
15. Jain, R.: Pneumonia detection in chest x-ray images using convolutional neural networks and transfer learning. *Measurement* **165**, 108046 (2020). <https://doi.org/10.1016/j.measurement.2020.108046>
16. Rubin, J., Sanghavi, D.: Large Scale Automated Reading of Frontal and Lateral Chest X—Rays Using Dual Convolutional Neural Networks (2018). arXiv preprint [arXiv:1804.07839](https://arxiv.org/abs/1804.07839)
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
18. Lakhani, P., Sundaram, B.: Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* **284**(2), 574–582 (2017)
19. Maji, P., Mullins, R.: On the reduction of Computational Complexity of Deep Convolutional Neural Networks. ICANN-2017 (2017)

# Automatic Fake News Detector in Social Media Using Machine Learning and Natural Language Processing Approaches



J. Srinivas, K. Venkata Subba Reddy, G. J. Sunny Deol,  
and P. VaraPrasada Rao

**Abstract** The definition of fake news is a cooked-up story with an objective to fool or to cheat people. The current research aims to detect fake news in social media like Twitter, Watsapp and Facebook by studying the responses of the proposed model on posts acquired from Reddit online news store. Automatic fake news detection is a complex activity as it involves the model to implement natural language processing concepts in-tandem with machine learning approaches. Two feature extraction algorithms, namely CountVectorizer (CV) and term frequency-inverse document frequency (TFIDF), were employed separately for extracting the most relevant features from the dataset. These features were fed to multinomial naive Bayes (MNB), random forest (RF), support vector classifier (SVC) and logistic regression (LR) classifiers for classifying fake news creating a total of eight classification models. A solitary CV-based model was considered as the baseline model for predicting fake news in r/theonion and r/nottheonion datasets. GridsearchCV was also implemented for finding the testing and training scores for the selected parameters. Out of these models, TFIDF with MNB achieved an accuracy of 79.05% and is considered as the best.

## 1 Introduction

According to a survey conducted by Gartner [1], by 2022 the volume of fake news circulated in the society will be more than genuine news. This can be attributed to the fact of the exponential growth of social media users. Fake news is generally defined as synthesised news containing misinformation, rumour and falsified facts

---

J. Srinivas (✉)  
SR University, Warangal, India

K. Venkata Subba Reddy · G. J. Sunny Deol  
Kallam Haranadha Reddy Institute of Technology, Guntur, India

P. VaraPrasada Rao  
Gokaraju Rangaraju Institute of Technology, Hyderabad, India

circulated over traditional media and even social media [2]. The general motive of a fake news propagator is to mislead readers, character assassinate an individual or to capitalise from sensationalism fake news which creates [3]. Attributes like spread rate, convenience and strong user base make social media the first choice of any fake news propagator. This causes a lot of chaos and distress among individuals as well as societies in a short span of time. Just detecting whether the news is fake or genuine will not be sufficient. To tackle this problem, early detection of fake news is necessary [4]. Existing methodologies are not so capable of detecting fake news and stop its rapid propagation in articles published in the internet, blogs, tweets, posts in social media apps like Watsapp and Facebook [5]. Click bait is also one form of fake news that lures the user by putting some attractive content and encouraging them to click on the news for gaining some offers or gifts [6].

A popular research proposed that in 2017, 67% of U.S. citizens above the age greater than 18 consumed news mainly from social media [7]. In comparison with genuine news, the fake news propagates relatively swifter, deeper into the society according to some researchers [8]. Therefore, it is necessary to detect and restrict the genesis and circulation of fake news through social media. Fake news detection is a challenging process to execute as it involves cross-checking the news item with credible third parties like newspapers, media houses and government agencies. Researchers can use the methodologies like artificial intelligence (AI) and natural language processing (NLP) for developing automatic models that can categorise news as fake or genuine.

This paper developed eight pipeline models by combing feature selection algorithms like CountVectoriser (CV) and term frequency-inverse document frequency vectoriser (TFIDFV) with machine learning algorithms like random forests (RF), multinomial naive Bayes (MNB), support vector classification (SVC) and logistic regression (LR). A baseline model by taking CV as the classifier was also considered. This paper attempts to differentiate between fake news and real news by developing a classification model trained on subreddit posts from r/theOnion and r/nottheonion datasets. The method of transforming words in documents into numbers so that the machine learning model interprets the words in the news articles is called vectorization. This a popular NLP technique used for extracting best features from the datasets.

The rest of the paper is organised as follows. Section 2 highlights the related work present in the literature. Section 3 gives an overview of the model architecture and briefly explains the components of the model. Section 4 will describe the experimental set-up. Section 5 will evaluate the developed model using performance metrics, and finally, Sect. 5 will list out the conclusions, limitations and future scope of the research demonstrated by this paper.

## 2 Related Work

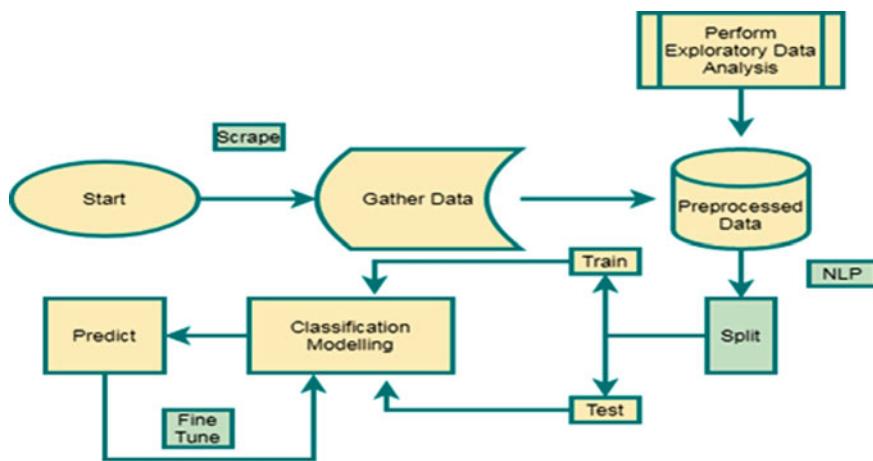
A variety of approaches have been developed in the past by many researchers to detect fake news automatically in social media [8]. Many studies conducted in the past stressed upon the fact that there is a need for automation of fake news classification rather than handcrafted techniques like crowdsourcing, hiring human employees [9, 10]. The authors of [11] proposed a novel process termed as influence mining using TextBlob NLP, SciPy tool kits. In this paper, a Bayesian ML model was developed, and it achieved an accuracy of 63.33% in predicting whether the news article is fake or genuine. A headline stance-based technique was proposed by a group of researchers for fake news classification using NLP techniques [12]. A substantial amount of research was done in the past on the topic of computer bots generating and propagating fake news on social media [13, 14]. The process of creating computer programs for generating and spreading rumours is very economical and safe for rumour mongers instead of hiring people to create fake news. Some researchers also studied the role of cyborg accounts in disseminating fake news in social media [15]. A new approach was developed that specifically investigated the credibility of the content published on Twitter social networking application using manual feature extraction techniques [16]. Regular expression-based rumour detection technique was developed for classifying information as fake or genuine on Twitter [17]. In another popular study, recurrent neural networks were employed to study and analyse the text-temporal relationship in Twitter posts for making accurate classifications about rumours and fake news [18]. Some researchers in the past have studied the comments related to fake news for analysing the structures of rumour cascades in Facebook posts [19]. The approach proposed in [20] presented an optimisation model for identification of spammers using matter in the posts and the network parameters. The next section of the paper presents the system design and explains the key components of the architecture.

## 3 System Design

In this section, system design is presented. This paper implemented the workflow system design technique for designing the system architecture. Data gathering, data pre-processing, developing the classification model and predicting are the various stages in the workflow pipeline.

### 3.1 Workflow Model

Figure 1 depicts the various stages of the workflow model that this paper followed for classifying fake news.



**Fig. 1** Workflow model for fake news classification

**Data Gathering** This paper used pushshift.io application programming interface wrapper for gathering data from Reddit posts. A total of 29,867 posts were scraped that included fifteen thousand posts from r/nottheOnion and 14,867 r/theonion datasets.

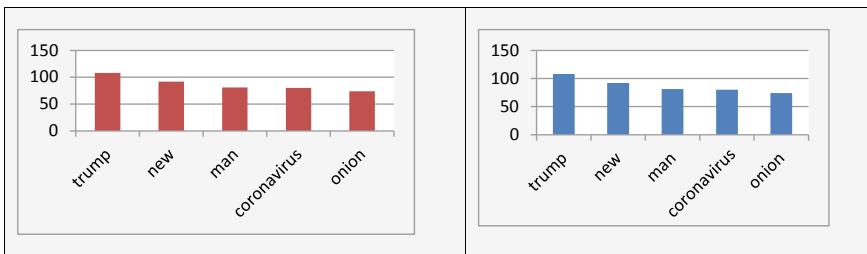
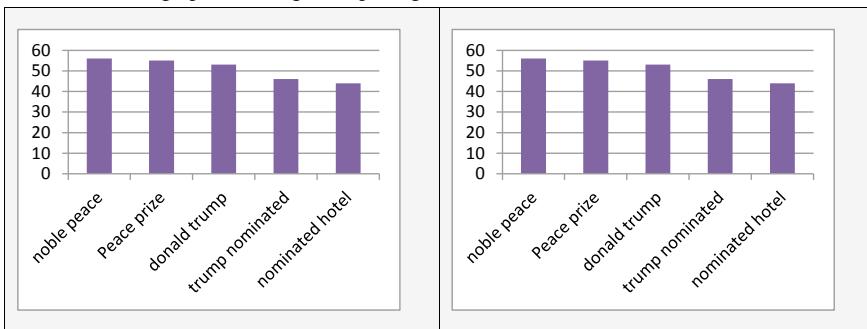
**Data Pre-processing** In this stage, duplicate records, punctuation marks and numbers were removed. Double spaces were converted to single space. Missing values were normalised, and the whole text was transformed into lower case. A total of 3592 duplicate records were found in the dataset. A clean r/nottheonion and r/theonion dataset was obtained as the output of this step.

**Exploratory Data Analysis (EDA)** EDA was performed on r/nottheonion and r/thonion datasets. This analysis helps in optimum design of the classification models. Generally, NLP techniques were implemented to perform the data analysis. For this purpose, the datasets will be concatenated into one dataset. The next step is to binarise the target, 1 is assigned to posts retrieved from r/theonion, and 0 is assigned to posts retrieved from r/nottheonion dataset. CV technique was used for this purpose. Analysis was performed to find out the top five unigrams, bigrams and stop words in the datasets taken. Table 1 represents the results of the EDA performed. The analysis suggests to remove the words ‘man’, ‘new’, ‘old’, ‘people’, ‘say’, ‘trump’, ‘woman’, ‘year’ from the dataset due to their high frequency in the posts. Tables 2 and 3 visualise the results obtained from the EDA performed.

**Train/Test Split** Usually, there are two methods for training and testing a model. In the first method, two separate datasets are used, one for training the model and the other for testing the model. In the second method, only one dataset is considered. This dataset is split into two segments, namely train data segment and test data segment. The train data segment is used for training the model, whereas the testing

**Table 1** Results of the exploratory data analysis performed using NLP

S. no.	n-gram class	Output
1.	Top 5 unigrams of r/onion	‘Trump’, ‘new’, ‘man’, ‘coronavirus’, ‘onion’
2.	Top 5 unigrams of r/nottheonion	‘Trump’, ‘man’, ‘peace’, ‘noble’, ‘says’
3.	Top 5 bigrams of r/onion	‘Onion presents’, ‘presents topical’, ‘topical episode’, ‘topical ep.’, ‘year old’
4.	Top 5 bigrams of r/nottheonion	‘Noble peace’, ‘peace prize’, ‘donald trump’, ‘trump nominated’, ‘nominated nobel’
5.	Common words in both datasets	‘Man’, ‘new’, ‘old’, ‘people’, ‘say’, ‘trump’, ‘woman’, ‘year’

**Table 2** Column graphs showing the top 5 unigrams and bigrams of r/theonion and r/nottheonion datasets**Table 3** Column graphs showing the top 5 bigrams of r/theonion and r/nottheonion datasets

data segment is employed to validate the model. This paper uses the train/test split technique. The dataset is split into two halves 70% for training and 30% for testing.

**Classification Model** In this phase of the workflow, the model is developed and trained with the dataset. Afterwards, it is tested with the test data segment. This paper created six different pipeline models for classifying fake news. Model 1, 2, 3 and 4 were created by a pipeline of CV and LR, CV and GNB, CV and SVC, CV and RF. Similarly, model 5, 6, 7 and 8 were created by using a pipeline of TFIID

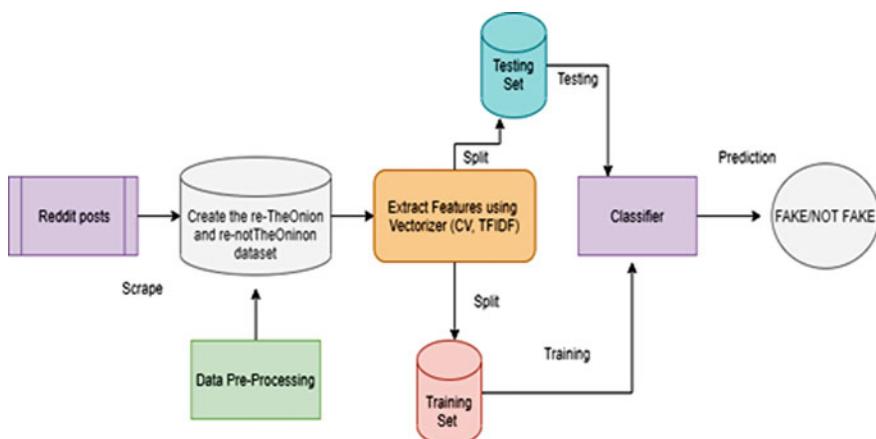
and LR, TFIFD and MNB, TFIFD and SVC, TFIFD and RF. These eight models were trained using the training segment of the datasets. A baseline CV model was also considered.

**Prediction** After vigorous training, the models were tested using the testing segment. The predictions were validated using performance metrics like accuracy, precision, recall, specificity and misclassification rate. This paper analysed the best model by comparing the performance metrics achieved by the models while classifying fake news over social media.

### 3.2 Process Flow

Figure 2 represents the system architecture of a fake news classifier with CV and TFIDF as feature extractors, respectively. The flow of process is listed in this section.

1. Scrape the data from Reddit posts into r/theonion and r/nottheonion datasets.
2. Apply data cleaning techniques over the gathered data repository.
3. Perform EDA using NLP techniques on the pre-processed data by combining the r/theonion and r/nottheonion posts into one dataset.
4. CountVectoriser feature extraction.
  - A. Apply CV technique and TFIDF on the dataset separately to extract the relevant features. Split the feature vector set into training and testing sets using the train/test split approach.
  - B. Now, train the classifiers LR, MNB, SVC and RF separately with the training set.



**Fig. 2** System architecture of the models using CV, TFIDF and classifiers

- C. After training is completed, test the models by feeding them with the testing set.
  - D. Find out the best model by analysing the performance of each model with the help of performance metrics achieved.
5. Identify the fake news detection model using the performance metrics.

### **3.3 Dataset**

Data was collected from Reddit. It is a social networking website where people publish information continuously. Reddit maintains two threads r/theonion and r/nottheonion. r/theonion contains satirical news articles, whereas r/nottheonion contains articles about true stories posted by people which look like posts from the r/theonion. This dataset contained the following attributes, namely title, subreddit, comments, author, # sub, score, domain and date. The title is used as the predictor, and the subreddit is the target.

### **3.4 Vectorisation**

The process of transforming a large corpus of text data into numbers is called vectorisation. This process helps in speeding up the training and testing procedures of the ML model. This paper employed two vectorisation techniques, namely CountVectorizer and term frequency-inverse document frequency for converting the words present in the text data into numbers. CV counts the number of times a word is appearing in a text document related to a text corpus. TFIDF is a statistical metric that reflects the relevance of a word to a document in a text corpus. Term frequency represents local relevance of a word by its appearance in a document. Inverse document frequency detects the signature words, whose frequency is relatively low across the documents.

### **3.5 Modelling the Classifiers**

In this paper, eight fake news detection models are created by combining two vectorisers and ML classifiers. Pipeline technique is used for modelling the fake news detectors. In pipeline technique, the intended methods are stacked together and executed one after the other. These methods are called as transformers. Table 4 lists out the various models that were created as part of research done by this paper. Here, the best model from both the groups is identified.

**Table 4** Proposed fake news detection models

Group	Name	Feature extractor	Classifier
A	Model 1	CountVectoriser	Logistic regression
	Model 2	CountVectoriser	Multinomial naive Bayes
	Model 3	CountVectoriser	Support vector classifier
	Model 4	CountVectoriser	Random forest
B	Model 5	Term frequency-inverse document frequency	Logistic regression
	Model 6	Term frequency-inverse document frequency	Multinomial naive Bayes
	Model 7	Term frequency-inverse document frequency	Support vector classifier
	Model 8	Term frequency-inverse document frequency	Random forest
C	Baseline	CountVectoriser	CountVectoriser

## 4 Experiments and Results Discussion

Initially, a baseline model is set up using CountVectoriser to find out the prediction score. The model achieved a prediction score of 53.57 and 46.42% on r/theonion and r/nottheonion datasets, respectively. These scores will be considered as the baseline for the rest of the models developed. Table 5 lists the results of GSCV method on the models created. After creating the pipeline and training them, GridSearchCV was employed to acquire the best parameter scores. The results clearly suggest that all the models are overfitting and clearly need fine-tuning of parameters. After fine-tuning the parameters of the model, test dataset was fed to the models. For each model, accuracy, precision, recall, specificity and misclassification rates were calculated based upon the confusion matrices generated for each models. Table 6 depicts the performance metrics achieved after feeding the test set. In terms of accuracy, Model 5 (TFIDF and LR) is the best, and Model 8 (TFIDF and RF) performs the worst.

**Table 5** Results of the GSCV on the models

Model	Best score	Training	Testing	Status
1	74.15	99.94	78.25	Overfit
2	75.74	99.52	79.20	Overfit
3	71.61	94.86	73.01	Overfit
4	71.45	99.52	73.33	Overfit
5	73.94	93.43	78.09	Overfit
6	71.02	85.48	75.07	Overfit
7	72.67	90.41	75.07	Overfit
8	71.45	99.52	73.33	Overfit

**Table 6** Performance metrics of the eight models

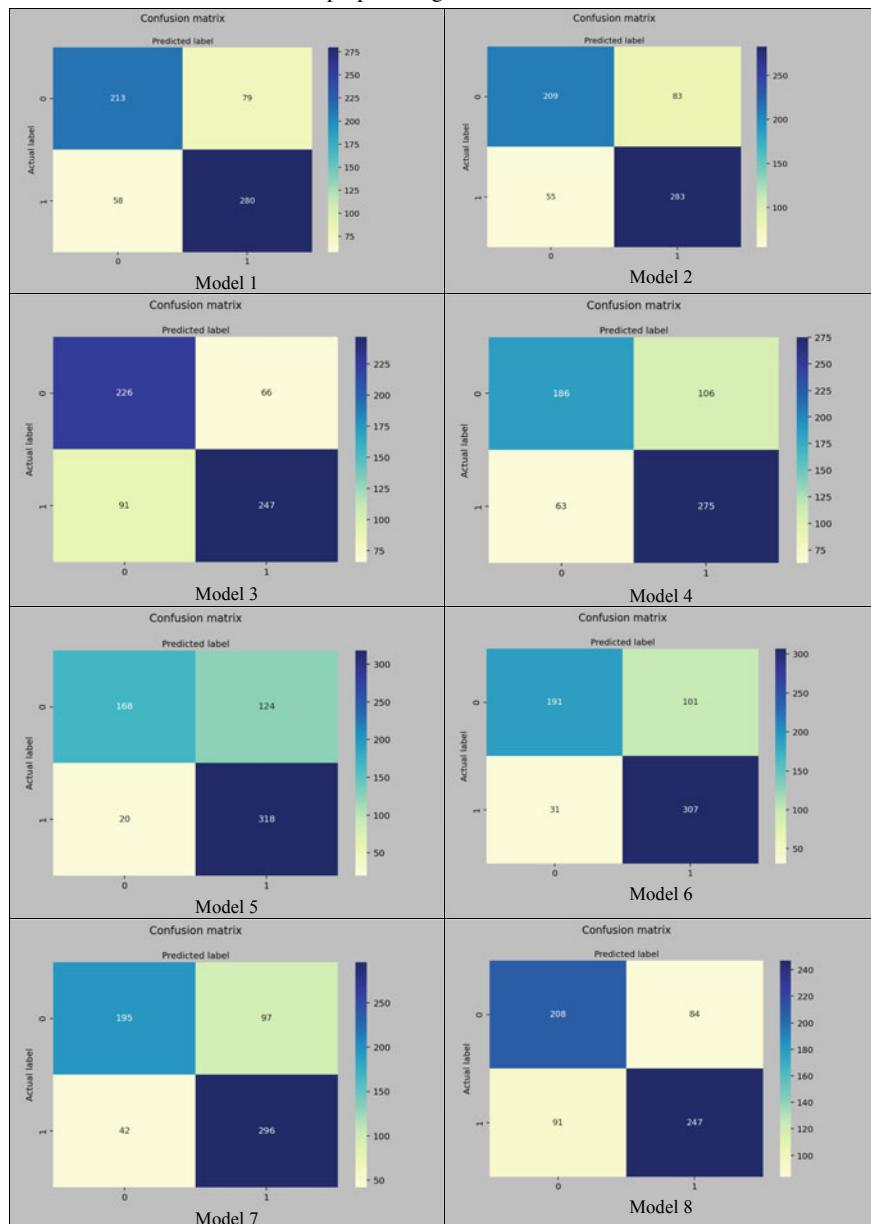
	Accuracy	Precision	Recall	Specificity	Misclassification rate
Model 1	78.25	77.99	82.84	72.95	24.33
Model 2	78.10	77.32	83.73	71.58	24.82
Model 3	75.08	78.91	73.08	77.40	25.78
Model 4	73.17	72.18	81.36	63.70	31.24
Model 5	79.05	75.25	90.83	65.41	25.68
Model 6	77.14	71.95	94.08	57.53	30.00
Model 7	77.94	75.32	87.57	66.78	26.28
Model 8	72.22	74.62	73.08	71.23	29.61

Similarly, Model 1 (CV and LR) achieves 77.99% precision rate and stands first while Model 6 (TFIFD and MNB) achieves 71.95% and stands last. On the other hand, when recall metric is considered, Model 6 (TFIFD and MNB) achieves the highest value of 94.08%, and Model 3 (CV and SVC), 8 (TFIDF and RF) achieve the lowest 73.08%. Model 1 achieves 72.95% specificity and stands first, whereas Model 6 achieves a low specificity of 63.70%. With a value of 24.33%, Model 1 has the best misclassification rate, and with a 31.24%, Model 4 has the worst misclassification rate. According to the analysis, Model 5 (TFIDF and LR) is the best model because of its accuracy

and satisfactory performance metrics. All the proposed models performed better than the baseline model. Table 7 depicts the confusion matrices achieved from the experiments conducted by this research.

## 5 Conclusions and Future Scope

This paper clearly explained the problems caused by fake news and ascertained the need of early detection of fake news in social media. For this purpose eight ML models were proposed. Various NLP techniques were implemented on the real-time dataset collected from the Reddit website. Eight confusion matrices were generated for calculating the accuracy, precision, recall, specificity and misclassification rate of the proposed models. Based upon the analysis, model created by pipeling TFIDF and MNB was the best with an accuracy of 79.05%. In future, the same work can be implemented on Facebook, Twitter and Watsapp datasets. In future, novel models that employ neural networks and deep learning techniques can be designed and applied for early fake news prediction in social media websites.

**Table 7** Confusion matrices of the proposed eight models

## References

1. Titcomb, J., Carson, J.: What exactly is it—and how can you spot it? *Fake News*. [www.telegraph.co.uk](http://www.telegraph.co.uk)
2. Thota, A., Tilak, P., Ahluwalia, S., Lohia, N. *Fake news detection: a deep learning approach*. *SMU Data Science Review* **1**(3), Article 10 (2018)
3. Liu, Y.: *Early Detection of Fake News on Social Media*. PhD Thesis, NJIT, USA (2019)
4. Allcott, H., Gentzkow, M.: *Social Media and Fake News in the 2016 Election*. Technical report, National Bureau of Economic Research (2017)
5. Bourgonje, P., Schneider, J.M., Georg Rehm from clickbait to fake news detection: an approach based on detecting the stance of headlines to articles. In: *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pp. 84–89 (2017)
6. <http://www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017-access-ed>. 10th September 2020
7. Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. *Science* **359**(6380), 1146–1151 (2018)
8. Conroy, N.K., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. *Association for Information Science and Technology* **52**(1), 1–4 (2015)
9. Chen, Y., Conroy, N.K., Rubin, V.L.: News in an online world: the need for an “automatic crap detector”. In: *Proceedings of ASIST 2015* (2015)
10. Kim, J., Tabibian, B., Oh, A., Schölkopf, B., Gomez-Rodriguez, M.: Leveraging the crowd to detect and reduce the spread of fake news and misinformation. In: *11th ACM International Conference on Web Search and Data Mining*, pp. 324–332 (2018)
11. Traylor, T., Straub, J., Gurmeet, Snell, N.: Snell classifying fake news articles using natural language processing to identify in-article attribution as a supervised learning estimator. In: *IEEE 13th International Conference on Semantic Computing* (2019). <https://doi.org/10.1109/icsc.2019.00086>
12. Bilal, G., Rosso, P., Rangel, F.: Stance detection in fake news a combined feature representation. In: *1st Workshop on Fact Extraction and Verification (FEVER)*, pp. 66–71 (2018)
13. Bessi, A., Ferrara, E.: Social bots distort the 2016 US Presidential election online discussion. *First Monday* **21**(11-7) (2016)
14. Ferrara, E., Varol, O., Davis, C., Menczer, F., Flammini, A.: The rise of social bots. *Commun. ACM* **59**(7), 96–104 (2016)
15. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of twitter accounts: are you a human, bot, or cyborg? *IEEE Trans. Dependable Secure Comput.* **9**(6), 811–824 (2012)
16. Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: *20th International Conference on World Wide Web*, pp. 675–684 (2011)
17. Zhao, Z., Resnick, P., Mei, Q.: Enquiring minds: early detection of rumours in social media from enquiry posts. In: *24th International Conference on World Wide Web*, pp. 1395–1405 (2015)
18. Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B.J., Wong, K.F., Cha, M.: Detecting rumors from microblogs with recurrent neural networks. 3818 (2016)
19. Cheng, J., Adamic, L.A., Kleinberg, J.M., Leskovec, J.: Do cascades recur? In: *25th International Conference on World Wide Web*, pp. 671–681 (2016, April)
20. Hu, X., Tang, J., Liu, H.: Online social spammer detection. In: *28th AAAI Conference on Artificial Intelligence*

# A Novel Method for Optimizing Data Consumption by Enabling a Custom Plug-In



Vijay A. Kanade

**Abstract** In an era of technology, digital marketing has evolved many folds. Digital marketing involves advertising delivered via digital channels such as search engines, websites, social media, email, and mobile apps. Digital marketing is a method opted by companies that endorse goods, services, and brands by using online media channels. In addition to digital marketing, personalization is another area of e-commerce that has taken the world of advertisement with storm. However, both these marketing gimmicks have impacted the bandwidth consumption of the Internet users significantly. Internet data usage has soared high in recent times as the users are blinded by digital marketing methods. The research paper proposes a novel custom plug-in, to be used on portable computing devices for keeping Internet data consumption in check, thereby eliminating the economic overhead for the end users.

## 1 Introduction

Personalization has been widely explored in outbound marketing, which is applied across all dynamic websites. The very objective of personalization is to deliver target advertisements that are customized for specific customers. Such personalization allows websites to evaluate the reason for the customer's visit to their webpage and in future target them in order to achieve high conversion goals, wherein the customer ends up buying the product facilitated via targeted advertisement.

The search window on the website alone provides a lot of customer centric information, which is used to deliver dynamic content on the websites. Say, an individual searches for "swimming pool nearby me." The individual will be shown a page with the header "swimming pool" and matching pictures plus texts. Now, if the customer looks for a "swimming pool along with the playground," he will be shown the same page, but the content will focus on the nearby playground. The images provided in the search results are linked using Google's AI technology, which has a good amount of accuracy, about 93%. This personalization of websites can further be linked to

---

V. A. Kanade (✉)  
Intellectual Property Research, Pune, India

customer's click behavior information or online activity characteristics along with customer profile information.

Once the customer's necessary data is collected, the advertisers need to make strategic decisions on what kind of product information is to be shown to the customer. Further, such targeted advertisements also consider groups of individuals, wherein demographic characteristics of diverse individuals are collected, and customers are delivered with profile-specific advertisement [1] (Fig. 1).

However, with such personalization comes the inevitable demand for bandwidth at the customer end—it may be a PC, smartphone, laptop, tablet, or any other device that the user may be using. When such personalized content is delivered to the customers, the dynamic nature of the content causes a heavy toll on the data consumption for the users. Say, the user searches for a YouTube video on his PC—YouTube not only provides the search video in its results but it also delivers the user-specific customized advertisements, along with related social media posts, trending news/blog articles, user's location information, and plenty more of dynamic content that gets updated in real time. These dynamic websites consume a lot of Internet data of the user, and



**Fig. 1** Dynamic website [2]

the user is forced to endure situations, wherein data is exhausted in minimum time. Therefore, there seems to be a long pressing need to contain the dynamic content of these dynamic websites by providing a customization option at the customer end, wherein the user is able to manually control and monitor its data usage on the Web content delivered via diverse vendors or e-commerce giants.

## 2 Dynamic Content

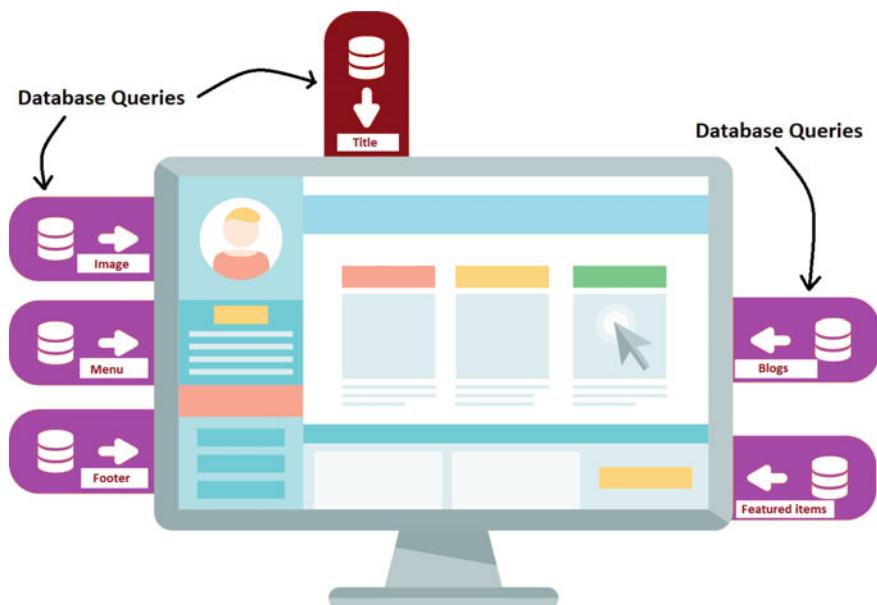
While it is true that smart content delights its customers, it plays a crucial role in burning out the data consumption for the customers. Let us understand what dynamic content really means.

Dynamic content creates a customized experience for the visitor at that very time instance. Common examples may include Amazon's recommendation engine, wherein the customers are recommended with additional products when the customer visits their website, based on the profile information or customer online behavior. Some other forms include personalization fields in emails that deliver customer-specific content. Consider, for example, a customer visits an e-commerce site for the first time. During this first visit, the customer browses around and clicks the "like" button on a few products and maybe purchases some products. Say, the same customer comes back and revisits the same e-commerce website a couple of weeks later, and the home site has now changed to say "Welcome back!" and further, the website also recommends items that the customer may like based on his history (Fig. 2).

The same dynamic content functions on cross-platforms. Consider a scenario where one morning a customer browses a cosmetics site for the first time and after a while the customer closes the website and opens up Facebook. The customer observes that all the ads surfacing on Facebook are from the cosmetic site that he just visited. Facebook chooses ads to show users based on their browsing history

**Fig. 2** Dynamic content [3]





**Fig. 3** Database-driven website

and interests. This implies that personalization happens over cross-platforms, and customer's activity is tracked all the while as he stays online.

### 3 Mechanism of Dynamic Content

For personalization to work, important factor is data collection, wherein customer's profile information (such as an email address, first name), shopping history, usage behavior data are collected by the websites. This data is organized and stored in database-driven websites with associated values—think of this as a filing system. The data collection works by scripts in a webpage's HTML that changes to make the page relevant to the user. Now, when the customer visits any website, that website then assesses the need of the page and shows the viewer content that is relevant to the user based on the collected data (Fig. 3).

### 4 Statistics for Internet Data Usage

As per the recent statistics, application vs approximate Internet data usage is disclosed in Table 1 [4]:

**Table 1** Application versus Internet data usage

S. no.	Application	$\approx$ Data usage
1	Web browsing	$\approx 60 \text{ MB/h}$
2	YouTube streaming (SD video quality) (a) YouTube content (high quality)	$\approx 6 \text{ MB/min or } 360 \text{ MB/h (480p)}$ (a) $\approx 150 \text{ MB/min or } 900 \text{ MB/h (1080p)}$
3	Netflix streaming (a) Low quality (b) Medium quality (c) High quality	$\approx 250 \text{ MB-1 GB/h}$ (a) $\approx 5 \text{ MB/min or } 300 \text{ MB/h}$ (b) $\approx 9 \text{ MB/min or } 450 \text{ MB/h}$ (c) $\approx 17 \text{ MB/min or } 1 \text{ GB/h}$
4	Social media (scrolling) (a) Facebook (video content)	$\approx 2 \text{ MB/min}$ (a) $\approx 120 \text{ MB/h or } 160 \text{ MB/h}$
5	Music streaming	$\approx 150 \text{ MB/h}$
6	Online gaming	$\approx 3 \text{ MB/h to a } 1 \text{ GB/h}$
7	FaceTime (Apple)	$\approx 90 \text{ MB per h}$

Now, the above data usage is specific to the application that is used by the user, thereby ruling out the data consumed by the dynamic content rendered via personalized ads, feeds, trending posts, etc. Now, if we imagine a situation where multiple data streams from varied sources run parallelly along with the intended application in the name of dynamic content, then it is evident that dynamic content is the real reason behind excessive data usage. In such cases, although the user does not perform any data-centric activity over the Internet but still the data consumption exceeds the intended data usage of the user—which leads to data overhead.

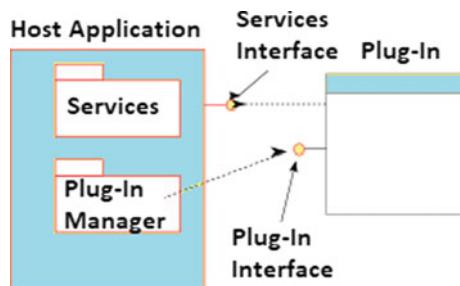
## 5 Custom Plug-In

The excessive data usage can be controlled by deploying a custom plug-in that keeps a check on data consumption. The disclosed custom plug-in acts as a piece of software that provides an add-on to a Web browser and gives the browser additional functionality, wherein the user or visitor is able to customize the window size on the website that he wishes to view. This plug-in allows the Web browser to display only the content within the customized window opted by the user (Fig. 4).

Consider an example where a user is viewing a YouTube video on a PC. As the user searches for the desired video on YouTube website/channel, the results displayed to the user include the list of videos that are related to the searched video. Now, the user can select any one of the videos from the list. As the user plays the video content, the background dynamic content such as targeted ads, e-commerce products, and social media posts comes into play and gets displayed parallelly on the website along with the video frame.

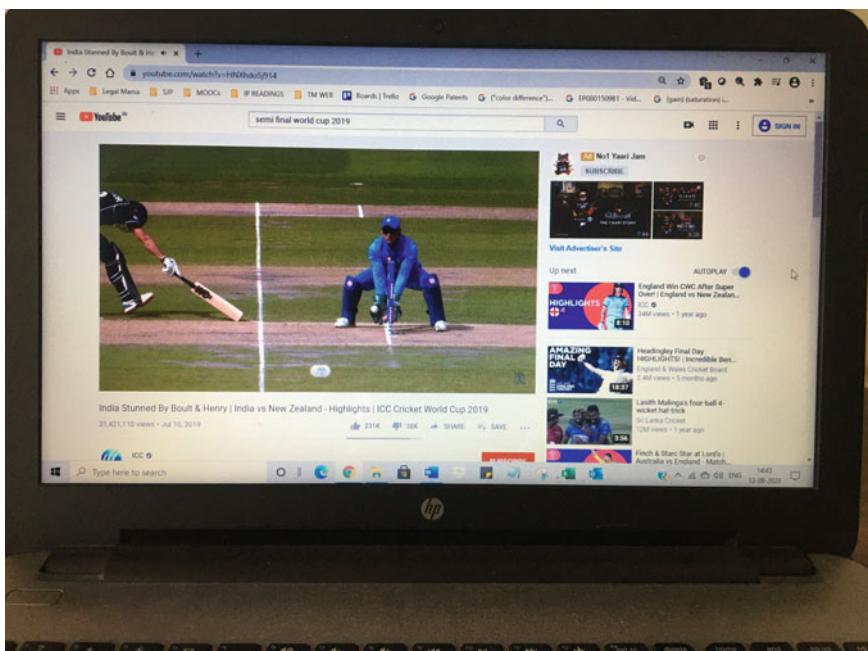
Now, the custom plug-in disclosed in the research allows the user to select a window frame on the website that he wishes to view. This window is manually

**Fig. 4** Generalized plug-in  
[5]

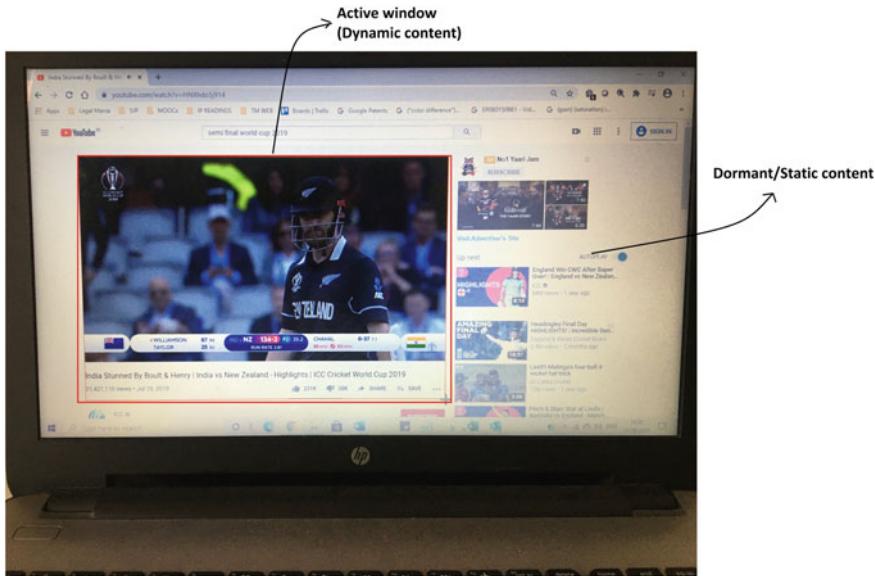


controllable, and the user can resize the window as per the need. The content played within the window frame acts as dynamic content, thereby allowing the user to view it in its original form, as streamed by the service provider. In addition, the content external to the window frame becomes dormant or static, thereby consuming zero data. Therefore, the custom plug-in puts a check on the data that is hogged by the background content such as ads, social media posts, and blogs that run simultaneously on the website along with the user intended content (Fig. 5).

The below snapshots disclose the custom plug-in framework for a YouTube video. Snapshot-I discloses a YouTube video running on a PC of the user.



**Fig. 5** Snapshot-I



**Fig. 6** Snapshot-II

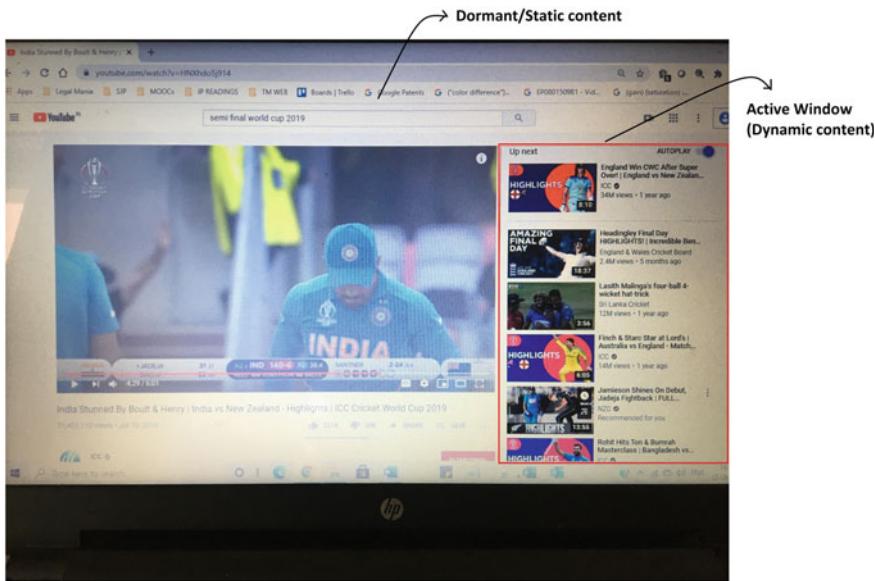
**Snapshot-II:** With the custom plug-in, the user can customize the window frame (i.e., active window) on the website as shown in the below image. The plug-in disables the content running external to the window. Thus, the content outside the active window becomes dormant or static, thereby saving the data usage.

**Snapshot-III:** Similar to Snapshot-II, Snapshot-III discloses the customized window active area along with dormant or static window where the data consumption is stalled (Figs. 6 and 7).

## 6 Advantages

The custom plug-in disclosed in the research proposal has following advantages over the conventionally running dynamic website model:

1. Reduction in Internet data consumption occurring due to personalization.
2. Custom plug-in is manually controllable, thereby allowing the user to resize the window frame as per user's choice or need.
3. Effective economic solution for the end customers.
4. Plug-in is compatible with any portable computing device such as PC, smartphone, and palmtop.



**Fig. 7** Snapshot-III

## 7 Conclusion

Digital marketing and personalization have played a significant role in the high volume of data traffic over the Internet. Data consumption is reaching new highs every year, with more smart devices cropping up every day and more users adopting such devices. The research paper proposes a unique solution for controlling the high data consumption, by employing custom plug-ins within user computing devices that essentially manage the flow of data traffic by providing manual controls in the hands of the end user.

## 8 Future Work

The custom plug-in may be enabled with an indicator that displays the data consumption within the customized window frame chosen by the user. This will allow the user to keep track of the data usage in real time.

**Acknowledgements** I would like to extend my sincere gratitude to Dr. A. S. Kanade for his relentless support during my research work.

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Declaration** We have taken permission from competent authorities to use the images/data as given in the paper. In case of any dispute in the future, we shall be wholly responsible.

## References

1. Manola, S.: What is Personalization in Digital Marketing. 27 May 2019
2. Approx World, Bansdroni, Kolkata, West Bengal, Dynamic Website Design
3. BTI Admin: Search Engine Optimization, Tips and Tricks. Dynamic Content: Making Big Data Work for Your Website, 22 February 2017
4. Seph, amaysim's tech geek: Internet Data Usage Guide: What Uses the Most Data?" 8 May 2020
5. Plug-in (computing), Wikipedia

# An Effective Mechanism for the Secure Transmission of Medical Images Using Compression and Public Key Encryption Mechanism



T. K. Ratheesh and Varghese Paul

**Abstract** In medical image applications, the diagnostic observations of a patient are transmitted to a remote location through internet. Since the information being transmitted is really sensitive as far as the diagnosis is concerned, the security of the image is a major concern. The size of the image being transmitted is another major concern in the scenario. We propose a methodology to mitigate problems relating to both security and size of image. The proposed mechanism first compresses the image to reduce its size using a lossless compression mechanism. The compressed image is then encrypted using the well-known public key cryptography algorithm ECC to ensure its security. The encrypted cipher image is transmitted to the remote center where the image undergoes decryption and decompression mechanisms to generate the actual image. The performance analysis shows that the system works well and is effective in ensuring the security of the image.

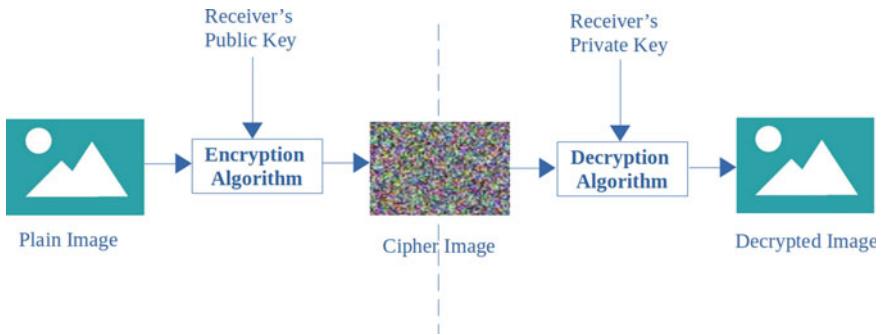
## 1 Introduction

Internet has become an inevitable part of human life these days. As the communication technology has improved a lot, the use of interactive multimedia in communications has increased to a great extent. Images occupy a greater fraction of multimedia communication, for example, military applications, medical application, security agencies, etc. These images may carry highly confidential and important information which is not supposed to get compromised or lost during its transit.

In medical image applications, a patient's medical diagnostic observations are transmitted from a medical institutions to a medical practitioner located in a remote place through Internet. Two major constraints that exist for transmitting medical images through Internet are its scale and privacy. Medical images are usually large sized files which require larger space and time for transmission. One solution to resolve this issue is the use of compression technique with which the size of the

---

T. K. Ratheesh (✉) · V. Paul  
Division of Information Technology, CUSAT, SoE, Kochi, India



**Fig. 1** Public key encryption

image can be reduced. There are mainly two approaches for image compression—lossless and lossy compression. As the medical images contain sensitive information the approach, we can adopt lossless compression where the size of the file is reduced without compromising its quality. The second constraint is the privacy of the image being transmitted. As the medical images contain confidential medical information, it needs protection against various security attacks. The confidentiality, integrity and authenticity of the image data must be protected during its transit. The popular technique for securing images in cryptography which involves an operation called encryption in which the image is transformed into a cipher image at the sender side. A reverse operation called decryption is done at the receiver side to retrieve the original image.

Public key encryption mechanism is one of the cryptographic mechanism for securing images which involves the use of two distinct keys for performing encryption and decryption operations. The process is shown in Fig. 1. The plain image is encrypted by the sender with the help of the public key of the receiver, forms the cipher image and is transmitted to the receiver. The receiver uses his own private key to decrypt the cipher image to generate the original plain image.

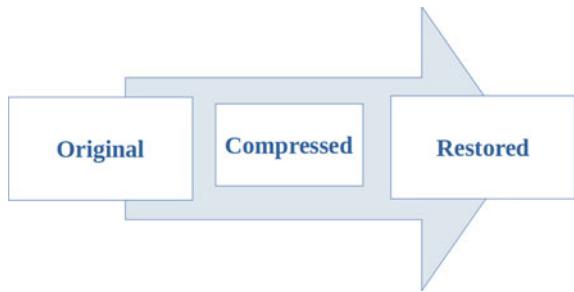
The popular public key algorithms for ensuring the confidentiality of data being transmitted are AES and ECC.

Lossless compression is a compression technique which reduces the size of a file with no loss of quality, i.e., lossless compression packs the data into smaller file without losing any data during the process of compression. The process is depicted in Fig. 2. It is seen from the figure that the original file is compressed to a smaller size, but upon decompression process, the restored image will be more or less the same as that of the original image.

In this paper, we propose a mechanism that compresses the image first to reduce its size and then encrypts the compressed image to protect its confidentiality. Operations of decryption and decompression are performed at the receiver side in order to regenerate the original image.

The rest of the paper is organized in six sections and is as follows. Section 2 describes the available systems in the literature regarding image security and

**Fig. 2** Lossless compression: original, compressed and restored



compression. Section 3 gives an overview of the popular public encryption algorithm ECC. Section 4 describes the proposed system. The performance analysis of the proposed system is described in Sect. 5. The paper concludes in Sect. 6.

## 2 Related Works

Several image encryption and compression techniques are presented in the literature. These techniques depict the need as well as the method of encryption and compression of image data. This section describes a few of them. Even though these methods are not specifically designed for medical images, it shows the scope and need of encryption and compression on images.

Narendra K. Pareek, Vinod Patidar and Krishan K. Sud have proposed a gray image encryption scheme [1] in which the image is partitioned into key-dependent dynamic blocks and then undergoes a key-dependent diffusion and substitution processes of 16 rounds. A mixing process is performed before the partitioning. The system is claimed to have high encryption rate.

The method proposed by Shihua Zhou, Bin Wang, Xuedong Zheng and Changjun Zhou used the concept of DNA computing [2]. They introduced two concepts named one-dimensional DNA cellular automata and T-DNA cellular automata. Reversible T-DNA cellular automata are defined and used for the encryption process. The authors claimed that the proposed method is capable of resisting several attacks like brute-force attacks, statistical attacks and differential attacks.

An elliptic curve and chaotic system-based image encryption scheme is proposed by Yuling Luo, Xee Ouyang, Junxiu Liu and Lvchen Cao [3]. The proposed system generates a scrambled image using the chaotic system concept. Elliptic curve and El Gamal algorithms are then used to improve the security of the image. DNA sequencing is also employed to generate the final cipher image. The authors claimed that the proposed method has high security and good efficiency when compared to the benchmarked system.

Alireza Arab, Mohammad Javad Rostami and Behnam Ghavami proposed an image encryption mechanism using chaotic system and AES encryption system [4]. Arnold chaos sequence is used for generating the encryption key. The encryption

is performed using the AES algorithm using the key generated with the chaotic sequence. The system is proved to be resistant to differential attacks, brute-force attack and statistical attacks.

Yinghua Li, He Yu, Bin Song and Jinjun Chen proposed an encryption mechanism for images for a cloud computing scenario [5]. The method is based on single-round dictionary and chaotic sequences. The system uses two chaotic functions. A unique dictionary is maintained for each image with which the images are encrypted adaptively and reconstructed.

Shaou-Gang Miaou, Fu-Sheng Ke and Shu-Ching Chen proposed a method for image compression in which JPEG-LS and interframe coding with motion vectors are combined [6]. In this method, the interframe coding is activated when and only when there found high interframe correlation. The proposed method achieves a compression gain of 13.3% and coding gain of 77.5% for MRI image sequence.

Ali Al-Fayadh, Abir Jaafar Hussain, Paulo Lisboa and Dhiya Al-Jumeily proposed an adaptive lossy image compression technique for the compression of medical drain images [7]. They used classified vector quantizer and singular value decomposition for compression of the image. The authors claimed that their system could regenerate the images with high PSNR compared to the bench marked technique.

Tushar Shinde have proposed an approach for image compression [8]. The approach uses an efficient clustering, fast direction-oriented motion estimation algorithm along with a minimum predictive cost image reordering scheme that gives a better level of compression.

Tony Leung, Michael W. Marcellin, Ali Bilgin have proposed an approach for visually lossless image compression mechanism [9] in which a visually lossless representation after adjustments to the window level is done and then compressed.

The systems proposed in [1–5] are designed for securing images using encryption methods where systems in [6–9] are designed for compressing images. All the papers mentioned above are pointing out the need for encrypting and compressing the images for transmission over an insecure and bandwidth-sensitive channel. The literature shows the need for a single system for ensuring security as well as for reducing size for transmission, and the proposed architecture concentrates on both of these aspects.

### 3 Elliptic Curve Cryptography

An elliptic curve  $E_P$  is a set of points that satisfies a cubic function in two variables over a finite field and are represented by the Weierstrass equation,

$$y^2 = x^3 + ax + b \pmod{p} \quad (1)$$

where  $a$  and  $b$  are two constants which satisfy  $4a^3 + 27b^2 \neq 0$ ,  $p$  is a prime [10].

Addition of two elliptic curve points  $P(x_1, y_1)$  and  $Q(x_2, y_2)$  resulting in  $R(x_3, y_3)$  is given as

$$x_3 = t^2 - x_1 - x_2 \quad (2)$$

$$y_3 = t(x_1 - x_3) - y_1 \quad (3)$$

where  $t = [(y_2 - y_1) / (x_2 - x_1)]$  if  $P \neq Q$  and  $t = [(3x_1^2 + a)/2y_1]$  if  $P = Q$ .

An interesting fact of elliptic curve is that the resultant point of the point addition will also be on the same elliptic curve.

By combining the point addition law on elliptic curve with the discrete logarithm operations, a public key cryptosystem could be established. Generation of keys for the communicating parties, the encryption process at the sender and the decryption process at the receiver can be defined because of the irreversible property of ECDLP.

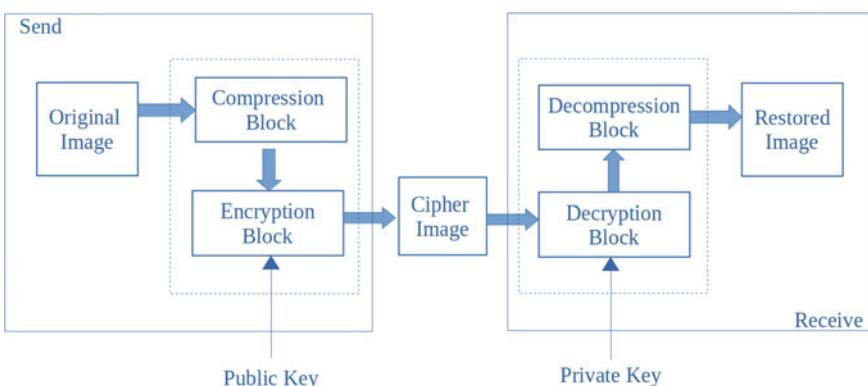
## 4 Proposed System

The proposed system is an effective mechanism to transfer images with solutions to the constants of scale and privacy. Figure 3 shows the overall architecture of the system.

The architecture is divided into two phases send phase which describes the processes a sender does before sending the plain image file and receive phase that describes the processes a receiver does when receiving the cipher image file.

### 4.1 Send Phase

The plain medical image is given to the compression block where the image undergoes compression to reduce its size. Since the medical images contain information



**Fig. 3** Proposed system architecture

which are very sensitive for medical diagnosis the compression method must ensure that no information is lost during the compression. So a lossless compression mechanism is used for reducing the size of the image. The compressed image is then given to encryption block. The encryption block encrypts the compressed image and forms the cipher image for transmission. The popular public key encryption method ECC is used for the purpose in which the public key of the receiver is used to generate the cipher image. The cipher image is then transmitted to the receiver.

## 4.2 *Receive Phase*

The receive phase in contrast to the send phase performs a series of operations to retrieve the plain image from the received cipher image. The cipher image undergoes a decryption process in the decryption block where the image is transformed to a pseudo plain image. The pseudo plain image is then processed by the decompression block where the pseudo image undergoes a decompression operation and generates the plain image as send by the sender.

## 5 Results and Discussion

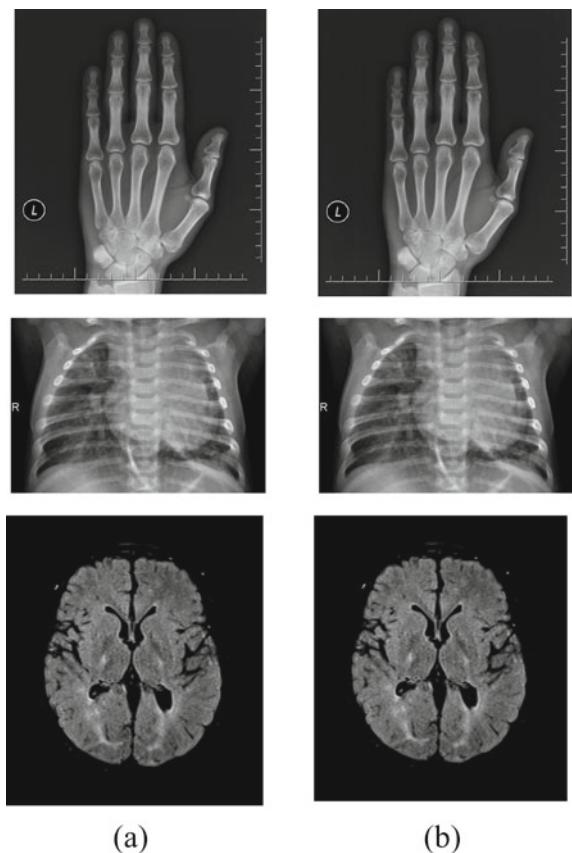
The simulation and result analysis is presented in this section. The proposed architecture is implemented in Python. The inputs and outputs of the system are shown in Fig. 4. The original image at sender side is shown in (a). The image after decryption and decompression at receiver is shown in (b). It is clear from the pictures that the receiver could regenerate the original image without errors.

### 5.1 *Performance Analysis*

Histogram analysis is used for illustrating the substitution and diffusion properties of the proposed method. The histograms of several test plain images and their corresponding regenerated images are shown in Fig. 5. The section (a) of the figure shows the histogram of the original plain image and (b) shows the histograms of the corresponding image regenerated at the receiver side.

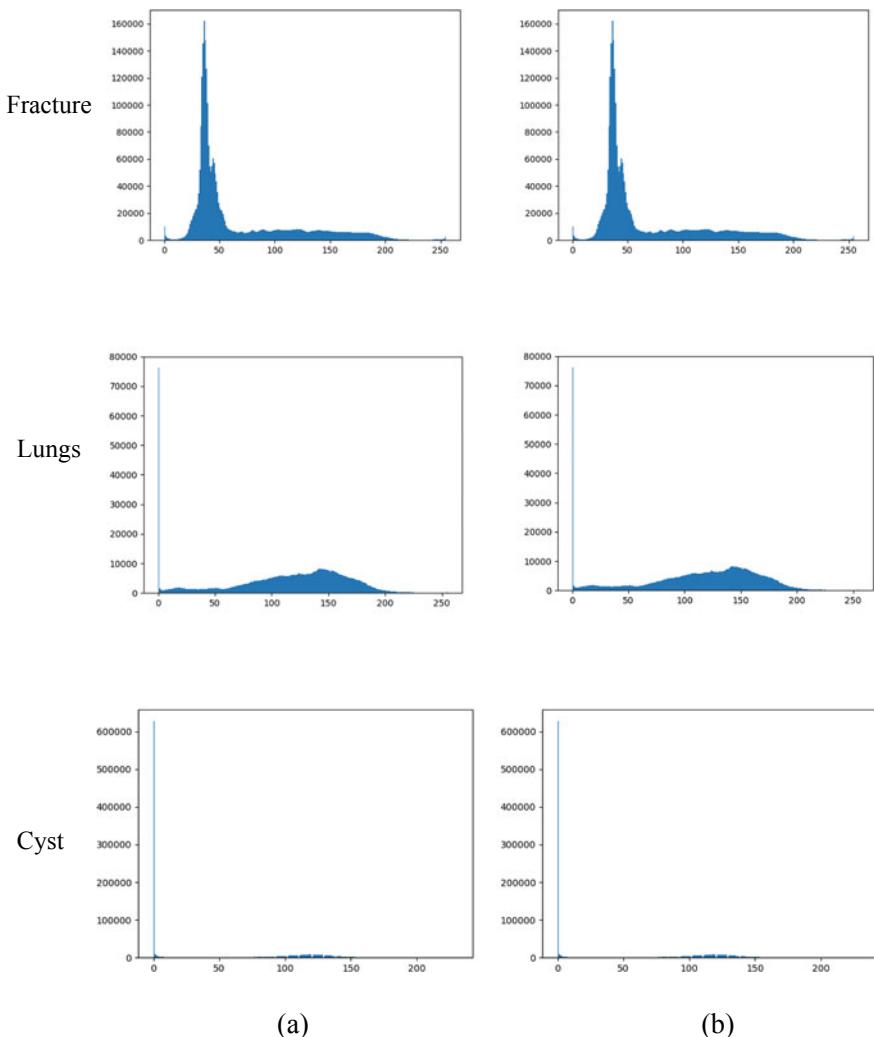
It can be concluded from the histogram analysis that the regenerated images have the same pixel distribution as that of the original image. The PSNR and MSE values calculated between the original and regenerated images show that both images are same and no noise has been introduced or no bits have missed after the compression and encryption process. Since the image is encrypted, the image does not provide any clue to attacker for performing any statistical attack on the proposed scheme. This makes the statistical attacks difficult in the images.

**Fig. 4** Plain and regenerated images



## 6 Conclusion

The transmission of content-sensitive images like medical images through the Internet is popular these days. As far as the medical images are concerned changes in content, it will lead to wrong diagnosis of diseases. So a system was proposed for the secure transmission of medical images using public key encryption mechanism. The system also uses a lossless compression method in order to reduce the size of the image before the transmission. The algorithm ECC is used for the encryption, and 7zr is used for compression. The performance analysis shows that the system works well and is effective in ensuring the security of the image. The histogram analysis, PSNR and MSR calculations, shows that the system is able to regenerate the data without any missing. Since the image is encrypted, it is secure and free from statistical attacks.



**Fig. 5** Histograms of original and regenerated images

## References

1. Pareek, N.K., Patidar, V., Krishan, K.: Diffusion–substitution based gray image encryption scheme. In: Digital Signal Processing, Elsevier, pp 1–8 (2013)
2. Zhou, S., Wang, B., Zheng, X., Zhou, C.: An image encryption scheme based on DNA computing and cellular automata. Hindawi Publishing Corporation, Discrete Dynamics in Nature and Society, vol. 2016 (2016)
3. Luo, Y., Xee, O., Liu, J., Cao, L.: An image encryption method based on elliptic curve ElGamal encryption and chaotic systems. IEEE Access **4** (2016)

4. Arab, A., Rostami, M.J., Ghavami, B.: An image encryption method based on chaos system and AES algorithm. Springer J. Supercomput. **75**, 6663–6682 (May 2019)
5. Li, Y., Yu, H., Song, B., Chen, J.: Image Encryption Based on a Single-Round Dictionary and Chaotic Sequences in Cloud Computing. Wiley, Hoboken (2019)
6. Miaou, S.-G., Ke, F.-S., Chen, S.-C.: A lossless compression method for medical image sequences using JPEG-LS and interframe coding. IEEE Trans. Inform. Technol. Biomed. **13**(5), 818–821 (2009)
7. Al-Fayadh, A., Hussain, A.J., Lisboa, P., Al-Jumeily, D.: An Adaptive hybrid Image Compression method and its application to Medical images. In: 2008 IEEE, pp. 237–240 (2008)
8. Shinde, T.: Efficient image set compression. In: 2019 IEEE, pp. 3016–3017 (2019)
9. Leung, T., Marcellin, M.W., Bilgin, A.: Visually lossless compression of windowed images. In: IEEE Data Compression Conference, IEEE Computer Society, p. 504 (2013)
10. Vasundhara, S., Durgaprasad, D.K.V.: Elliptic curve cryptosystems. Math. Comput. **48**(177), 203–209 (1987)

# A Systematic Survey on Radar Target Detection Techniques in Sea Clutter Background



R. Navya and R. Devaraju

**Abstract** Sea clutter is an unwanted return or echo signal which exhibits non-linearity and disarray and is random in nature. The reliable way of detecting the moving targets in the presence of clutter background has always been a setback in radar signal processing. The unwanted sea clutter greatly influences the target detection, which biases its characteristics and increases the difficulty of radar detection. With the development of advanced radar processing methods, the sea clutter can be suppressed considerably and targets can be detected reliably. In this paper, we report a review on current development in clutter suppression, clutter modeling based on the statistical modeling of sea clutter, which covers amplitude features of sea clutter, i.e., Rayleigh distribution, Log-Normal distribution, Weibull distribution, and K-distribution. The current developments in radar target identification methods are also conferred, within the framework of these clutter models. This review article provides a systematic comparison of various methods to suppress the clutter and to detect the target in a sea clutter environment.

## 1 Introduction

The target detection in marine environment has been one of the major challenges in radar technology due to presence of sea clutter. Research on sea clutter has received more attention in the study of ocean aspects. The characteristics of sea clutter and the target can be used to distinguish the false echo, which is used in the military and non-military applications. Radar is widely used for Coastal and National Security, which includes detection of low-flying, submarine periscopes, missiles, aircraft, small marine vessels and small pieces of ice, etc. These Radars are also used for marine traffic management, marine navigation, Finding Oil spilling,

---

R. Navya (✉) · R. Devaraju  
Dayananda Sagar University, Bangalore 560068, Karnataka, India  
e-mail: [navya-ece@dsu.edu.in](mailto:navya-ece@dsu.edu.in)

R. Devaraju  
e-mail: [devaraju-ece@dsu.edu.in](mailto:devaraju-ece@dsu.edu.in)

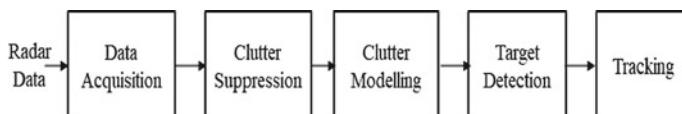
Detecting Marine debris, Sea patrolling, detection of small fishing vessels, and harbor surveillance.

Radar that are operating in a marine environment consequently experience backscattering, which is usually referred to as sea clutter or sea echo. The backscatter from the sea is an unwanted echo that may disrupt the radar's operation. Sea clutter characteristics vary widely depending on predominant conditions of environment and geographical location. By understanding the clutter characteristics, a suitable signal processing strategy has to be developed by radar system designers to revise performance under different conditions. This well-defined statistical model developed has to eliminate the clutter return is an imperative step. The temporal and spatial properties of the sea echo should be incorporated while designing the models, by considering the wide range of viewing geometries for different weather conditions.

Traditionally, modeling of sea clutter has been a random presumptive method, with a radar range resolution unit being a model basis of the stochastic method, for high grazing angles, low-resolution and large number of scattered cells in traditional radars which exhibits range resolution units that have the characteristics of random transformation. For these types of sea clutter Rayleigh distribution is frequently used to describe amplitude distribution. However, with a lot of developments in radar technologies like high-frequency, high-resolution, and low-grazing angles, and sea clutter exhibits non-Gaussian, long trailing phenomenon with significant deviations from Rayleigh distribution. For sea clutter with these characteristics Weibull, Log-Normal, and K-distribution are used to describe sea clutter amplitude distribution.

## 2 General Model Blocks of Target Detection

In general, the implementation blocks of target detection as shown in Fig. 1, involves the raw data acquisition from radar, then applying clutter suppression methods to achieve better signal-to-noise ratio. The processed data is streamed through different statistical clutter modeling techniques to differentiate the target from clutter influence. By applying the required threshold on the processed data target can be detected and tracked.



**Fig. 1** Implementation blocks of target detection

### 3 Review of Related Work

A literature survey is carried out on various methods for suppressing the sea clutter to detect the target. Various procedures, techniques are being used by many researchers for detecting slow and fast-moving targets by suppressing sea clutter. In this section, we summarize the methods used by various authors, with the merits and demerits to improve the target detection more efficiently under the clutter background, and comparisons of various review work are as shown in Table 1.

Parsa [1], proposed a method for detecting fast-moving targets by suppressing the sea clutter using a non-coherent X-band radar, by considering antenna rotation as the main criteria. A large set of data was collected from two X-band radars by operating Radar#1 with more antenna rotation (rpm) and then Radar#2 which are horizontally polarized [2] simultaneously and installed on a moving vehicle. By using recorded data, the two antenna rotation speeds for a fast motile target are analyzed using a scan-to-scan integration procedure. On analysis, the performance data was obtained on detection of the target by considering its antenna speed, range, and the number of scans by scan averaging method. By considering the different scanned averaged data, the probability of detection was obtained. So the author concludes increasing the rotation speed of the antenna, improves the performance of detection of fast-moving targets by integrating the fixed number of scans in moderate sea state.

Liu et al. [3] elaborate sea clutter production and its effect. For the suppression of sea clutter and target detection, the authors proposed a new  $\alpha$ - $\beta$ - $\gamma$  filter. Analysis of performance with the real data is provided and standards to choose the parameters are given. Further, they list the limiting factors which severely affect the detection capability of radar such as operating frequency, polarization mode, antenna visual angle, wind direction, and sea state.

Authors claim that traditional stochastic modeling of sea clutter like Rayleigh distribution, Weibull, Log-Normal, and K-distribution would not provide much support in understanding of physical or analytical characteristics. So to understand the sea clutter features, characteristics, authors use the fractal concept for description and modeling of sea surface roughness [4, 5]. The confused state behavior of sea clutter is challenging to design a new filter for target detection. Further includes techniques like extinction pulse and time-frequency. The neural network and wavelet-based approach were also considered. Using X-band non-coherent radar the cluttered data is recorded, which is installed in the East China Sea was used for analysis. This is an azimuth and range surveillance radar and concerning each azimuth repetition pulse data has been recorded. The proposed technique is a Kalman filter based on new  $\alpha$ - $\beta$ - $\gamma$  filter, to separate clutter from a target in the frequency domain instead of designing a clutter model adaptable to every situation of varying sea and weather conditions. It is observed that using low pass filter theory the low-frequency component creates a large-scale structure of sea clutter and it governs sea echo for a maximum portion, then that of the high-frequency component. As a result of detailed analysis, the bandwidth and signal-to-clutter ratio improvement are affected by three parameters  $\alpha$ ,  $\beta$ , and  $\gamma$ . A system function of the filter has been evaluated for the optimal parameter

**Table 1** Comparison of review work

Reference	Aim	Contribution	Future scope/Research direction
1	Analysis based on a fast-moving target using scan-to-scan integration by the sigma S6 radar processor	When the target looks stationary above 0.8 for an aerial target, by using scan-to-scan integration the detection efficiency is increased	When there is a strong sea clutter region the target detectability decreases for (0–2 Nm) range
2	$\alpha$ - $\beta$ - $\gamma$ filter is used to decrease the sea clutter influence on the target	Kalman filter is used for clutter suppression and target detection with the best accuracy	High-frequency components are not eliminated; low-frequency large-scale structure of sea clutter is only eliminated Small targets immersed in sea clutter with negative power are not eliminated. This is a flaw to be amended
3	The statistical distributional modeling of sea clutter at very low-grazing angles used to fit the data	Log-logistic distribution fits the data to the best and with a good local fit	The commonly used data fitting results showed that distribution like Weibull, Log-Normal underfit the tail region by giving an underestimated detection threshold If the target is non-fluctuating the probability of detection varies with less SCR, while fluctuating target with high signal-to-noise ratio (SCR)
4	CFAR detection using orthogonal projection	On experimental analysis, the total time spent to suppress clutter is 0.016 s using orthogonal projections then SVD	Clutter is better estimated and suppressed in HH polarization than in VV polarization mode
5	Correlation characteristics, amplitude characteristic vectors and Hurst exponent in the fractional Fourier transform are considered to distinguish the target and sea clutter	Extreme learning machine is effective method when compared with SVM	ELM can be applied only for lower sea state For higher sea state, the accuracy rate is less

(continued)

**Table 1** (continued)

Reference	Aim	Contribution	Future scope/Research direction
6	Panjer Probability Hypothesis Density filter is used to classify the true object from clutter True targets are recognized by their motion (speed) and measurement models	Tracking accuracy is maintained for temporally sparse radar scans	Tracking error is more when the number of scanning interval is increased
7	Eigenvalue-based detection scheme used to suppress clutter with short pulse	To distinguish the target from the clutter, maximum eigenvalue of the covariance matrix is adopted	The method is applied only when the pulse train length is small
8	CFAR- Temporal sequence of consecutive navigation radar imaging method is used to suppress the clutter	X-band temporal sequences of navigation radar images were exploited to describe the proposed method	Average Detection precision is 72%

combination of  $\alpha = 0.01$ ,  $\beta = 0.0033$ ,  $\gamma = 0.6$ . The experimental result conducted on one-way and round scanning real data shows that a large part of the strong sea clutter with high power is suppressed, which resulted in a better distinction between target and sea clutter. In this method, small targets can submerge within the sea, since experimental results showed that both the power of clutter and target has been reduced.

Song et al. [6] aims on finding the statistical modeling of sea clutter and to identify best-fitting distribution with emphasis on statistical modeling of sea clutter. Authors used the clutter data from X-band high-resolution coastal radar operating at low-grazing angles for statistical modeling [7] of sea clutter. Since detection probability depends on the region of distribution and the tail region mainly contributing to a threshold.

Authors derive a common method for calculating the expected detection probability to evaluate how detection probability calculation is affected by global fit. By using maximum likelihood estimation and least-squares estimation the recorded data is fitted to the following distributions like Weibull, K, Log-Normal, Pareto + noise, K + Rayleigh, KK [8] distribution, and also Log-logistic distribution. The fitting results exhibit that Weibull, K, and Log-Normal distribution fits the clutter distribution at the tail region, which results in an underrated threshold for detection. The experimental result shows that the Log-logistic distribution is ideal to model the entire region of sea clutter, while the lately developed K + Rayleigh distribution, which is superlative for thermal noise, fits the tail region to best. Authors also demonstrate that for non-fluctuating targets, considering the global fit the clutter distribution has a constant

impact on the expected probability of detection calculation. For fast-moving targets with a high SCR, the detection probability is less.

Yang et al. [9] propose an Orthogonal Projection (OP) scheme to detect the target by suppressing the sea clutter. The authors combine OP with a Constant False Alarm Rate (CFAR) of cell averaging to design a new detector. They conduct experiments to compare and demonstrate the clutter suppression averaging and the complexity of OP with singular value decomposition (SVD). To experiment on the data, the authors construct the clutter subspace using the data vectors of cells under tests (CUT) neighboring range cells. The observed signal of CUT is projected to the orthogonal subspace of clutter by suppressing the sea clutter of the CUT.

Authors investigate the methods suggested by scholars to improve the signal-to-clutter ratio (SCR). They emphasize mainly on suppressing clutter rather than increasing target echo. The singular value decomposition (SVD), block-adaptive filter, adaptive fractional Fourier transform, knowledge-aided reconstruction, and range distribution [10] via feature-based detector are some of the clutter suppression techniques discussed by the authors. McMaster University collected experimental sea clutter data from IPIX radar, which is available online for download and analysis. The operating frequency of IPIX radar is 9.39 GHz and it operates in staring mode when the radar is stationary and looks at the target scene with pulse repetition frequency of 1 kHz. The range resolution of the sampled data is 30 m which is sampled at every 15 m. They used sea clutter data for experimental analysis in the MATLAB environment.

The computational efficiency is found to be about 0.016 s for orthogonal projection and 3.047 s for singular value decomposition to perform clutter suppression. The experimental result indicates that OP methods computation efficiency is better than that of SVD. The like-polarized (HH, VV) signals are more preferred than the cross-polarized (HV, VH). More specifically clutter suppression and target estimation are better when HH polarized signals are used in most cases. Radar signal processing mostly uses sea clutter modeling to suppress the echoes, reflections by the sea to find the target. and using OP detection performance is improved. OP has the advantage of easy implementation while combining it with cell averaging CFAR is simple and its computational complexity is much less than that of SVD.

Jing et al. [11] proposed a new algorithm called Extreme Learning Machine [ELM] for target finding in presence of sea clutter by classifying its features. Firstly, the authors analyze the problem by considering the correlation characteristics of sea clutter which is further divided into temporal correlation and spatial correlation. But temporal correlation for target detection mainly uses a feature vector and clutter fluctuation features are returned on the time measurement in the same resolution unit. Secondly, the amplitude characteristics of sea clutter, an important part of statistical characteristics, are considered and analyzed using K-distribution. The most extensive range of fitting parameters is exhibited by K-distribution, so that target and unwanted echo can be effectively separated. Another characteristic of sea clutter that differentiates that from the target is fractal property which possesses self-similarity properties of target used to increase the efficiency. Hurst exponent in the fractional Fourier transform (FrFT) domain [12, 13] is another method that increases the signal-to-clutter

ratio (SCR) to achieve the target detection effectively. Finally, the extreme learning machine (ELM) is used in pattern classification of these features and concluded that the extreme learning machine can separate the target echo from sea echo very efficiently.

Schlängen and Charlish [14] discuss an approach that concurrently tracks the echo from the object and correlates clutter echoes. The target and clutter are distinguished, based on their profile history in precise dynamic models. To conduct experimental analysis, authors use the dataset from the finest trial database conducted by the Council for Scientific and Industrial Research (CSIR), which is recorded at a coastal area close to Cape Agulhas in South Africa in 2006. For the target like small vessels, the Radar cross section of the target is very small compared to that of sea clutter. As a result, intensity-based clutter modeling may either remove a lot of false targets or accept a lot of clutter echo by simple thresholding.

Authors use a Bayesian approach for the classification of target and clutter. In this approach, the processor interprets the related false returns to a separate class of objects in parallel to correlation to target. Plot extraction is achieved, firstly by applying a rough intensity threshold across each range cell, then creating a connected component followed by extracting coordinates from each component. After accepting more false detections, classification of clutter and target is done by considering target dynamics for multiple intervals. The Bayesian multi-object filter is used to estimate two groups of objects simultaneously, i.e., objects of type target (vessel) and clutter (wave). Classification between both the clusters is based on previous information of origin-motion model, probability of detection existence, and measurement accuracy. Panjer Probability Hypothesis Density (PHD) filter [15] is selected for multi-target classification. Two independent object groups are estimated and the same group of measurements is used for updating and correcting steps.

Authors conclude that suppressing the echo in a frame-by-frame manner may lead to failed target detection in case of low RCS targets or more false alarms in case of low threshold. By using a Panjer Probability Hypothesis Density (PHD) filter for classification between two populations, i.e., target object and clutter object, the number of false tracks is significantly reduced.

Zhao et al. [16] propose a detection scheme to identify moving objects or targets in sea clutter background for short-pulse operating radars. The proposed scheme is a combined algorithm, combining an eigenvalue with sub-band decomposition. They used a discrete Fourier transform-modulated-based filter bank for sub-band decomposition which can effectively suppress sea clutter and also increases the coherent integration time. The target spectrum can be separated from the clutter spectrum by transforming the original scenario into a scene where the target spectrum overlays with that of clutter. An eigenvalue detector cascade is performed in each sub-band to detect the target.

The advantage of using the linear-phase DFT modulated finite-impulse response (FIR) filter bank is, it maintains the value of the phase structure of the target echo unaffected and it has high stop-band rejection. Further, the authors compare by determining the frequency-domain characteristics of each finite-impulse filter separately and the Discrete Fourier Transform (DFT) modulated method. They found that the

Discrete Fourier Transform (DFT)-modulated method is simple to implement and can reduce the consumption of hardware and computation load.

In the target sub-band, the usage of a non-coherent technique which uses amplitude difference is more advantageous for detection. This is mainly because of the non-obvious phase difference produced due to the overlapping of the target and the clutter spectrums. The covariance matrices maximum eigenvalue is embraced to distinguish the target from the clutter because the maximum eigenvalue will reflect the signal intensity and it seizes the signal correlation very well. Authors prove with the simulation outcomes that the proposed cascade algorithm with an eigenvalue detector gives better detection of targets in short-pulse operation. It is also better than the adaptive normalized matched filter (ANMF) [17] and sub-band decomposition-based adaptive normalized matched filter (SANMF) methods when the short-term coherent accumulation and availability of radar reference cells are limited.

Ding et al. [18] proposed a method for sea clutter suppression by using sensors designed for radar navigation to obtain a temporal sequence of radar images, which specifically works in X-band and they used in coherent radars of horizontal polarization. By using a time-based sequence of navigation images for the CFAR method to reduce clutter influence on target, the effective 5 steps which include raw data acquisition from radar, noise suppression, calculation of threshold, applying the threshold, and removal of a false alarm from detection were proposed to reduce the effect of sea clutter on target. In data acquisition or cumulative steps, firstly,  $k$  ( $k > 1$ ) temporal repeated radar images are improved to increase the target signal strength to separate that from sea clutter. Secondly, in background depression, sea clutter intensity variation is fitted with range after every scan of the cumulative image and then the sea clutter strong intensity is detracted from the cumulative image for every pixel. In threshold calculation, using Constant False Alarm (CFAR) [19] the detection of a target is done using an adaptive algorithm method to distinguish target returns from noise in radar systems, clutter, and interference. By considering the probability density function and its probability distribution function [20], CFAR is calculated on giving the threshold value. The Probabilistic Neural Networks (PNN) model is used to approximate the sea clutter probability density function in radar images. Finally, the resulting image obtained a greater threshold as compared with the reference threshold, then it is considered as a target. Finally, false alarm is removed to obtain the associated size of the area of possible target by flood fill algorithm. An experimental test was carried out using X-band navigation radar images with approximately 250 temporal sequences, 7.5 m is considered as distance resolution, and 0–360° of azimuth angle variation. After conducting the experimental test, the proposed method is capable of reducing the sea clutter significantly from navigation images of radar and to detect the ship efficiently which is immersed in the sea clutter. Finally, the authors conclude based on experimental results that 72% of detection accuracy is obtained.

## 4 Conclusion

In this paper, we have reviewed recent literature on sea clutter suppression techniques used for radar target detection in the presence of sea clutter background. Non-coherent X-band radars are used more commonly for suppressing the sea clutter. Signals with two orthogonal components are chosen and like-polarized (HH, VV) signals are preferred than the cross-polarized (HV, VH). More specifically clutter suppression and target estimation are better when HH polarized signals are used in most cases. Radar signal processing mostly uses sea clutter modeling to suppress the echoes, reflections by the sea to find the target. From the prospect of the statistical properties, the measured data are fitted to a suitable model to obtain the desired target spot. Distribution information processing and high-detectable adaptive technology have to be combined by cascading the base algorithms such as K + Rayleigh distribution, sub-band decomposition with an eigenvalue. Finally, issues with existing models and further research challenges were presented to provide guidelines for future research trends.

## References

1. Parsa, A.: Fast moving target detection in sea clutter using non-coherent X-band radar. IEEE Natl. Radar Conf. Proc. **2**, 1155–1158 (2014). <https://doi.org/10.1109/RADAR.2014.6875770>
2. Parsa, A., Hansen, N.H.: Comparison of vertically and horizontally polarized radar antennas for target detection in sea clutter—An experimental study. In: IEEE Radar Conference, Atlanta, GA, pp. 0653–0658 (2012). <https://doi.org/10.1109/RADAR.2012.6212220>
3. Liu, J., Meng, H., Wang, X.: Radar sea clutter suppression and target detection with  $\alpha$ - $\beta$ - $\gamma$  filter. In: 9th International Conference on Signal Processing, Beijing, pp. 2376–2379. [https://doi.org/10.1109/ICOSP.2008.4697627\(2008\)](https://doi.org/10.1109/ICOSP.2008.4697627(2008))
4. Martorella, M., Berizzi, F., Mese, E.D.: On the fractal dimension of sea surface backscattered signal at low grazing angle. IEEE Trans. Anten. Propag. **52**, 1193–1204 (2004)
5. Berizzi, F., Dalle-Mese, E., Martorella, M.: A sea surface fractal model for ocean remote sensing. Int. J. Remote Sens. **25**, 1265–1270 (2004)
6. Song, D., Sarikaya, T.B., Serkan, S.T., Tharmarasa, R., Sobaci, E., Kirubarajan, T.: Heavy-tailed sea clutter modeling for shore-based radar detection. IEEE Radar Conf. Radar Conf. pp. 1504–1509 (2018). <https://doi.org/10.1109/RADAR.2018.8378789>
7. Hu, J., Tung, W., Gao, J.: Detection of low observable targets within sea clutter by structure-function based multifractal analysis. IEEE Trans. Anten. Propag. **54**(1), 136–143 (2006)
8. Rosenberg, L., Crisp, D.J., Stacy, N.J.: Analysis of the KK-distribution with medium grazing angle sea clutter. IET Radar, Sonar Navig. **4**(2), 209–222 (2010)
9. Yang, Y., Xiao, S.P., Wang, X.S.: Radar detection of small target in sea clutter using orthogonal projection. IEEE Geosci. Remote Sens. Lett. **16**(3), 382–386 (2019). <https://doi.org/10.1109/LGRS.2018.2875705>
10. Shi, Y., Xie, X., Li, D.: Range distributed floating target detection in sea clutter via the feature-based detector. IEEE Geosci. Remote Sens. Lett. **13**(12), 1847–1850 (2016)
11. Jing, W., Ji, G., Liu, S., Wang, X., Tian, Y.: Target detection in sea clutter based on ELM. In: Li, J. et al. (eds.) Wireless Sensor Networks. CWSN 2017. Communications in Computer and Information Science, vol. 812. Springer, Singapore (2018). [https://doi.org/10.1007/978-981-10-8123-1\\_3](https://doi.org/10.1007/978-981-10-8123-1_3)

12. Hu, J., Tung, W.-W., Gao, J.: Detection of low observable targets within sea clutter by structure function based multifractal analysis. *IEEE Trans. Anten. Propag.* **54**(1), 136–143 (2006). <https://doi.org/10.1109/TAP.2005.861541>
13. Chen, X., Guan, J.: A fast FRFT based detection algorithm of multiple moving targets in sea clutter. In: IEEE Radar Conference, Washington, DC, pp. 402–406 (2010). <https://doi.org/10.1109/RADAR.2010.5494587>
14. Schlangen, I., Charlish, A.: Distinguishing small targets from sea clutter using dynamic models. In: IEEE Radar Conference, pp. 1–6 (2019). <https://doi.org/10.1109/RADAR.2019.8835683>
15. Schlangen, I., Delande, E.D., Houssineau, J., Clark, D.E.: A second-order PHD filter with mean and variance in target number. *IEEE Trans. Signal Proces.* **66**(1), 48–63 (2018)
16. Zhao, W., Chen, Z., Jin, M.: Subband maximum eigenvalue detection for radar moving target in sea clutter. *IEEE Geosci. Remote Sens. Lett.* 1–5 (2020). <https://doi.org/10.1109/lgrs.2020.2971589>
17. Shui, P.-L., Shi, Y.-L.: Subband ANMF detection of moving targets in sea clutter. *IEEE Trans. Aerosp. Electron. Syst.* **48**(4), 3578–3593 (2012)
18. Ding, X., Huang, W., Zhou, C., Chen, P., Liu, B.: Ship detection in presence of sea clutter from temporal sequences of navigation radar images. *MIPPR, Autom. Target Recognit. Image Anal.* **7495**, 74954I (2009). <https://doi.org/10.1117/12.832882>
19. Hu, W., Wang, Y., Wang, S., et al.: A robust CFAR detector based on ordered statistics. In: Proceedings of 2006 CIE International Conference on Radar, pp. 1–4 (2006)
20. Jiang, Q., Aitnouri, E., Wang, S., Ziou, D.: Automatic detection for ship target in SAR imagery using PNN-model. *Can. J. Remote Sens.* **26**(4), 297–305 (2000). <https://doi.org/10.1080/07038992.2000.10874780>

# An Ensemble Model for Predicting Chronic Diseases Using Machine Learning Algorithms



B. Manjulatha and Suresh Pabboju

**Abstract** Correct diagnosis of a disease plays a vital role in today's environment. Diabetes and liver are one such chronic diseases which are the most hazardous ailment that affects a large number of individuals which may lead to death. Machine learning algorithms help to predict the diseases early which saves many lives of mankind in the world. Datasets like Pima Indian diabetes dataset (PIMA), Indian liver patient data (ILPD) and cardiovascular disease (CVD) are taken from UCI repository to compare the results by applying various well-known algorithms. Every algorithm gives its output independently but knowing the highest accuracy is difficult as each algorithm gives different result, i.e., may be less or more according to their dimensions. In this system, accuracy is increased by combining individual algorithms such as decision tree, SVM, logistic regression, ANN, random forest classifier, KNN to construct an ensemble hybrid model which gives more accurate, accuracy.

## 1 Introduction

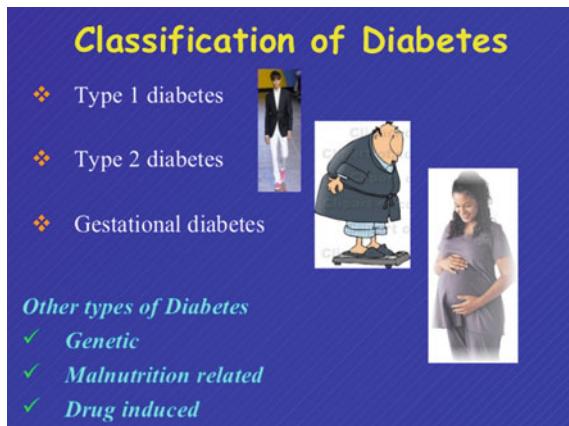
Detecting the disease is the first and most critical step in today's infectious disease control world. Some people do not have any symptoms or they did not even recognize them for few chronic diseases, and they will assume that they are healthy. Diabetes Mellitus [1] is a chronic disease which causes high blood sugar. It is of two types, namely Type 1 diabetes, Type 2 diabetes and gestational diabetes. Type-2 diabetes is most dangerous which has a huge increase in blood glucose and may lead to heart disease and even death. Genes may play a major role, i.e., family members share their genes. If it is not treated that make them more likely to get diabetes properly within specified time, it may damage our nerves, eyes, kidneys and other organs of our body. Even youngsters are suffering from this hazardous disease without the knowledge

---

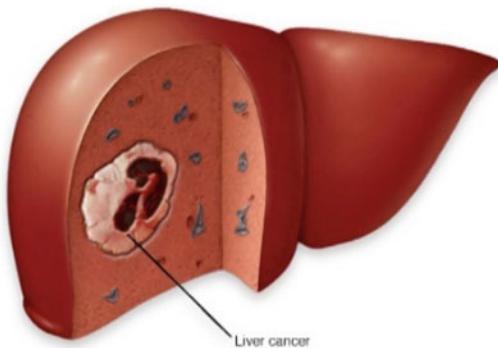
B. Manjulatha (✉)  
OU Scholar, Hyderabad, India  
e-mail: [manjulatha@vbithyd.ac.in](mailto:manjulatha@vbithyd.ac.in)

S. Pabboju  
Information Technology, CBIT, Hyderabad, India

of the person depending upon their weight, genes, eating habits and many more. So, here machine learning plays a crucial role in predicting the disease early so that it can stop deaths of many people.

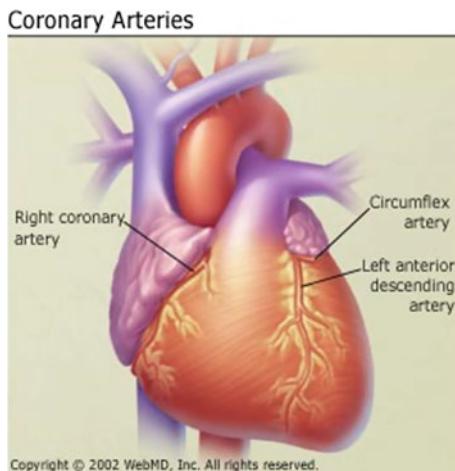


Liver [2] is moon-shaped organ which weighs around 3 lb in adults. It manufactures proteins, fats and carbohydrates by eliminating waste products. It is majorly caused by consuming excess alcohol and smoking. Types of liver disease are hepatitis, cirrhosis, liver cancer, liver failure etc. Symptoms of this disease may include edema, fatigue, jaundice, etc. [3].



© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH. ALL RIGHTS RESERVED.

Cardiovascular disease [4] is otherwise called as heart disease. It is the major organ in our body which pumps blood to all parts of the body. The major risk factors of this disease include obesity, hypertension and consumption of tobacco, eating unhealthy diet and smoking and results in death.



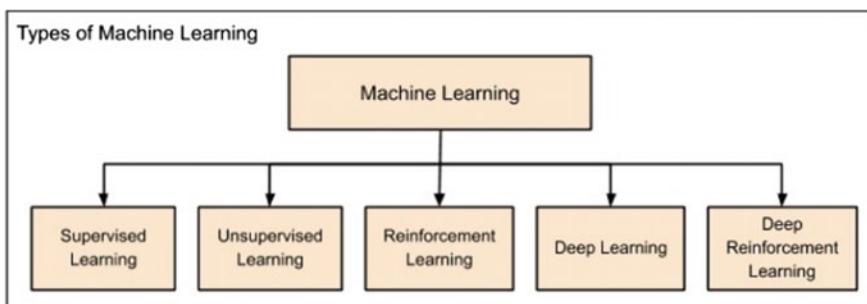
Copyright © 2002 WebMD, Inc. All rights reserved.

These three chronic diseases cause a great damage to our body if not detected early. The main objective of this paper is to predict the diseases early in order to increase the life span of every individual. Current research also confirmed that there will be huge recovery chances if diseases are discovered early.

With the rise in big data machine learning has become a key technique for solving disease detection, face recognition, etc. It is a tool to collect the data, analyze and finally predict. Section 1 of this paper is Introduction; Sect. 2 describes machine learning algorithms. Section 3 discusses about related work of various researchers. Section 5 describes the datasets used and its results. At last, Sect. 6 concludes with conclusion and future scope.

## 2 Machine Learning Algorithms

Machine learning algorithms are categorized into five types.



Most widely supervised algorithms are linear regression, logistic regression, random forest, gradient boosting, support vector machine, decision tree, nearest neighbor, etc.

Unsupervised algorithms are K-means clustering, PCA, association rule, etc.

Reinforcement learning algorithms are Q-learning, temporal difference, Monte Carlo tree search, asynchronous actor-critic agents.

### 3 Related Work

Analysis done by various authors by using different machine learning algorithms

Author	Dataset	Algorithm/methodology used	Conclusion /future scope
Kamrul Hasan [5]	PIMA	Hybrid model ( $AB + XB$ )	Hybrid algorithm gives high accuracy
Roopa [6]	PIMA	Principal component analysis	Achieved highest accuracy of 82%
Komal Kumar [7]	CVD	KNN, RF, LR, SVM and DT	Accuracy is 85%
Yahyaoui [8]	PIMA	DT, SVM and RF	High accuracy and automatic extraction can be done for better fitting model
Neelaveni [9]	Alzheimer	DT, SVM	Future work results in predicting the disease using brain MRI scans
Bhavana [10]	PIMA	KNN, Naïve Bayes, Random Forest, J48	Hybrid algorithm gives high accuracy
Al-Zebari [11]	PIMA	Decision tree, logistic regression, SVM, KNN, Discriminant Analysis	Achieved highest accuracy of 77%
Kumar [12]	MPRLPD and ILPD liver datasets	SMOTE algorithm	Performs better accuracy
Sai Prakash [13]	Heart disease	Random forest, vector support, logistic regression and XG-Boost	SVM and logistic regression performs high accuracy
Singh [14]	Heart disease	KNN, SVM, LR, DT	KNN perform high accuracy
Dahiwade [15]	Symptoms of the patients	KNN, CNN	CNN performs high accuracy with 84.5%

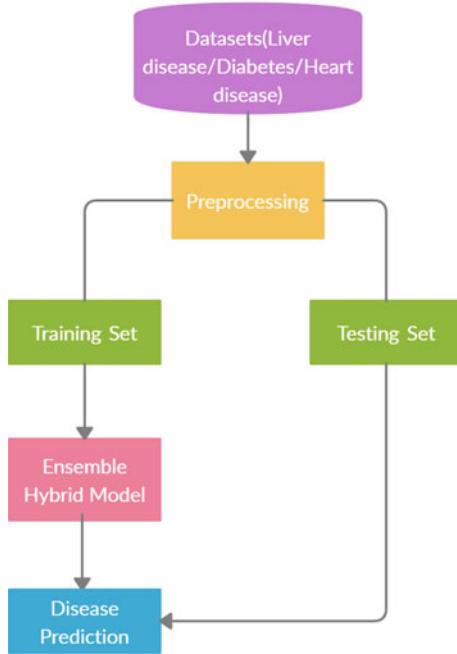
(continued)

(continued)

Author	Dataset	Algorithm/methodology used	Conclusion /future scope
Harimoorthy [16]	Diabetes, heart, kidney diseases	SVM-radial bias	Performs better accuracy for all disease types
Pethunachiyar [17]	Diabetes	SVM-linear kernel	Achieved 100% accuracy
Verma [18]	Skin disease	Bagging and boosting techniques	SVM provides best result of 99.67%
George Amalarethinam [19]	Diabetes	Data mining techniques	Performs better accuracy
Kalipe [20]	Malarial disease	Machine learning and deep learning techniques	Deep learning approach performs better accuracy

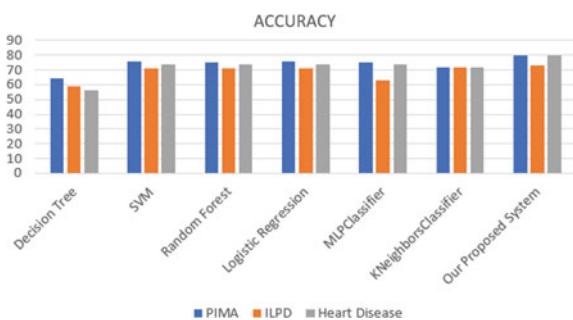
## 4 Methodology

Initially we take disease dataset from UCI or Kaggle repository. After that pre-processing of data is done which cleans the data without any noisy data and missing data. Missing values can be removed by applying various pre-processing techniques. Here, standard scaler is used. Then applying machine learning algorithms to predict the disease accurately. The dataset is being splitted into train data and test data in the ratio of 80:20. The proposed work is carried out in Google colab which is a free online cloud based jupyter notebook environment that allows to perform various machine learning and deep learning models on CPUs, GPUs and TPUs and doesn't require any software to be installed In this study, three chronic diseases like ILPD, PIMA, Heart disease are used for building and testing the models.



## 5 Experimental Results

For all experiments, training set is taken as 80%, and testing test is taken as 20%. The performance parameters for the proposed system are taken as accuracy, ROC AUC score, precision, recall, F1-score, accuracy on training set and testing set (Tables 1, 2, 3 and 4).



**Table 1** Classification report for PIMA dataset

Classifier	Classification report for PIMA dataset						ROC AUC score	
	Precision		Recall		F1-score			
	0	1	0	1	0	1		
Decision tree	0.70	0.48	0.78	0.37	0.74	0.42	0.57	
SVM	0.78	0.71	0.88	0.54	0.83	0.61	0.70	
Random forest	0.76	0.7	0.89	0.48	0.82	0.57	0.68	
Logistic regression	0.78	0.71	0.88	0.54	0.83	0.61	0.70	
MLPClassifier	0.76	0.70	0.89	0.48	0.82	0.57	0.68	
KNN	0.59	0.62	0.46	0.73	0.52	0.57	0.68	
Our proposed system	0.5	0.69	0.89	0.46	0.82	0.56	0.68	

**Table 2** Classification report for liver dataset

Classifier	Classification report for liver disease dataset						ROC AUC score	
	Precision		Recall		F1-score			
	0	1	0	1	0	1		
Decision tree	0.32	0.72	0.36	0.68	0.34	0.7	0.52	
SVM	0	0.71	0	1	0	0.83	0.5	
Random forest	0.75	0.73	0.64	0.82	0.69	0.77	0.73	
Logistic regression	0	0.71	0	1	0	0.83	0.5	
MLPClassifier	0.3	0.71	0.21	0.8	0.25	0.76	0.5	
KNN	0.36	0.74	0.36	0.74	0.36	0.74	0.55	
Our proposed system	0.38	0.73	0.24	0.84	0.3	0.78	0.54	

**Table 3** Classification report for heart disease dataset

Classifier	Classification report for heart disease dataset						ROC AUC score	
	Precision		Recall		F1-score			
	0	1	0	1	0	1		
Decision tree	0.53	0.57	0.36	0.73	0.43	0.64	0.54	
SVM	0.75	0.73	0.64	0.82	0.69	0.77	0.73	
Random forest	0.75	0.73	0.64	0.82	0.69	0.77	0.73	
Logistic regression	0.75	0.73	0.64	0.82	0.69	0.77	0.73	
MLPClassifier	0.75	0.73	0.64	0.82	0.69	0.77	0.73	
KNN	0.59	62	0.46	0.73	0.52	0.67	0.59	
Our proposed system	0.5	0.69	0.89	0.46	0.82	0.56	0.68	

**Table 4** Accuracy of datasets

Classifier	PIMA	ILPD	Heart disease
Decision tree	64	59	56
SVM	76	71	74
Random forest	75	71	74
Logistic regression	76	71	74
MLPClassifier	75	63	74
KNeighborsClassifier	72	72	72
Our proposed system	80	73	80

## 6 Conclusion and Future Scope

This study gives a comparative result of machine learning algorithms [21] for diabetes, liver disease and heart disease dataset. The results show that the classification accuracy is reached to 80% for ensemble hybrid model. In the future, we would like to improve the accuracy by applying deep learning algorithms.

## Reference

1. <https://www.mayoclinic.org/diseases-conditions/diabetes/symptoms-causes/syc-20371444>.
2. <https://www.healthline.com/health/human-body-maps/liver#structure>.
3. <https://www.mayoclinic.org/diseases-conditions/liver-cancer/symptoms-causes/syc-20353659>.
4. <https://www.webmd.com/heart-disease/guide/diseases-cardiovascular#1>.
5. Kamrul Hasan, M.D.: Diabetes prediction using ensembling of different machine learning classifiers. In: IEEE-Special Section on Deep Learning Algorithms for Internet of Medical Things, vol. 8 (2020)
6. Roopa, H.: A Linear model Based on Principal Component Analysis for Disease Prediction, vol. 7. IEEE Access. <https://doi.org/10.1109/ACCESS.2019.2931956> (2019)
7. Komal Kumar, N.: Analysis and prediction of cardio vascular disease using machine learning classifiers. In: 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE (2020)
8. Yahyaoui, A.: A Decision Support System for Diabetes Prediction Using Machine Learning and Deep Learning Techniques. IEEE (2019)
9. Neelaveni, J.: Alzheimer disease prediction using machine learning algorithms. In: 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS). IEEE (2020)
10. Bhavana, N.: A review of ensemble machine learning approach in prediction of diabetes diseases. Int. J. Fut. Revolut. Comput. Sci. Commun. Eng. **4**(3), 463–466. ISSN: 2454-4248 (2018)
11. Al-Zebari, A.: Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection. IEEE (2019)
12. Kumar, P.: Early detection of the liver disorder from imbalance liver function test datasets. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **8**(4). ISSN: 2278-3075 (2019)
13. Sai Prakash, C.: Data Science Framework—Heart Disease Predictions, Variant Models and Visualizations. IEEE (2020)

14. Singh, A.: Heart disease prediction using machine learning algorithms. In: 2020 International Conference on Electrical and Electronics Engineering (ICE3-2020) (2020)
15. Dahiwade, D.: Designing disease prediction model using machine learning approach. In: Proceedings of the Third International Conference on Computing Methodologies and Communication (ICCMC 2019). IEEE Xplore Part Number: CFP19K25-ART; ISBN: 978-1-5386-7808-4 (2019)
16. Harimoorthy, K.: Multi-disease prediction model using improved SVM-radial bias technique in healthcare monitoring system. *J. Amb. Intell. Humanized Comput.* (2020)
17. Pethunachiyar, G. A.: Classification of diabetes patients using kernel based support vector machines. In: 2020 International Conference on Computer Communication and Informatics (ICCCI-2020), 22-24 Jan 2020. Coimbatore, INDIA (2020)
18. Verma, A.K.: Skin disease prediction using ensemble methods and a new hybrid feature selection technique. *Iran J. Comput. Sci.*(2020)
19. George Amalarethinam, D.I.: Prediction of diabetes mellitus using data mining techniques: a survey. *Int. J. Appl. Eng. Res.* **10**(82). ISSN 0973-4562 (2015)
20. Kalipe, G.: Predicting malarial outbreak using machine learning and deep learning approach: a review and analysis. In: International Conference on Information Technology (ICIT), IEEE (2018)
21. Manjulatha, B., Pabboju, S.: Big data analytics and its applications. *Int. J. Recent Technol. Eng. (IJRTE)* **8**(4), 10928–10931. ISSN: 2277-3878, Nov 2019

# COVID-19 Face Mask Live Detection Using OpenCV



Anveshini Dumala, Anusha Papasani, and Sireesha Vikkurty

**Abstract** Severe acute respiratory syndrome coronavirus (SARS-CoV) is recognized, and very first person infected is from the Guangdong province of southern China in 2002 while the virus that causes COVID-19 (Corona VIrus Disease-2019) is known as SARS-CoV-2. World Health Organization (WHO) named it as “COVID-19” on February 11, 2020. Currently, the COVID-19 has frightened the whole world of human beings and pushed into the pandemic. This coronavirus affects the respiratory system by entering into the human body through the droplets of saliva and mucus. It takes 14 days to observe the symptoms of the virus attack. In the meantime, the virus affected person may spread the virus to the coexisting people in the abode unknowingly. Also, it takes 48 h to confirm if a person is virus attacked after the test sample is collected. So, there is a serious need to wear a face mask that covers the nose and mouth besides maintaining the social distance to break the chain of massive increase. This paper attempts to detect if an individual wears a mask, using OpenCV. The accurate identification of landmarks of face in the image is an imperative challenge. Being instinctive it is simple for a human to detect the object, but it took years of research to raise the accessibility of quality datasets and a remarkable progress. The purpose of the paper is identifying the count of faces with the mask in the image and count of faces without a mask on live webcam.

## 1 Introduction

The word “image segmentation” in computer vision is dividing the image into pixels with respect to color, pattern, etc. These groups of pixels are regularly termed as *super-pixels*, whereas the goal of “instance segmentation” is to identify specified objects in the image and highlight the object of interest. In contrast to the semantic

---

A. Dumala (✉) · A. Papasani

Vignan's Nirula Institute of Technology and Science for Women, Guntur, Andhra Pradesh, India

S. Vikkurty

Vasavi College of Engineering, Hyderabad, Telangana, India

segmentation, ‘instance segmentation’ does not aim at labeling pixel in the image. The precise identification of definite facial features and landmarks is an initial step which simplifies the complicated image analysis like face detection, expression, age estimation, gender categorization [1, 2], facial paralysis, palsy, and sleep apnoea [3, 4]. Several researchers have investigated if any association [5] lies in between facial landmarks and the sleep apnoea masks [6].

## 2 Proposed Methodology

The face mask recognition in OpenCV has a trainer and a detector. It is even possible to create and train our own classifier for any applications related to the cars, plane, etc. OpenCV already contains many pre-trained classifiers for face, eyes, etc. Now, in this proposed method, we detect the face wearing a mask. Those XML files are related to mask faces stored in opencv/data/haarcascades/folder. The proposed model has three different steps.

- a. Defining the objective.
- b. Selecting a suitable dataset.
- c. Mark the regions of interest in the image.

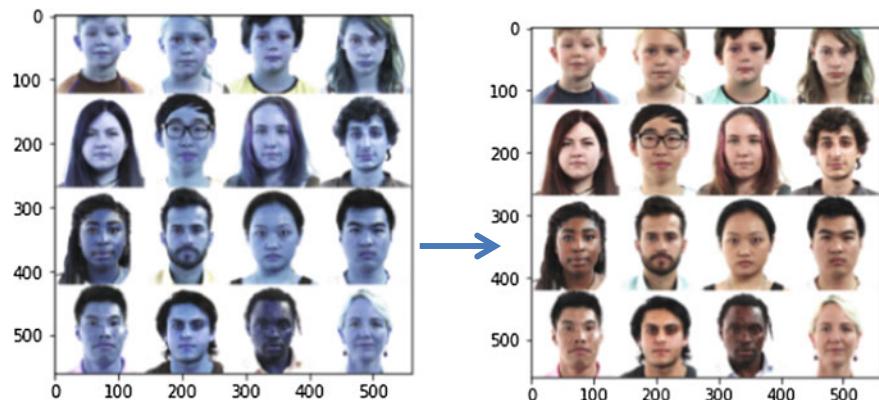
### Defining the objective

The objective of the proposed model is to identify the nose and mouth in the facial image so that the model identifies the count of faces that have no face mask.

### Selecting suitable dataset

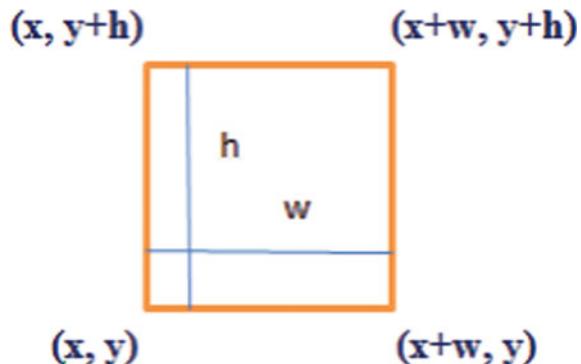
The proposed model considers the live data using webcam. Video is the collection of frames. For this, the images should be collected from the live web video and import the images into OpenCV. The process of importing the images has the following steps involved. Read in the image using the imread function and import the necessary libraries. Use the multi-scaling factors, the image can be resized as required for the purpose of analysis.

**OpenCV** is one of the libraries available in Python that deliberately solves the computer vision problems. The method cv2.rectangle() is used to mark a square or a rectangle around the object in the image. It has different arguments used as cv2.rectangle (image, start\_point, end\_point, color, thickness) where “**image**” is the image on which rectangle/square is marked as shown in Fig. 2. “**start\_point**” is the initial coordinates of rectangle. The coordinates are the tuples of X coordinate value and Y coordinate value. **end\_point** is the ending coordinates of rectangle. **Color** is the color of border line of rectangle. For BGR, a tuple, e.g., (255, 0, 0) is passed for blue color as shown in Fig. 1.



**Fig. 1** Conversion of the grayscale image into color image

**Fig. 2** Coordinates to form a rectangle



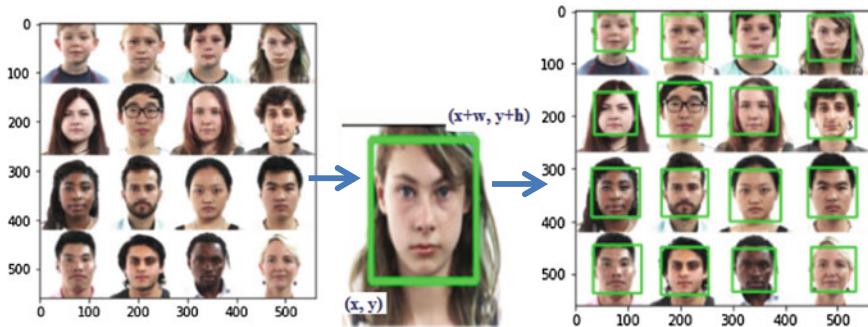
## 2.1 Extract the Region of Interest

All the faces in the image are identified using the coordinates  $(x, y)$ . Using those coordinates, a square is formed as shown in Figs. 2 and 3. In the square region of the image, the other facial landmarks are also identified.

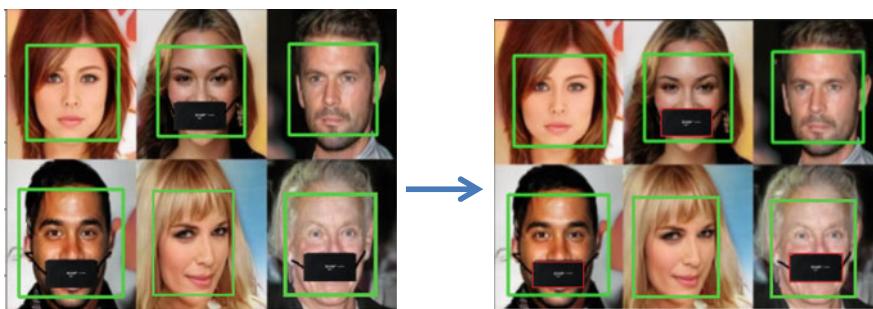
In this paper, the proposed model mainly focuses on the nose and mouth. The model identifies the nose and mouth in the square region. If the nose and mouth are covered with the mask, then it identifies the mask region as shown in Fig. 4.

## 3 Results and Analysis

Initially, the model is proposed to identify the face in the image. Later, the proposed model is enhanced that it takes the images from the live stream from the webcam



**Fig. 3** Identification of face in the image



**Fig. 4** Identification of **a** face green and **b** mask red in the image

and identifies the count of faces without a mask. The model displays the message “without mask-\*.”

In [Fig. 5](#), the proposed model identified the nose and mouth of faces in the image and gives the count “without mask-1.” The faces with the mask are not identified.

In [Fig. 6](#), the proposed model identified the nose and mouth of faces in the image and gives the count “without mask-2.” The faces with the mask are not identified.

In [Fig. 7](#), the proposed model identified the nose and mouth in the faces in the image and gives the count “without mask-3.” The faces with the mask are not identified.

## 4 Conclusion and Future Scope

This paper attempts to detect if an individual wears a mask, using OpenCV. The accurate identification of landmarks of face in the image is an imperative challenge. Being instinctive it is simple for a human to detect the object, but it took years of research to raise the accessibility of quality datasets and a remarkable progress. The

**Fig. 5** Identification of “faces without a mask”



**Fig. 6** Identification of “faces without a mask”



purpose of the paper is identifying the count of faces with the mask in the image and count of faces without a mask on live webcam. In the future, a framework can be designed that embeds sensors on the inside of a mask or develop a module that can be attached to any over-the-counter mask.

**Fig. 7** Identification of “faces without a mask”



## References

1. Chellappa, R., Wu, T., Turaga, P.: Age estimation and face verification across aging using landmarks. *IEEE Trans. Inf. Forensic Secur.* **7**(6), 1780–1788 (2012)
2. Biswaranjan, K., Devries, T., Taylor, G.W.: Canadian Conference on Computer and Robot Vision, pp. 98–103. IEEE Montreal (2014)
3. Jowett, N., Dusseldorp, J., Hadlock, T.A., Guarin, D.L.: A machine learning approach for automated facial measurements in facial palsy. *JAMA Facial Plast. Surg.* **20**(4), 335 (2018)
4. Guoliang, X., Jiaxin, S., Anping, S., Xuehai, D., Gang, X., Wu, Z.: Assessment for facial nerve paralysis based on facial asymmetry. *Australas. Phys. Eng. Sci. Med.* **40**(4), 851–860 (2017)
5. Balaei, T., Sutherland, K., Cistulli, A.P., de Chazal, P.: Automatic detection of obstructive sleep apnea using facial images. In: International Symposium on Biomedical Imaging, pp. 215–218 (2017)
6. McEwan, A., Johnston, B., de Chazal, P.: Semi-automated nasal PAP mask sizing using facial photographs. In: International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 1214–1217 (2017)

# Chest X-Ray Image Analysis of Convolutional Neural Network Models with Transfer Learning for Prediction of COVID Patients



M. Shyamala Devi, P. Swathi, N. Pavan Kumar, Ravi Varma Tungala,  
Saranya Vivekanandan, and Priyanka Moorthy

**Abstract** The COVID-19 pandemic grounds a major outbreak around the world, having a severe impact on the health and life of many people globally. The critical step in fighting COVID-19 is the capability to identify the infected patients during initial stages and situate them under extraordinary care. Detecting COVID-19 disease from radiography and radiology images is one of the effective way to detect the infected patients. Inspired by the chest radiograms of patients infected with COVID-19, we study the application of machine learning and convolutional neural network models to detect COVID-19 patients from their chest radiography images. We first prepare a dataset of chest X-rays from the publicly available ieee8023/covid-chestxray-dataset. This paper aims to provide the following contributions. Firstly, the dataset is preprocessed and segregated as healthy, COVID-19, bacterial pneumonia and viral pneumonia. Secondly, the dataset is processed to form the initial trained layers of base model and is fitted with several convolutional neural network models like VGG, ResNet, Xception and DenseNet to extract the high-level general features. Thirdly, the base model of several convolutional neural network models are added to the custom layers developed with transfer learning deep learning approach to

---

M. Shyamala Devi (✉) · P. Swathi · N. Pavan Kumar · R. Varma Tungala  
Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of  
Science and Technology, Chennai, Tamil Nadu, India  
e-mail: [shyamaladev@veltech.edu.in](mailto:shyamaladev@veltech.edu.in)

P. Swathi  
e-mail: [vtu10552@veltechuniv.edu.in](mailto:vtu10552@veltechuniv.edu.in)

N. Pavan Kumar  
e-mail: [vtu10513@veltechuniv.edu.in](mailto:vtu10513@veltechuniv.edu.in)

R. Varma Tungala  
e-mail: [vtu8975@veltechuniv.edu.in](mailto:vtu8975@veltechuniv.edu.in)

S. Vivekanandan  
Infologia Technologies, Chennai, Tamil Nadu, India

P. Moorthy  
RedBlackTree, Chennai, Tamil Nadu, India

analyze the performance of prediction of COVID-19 patients. Fourth, the performance of the convolutional neural network models along with transfer learning is analyzed with the metrics like model loss, precision, accuracy, recall and F-Score. The project is implemented with Python in Spyder under anaconda navigator. Experimental results show that Xception CNN model is found to experience 95% accuracy with the available chest X-ray images.

## 1 Literature Review and Shortcomings

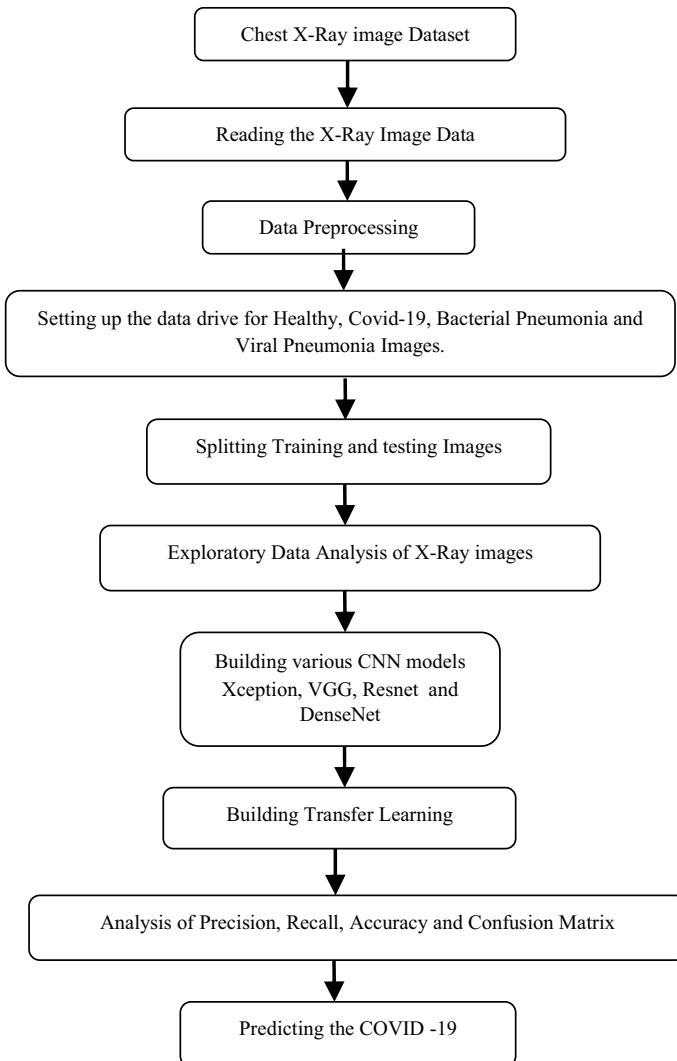
CXR image classification approaches in medical image analysis [1] were done to analyze the cardiomegaly using deep convolutional activation feature, and advancing the CAD is needed to improve the quality of radiologists' tasks. Validation and adoption of early developed neural networks are used for the classification of chest X-ray images [2] by achieving 93% precision. CAD system [3] for disease prediction (pulmonary nodules, tuberculosis and interstitial lung diseases), focusing on basic algorithm principles of the algorithm, and the results show that manifestations of tuberculosis and cavities with accuracy ranging from 92 to 95%. The techniques of bone suppression with hand-crafted features could have errors that affect the classification performance [3]. The neural network models [4] are trained and tested on chest X-ray database with CNN having accuracy of 92.4%. The different CNN architectures on the NIH "ChestX-ray 14" are done, performing the comparison of manual ground truth labels versus NLP labels which is unrealistic due to the unavailability of annotation [5].

Various problems based on lung diseases that affects children are provided with the diagnostic tool that classifies a chest X-ray image that shows if it is under the normal or pneumonia category [6]. The process of CNN algorithm on a chest X-ray dataset is attempted to classify pneumonia [7]. It requires further transfer learning, training a fine-tuned deep neural network and stabilizing training process. The optimal solution for classifying abnormal and normal chest X-ray images [8] is achieved with the substantive features provided by DenseNet followed by optimal hyperparameter values of SVM classifier. The medical imaging analysis involves with COVID-19, including image acquisition, segmentation, diagnosis and follow-up along with the training and testing of AI algorithms are still not efficient and quality of datasets are not sufficient [9]. Support vector machine [10] methodology is used for detecting the coronavirus infected patient using X-ray images. Mask-RCNN is used for global, local features for pixel-wise segmentation for identification and localization of pneumonia in chest X-rays images [11].

## 2 Overall Architecture

### 2.1 Dataset Preparation

The dataset of chest X-rays from the publicly available ieee8023/covid-chestxray-dataset is used for implementation. The overall workflow is shown in Fig. 1. The contributions of this paper are given below.



**Fig. 1** Overall workflow of this paper

- (i) Firstly, the dataset is preprocessed and segregated as healthy, COVID-19, bacterial pneumonia and viral pneumonia.
- (ii) Secondly, the dataset is processed to form the initial trained layers of base model and is fitted with CNN models like VGG, ResNet, Xception and DenseNet to extract the high-level general features.
- (iii) Thirdly, the base model of CNN models are added to the custom layers developed with transfer learning approach to analyze the performance of prediction of COVID-19 patients.
- (iv) Fourth, the performance of the CNN models along with transfer learning is analyzed with metrics model loss, precision, accuracy, recall and F-Score.

### 3 Convolutional Neural Network

#### 3.1 CNN Models

The convolutional neural network models used in this paper are VGG, ResNet, Xception and DenseNet to extract the high-level general features. The VGG model was proposed by Karen Simonyan and Andrew Zisserman of the Visual Geometry Group (VGG) Laboratory of Oxford University. VGG19 is one model from CNN that contains 19 layers comprising of 16 convolution layers, 3 fully connected layer, 5 MaxPool layers and 1 SoftMax layer. This method uses spatial padding to preserve the image spatial resolution. The ResNet is the residual neural network that is designed to overcome the vanishing gradient problem and consists of group of convolutional layers together with skip connections followed by average pooling and finally ended up with fully connected output layer. In DenseNet, all the convolutional layers are connected directly with each other. Here, the input of every convolution layer consists of previous layer feature maps, and the output is directed to forthcoming layer, where the depth concentration of the image is aggregated with the feature maps. There is no intermediate ReLU nonlinearity associated with the modified depthwise separable convolution in the Xception CNN model. It is 71 layer deep CNN model that has pointwise convolution followed by a depthwise convolution.

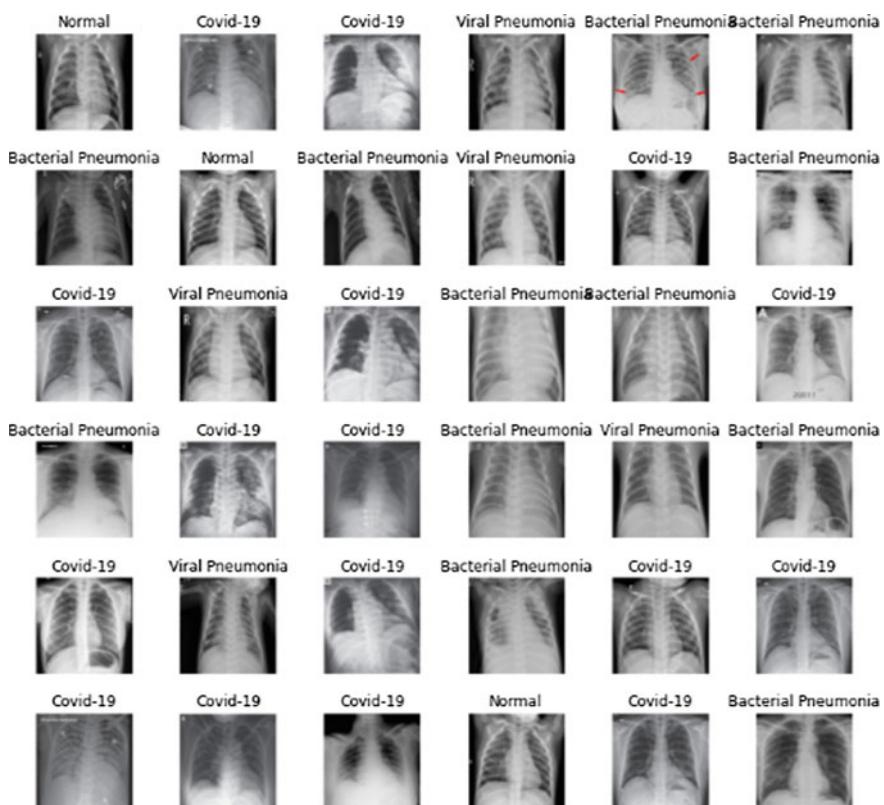
#### 3.2 Dataset Exploratory Analysis

The extracted dataset is fragmented with separated folders for testing and training data with chest X-ray images of folder containing 0: ‘normal,’ 1: ‘COVID-19,’ 2: ‘bacterial pneumonia,’ 3: ‘viral pneumonia’ and is shown in Fig. 3.

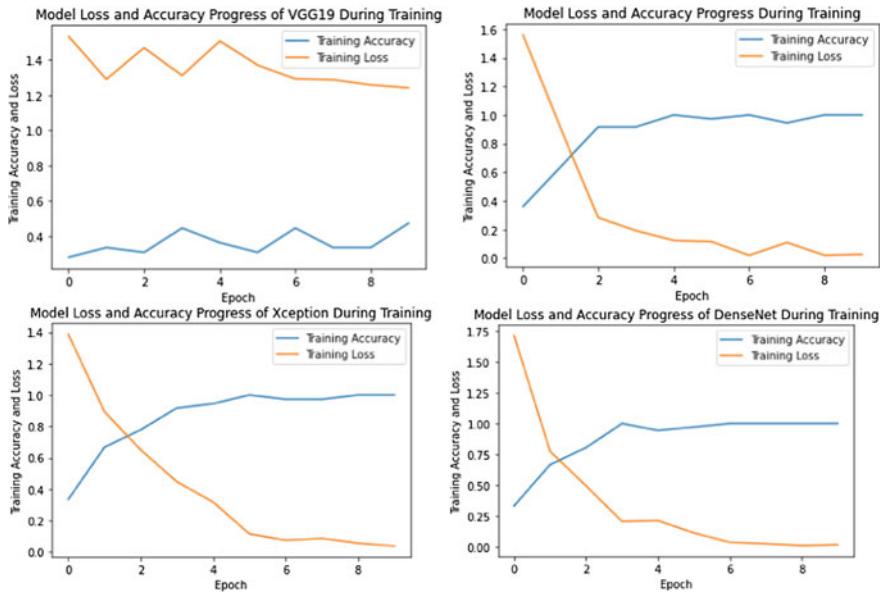
## 4 Results and Discussion

### 4.1 Implementation Setup

The dataset of chest X-ray images is fragmented with the required implementation and is arranged with separate folder for normal, COVID, viral infected and bacterial infected patients. The dataset is loaded, and the analysis is extracted to find the number of images. The image generator is used to generate tensor images data and to perform normalization process like shuffling and image resizing. The 20% of the dataset is used for cross-validation. The labeling of names is done for all the images in the training dataset, and the grid of images is displayed and is shown in Fig. 2.



**Fig. 2** Dataset information with 360 images belonging to four classes as 0: ‘normal,’ 1: ‘COVID-19,’ 2: ‘bacterial pneumonia,’ 3: ‘viral pneumonia’



**Fig. 3** Model loss and accuracy progress [top] (left) VGG (right) ResNet [bottom] (left) Xception (right) DenseNet

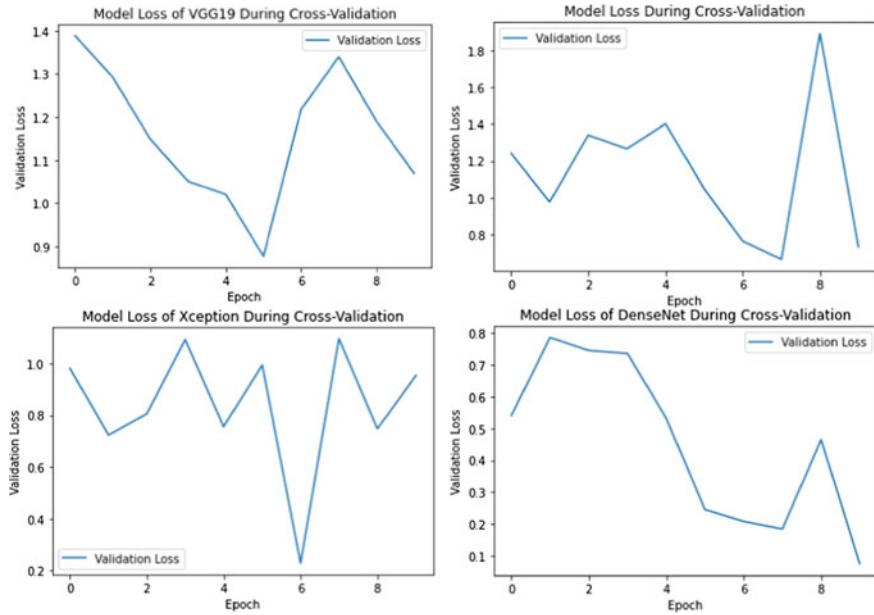
## 4.2 Performance Analysis

The base layer are designed with CNN models like VGG ResNet, Xception and Densenet. The base layer model is imported with the pre-trained weights, and the first layers are used to extract the high-level general features of the image. The last layers are used to perform the classification. Transfer learning is performed for the testing dataset. The base model is trained with specified CNN models with the pre-trained weights, and the same functionality is enforced for the newly given dataset to perform like the trained model.

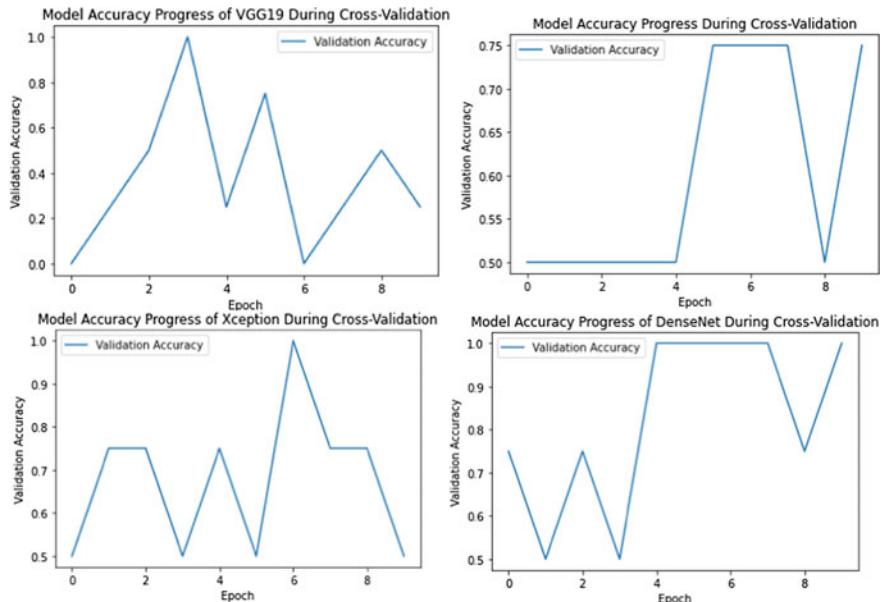
The model loss and accuracy progress of the VGG ResNet, Xception and Densenet CNN models is shown in Fig. 3. The model loss analysis of VGG ResNet, Xception and Densenet CNN models is shown in Fig. 4.

The model accuracy analysis of VGG ResNet, Xception and Densenet CNN models is shown in Fig. 5.

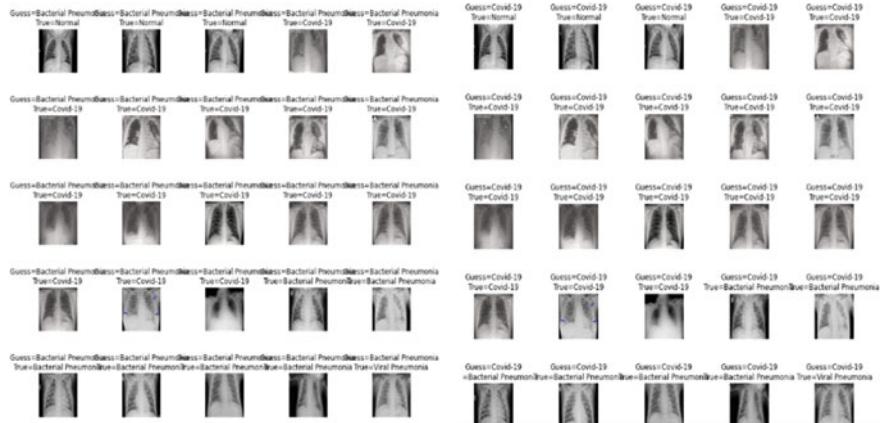
The Prediction details of the VGG ResNet, Xception and Densenet CNN models is shown in Figs. 6 and 7. The performance analysis is shown in Table 1.



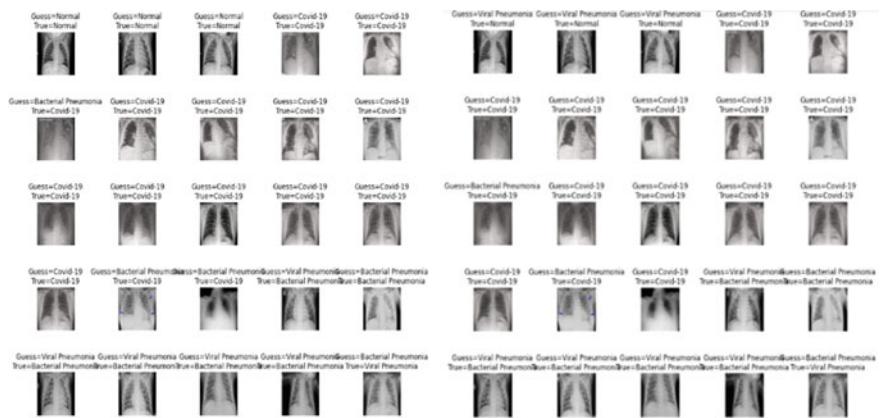
**Fig. 4** Model loss analysis [Top] (left) VGG (right) ResNet [Bottom] (left) Xception (right) DenseNet



**Fig. 5** Model accuracy analysis [Top] (left) VGG (right) ResNet [Bottom] (left) Xception (right) DenseNet



**Fig. 6** Prediction of CNN models (left) VGG (right) ResNet



**Fig. 7** Prediction of CNN models (left) Xception (right) DenseNet

**Table 1** Performance analysis of CNN model toward COVID prediction

Features	VGG	Resnet	Xception	DenseNet
Total params	20,024,384	23,587,712	20,861,480	18,321,984
Trainable params	20,024,384	23,534,592	20,806,952	18,092,928
Non-trainable params	0	53,120	92,528	229,056
Test accuracy	0.54838	0.633870	0.97838	0.831612
Precision	0.60	0.61	0.97	0.81
Recall	0.52	0.60	0.96	0.83
F-Score	0.55	0.58	0.92	0.82
Accuracy	0.51	0.59	0.95	0.80

## 5 Conclusion

This paper attempts to explore the chest X-ray image feature analysis by designing the convolutional neural network model as the base model. The last layers of the model are used to perform the classification task. The base models are trained with predefined weights, and the functionalist of the base model designed with VGG, ResNet, DenseNet and Xception is then transferred to the new target model to achieve the transfer learning. Experimental results show that Xception CNN model has the high precision of 0.97, recall of 0.96, F-Score of 0.92, accuracy of 95% and test accuracy of 97% toward prediction of COVID-19 patients.

## References

1. Aliman, S., Ismail, A., Rahmat, T.: Chest X-rays image classification in medical image analysis. *Appl. Med. Inf.* **40**, 63–73 (2018)
2. Asmaa, A., Mohammed, M., Abdelsamea, M.M.G.: Classification of COVID-19 in chest X-ray images using DeTraC deep convolutional neural network. *Appl. Intell.* 1–9 (2020)
3. Qin, C., Yao, D., Shi, Y., et al.: Computer-aided detection in chest radiography based on artificial intelligence: a survey. *BioMed. Eng.* **17**, 113 (2018)
4. Abiyev, R.H., Ma'aithah, M.K.S.: Deep convolutional neural networks for chest diseases detection. *J. Healthcare Eng.* 1–11. Article ID 4168538 (2018). <https://doi.org/10.1155/2018/4168538>
5. Tang, Y.-X., Tang, Y.-B., Peng, Y., Yan, K., Bagheri, M., Bernadette, A.R., Brandon, C.J., Lu, Z., Han, M., Xiao, J., Summers, R.M.: Automated abnormality classification of chest radiographs using deep convolutional neural networks. *npj Digit. Med.* **3**, 70 (2020)
6. Tobias, R.R.N.M.I., et al.: CNN-based deep learning model for chest X-ray health classification using tensorflow. In: 2020 RIVF International Conference on Computing and Communication Technologies (RIVF), Ho Chi Minh, Vietnam, pp. 1–6 (2020)
7. Yadav, S.S., Jadhav, S.M.: Deep convolutional neural network based medical image classification for disease diagnosis. *J. Big Data* **6**, 113 (2019)
8. Varshni, D., Thakral, K., Agarwal, L., Nijhawan, R., Mittal, A.: Pneumonia detection using CNN based feature extraction. In: IEEE International Conference on Electrical, Computer and Communication Technologies, Coimbatore, India, pp. 1–7 (2019)
9. Shi, F., Wang, J., Shi, J., Wu, Z., Wang, Q., Tang, Z., He, K., Shi, Y., Shen, D.: Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* (2020). <https://doi.org/10.1109/RBME.2020.2987975>
10. Sethy, P.K., Gehera, S.K., Ratha, P.K., Biswas, P.: Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. *Int. J. Math. Eng. Manag. Sci.* 643–651 (2020). <https://doi.org/10.33889/IJMEMS.2020.5.4.052>
11. Meraj, S.S., Yaakob, R., Azman, A., Rum, S.N.M., Nazri, A.S.A., Zakaria, N.F.: Detection of pulmonary tuberculosis manifestation in chest X-rays using different convolutional neural network (CNN) models. *Int. J. Eng. Adv. Technol. (IJEAT)* **9**(1) (2019). ISSN: 2249-8958

# Predicting Customer Loyalty in Banking Sector with Mixed Ensemble Model and Hybrid Model



Jesmi Latheef and S. Vineetha

**Abstract** Customer Relationship Management systems are used to enable organizations to acquire new customers, develop a continuous relationship with them, and increase customer retention for more profitability. Customer Loyalty is also known as Customer Churn. The main intention of churn prediction is to classify and find customers into churker and non-churker. A churned customer means there is more chance that the customer is about to leave the organization. So in order to find the churn customers will give more benefits to the organization. Thus, churn prediction can avoid the loss of revenue by retaining the existing customers. There are several techniques available with ensemble and hybrid models. This paper aims to predict customer loyalty in banking sector with a novel method named mixed ensemble model and hybrid model. Ensemble acts as a wrapper for group of machine learning or deep learning methods. This paper proposes two methods to predict customer churn using ensemble method with a mixed group containing XGB Classifier, LightGBM Classifier, and MLP model. And also build a hybrid model with the combination of Multilayer Perceptron (MLP) model and Convolutional Neural Network (CNN) model. Toward this, churn data of banking sector is used and build the systems then compare the performance of two. Thus, the system with more accuracy is termed more useful for organizations to find the customers with more chances to become churn. The results of experiments showed that the two proposed systems for churn prediction perform with an accuracy of 86% to 87%.

## 1 Introduction

Organizations in the competitive market mainly believe in the gains which come from their valuable customers. So CRM (Customer Relationship Management) includes

---

J. Latheef (✉) · S. Vineetha  
Rajiv Gandhi Institute of Technology, Kottayam, India  
e-mail: [jesmilatheef@gmail.com](mailto:jesmilatheef@gmail.com)

S. Vineetha  
e-mail: [svineetha@rit.ac.in](mailto:svineetha@rit.ac.in)

the concepts like customer acquisition, maintenance, and satisfaction. As from the previous studies, it is clear that acquiring new customer is more expensive than maintaining the old customers. Customer churn is an elementary obstacle for enterprises and it is interpreted as the mislaying of customers because they move out to other services. By predicting customer churn behavior in advance, it gives a high valuable insight in terms of revenue in order to retain and increase their customer base. To minimize customer churn, the company or organization should be able to predict the conduct of customer correctly and establish relationships between customer attrition and keep factors under their control.

Moreover, Churn prediction is a task of binary classification, which differentiates churners from non-churners. In order to reduce customer churn rate, one must follow the steps, and thus, loss of revenue can also be reduced. Hence, the aim of customer churn prediction is to detect customers with high tendency to leave. For that, the industry must know the reasons, which can be reflected from the gathered data. There are several attempts existing for churn prediction in banking sector using machine learning techniques like unsupervised, semi-supervised, and supervised. Deep learning models can also be very effective for this job [1, 2]. In this paper, a mixed ensemble model was used from the family of Ensemble algorithms and a hybrid model is presented for a more accurate prediction of customer churn. Therefore, in this paper, two different combination methods are examined in terms of customer churn prediction.

The remainder of this paper is organized as follows. Section 2 reviewed the literature related to customer churn and modeling technique used in this paper. Section 3 describes the system architecture, and Sect. 4 presents the implementation and experimental results. Finally, conclusion is provided in Sect. 5.

## 2 Literature Survey

Machine learning techniques have been widely used for evaluating the probability of customer to churn. Most of the literatures are discussed various machine learning algorithms, deep learning algorithms, and a brief of the literatures is describing here.

### 2.1 Data Mining Techniques

In order to develop effective and accurate customer churn prediction model, many data mining methods have been used. Data mining is used for hidden, valid, and potentially useful pattern extraction in datasets. However, it is all about finding previously unknown relationships among the data. Also data mining can be used for prediction which involves using some variables or fields in the database to predict unknown or future values of other variables of interest. Data mining methods include classification, clustering, regression, and so on [3–13].

## 2.2 Ensemble Algorithms

Ensemble learning helps improve machine learning results by combining several models. This approach allows the production of better predictive performance compared to a single model. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking) [14–17].

## 2.3 Artificial Neural Network

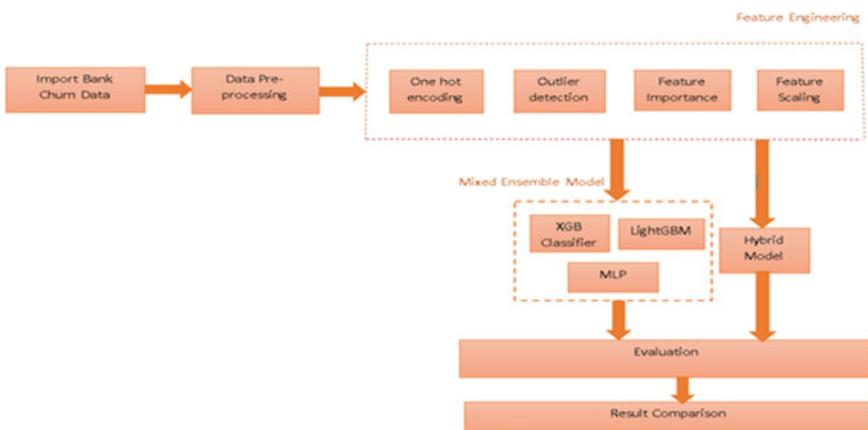
Predict customer churn in a financial institution using Multilayer Perceptron of Artificial Neural Network architecture. In [18] churn and non-churn customers were predicted using ANN and different performance metrics are used for finding the effectiveness of the system. MLP is an architecture of the artificial neural network that consists of multiple layers where each layer is fully connected with the next layer in a feed-forward direction. The first layer and the last layers represent the inputs (independent variables) and outputs (target variables) of the system, respectively. Connections between the nodes are represented as weights. The more the numbers of hidden layers in the network, the more the complexity of the network [19].

## 2.4 Hybrid Models

A hybrid model is termed as a combination of two or more basic models. There are several attempts on churn prediction using this hybrid model. In [20], it uses a combination of bagging, boosting and LOLIMOT algorithms and is named as Ordered Weighted Averaging (OWA) technique. Another hybrid method is presented that predicts customers churn more accurately, using data fusion and feature extraction techniques. Two algorithms, LOLIMOT and C5.0, were trained then the outputs of the individual classifiers were combined with weighted voting [21–23].

# 3 System Architecture

The proposed system architecture of churn prediction using mixed ensemble model and hybrid model are discussed. Figure 1 shows the entire architecture of the system. The system begins with importing the churn data of bank and the following steps are performing.



**Fig. 1** Overview of system architecture

### 3.1 Data Pre-processing

The first step is to clean and arrange the raw data. That is removal of null value, irrelevant feature removal and also find statistical summary of the dataset. In this step, identify and remove irrelevant features that effect the algorithm. Find any columns that have a single unique value. A feature with only one unique value cannot be useful for machine learning because this feature has zero variance. Thus, irrelevant feature can be regarded as noise whose presence will affect the final result adversely.

### 3.2 Feature Engineering

To improve the performance of the models that going to create, it is important to do feature engineering. In this step, preparing the dataset to make compatible with the machine learning algorithm requirements.

### 3.3 Model Creation

After completing the above steps the data is ready for modeling. In this paper, proposed a mixed ensemble algorithm and a hybrid model for customer churn prediction in banking sector.

As ensemble learning helps improve machine learning results by combining several models, that is try to ensemble both machine learning technique and deep

learning technique. Here, the deep learning technique used is specific to MLP (Multi-layer Perceptron). This approach allows the production of better predictive performance compared to ordinary ensemble model and all other single models. By this way, combined the predictions from multiple machine learning algorithms and one deep learning algorithm. Voting classifier is not an actual classifier but a wrapper for set of different ones that are trained and evaluated in parallel in order to exploit the different abnormalities of each algorithm. Here, combined different algorithms as ensemble and train the model, then to predict the final output. The final output on a prediction is taken by majority vote according to two different strategies of either soft voting or hard voting. In this model, soft voting is used. The multiple algorithms that are going to ensemble are of the following:

1. eXtra Gradient Boost (XGB)
2. LightGBM
3. Multilayer Perceptron (MLP)

Next model used is a kind of hybrid model. A hybrid model is simply integrating or combining different individual prediction models and that leads to overcoming certain limitations with the use single model and also improve the performance of the system. In this paper, integrating two or more deep learning models by cumulating the results of the models and predict the final output based on the result. Here, proposed to make a hybrid model with cumulating the outputs of Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN).

### ***3.4 Evaluation of Models***

Evaluation models are used to quantify model performance. In case of supervised learning problems, such as classification and regression, they focus on evaluation metrics. There are several evaluation metrics exist. Among them, confusion matrix and ROC curve-based evaluations are the most common methods. They are used for analyzing the performance of the created model.

## **4 Implementation and Evaluation**

The system is implemented in Colab platform provided by Google and Pyhton3 is used for implementation. The proposed model is deployed by selecting a dataset which should be simple and understandable, that contains information about customer information of banking sector and should be an adaptable dataset for churn prediction of banking sector. So importing a bank churn modeling dataset from Kaggle with 14 features and 10,000 records. The system starts with data pre-processing. In the pre-processing step, null values and irrelevant features were removed. Irrelevant features can be removed based on checking unique characters

**Table 1** Accuracy and AUC-ROC of different models with mixed ensemble model

Model	Accuracy	AUC-ROC
XGB	86	0.87
LightGBM	85	0.87
MLP	83	0.83
Mixed ensemble	86	0.87

of the corresponding category. In this work, checked for three unique characters and if the feature has greater than 3 unique characters, it is not relevant for the model and it should be removed. Here, to know how the correlation of features, Correlation Matrix is used. After understanding the trends and distribution of data, it is time to prepare the data for modeling, this is done by feature engineering.

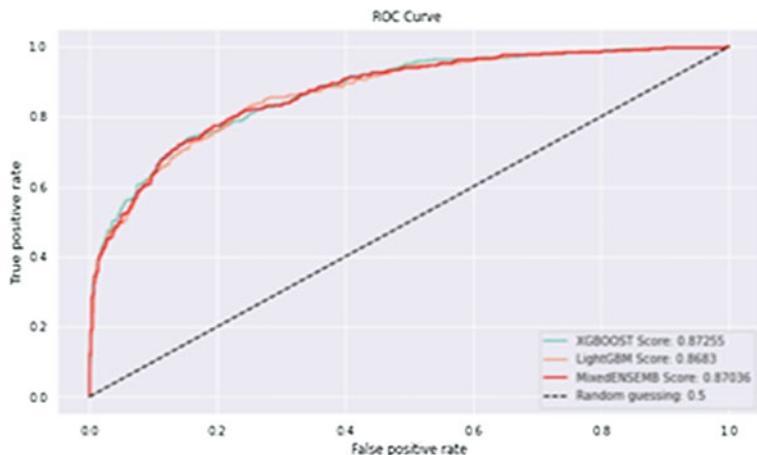
Next step is to make the numerical feature that is different in terms of ranges they are represented. After completing the data preparation next is the modeling phase. In the proposed system, modeling includes two modules. First, a mixed ensemble model was created and then implemented second model which is a hybrid model. Table 1 shows a comparison of individual model accuracies and mixed ensemble model accuracy. From the table, it is clear to imply that the mixed ensemble model predicts the class based on the probabilities. Hence, the overall accuracy of the mixed ensemble model 86%.

In order to increase the accuracy, the next model in our system is a hybrid model. Here, building a hybrid model with the integration of Convolutional Neural Network (CNN) and Multilayer Perceptron (MLP). CNN takes inputs in the form of matrices and vectors, and thus, converting the inputs into three-dimensional format will give a higher accuracy. In this model, output of CNN model is cumulating with the output of MLP model. Combing the model will get benefits of both models. Thus, the model is trained and validated, and it shows an accuracy of 87%.

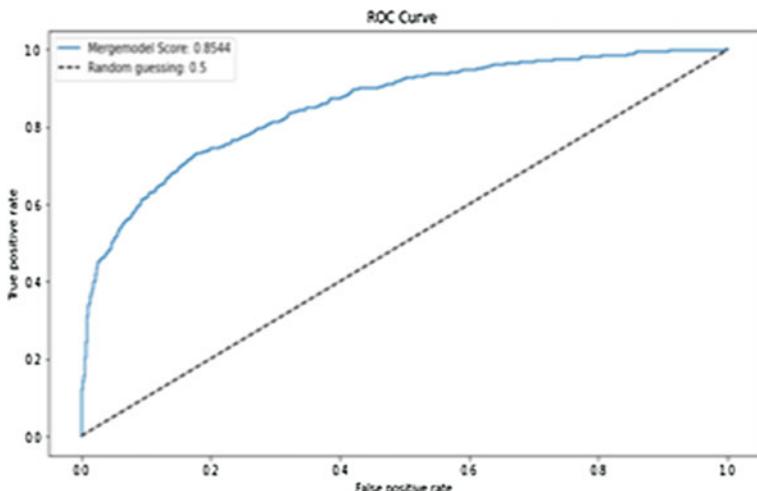
To evaluate the system confusion matrix and AUC—ROC curves were used. Confusion matrix is a summary of prediction results of the classification problem. Classification rate or accuracy of the system can be evaluated from the confusion matrix. In AUC—ROC curves, the two parameters true positive rate (TPR) and false positive rate (FPR) are plotted. Figure 2 shows the ROC curves of individual models and mixed ensemble model. In the figure, the area under highlighted (red) color curve shows the performance efficiency of the ensemble algorithm. Figure 3 shows the AUC-ROC curve of the hybrid model, and it points out the measure of model performance.

## 5 Conclusion and Future Work

In this system, the churn data of banking sector is used to predict whether the customer is churned or not by using voting classifier of mixed ensemble method. Voting classifier is one of the most powerful techniques for wrapping the methods, because in



**Fig. 2** ROC curves of individual models and mixed ensemble model



**Fig. 3** ROC curves of individual models and hybrid model

case of a classification problem in machine learning, it is really good to make use of multiple models before taking any decision. The main advantage of ensemble is, if the data is dynamic then the accuracy of different learning algorithms may vary. So by using ensemble technique, it will consider the predictions of collection of algorithms and take decisions based on that. Deep learning techniques are also more accurate for these kinds of problems. Hence, the novel method of mixed ensemble model have greater performance than ordinary ensemble models. Along with this model, a hybrid model is created, and the experimental results indicate that the hybrid model outperforms other models in terms of prediction accuracy. Therefore, it concluded

that the hybrid model by combining CNN and MLP can perform better than the baseline models and the mixed ensemble model. In addition, the CNN + MLP hybrid model performs more stable than the other models. Thus, customer loyalty or churn can be effectively find using mixed ensemble modeling and hybrid model.

For future work, the models can be extended to other prediction techniques such as fuzzy, genetic algorithms, and ANFIS. Also new aspects of experimental tuning can be introduced such as multiple objective optimizations, explanatory variables selection, transformation process, and also other evaluation measures.

**Acknowledgements** I would like to acknowledge the contribution and support from the Computer Science and Engineering Department of Rajiv Gandhi Institute of Technology, Kottayam.

## References

1. Sabbeh, S.F.: Machine-learning techniques for customer retention: a comparative study. *Int. J. Adv. Comput. Sci. Appl.* **9**(2) (2018)
2. Khan, A.A., Jamwal, S., Sepehri, M.M.: Applying data mining to customer churn prediction in an Internet Service Provider. *Int. J. Comput. Appl.* **9**(7) (2010)
3. Dalvi, P.K., Khandge, S.K., Deomore, A., Bankar, A., Kanade, V.A.: Analysis of customer churn prediction in telecom industry using decision trees and logistic regression. In: 2016 Symposium on Colossal Data Analysis and Networking (CDAN), Indore, pp. 1–4 (2016)
4. Wu, L., Li, M.: Applying the CG-logistic regression method to predict the customer churn problem. In: 2018 5th International Conference on Industrial Economics System and Industrial Security Engineering (IEIS), Toronto, ON, pp. 1–5 (2018)
5. Bharadwaj, S., Anil, B.S., Pahargarh, A., Pahargarh, A., Gowra, P.S., Kumar, S.: Customer Churn prediction in mobile networks using logistic regression and multilayer perceptron (MLP). In: 2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT), Bangalore, India, pp. 436–438 (2018)
6. Islam, M., Habib, M.: Data mining approach to predict prospective business sectors for lending in retail banking using decision tree. *arXiv preprint arXiv.1504.02018* (2015)
7. Kumar, G.R., Tirupathaiah, K., Krishna Reddy, B.: Client Churn prediction of banking and fund industry utilizing machine learning techniques. *IJCSE* **7**(6), 842–846 (2019)
8. Kumar, A.S., Chandrakala, D.: An optimal churn prediction model using support vector machine with Adaboost. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2**(1), 225–230 (2017)
9. Mahajan, D., Gangwar, R.: Improved customer Churn behaviour by using SVM. *Int. J. Eng. Technol.* 2395–0072 (2017)
10. Kumar, S., Viswanandhne, S., Balakrishnan, S.: Optimal customer Churn prediction system using boosted support vector machine. *Int. J. Pure Appl. Math.* **119**(12), 1217–1231 (2018)
11. ApurvaSree, G., Ashika, S., Karthi, S., Sathesh, V., Shankar, M., Pamina, J.: Churn prediction in Telecom using classification algorithms. *Int. J. Sci. Res. Eng. Dev.* **5**, 19–28 (2019)
12. Prajapati, D., Dubey, R.K.: Analysis of customer Churn prediction in telecom sector using random forest.
13. Jaisakthi, S.M., Gayathri, N., Uma, K., Vijayarajan, V.: Customer Churn prediction using stochastic gradient boosting technique. *J. Comput. Theor. Nanosci.* **15**(6–7), 2410–2414 (2018)
14. Mishra, A., Reddy, U.S.: A comparative study of customer churn prediction in telecom industry using ensemble based classifiers. In: 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, pp. 721–725 (2017)
15. Idris, A., Khan, A.: Prediction system for telecom using filter wrapper and ensemble classification. *Comput. J.* **60**(3), 410430 (2017)

16. Vijaya, J., Sivasankar, E.: Computing efficient features using rough set theory combined with ensemble classification techniques to improve the customer churn prediction in telecommunication sector. *Computing* **100**(8), 839–860 (2018)
17. Wang, X., Nguyen, K., Nguyen, B.P.: Churn prediction using ensemble learning. In: Proceedings of the 4th International Conference on Machine Learning and Soft Computing (2020)
18. Amuda, K.A., Adeyemo, A.B.: Customers Churn prediction in financial institution using artificial neural network. *arXiv preprint arXiv: 1912.11346* (2019)
19. Khan, Y., et al.: Customers Churn prediction using artificial neural networks (ANN) in Telecom Industry. Editorial Preface from the Desk of Managing Editor 10.9 (2019)
20. Basiri, J., et al.: A hybrid approach to predict churn. In: 2010 IEEE Asia-Pacific Services Computing Conference. IEEE (2010)
21. Hemalatha, P., Amalanathan, G.M. (2019). A hybrid classification approach for customer churn prediction using supervised learning methods: banking sector. In: 2019 International Conference on Vision Towards Emerging Trends in Communication and Networking (ViTECoN) (pp. 1–6). IEEE, Mar 2019
22. Tsai, C.-F., Lu, Y.-H. (2009). Customer churn prediction by hybrid neural networks. *Exp. Syst. Appl.* **36**(10), 12547–12553 (2009)
23. Jamalian, E., Foukerdi, R.: A hybrid data mining method for customer churn prediction. *Eng. Technol. Appl. Sci. Res.* **8**(3), 2991–2997 (2018)

# Design Patterns and Microservices for Reengineering of Legacy Web Applications



V. Dattatreya, K. V. Chalapati Rao, and M. Raghava

**Abstract** Financial management system (FMS) is a software application that enables any organization to manage its income and expenses with the goal of maximizing profits and making the system long-lasting. FMS performs business activities like invoicing and billing with minimal accounting errors. In this research work, we have taken up Andhra Pradesh Government legacy FMS package as a case study for the diagnosis of its rudimentary code and its characterization for the poor performance. Our investigation revealed various design anomalies that are spread across multiple modules of FMS. In order to fix the performance problems, we have adapted the Mikado graphs-driven sprint methodology and reengineering approach to redesign the codebase. The fundamental goal of this research is to identify and incorporate suitable design patterns to improve the reliability and extendability parameters. The other major contribution of this research work is about modernizing the FMS through refactoring and restructuring. Strangler fig pattern-driven microservices architecture is implemented, and the resulting product is referred to as pattern-oriented FMS(PoFMS). The solution offered by this research work reduces the CAPEX and OPEX components of the FMS project and improved its business capabilities.

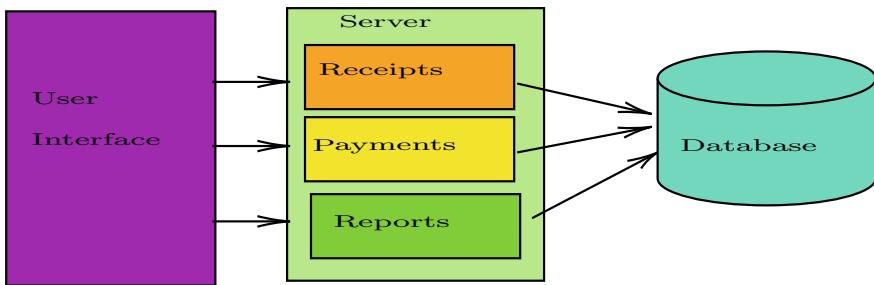
## 1 Introduction

Many business organizations are still driving their operations through Web applications developed based on outdated technologies and older versions of software libraries. Legacy systems have large codebase [1] and clumsy functionalities. A small modification in one module may lead to possible unknown changes in other modules that end up with bugs and erroneous results. Most of the legacy systems do

---

V. Dattatreya · K. V. Chalapati Rao · M. Raghava (✉)  
CVR College of Engineering, Hyderabad, India  
e-mail: [raghava.m@cvr.ac.in](mailto:raghava.m@cvr.ac.in)

V. Dattatreya  
e-mail: [v.dattatreya@cvr.ac.in](mailto:v.dattatreya@cvr.ac.in)



**Fig. 1** Legacy FMS

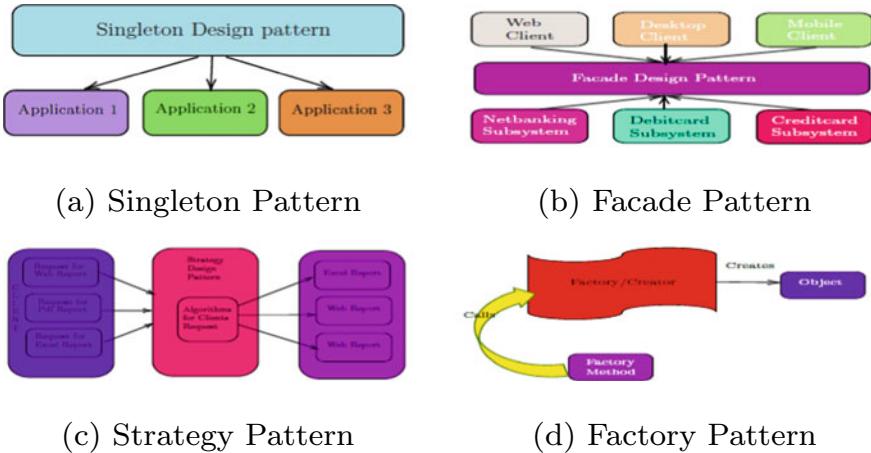
not even have proper design and documentation and operate under special environment. Moreover, such legacy systems do not encourage diversified data persistence models. However, conversion of legacy system data into the most recent data types is also time consuming and almost impractical. Software systems often depend upon third-party APIs to feature the latest capabilities. Thus, migrating a legacy system that does not live up to the expectations of the modern requirements of the stakeholders essentially demands an intelligent methodology and a disciplined approach. In addition, the end product must be built from small codebase and must offer a luxury to include future functional and architectural reorganizations. Reengineering plays a significant role in transformation of legacy systems. Nevertheless, the legacy software systems have played a role in catering to business requirements for various industry sectors over the past few decades, and abrupt deprecation of them leads to financial losses to the organization. Hence, reengineering of legacy software which consumes smaller time-frames is critical for a success. As opposed to discarding the existing software, a solution must be identified that allows us to fit the existing software into the new system and slowly replace the old system. In this context, reengineering plays a significant role in transformation of legacy system. Reengineering is defined as examining the existing software in order to figure out its specifications, with subsequent modifications to develop a new version of the improved software. In this work, we present the diagnosis of legacy code and apply intuitions into reengineering the codebase and refactoring of user interfaces. The combined effect of design patterns with agile methodology helped us to redesign and develop financial management system (FMS)—an enterprise application, within shorter times. This FMS is currently deployed on Andhra Pradesh Comprehensive Financial Management System (APCFMS). Legacy APCFMS application, shown in Fig. 1, had been developed to implement the client–server model comprising three modules, namely receipts, payments and reports. In all the three modules, there is a communication between client and server in turn with the database and resulting in the 3-tier architecture. Each module establishes an independent database connection to pull and aggregate the data and displays the response results. The following sections detail the pitfalls and functional vagaries of the system and address them through the proposed combined approach of design patterns and agile methodologies.

## 2 Performance Scaling Using Singleton Pattern-Based Reliable Database Connections

Designing a Web application typically uses a database connection object. In Legacy Andhra Pradesh Financial Management Systems (APFMS), the department-specific applications that are accessing the associated data models could create independent database connections on to different data models. In a regular scenario, multiple applications could connect to the same database technology with their own database schema in isolation leading to memory overhead and reduced system performance. This design overhead is addressed by implementing singleton pattern at database connection level which can be shared by multiple applications. The singleton pattern [2] is a design pattern that controls the creation of a database connection object. Thus, FMS ensures that one instance of a singleton database connection object is shared by multiple applications and implemented in a single sprint, by using private constructor and synchronized method. The role of singleton pattern is to create a database instance that can be shared by all the Web pages designed in a codebase. In order to implement this pattern, the connection manager first checks whether the connection pointer is null or not. If it is found to be null, then only it creates database object on the fly by using private constructor. This design philosophy is referred to as lazy instantiation and ensures a reliable database connection for each application. This is shown in Fig. 2a.

## 3 Providing Unified Interface for Multiple Remittance Options Using Facade Pattern

Legacy FMS allowed only one remittance option through bank. The recent e-commerce standards expect multiple payment options such as net banking, credit card and debit card through a well-defined payment gateway. The code redesign is implemented through a facade pattern with unified interface [3] for different options for payment systems. The users can select any one of the payment options, and based on the payment option, the corresponding interface appears. After completion of the task, we process the payment details, and the amount is credited into APCFMS account and is communicated to user as a confirmation of payment. Facade design pattern follows the principle called composition rule/compose programs in order to connect to other programs. The architectural diagram shown in Fig. 2b highlights the seem identification and implementation of new facade pattern.



**Fig. 2** Adapted patterns into FMS

#### 4 Saving the Result Sets in Various File Formats Using Strategy Pattern Ensuring Extendibility

FMS was lacking reporting tools to generate reports for a specific purpose in different formats. We designed a reporting tool for PFMS containing three types of formats, namely Excel sheets, PDF files and Web forms by applying strategy pattern. Considering first Web view, the database results based on the user's request are displayed in HTML file with database fields as column names in HTML Table, and the database records are displayed as rows. Next one is PDF view, in which we generate a general PDF template using PDF engine. The SQL query results are inserted into PDF file and displayed as PDF document in Web browser. In Excel sheet, the user selects some columns, and based on his selection, we retrieve data from the database and create new Excel file using Excel engine. This engine dynamically creates a general template and inserts the results into Excel sheet and displays it on the Web browser with open and download options. Strategy pattern [4] for PFMS is shown in Fig. 2c.

#### 5 Keeping Only One Database Active Using Factory Method Ensuring Extendibility

The Web application accesses various database tables by executing SQL scripts, stored procedures and retrieves the results in the form of records. Subsequent to handling race conditions, the resulting inconsistencies in the state of database factory methods are implemented while creating database objects. This pattern helped to maintain one database connection active, in the context of the application irrespec-

```

package receiptsmodule;
import java.sql.Connection;
import java.sql.DriverManager;
import java.sql.ResultSet;
import java.sql.SQLException;
import java.sql.Statement;
import java.util.ArrayList;
//This class defines the implementation of Singleton Design Pattern
public class ReceiptsDB {
    private static ReceiptsDB receiptsDB=null;
    private ReceiptsDB() {
    }
    synchronized static ReceiptsDB getObject() {
        if(receiptsDB==null){ receiptsDB=new ReceiptsDB(); }
        return receiptsDB;
    }
    final String DB_URL=
    "jdbc:postgresql://127.0.0.1:5432/
    test_receipts";
    ...
}

final String User="postgres";
final String Password="password";
public Connection getConnection(){
    Connection connection = null;
    try {
        Class.forName("org.postgresql.Driver");
        connection = DriverManager.
        getConnection(DB_URL,User,Password);
        >new SQLException e)
    } catch (Exception e) {
        e.printStackTrace();
    }
    return connection;
}

```

(a) Singleton Pattern Java Code

```

code snippet for Factory Design Pattern
public class FactoryDB {
    ...
    public static ReceiptsDB receiptsDB;
    public static ReportsDB reportsDB;
    public static PaymentsDB paymentsDB;
    public static Object getByNameOrID(String database) {
        Object obj=null; if(database.equals("receipts")){
            ...
        }
        else if(database.equals("reports")){
            ...
        }
        else if(database.equals("payments")){
            ...
        }
        return obj;
    }
}

```

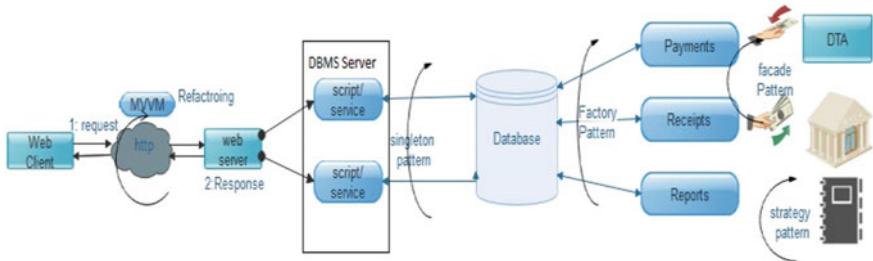
(b) Factory Pattern Java Code

Fig. 3 Adapted design patterns Java code

tive of database objects created. The factory design pattern [5] generates a generic database object. The database factory class is to instantiate the specific database connection object. Factory design pattern instantiates the concrete database provider dynamically based on application, thereby ensuring extendibility. The factory method design pattern is shown in Fig. 2d. Singleton design pattern Java code and factory design pattern Java code are shown in Fig. 3a, b.

## 6 Integration of Agile Programming Into Pattern-Oriented FMS

Combining agile programming and design patterns results in a synergy with their integration being more effective than the sum of the two. One of the essential steps of agile methodology is refactoring [6]. The refactoring feature supports the developers without modifying the external behaviour, but it improves the overall performance of the application by changing the internal behaviour. It is a way to clean up the

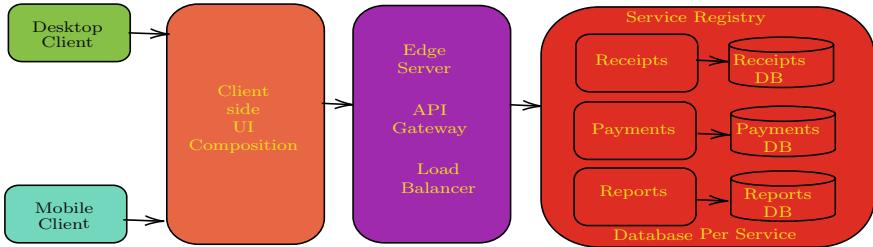


**Fig. 4** Mikado graph for FMS

code that reduces the occurrence of bugs. A good design comes first and then coding. Refactoring is contrary to this approach. With refactoring, we can find that the balance of work changes in terms of interfaces. The ensuing interaction gives rise to a program with a design that produces good results as development progresses. In APFMS, the rudimentary user interface implemented using legacy software libraries is refactored using Mikado graphs [7]. In receipts home page, all the fields related to departmental codes are replaced with a string, and all the field data are separated by using regular expressions. This is shown in Fig. 4.

## 7 Microservices

Microservices are autonomous and small, and their purpose is clearly defined. Microservices are very popular architectural patterns nowadays as a viable alternative to monolith applications. In microservices, the software application is divided into components that can be developed, tested and deployed independently [8]. Since they communicate via messaging, they are not dependent on the same programming language. So, developers can use the programming language that they are most familiar with and they can communicate with each other without difficulty. Since teams are working on smaller applications and are more focused on specific problem domains, their projects tend to be more agile. The development teams can iterate faster, address new features on a shorter schedule, add new features and turn around bug fixes immediately. Another advantage of microservice is to deploy any service independent of others. This feature leads to faster implementation of new features. A further advantage is scalability, since new functionality can be added on demand. Finally, easy maintainability is also another feature for microservices, since developers can modify or replace a microservice with a new one without effecting the entire application.



**Fig. 5** Microservices for PoFMS

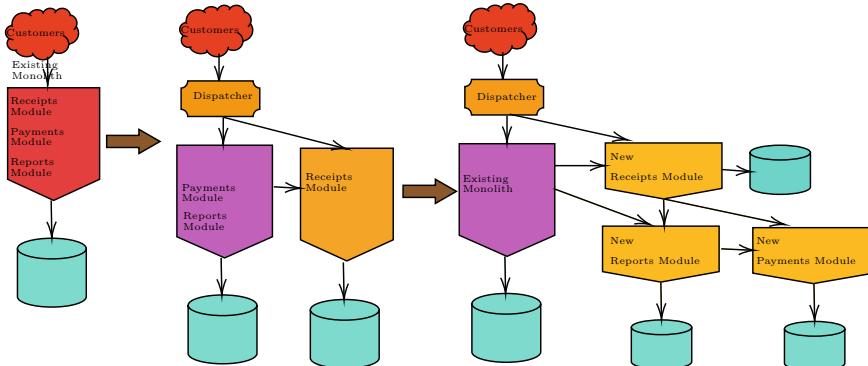
## 8 Migrating Legacy Financial Management Systems into Modernization

Let us consider PoFMS application that takes care of receipts from customers, payments to all the pensioners and generating reports. The application consists of several components including the user interface (UI) for customers, along with some backend services such as withdrawing the amount from their accounts. In PoFMS, there is no separate mechanism for write and read operations in a synchronized manner to update the reports module. To achieve this, we use some set of microservices. The flow diagram of PoFMS microservice architecture is shown in Fig 5.

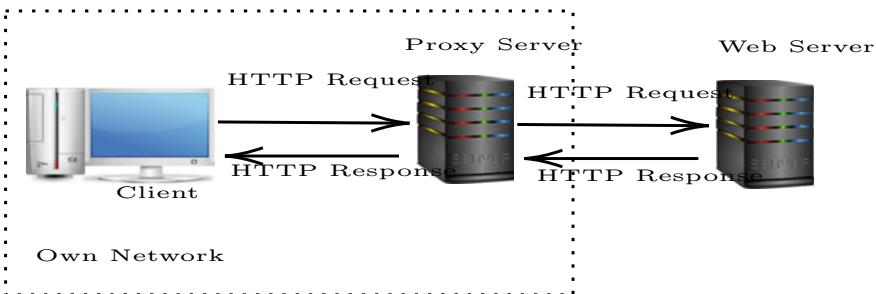
## 9 Strangler Fig Application

When the system has a technique of using rewritings of data frequently, is called strangler fig application [9]. Martin Fowler proposed this pattern, inspired by a some type of fig that seeds itself in the upper branches of trees. The fig then grows subside towards the ground to take root, gradually wrapping the original tree. The existing tree becomes originally a support structure for the new fig, and if taken to the final stages, the original tree will be losing its life, leaving the new fig. The new fig is now self-supporting fig in its place. In the context of modern Enterprise applications, the objective here is to have a new system to wrap all the legacy functionalities into the features of modern application. The idea is that legacy and modern application can coexist, giving the new system time to grow and potentially replace the old system. The vital benefit to this pattern is that it supports by allowing incremental migration to a new system. We execute this methodology for our software, to not only take incremental steps, but also ensure that each step is simply reversible. We will consider the nature of the calls being made into the existing system. A protocol such as HTTP is very amenable to redirection. The strangler fig pattern is shown in Fig. 6.

The architecture of HTTP reverse proxy [10] is shown in Fig. 7.



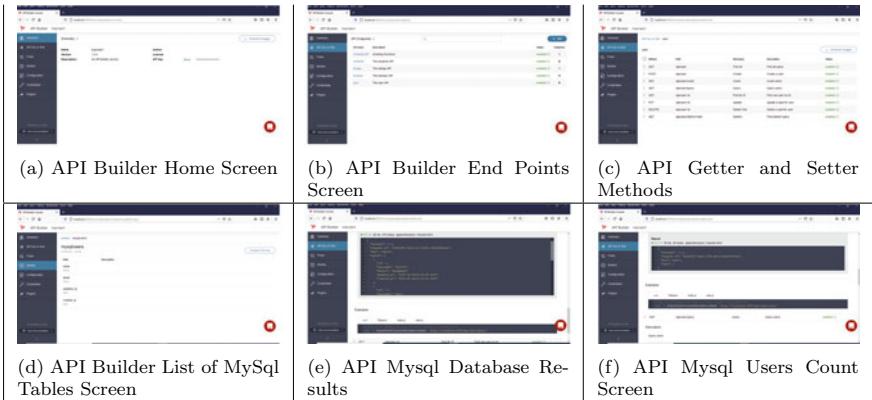
**Fig. 6** Strangler fig pattern



**Fig. 7** HTTP reverse proxy architecture

## 10 Results

The reengineering of FMS into a microservice-based application is realized by API Builder. The functional modules of PoFMS are separated into three different microservices of independent databases, and their interaction is significantly promoted by the API gateway. The main screen of API is shown in Fig. 8a. The API endpoints screen is shown in Fig. 8b. The API builder getter and setter for models is shown in Fig. 8c. The API Builder Mysql Tables are shown in Fig. 8d. API Database Table Count refers to the number of rows in a table as shown in the Fig. 8e. API Database Table Count the number of Rows in a Table is shown in Fig. 8f.



**Fig. 8** Microservices using API gateway

## 11 Conclusions and Future Work

In legacy FMS, the department-specific applications that are accessing the associated data models could create independent database connections on to different data models. Multiple applications could be connected to the same database technology with their own database schemas in isolation, thereby leading to memory overhead and reduced system performance. This design overhead is addressed by implementing singleton pattern. Thus, FMS ensures that one instance of a singleton database connection object is shared by multiple applications implemented by using private constructor and synchronize method. The factory design pattern generates a generic database object. The database factory class is to instantiate the specific database connection object. Legacy FMS was lacking reporting tools to generate reports for a specific purpose in different formats. We designed a reporting tool for APFMS containing three types of formats, namely Excel sheets, PDF files and Web forms by applying strategy pattern. Further, the server side business logic is reengineered with microservice architecture based on strangler fig pattern. The resulting application has got three services, namely receipts, reports and payments with their independent databases. In future work, a few more services such as event sourcing and cloud deployment are planned.

**Acknowledgements** I thank V.M.Rayudu, Senior Operations Manager, APCFMS for giving his immense support in implementing this project.

## References

1. Andhra Pradesh Comprehensive Financial Management Systems. <https://cfms.ap.gov.in/>
2. Lyon, D., Castellanos, F.: The parametric singleton design pattern. *J. Object Technol.* **6**, 13 (2007)
3. Schmidt, D.C: Wrapper facade, a structural pattern for encapsulating functions within classes. *CReport Magazine* (1999)
4. Christopoulou, A., Giakoumakis, E.A., Zafeiris, V.E., Soukara, V.: Automated refactoring to the strategy design pattern. *Inf. Softw. Technol.* **54**, 12021214 (2012)
5. Ireno, D.R.: Dynamic factoryNew possibilities for factory design pattern. In: 28th European Conference on Modelling and Simulation, 2730 May 2014. ISBN 978-0-9564944-8-1
6. Hodgetts, P.: Refactoring the development process: experiences with the incremental adoption of agile practices. In: Agile Development Conference (2019)
7. Birchall, C.: Re-engineering Legacy Software. Manning Publications, Shelter Island, NY (2016)
8. Lewis, J., Fowler, M.: Microservices. <https://martinfowler.com/articles/microservices.html> (2017)
9. Newman, S.: Monolith to Microservices-Evolutionary Patterns to Transform Your Monolith. O'reilly Media Inc., Sebastopol, CA (2019)
10. Yan, F., Wang, Y.: A security web gateway based on HTTP reverse proxy. *DEStech Trans. Eng. Technol. Res.* (2017)

# A Comparative Study on Single Image Dehazing Using Convolutional Neural Network



Poornima Shrivastava, Roopam Gupta, Asmita A. Moghe, and Rakesh Arya

**Abstract** In many computer vision tasks, dehazing is an important preprocessing step such as in various application areas of tracking and object recognition. Dehazing is the task of removing haze content or haze effect from the image under consideration. Nowadays, learning-based methods are used widely for dehazing. Among them, a convolutional neural network is one such method. After the successful use of CNN in different fields of image processing, it is being applied with modifications in various algorithms for dehazing as well. Deep CNNs are good at obtaining feature information automatically in a data-driven manner. In this paper, we have discussed various methods that use CNN for dehazing and their performance is measured by PSNR and SSIM.

## 1 Introduction

The weather conditions like snow, rain, mist, haze, hail that occur in day-to-day life play a major role in deteriorating the image quality [1]. These particles suspended in the atmosphere, either scatter the light or absorb the reflected air light. Due to this scattering, the outdoor images captured in foggy or hazy weather suffer from unsatisfactory visual effects like very low visibility distance, color paling of objects, and less contrast [2] of images as shown in Fig. 1. Many applications like remote sensing, driving assistance, visual surveillance smartphone camera, and autonomous

---

P. Shrivastava (✉)

Department of Electronics and Communication, UIT, RGPV, Bhopal, India

R. Gupta · A. A. Moghe

Department of Information Technology, UIT, RGPV, Bhopal, India

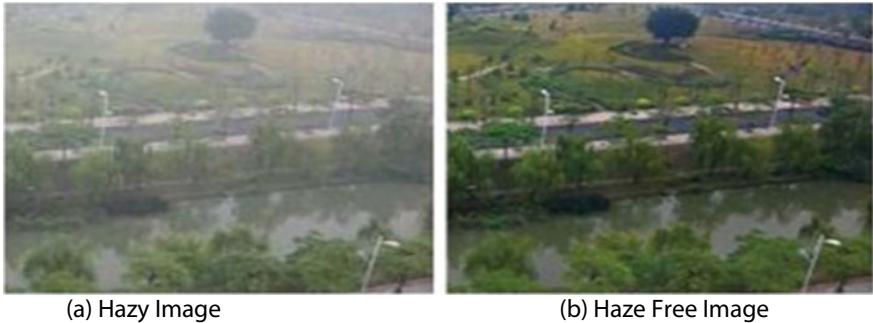
e-mail: [roopamgupta@rgtu.net](mailto:roopamgupta@rgtu.net)

A. A. Moghe

e-mail: [Aamoghe@rgtu.net](mailto:Aamoghe@rgtu.net)

R. Arya

Madhya Pradesh Council of Science and Technology, Bhopal, India



**Fig. 1** Deep hazy and haze-free image [6]

driving [1, 2] strictly require clear images and removal of haze that is why dehazing has become an important research area. For dehazing an image, the formation of a hazy image is taken into consideration so, in atmospheric optics [3], the problem of haze formation is deeply studied [4]. By the combination of the atmospheric light and scene radiance, a hazy image is formed [5]. As the scene depth keeps on increasing, more atmospheric light and less scene radiance are captured by the camera.

Researchers have proposed many dehazing methods in the past and based on problem formulation, these are categorized into (i) restoration-based approaches (ii) enhancement-based approaches, and (iii) fusion-based methods, some of the traditional image enhancement techniques are histogram equalization, Retinex algorithms, wavelet transform [2]. Due to haze, image contrast degrades, so the above methods try to improve it. In fusion-based methods, two images that are contrast enhancement and white balance type are derived for fusion and used when a hypothesis or prior is invalid [4, 7]. Due to the absence of a physical model and not considering the degradation process of the image, these methods do not remove fog completely. Hence, the restoration-based methods put forth to consider the degradation process, and by inverting it, a clean image is obtained [2].

Dehazing methods according to its features are classified into [a] single image [b] multi-images [1]. Earlier most of the restoration methods were based on reference models similar to physical or geometrical models which require information from user apart from additional information. These methods fail when wrong input is given by the user or when the information is inaccurate [7]. Other limitations are that it required additional information and tools and all these occur due to the lack of technology and hardware device [1]. So multiple image-based methods were proposed in which more than one ordinary camera is required and the methods under this are classified into three categories that are RGB/near-infrared images, different weather conditions-based, and different polarization degrees-based[7]. It is difficult to obtain all this additional information, and in practice, these methods are typical. Thus much work is done on haze removal on a single image due to its flexibility and practicality [2]. Most of the methods for single image fog removal are based on the atmospheric scattering model (ASM) [7] proposed by [8] shown in Eq. (1).

$$I(x) = J(x)t(x) + [1 - t(x)]A \quad (1)$$

where  $I(x)$  describes the image with haze content,  $J(x)$  is the radiance of scene, [7] which is to be retrieved from the foggy image that is  $I(x)$ , [5, 7] the haze removal process shown in Eq. (2) is derived from Eq. (1) [5]:

$$J(x) = I(x) - A(x)(1 - t(x))/t(x) \quad (2)$$

where  $A$  is considered as horizon radiance, it is the illumination created by atmospheric scattering,  $t(x)$  is its medium transmission, depth of scene point  $x$ , and it shows the amount of radiance of the scene that reaches at the camera without being dispersed, as shown in Eq. (3).

$$t(x) = e - \beta \cdot d(x) \quad (3)$$

where  $\beta$  is a scattering coefficient [2, 7]. “Dehazing processes can be further decomposed into three subproblems (1) Compute atmospheric light  $A(x)$ , (2) Forecast transmission  $t(x)$ , (3) Retrieve scene radiance  $J(x)$ ” [1, 4]. This model considers that hazy weather is the reason for the degradation of images. This model is a collection of mathematical equations with two parameters, i.e.,  $A(x)$  and  $t(x)$  [9]. Some of the methods of prior and learning-based used this type of model for dehazing.

## 2 Multiple Images-Based Dehazing Methods

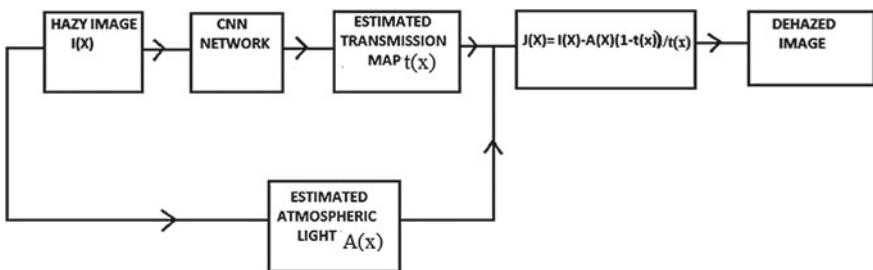
Some of the work in multiple image-based methods are Narasimhan [9] proposed a chromatic model of a numerous image approach in which two or more than two images of a similar type of scene in distinct haze situation are used to remove effects of weather, and it recovered the scene color by additional information provided by users under different weather conditions after investigating visual manifestations. Schechner [10] Schwartz [11] proposed a dehazing method based on polarization to estimate the depth of the image using two or more than two images taken from distinct angles by a polarizer. Chen [12] proposed a technique in which they estimated the depth of the image by using image pairs taken on a sunny and foggy day. Schaul [13] and Feng [14] restored an image by obtaining an infrared image and an RGB image at the same time. [7].

## 3 Single Image Dehazing by Prior-Based Methods

He et al. [15] put forward a dark channel prior (DCP) which declares that in a hazy image, there exists no less than one color channel for each pixel which should have a low value of intensity. Tang et al. [16] put forth a new method of machine learning for

haze removal in which a regression model gets trained by four multiple handcrafted features, i.e., local maximum saturation, local maximum contrast, dark channel, and a regression model for transmission prediction. Luan et al. [17] proposed a method that combines seven types of haze features by using the support vector regression (SVR). Galdran et al. [18] improved the scene contrast by an extended perception-inspired variational framework. Fattal [19] put forth a method for clear images that utilized local color lines prior to independent component analysis (ICA). Berman et al. [20] worked on a non-local haze-line prior technique. Zhu et al. [21] worked on color attenuation prior to single image dehazing. Kratz and Nishino [22] estimated the transmission by assuming that depth and albedo are statistically independent and formulated a factorial Markov random field. Sulami et al. [23] estimated an appropriate global constant atmospheric light vector by applying the color lines prior. Wang and Fan [24] approached to mix details of multi-level chromaticity priors by proposing a method called multiscale depth fusion (MDF) with “local Markov regularization” [1]. Meng et al. [25] dealt with the boundary constraint to efficiently remove the haze contextual regularization. All these priors have been used for the estimation of the transmission coefficient for dehazing and for achieving impressive performance [1]. Some images often do not meet for each prior, and some hazy images, it is not sufficient to capture intrinsic attributes and also have limitations like “halos, distortion of colors, underestimating the thickness of haze content, etc. [5].

System framework of Fig. 2 gives an idea about various methods based on CNN’s basic dehazing method. It mainly includes the following steps. (1) Take a hazy image and then obtain global atmospheric light from it, (2) Build a modified network using the CNN model, (3) Train the network, optimize the network parameters, for prior knowledge of image dehazing, (4) Use a model based on CNN, i.e., convolutional neural network for the estimation of transmission map, (5) For training purposes, build the data set required, (6) Input a natural close and long-range hazy image and test the standard of the output image. Finally, a clear image is recovered.



**Fig. 2** A de-hazed model by atmospheric scattering method by simple CNN-based method

## 4 Work on DeHazing on Deep Neural Networks

Several works on CNN have been done but they were not capable to capture the intrinsic attributes of hazy images [3] so modifications have been done in the CNN network to get the desired result. Some of the recent works done in this area are as follows.

Wang et al. [1] put forward an atmospheric illumination prior (AIP) network which is an image-to-image mapping, and it considered the luminance channel of YCrCb color space including a fusion network and dehazing network in which the dehazing network recognizes features of haze and restores the lacking details in luminance channel, and in fusing network, a different attribute of three channels is fused in the color space. According to them, the atmospheric illumination has a major effect on the luminance channel than on the chrominance channel but limited in a way that in YCrCb color space, unbalanced problems  $\Delta H(x)$  have not been considered. Ling [2] proposed a deep network using CNN for calculating the transmission map of three channels that automatically learn features of haze relevant and in a joint manner used a local patch color channels of RGB type and discover that the most informative haze pertinent features are three color channels and information of local spatial type. Huang et al. [5] put forward a hybrid model based on a single image that combines the supervised and unsupervised model called Deeptrans Map to get the correct transmission map. Unsupervised learning leads to the learning of the hazy features showing details of a hazy image that contains image color, brightness, and structure and is used for learning the extraction of features. With haze-relevant features of multiscale type, image details can be represented by modifying the number and size (i.e., scale) of invisible layers and transmission map is achieved through supervised learning but this method does not apply to unevenly distributed haze images and night time hazy images. You et al. [26] proposed CNN-RNN that is a convolutional neural network and recurrent neural network which acts as a bridge to connect coarse type-to-fine type module in which CNN takes control of the view of global, which is local-to-global and RNN captures the local view, which is global-to-local. Jinjiang et al. [27] proposed a deep CNN using a residual network where atmospheric light is not estimated for dehazing. It is subdivided into two network parts wherein the first part of the transmission map is estimated by using a foggy image as input, and in the second part, haze is removed by using a ratio of the hazy or foggy image and transmission map as the input. Wang et al. [28] put forward the “Deep Residual Haze Network (DRHNet), which is a generalized end-to-end dehazing model because it can be used for the de-raining purpose as well. Here the image is restored from haze by subtracting the foggy or image from the learned negative residual map. To effectively aggregate the contextual information, a context-aware feature extraction module is proposed. To accelerate its convergence and to improve its representation ability, a nonlinear activation function is proposed, which is called reverse parametric rectified linear unit (RPReLU). Li et al. [29] reconstructed the latent de-haze image through perception-inspired haze removal subnetwork that is deep CNN for a single image, and then in another subnetwork, refinement was done to improve the properties like

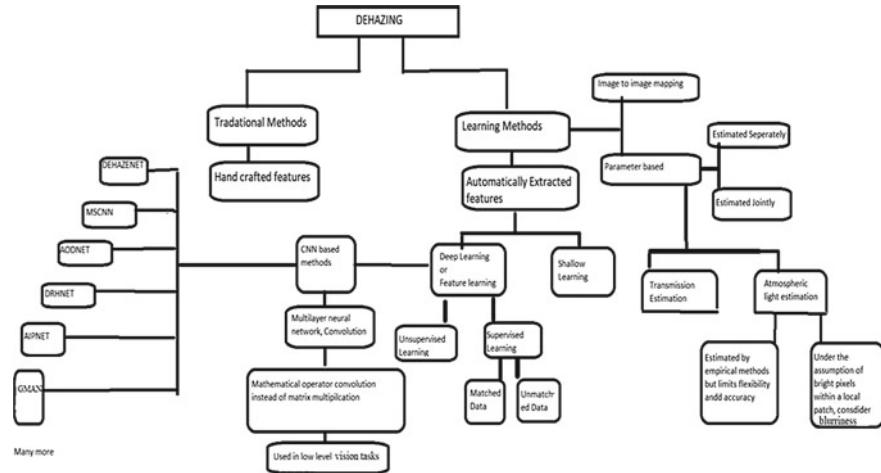
color and contrast of the outcome by optimizing the joint multiterm loss. Cameron Hodges et al. [30] proposed a CNN model which was trained from unmatched data with the help of Siamese network architecture to learn haze-relevant features in which two learning-based architectures, i.e., dehazing and discriminator subnetwork are combined, trained, and tuned for better dehazing performance and if the performance of dehazing network is good then discriminator will give low output.

Ren et al. [31] proposed a network that maps between images of haze content and their transmission maps. It is a multiscale deep or intense neural network consisting of two nets that are, a coarse-scale type net in which a holistic transmission map is calculated and another one is fine-scale type net which locally redefines the de-haze results and its edges. Liu et al. [32] proposed a non-optimized generic method for haze or fog removal from the image which is a simple CNN model based on residual learning because this technique does not need the knowledge of parameters of ASM that is transmission map  $t(x)$  and atmospheric light  $A(x)$  and still, it overcomes the limitations of excessive edge sharpening and color darkening. A CNN-based encoder-decoder network is used with residual learning on local and global levels. Xiao et al. [33] proposed a new haze layer-based model through end-to-end mapping in which a hazy image from the hazy layer is de-hazed to get a residual image using CNN based network. Yin et al. [34] proposed a new “parallel spatial/channel-wise attention block” for haze removal which is applied to the encoder’s end to guide the decoder and reconstruct clear images. The important spatial position of features is highlighted by the spatial attention module which is connected in parallel with the channel-wise attention module and with the help of residual learning, the model’s dehazing capability is further enhanced but the performance for night time hazy images was not satisfactory. Haouassi et al. [35] proposed a two-stage dehazing network in which the first stage labeled as a-Est calculates an exact atmospheric light algorithm that includes a quadtree decomposition and blurry effect due to hazy image and the second stage further consists of two subnetworks, in which one is used for computing rough transmission maps (CMCNNtr) and another one is for its refinement (CMCNNt). It considers atmospheric light along with transmission map estimation.

## 5 Discussion and Analysis

There are various applications of deep learning techniques in image processing like contrast-enhancing, deblurring, de-raining, enhanced resolution, and so on. Similarly, with the popularity of deep learning in the area of image denoising, semantic segmentation, object detection, and image classification/annotation, researchers tried to use CNN for dehazing using a single image.

Here various methods of dehazing have been discussed. Much emphasis has been given to learning-based single image dehazing methods. The analysis of methods and algorithms describe several advantages and limitations qualitatively and quantitatively. The platform used in these methods is MATLAB, PYTHON, TensorFlow,



**Fig.3** Different stages comes under dehazing

etc. Some of the deep learning methods are based on end-to-end dehazing techniques, i.e., mapping between a hazy image to a de-hazed image without estimating any additional parameters. Another one is based on the usage of additional parameters for the removal of haze. So deep neural network methods based on CNN that are discussed in Table 2 are categorized into two classes, and based on Fig. 3, many steps need to be considered for the choice of proper dehazing model. While using neural networks for dehazing purposes, some of the problems that occur in the original scene are the loss of structural integrity, and due to this, the value of SSIM decreases which leads to unrealistic visual results [36, 37].

Results from Table 1 in [28] shows that DRHNET [28] performs better than some prior-based methods and some popular deep or intense learning-based methods. It is based on residual learning in which mapping is quite straightforward. Prior-based methods that consider the physical model as a basis sometimes perform similar to learning-based methods, and their performance is comparable. Some algorithms in deep learning-based methods performed well in the restoration of a hazy image using CNN within the values of the parameter of PSNR, SSIM. This shows the sign of good restoration. Another method is an end-to-end generalized CNN which does not require knowledge of the ASM. Here the estimation of parameters is completely avoided and has the benefits of overcoming some of the common pitfalls, like darkening of color and excessive sharpening of edges but for improving performance further, various refinements can be made [32]. Multiscale convolutional networks (MSCNN) are trainable deep CNNs which are also an image-to-image system. It is considered a YCrCb color space that can automatically detect foggy regions, can restore the lacking texture information, lightweight system framework, and enhances visual contrast with the natural color of the scene but the unbalanced problem in YCrCb color space was not studied [1]. An alternative to generative adversarial

**Table 1** Result of quantitative evaluation in average PSNR (dB) and SSIM [28]

	DCP	BCCR	GRM	CAP	NLD	DEHAZENET	MSCNN	AODNET	GMAN	DRHNET
PSNR	18.87	17.87	20.44	21.31	18.53	22.66	20.01	21.01	27.94	31.39
SSIM	0.794	0.770	0.823	0.824	0.702	0.833	0.791	0.837	0.897	0.974

networks is the Siamese network (discriminator) which has also been discussed. The advantage with GAN-based method is that it removes more haze but fine details are not covered, like in parameter estimation of ASM. Methods based on feature learning mostly learn haze related features of low level [30]. All these networks are end-to-end type. Better results can be obtained from deep neural networks, and in this direction, there is much to be explored further.

## 6 Conclusion

Most of the techniques discussed in this paper remove haze by taking into consideration the calculation of the depth map and transmission map which are complex parameters to be computed properly. Due to this, the computational complexity and time of these algorithms increase. Some of the techniques under CNN-based model do not use any sort of parameter estimation, and they directly mapped the hazy image to the dehazy image. This has reduced computational complexity. This paper discusses the advantages and shortcomings of these dehazing methods both quantitatively and qualitatively. PSNR and SSIM are the comparative metrics used here. The analysis shows that some of the intense learning-based methods of single image haze removal technique are residual multiscale convolutional networks, trainable model end-to-end type, the system architecture of image-to-image, YCrCb color space considered [1] type, DRHNET, AIPNET, and AODNET performed in the direction of a better outcome in terms of PSNR and SSIM in the synthetic hazy image and sometimes in real images also.

## References

1. Wang, A., Wang, W., Liu, J., Gu, N.: Image-to-image single image dehazing with atmospheric illumination prior. *IEEE Trans. Image Process.* **28**(1), 381–393 (2019)

**Table 2** Recent of dehazing methods its advantages and disadvantages

Methods	Advantages	Limitations
Multiscale convolutional networks, trainable model end-to-end type, the system architecture of image-to-image, YCrCb color space considered [1]	Hazy regions identified automatically, deficient texture information restored, the lightweight system increases contrast, lacking texture information. scene intrinsic color is restored	Unbalance problem in YCrCb color space, more optimizations to be done
Three-channel transmission map [2]	Information of color and local spatial, video dehazing	The color difference from input frames

(continued)

**Table 2** (continued)

Methods	Advantages	Limitations
Hybrid model, combined learning of unsupervised and supervised [5]	Transmission estimation model's accuracy and robustness improved	There are still many problems to be further addressed
CNN-RNN, recurrent neural network[26]	Limited scenario predicts absolute visibility, performs well in predicting relative visibility for different situations, model accurately adjusted in case of less data where true visibility data are available less	Fine-tuning on the model of natural images is slightly effective due to the non-encoding of visibility information
Deep CNN, residual-based, atmospheric light value trained by the residual network, $t(x)$ through deep CNN.[27]	Atmospheric light estimation is avoided and improves the dehazing efficiency, no color distortion, image blur	To improve performance, the network training strength should be increased
Negative residual map, end-to-end model, context-aware feature extraction [28]	Used in both de-raining and dehazing	To handle white scenes, it does not consider particular parts
Deep end-to-end type, PDR-Net, i.e., perception-inspired dehazing subnetwork, new refinement subnetwork,[29]	Used in high-level vision problems	Should improve performance, accelerates network convergence
Trained from unmatched images, Siamese network architecture deep CNN model, to learn hazy features, end-to-end method[30]	Compare unmatched images for learning hazy features that are not changeable to some of the image variable like scene, lightening, and pose	Subjective testing by human observers, resources required to use these methods
End-to-end generalized model, CNN, ASM is not used, and estimation of the parameter is completely avoided [32]	Overcome some common pitfalls, like darkening of color, excessive sharpening of the edge	Optimize the design of the network is not done, other refinements can be made further
Residual images, end-to-end mapping, convolutional neural network, guided filter, learn directly residual image[33]	High learning rate, low computation, consume less computation time, speedy convergence process.	Should improve running speed for a common computing platform
An end-to-end model, parallel spatial/channel-wise attention block, pyramid pooling with encoder-decoder [34]	Good accuracy, visual results, competitive running time performance	Unsuitable for night hazy images
The two-stage system, i.e., “A-Est”, cascaded multiscale CNN[35]	Efficient estimation of AL and $t$ , accurate algorithm for ASM	

2. Ling, Z., Fan, G., Gong, J., et al.: Learning deep transmission network for efficient image dehazing. *Multimed. Tools Appl.* **78**, 213–236 (2019)
3. Timofeev, Y.M., Vasilev, A.V.: *Theoretical Fundamentals of Atmospheric Optics*. Cambridge International Science Publishing, 200
4. Song, Y., Li, J., Wang, X., Chen, X.: Single image dehazing using ranking convolutional neural network. *IEEE Trans. Multimedia* **20**(6), 1548–1560 (2018). <https://doi.org/10.1109/TMM.2017.2771472>
5. Huang, J., Jiang, W., Li, L., Wen, Y., Zhou, G.: DeeptransMap: a considerably deep transmission estimation network for single image dehazing. *Multimedia Tools Appl.* **78**, (2018). <https://doi.org/10.1007/s11042-018-6536-x>
6. Wang, W., Yuan, X.: Recent advances in image dehazing. *IEEE/CAA J. Automatica Sinica* **4**(3), 410–436 (2017)
7. Yunan, L., Miao, Q., Liu, R., Son, J., Quan, Y., Huang, Y.: A multi-scale fusion scheme based on haze-relevant features for single image dehazing. *Neurocomputing* **283**(2018), 73–86 (2018)
8. McCartney, E.J.: *Optics of the Atmosphere: Scattering by Molecules and Particles*. Wiley, New York (1976)
9. Narasimhan, S.G., Nayar, S.K.: Interactive (de)weathering of an image using physical models. In: *IEEE Workshop on Color and Photometric Methods in Computer Vision. Conjunction with ICCV* (2003)
10. Schechner, Y.Y., Narasimhan, S.G., Nayar, S.K.: Polarization-based vision through the haze. *Appl. Opt.* **42**(3), 511–525 (2003)
11. Shwartz, S., Namer, E., Schechner, Y.Y.: Blind haze separation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, pp. 1984–1991. IEEE, (2006)
12. Chen, G., Wang, T., Zhou, H.: A novel physics-based method for restoration of foggy day images. *J. Image Graph.* **5**(13), 887–893 (2008)
13. Schaul, L., Fredembach, C., Süsstrunk, S.: Color image dehazing using the near-infrared. In: *IEEE International Conference on Image Processing*, pp. 1629–1632 (2009)
14. Feng, C., Zhuo, S., Zhang, X., Shen, L., Süsstrunk, S.: Near-infrared guided color image dehazing. In: *IEEE International Conference on Image Processing*, pp. 2363–2367 (2013)
15. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp 1956–1963 (2009)
16. Tang, K., Yang, J., Wang, J.: Investigating have relevant features in a learning framework for image dehazing. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2995–3002 (June 2014)
17. Luan, Z., Shang, Y., Zhou, X., et al.: Fast single image dehazing based on a regression model. *Neurocomputing* **245**, 10–22 (2017)
18. Galdran, A., Vazquez-Corral, J., Pardo, D. et al.: A variational framework for single image dehazing. In: *Proceedings of European Conference on Computing Vision (ECCV)*, pp. 259–270 (Sep. 2014)
19. Fattal, R.: Dehazing using color-lines. *ACM Trans. Graphics* **34**(1), 13:1–13:14 (2014)
20. Berman, D., Treibitz, T., Avidan, S.: Non-local image dehazing. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1674–1682 (2016)
21. Zhu, Q., Mai, J., Shao, L.: A fast single image haze removal algorithm using color attenuation prior. *IEEE Trans. Image Process.* **24**(11), 3522–4353 (2015)
22. Kratz, L., Nishino, K.: Factorizing scene albedo and depth from a single foggy image. In: *IEEE International Conference on Computer Vision*, pp. 1701–1708 (Sept. 2009)
23. Sulami, M., Geltzer, I., Fattal, R., Werman, M.: Automatic recovery of the atmospheric light in hazy images. In: *IEEE International Conference on Computational Photography* (2014)
24. Wang, Y., Fan, C.: Single image defogging by multiscale depth fusion. *IEEE Trans. Image Process.* **23**(11), 4826–4837 (2014)
25. Meng, G.F., Wang, Y., Duan, J., Xiang, S., Pan, C.: Efficient image dehazing with boundary constraint and contextual regularization. In: *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, pp. 617–624 (2013)

26. You, Y., Lu, C., Wang, W., Tang, C.: Relative CNN-RNN: learning relative atmospheric visibility from images. *IEEE Trans. Image Process.* **28**(1), 45–55 (2019). <https://doi.org/10.1109/TIP.2018.2857219>
27. Li, J., Li, G., Fan, H.: Image dehazing using residual-based deep CNN. *IEEE Access* **6**, 26831–26842 (2018). <https://doi.org/10.1109/ACCESS.2018.2833888>
28. Wang, C., Li, Z., Wu, J., Fan, H., Xiao, G., Zhang, H.: Deep residual haze network for image dehazing and deraining. *IEEE Access* **8**, 9488–9500 (2020). <https://doi.org/10.1109/ACCESS.2020.2964271>
29. Li, C., Guo, C., Guo, J., Han, P., Fu, H., Cong, R.: PDR-Net: perception-inspired single image dehazing network with refinement. *IEEE Trans. Multimedia* **22**(3), 704–716 (2020). <https://doi.org/10.1109/TMM.2019.2933334>
30. Hodges, C., Bennamouna, M., Rahmani, H.: Single image dehazing using deep neural networks. *Pattern Recogn. Lett.* **128**, 70–77 (2019)
31. Ren, W., Pan, J., Zhang, H., et al.: Single image dehazing via multi-scale convolutional neural networks with holistic edges. *Int. J. Comput. Vis.* **128**, 240–259 (2020). <https://doi.org/10.1007/s11263-019-01235-8>
32. Xiao, J., Shen, M., Lei, J., Zhou, J., Klette, R., Sui, HaiGang: Single image dehazing based on learning of haze layers. *Neurocomputing* **389**(2020), 108–122 (2020)
33. Liu, Z., Xiao, B., Alrabeiah, M., Wang, K., Chen, J.: Single image dehazing with a generic model-agnostic convolutional neural network. *IEEE Signal Process. Lett.* **26**(6), 833–837 (2019). <https://doi.org/10.1109/LSP.2019.2910403>
34. Yin, S., Wang, Y., Yang, Y.-H.: A Novel Image dehazing Network with a Parallel Attention Block. *Pattern Recogn.* **102**, 107255 (2020)
35. Haouassi, S., Wu, D.: Image dehazing based on (CMTnet) cascaded multi-scale convolutional neural networks and efficient light estimation algorithm. *Appl. Sci.* **10**, 11 (2020)
36. Teixeira Gonçalves, L., de Oliveira Gaya, J.F., Lilles Drews Junior, P.J., da Costa Botelho, S.S.: GuidedNet: single image dehazing using an end-to-end convolutional neural network. In: 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 79–86. Parana (2018). <https://doi.org/10.1109/SIBGRAPI.2018.00017>
37. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* **13**(4), 600–612 (2004)

# ***Plasmodium falciparum* Detection in Cell Images Using Convolutional Neural Network**



**Smaranjit Ghose, Suhrid Datta, C. Malathy, and M. Gayathri**

**Abstract** Malaria is one of the major burdens to global health, which results in countless deaths every year. It is caused by a group of *Plasmodium* parasites which spreads through the bite of the female anopheles mosquito. The infected mosquito first bites the host, and the parasite enters the bloodstream which proceeds to go to the liver. From the liver, the parasites grow and multiply in the red blood cells. The infected red blood cells eventually burst and release more parasites. The diagnosis of malaria is done by doing a blood test where the count of the parasite is found by examining thin blood smears under a microscope. This method is also used for testing drug resistance, measuring drug effectiveness, and classifying disease severity. Microscopic diagnostic methods are cumbersome and thus require a lot of skill and experience to execute. In this study, we propose the use of a deep convolutional neural network to detect the presence of red blood cell images. This would assist in automating the detection of malaria from red blood cell images and aid in early diagnosis.

## **1 Related Works**

Unsupervised and supervised learning methods can be used for performing cell image analysis. Supervised learning methods such neural networks were used for making a classification tool to differentiate between parasitized and uninfected cell images [1]. Machine learning methods were used for content-based image retrieval and to discover new clinicopathological relationships from pathological images [2]. K-nearest neighbours are used for both classification and regression tasks, this method was used

---

S. Ghose · S. Datta (✉) · C. Malathy · M. Gayathri

Department of Computer Science and Engineering, SRMIST, Kattankulathur, India

e-mail: [smaranjitghose@protonmail.com](mailto:smaranjitghose@protonmail.com)

C. Malathy

e-mail: [cmalathyc@srmist.edu.in](mailto:cmalathyc@srmist.edu.in)

M. Gayathri

e-mail: [gayathrm2@srmist.edu.in](mailto:gayathrm2@srmist.edu.in)

for classifying and diagnosing malaria from cell images from a given set of features, and other methods such as linear regression and decision trees can also be used for diagnosing malaria [2]. Support vector machine (SVM) is used for both classification and regression problems, and therefore, SVMs were used for classifying and diagnosing of malaria by using features as input that were extracted from the images [3]. Resnet50 is a convolutional neural network that consists of 50 layers, and it can be used as an image classification model to classify between parasitized and uninfected cell images [4]. VGG19 is a 19-layer deep convolutional neural network which is made up of 19 layers and can be used to make an image classification model to classify between parasitized and uninfected images [5]. Restricted Boltzman machine is a deep belief network that is an undirected bipartite graph model, and this can be used for the detection of parasites from the thin blood smears [6].

## 2 Introduction

Malaria is caused by the protozoan parasites belonging to the *Plasmodium* genus [7]. The carriers of this parasite are female anopheles mosquito. The mosquito bites the host, which infects the red blood cells. According to the World Health Organisation, almost 3.2 billion people over 95 countries are at a risk of being infected by the parasite and suffering from malaria [8]. The most common types of *Plasmodium* parasites are as follows:

- *Plasmodium falciparum*: This group of protozoan parasites is found in Africa and is the cause of most of the deaths caused by malaria.
- *Plasmodium vivax*: This group of protozoan parasites is mostly found in South America and Asia. The symptoms are much milder as compared to *Plasmodium falciparum*, but these groups of *Plasmodium* parasites generally stay in the liver for up to 3 years.
- *Plasmodium ovale*: This group of protozoan parasites is found in the western part of Africa and generally lives in the liver for a long time without producing any type of symptoms.
- *Plasmodium malariae*: This group of protozoan parasites has an incubation period of 16–59 days. They cause chronic infection which in severe cases last a lifetime.
- *Plasmodium knowlesi*: This group of protozoan parasites is generally found in southeast Asia and causes very high levels of parasitemia which proves to be fatal.

The early diagnosis of malaria is important as it leads to proper treatment and prescribing of proper medications. The current procedure of diagnosing malaria is performed by a microscopic examination in which blood smears are inspected and checked for the presence of infected erythrocytes [9]. At first, the bloodstains are stained using Giemsa staining. The staining method is used for highlighting the parasites, white blood cells, and also the platelets. This is then examined under a microscope to check for the shape, size, and characteristics of the red blood cells. This procedure is also used for testing drug resistance, drug effectiveness [10],

and also, for knowing the severity of the disease. The diagnosis procedure is not standardized and therefore is heavily dependent on the experience and skill level of the pathologist. In rural areas where clinics and testing centres are in scarcity, there is always a dire need for pathologists for performing the tests required [3].

Therefore, in this study, we propose a deep learning [11]-based solution for the detection of malarial parasites, *Plasmodium falciparum*, in red blood cell images. This is done by using a deep convolutional neural network (DCNN) for classifying blood cell images containing and not containing *Plasmodium falciparum* parasite.

### 3 Methods

#### 3.1 Dataset Description

For our study, we used the dataset provided by the National Institutes of Health (NIH) [12]. The dataset consists of red blood cell images of which 13,779 images were parasitized blood cells and 13,779 images were uninfected cells. The images are of segmented cells of thin blood smear slides. All images were Giemsa-stained [13], and this staining method is used for the thinning of both thick and thin smeared blood for malaria screening. The infected red blood cells contained *Plasmodium falciparum* which is a common malarial parasite [14]. The images in 1 and 2 are two sample images of infected and uninfected red blood cell images from the dataset.

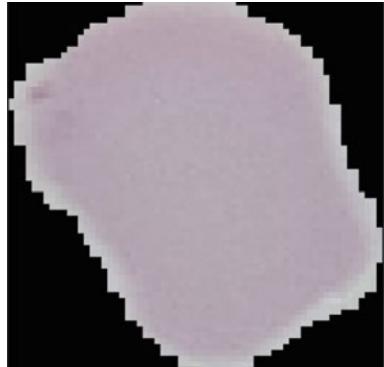
#### 3.2 Model Architecture

Deep convolutional neural networks (DCNNs) [15] are a class of neural networks. DCNNs are used for various image-related tasks such as image classification, image

**Fig. 1** Infected cell image



**Fig. 2** Uninfected cell image



segmentation and also to perform object detection. They consist of three layers which are convolutional, pooling, and fully connected layers. The convolutional layers perform edge detection, and therefore, DCNNs are suitable for the use of images. The convolution layer is followed by a pooling layer, and an activation function is used for introducing nonlinearity. The activation function is used after each convolution layer. The results obtained from the convolution and pooling layers are then fed to the fully connected layer, and softmax is used for classifying the images according to labels.

To make use of knowledge obtained from training a DCNN on Imagenet [16], we make use of transfer learning. Transfer learning allows us to leverage the feature maps that were obtained from training the DCNN on millions of images. At first, the layers of the model are taken and are frozen. The layers are frozen to prevent the loss of any knowledge that may happen during any future training. Then, the new trainable layers are added on the top of frozen layers to perform predictions.

We make use of EfficientNet B1 [17] for our experiment. The architecture of EfficientNet optimizes flops by using a multi-neural search architecture. The convolution layers of EfficientNet are divided into two parts, namely pointwise convolution and depthwise convolution, and this helps in reducing calculation time while having a minimum loss for accuracy. The MBconv block in EfficientNet first extends to the channels of the images and then compresses them which results in lesser number of skipped connections.

EfficientNet also utilizes compound scaling, and in compound scaling, the length, breadth, and width of the network are increased with respect to the baseline architecture of EfficientNet as seen in Fig. 3. The layers in EfficientNet have been increased by keeping a fixed constant ratio, and this helps in increasing the accuracy of the model.

At first, the EfficientNet B1 is instantiated which is pre-loaded with weights trained on ImageNet. After that, we freeze the convolutional base, and this prevents the weights in a given layer from being updated during training. Further, we add layers on the top to perform predictions according to the labels of the image. The first layer that is added on top is an average pooling layer where the input is downsampled by

**Fig. 3** Synopsis of the proposed architecture

Stage	Operator	Resolution	Channels	Layers
1	Conv3x3	224x224	32	1
2	MBConv1	112x112	16	1
3	MBconv6	112x112	24	1
4	MBconv6	56x56	40	2
5	MBconv6	28x28	80	3
6	MBconv6	14x14	112	3
7	MBconv6	14x14	192	4
8	MBConv6	7x7	320	1

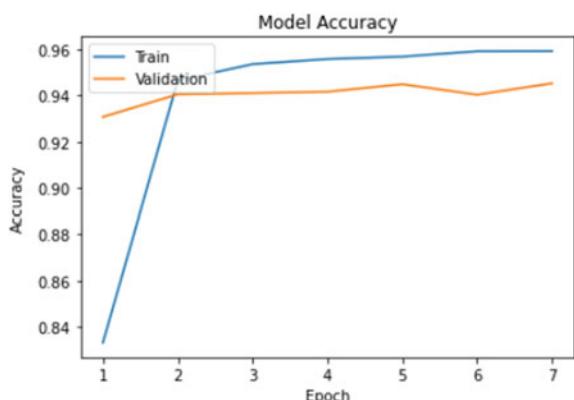
a kernel of size  $4 \times 4$ . The output obtained from the average pooling layer is fed to a flatten layer for converting the feature values in the matrix into vectors. The vector values are then passed through a dense layer. In the last layer, a two-way softmax activation is used for classifying the images into two classes.

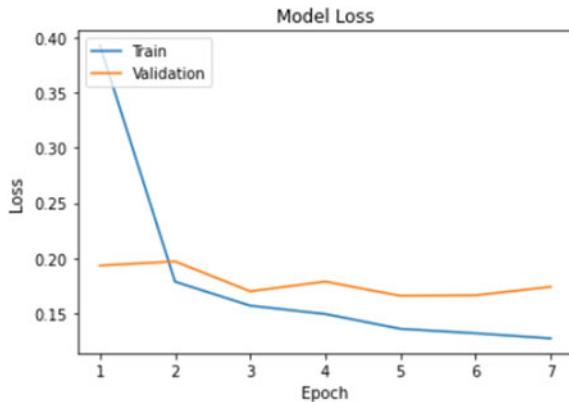
The table in the 3 shows the baseline architecture of EfficientNet. The operator column shows the exact orientation of blocks, the resolution represents the input resolution that is going to be utilized by the blocks, the channels represent the number of output channels of the blocks, and the layers represent the number of times the blocks were repeated.

## 4 Result

We used Tensorflow 2.0 as the framework for training our model. The batch size of the model was set to 32, and the learning rate of 0.00001 was used. Steps per epoch is a function of training length and batch size, and this was set to 5000. After training the model for a cycle of 30 epochs, a training accuracy of 96% and a test accuracy of 94% were obtained. The images in 4 and 5 are the accuracy and loss curves obtained from the model.

**Fig. 4** Accuracy curve



**Fig. 5** Loss curve

To assess the performance of our model, we calculated the following scores:

- True positives (TP) are defined as the cell images that were predicted to be malaria positive were actually malaria positive.
- False positives (FP) are defined as the cell images that were predicted to be malaria positive with malaria were actually malaria negative.
- True negatives (TN) represent that the images of the malaria that were malaria negative are actually malaria negative.
- False negatives (FN) represent that the cell images that were malaria negative are actually malaria positive.
- Precision: It represents number of correct malaria positive predictions. The precision score obtained was 0.9.

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

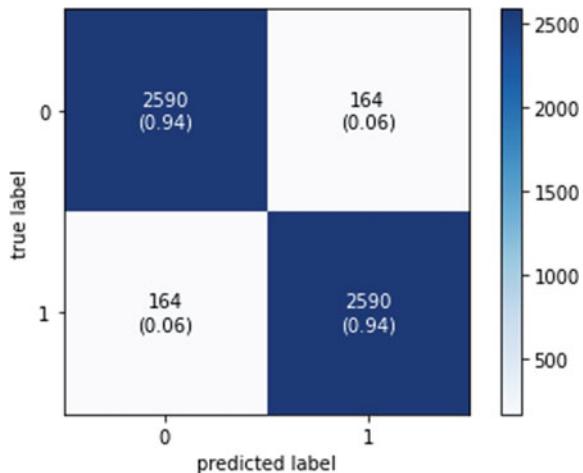
- Recall: It gives the number of correct malaria negative predictions. The recall score obtained was 0.9.

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

- F1 score: The F1 score is defined function of precision and recall. This helps us to find the number of instances our model was accurately able to classify without missing a significant number of instances. The F1 score obtained was 0.98.

$$\text{F1 score} = 2.[(\text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})]$$

- Specificity: It refers to the percentage of cell images that had cells that were actually classified as uninfected. The score obtained was 90%.
- Sensitivity: It refers to the percentage of cell images that were malaria positive were actually classified as malaria positive. The score obtained was 90%.

**Fig. 6** Confusion matrix

Furthermore, the confusion matrix was also calculated as seen in Fig. 6 where label 0 is malaria positive, and 1 is malaria negative

## 5 Conclusion

In this endeavour, we present a deep learning approach to identify the presence of *Plasmodium falciparum* in the cell images. In the current procedure, it requires to count the number of infected red blood cells with the help of a microscope. This makes the process very time consuming and cumbersome to execute. This same process is even used for testing the effectiveness of any kind of medicine of malaria and also for checking the severity of malaria. We used the dataset provided by the NIH which comprised of segmented cells from the thin blood smear slide images and were magnified using a microscope. After that, we trained a classification model on the dataset and achieved a test accuracy of 94% and a validation of 95%. Additionally, we observed a recall of 0.9 and a precision score of 0.9. Thus, our work proves that convolutional neural networks can be used for the detection of malaria and hope that in future assist pathologists.

## References

1. Kan, A.: Machine learning applications in cell image analysis. *Immunol. Cell Biol.* **95**(6), 525–530 (2017)
2. Komura, D., Ishikawa, S.: Machine learning methods for histopathological image analysis. *Comput. Struct. Biotechnol. J.* **16**, 34–42 (2018)

3. Das, D.K., Ghosh, M., Pal, M., Maiti, A.K., Chakraborty, C.: Machine learning approach for automated screening of malaria parasite using light microscopic images. *Micron* **45**, 97–106 (2013)
4. Reddy, A.S.B., Juliet, D.S.: Transfer learning with ResNet-50 for Malaria cell-image classification. In: 2019 International Conference on Communication and Signal Processing (ICCSP), pp. 0945–0949. IEEE (2019)
5. Var, E., Tek, F.B.: Malaria parasite detection with deep transfer learning. In: 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 298–302. IEEE (2018)
6. Bibin, D., Nair, M.S., Punitha, P.: Malaria parasite detection from peripheral blood smear images using deep belief networks. *IEEE Access* **5**, 9099–9108 (2017)
7. Coluzzi, M.: Malaria vector analysis and control. *Parasitol. Today* **8**(4), 113–118 (1992)
8. Organization, W.H.: Malaria Microscopy Quality Assurance Manual-Version 2. World Health Organization (2016)
9. APPENDIX A Microscopic Procedures for Diagnosing Malaria
10. Bloland, P.B., Organization, W.H., et al.: Drug Resistance in Malaria. World Health Organization (2001)
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**(7553), 436–444 (2015)
12. Kaggle Malaria dataset
13. Umlas, J., Fallon, J.N.: New thick-film technique for malaria diagnosis. *Am. J. Trop. Med. Hyg.* **20**(4), 527–529 (1971)
14. Tilley, L., Dixon, M.W., Kirk, K.: The *Plasmodium falciparum*-infected red blood cell. *Int. J. Biochem. Cell Biol.* **43**(6), 839–842 (2011)
15. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), 2278–2324 (1998)
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition (pp. 248–255). IEEE (2009)
17. Tan, M., Le, Q.V.: Efficient net: rethinking model scaling for convolutional neural networks (2019). arXiv preprint [arXiv:1905.11946](https://arxiv.org/abs/1905.11946)

# An Online Path Planning with Modified Autonomous Parallel Parking Controller for Collision Avoidance



Naitik M. Nakrani and Maulin M. Joshi

**Abstract** In recent years, investigation of algorithms for autonomous vehicle design is one of the significant fields of research. Autonomous parking is a challenging issue, especially where traffic density is high, and people have to opt for parallel parking than garage parking. The design of an autonomous parking controller generally does not take into consideration nearby obstacles presence while maneuvering. The addition of an online path planning module can significantly improve the usability of such a controller for real-time applications. In this paper, an angle correction module is introduced that improves the standalone parking controller using ultrasonic sensor information to generate optimized path planning and avoid collision. Simulations results show the efficacy of the proposed work and support the design of the angle correction module. This type of intelligent correction module will surely ease the parking problem in a dense area and make parking safer, being collision-free solutions.

## 1 Introduction

Parallel parking is one of the most challenging issues of maneuvering for the human driver. The design of an autonomous parallel parking system for vehicles is even challenging because of many uncertainties present in the real world. It is tough to learn human-like expertise for machines. Though research work is available on designing a parallel parking controller, clear scope exists for a fully autonomous parking algorithm using a machine learning approach that can mimic human intelligence.

In the literature generally, autonomous parking is termed as a standalone problem where inputs are given to the controller from the current state, i.e., orientation or location. It starts from a preplanned starting point and finishes within the parking

---

N. M. Nakrani (✉)  
Uka Tarsadia University, Bardoli, India

M. M. Joshi  
Sarvajanik College of Engineering and Technology, Surat, India  
e-mail: [maulin.joshi@scet.ac.in](mailto:maulin.joshi@scet.ac.in)

slot with endpoint and target orientation. However, the addition of extrasensory information to such a system may promise safety and intelligence. Our work is primarily focused on the improvement of typical path planning with the help of additional ultrasonic sensors' distance information. With this, a collision-free parallel parking system can be visualized.

The most important part of the autonomous parking system is proper path planning along which vehicle can move and reach to parking space. Many authors [1–13] have discussed offline and online path planning. They have trained their system with different types of curvilinear trajectory with non-holonomic constraints. Generally, online path planning is a better choice because of its ability to adapt with continuous environment sensing. Once path planning is properly designed, the next task is to track along through path. Such steering action generally gets completed with machine learning algorithms. A fuzzy logic theory provides a good choice as it has the capacity to translate human linguistic intelligence into a rule base. In [14, 15], a fuzzy-based parallel parking controller is developed and implemented with fifth-order polynomial path planning. They have simulated reference path with multiple start points and tested their fuzzy rule base for forward and reverse manner parking. However, the drawback is that environment is assumed free from any possible collision with any other objects. These make parking feasible with only fixed initial starting points for parking and collision free. On the other hand, the practical world may have nearby parked vehicles, infrastructure walls besides parking space. Such existing systems can be improved with the integration of sensor information.

In our earlier work [16, 17], we discussed the autonomous parallel parking controller and sensor-based navigation module for car-like mobile robots. In this paper, a new steering angle correcting module is attached with a typical parking controller which can start parking from any point and can reach to parking space without collision and completes its parking once a feasible initial point for parking has arrived. For environment sensing, ultrasonic, infrared, vision, sonar, LIDAR, etc. sensors are used. Among them, ultrasonic sensing is an accurate and low-cost solution. It also reduces the computational burden to controllers, LIDAR, etc. In our work, the ultrasonic sensor information is used by our angle correction module to make intelligent decisions.

The organization of this paper is as follows: Typical path planning-based parking controller limitation is discussed in Sect. 2. In Sect. 3, a distance-based angle correction controller is introduced. The novel part of our work is also discussed in this section. In Sect. 4, various scenarios are simulated in a confined environment to visualize the application of work. Finally, the conclusions are given in Sect. 5.

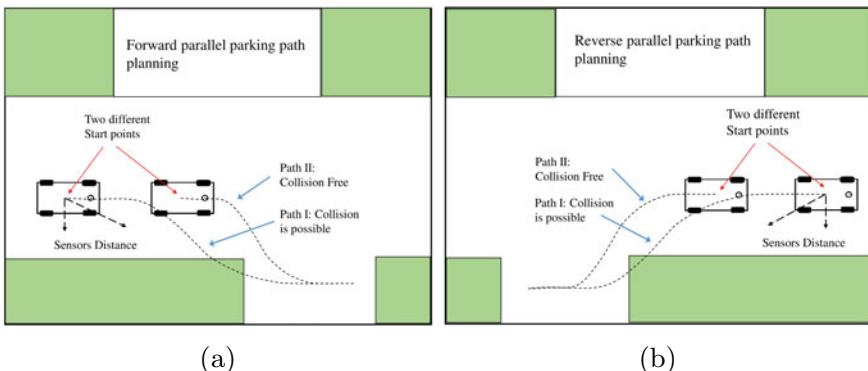
## 2 Problem Description

Parallel parking is a type of autonomous parking where vehicle parks itself parallel to the roadside, curb, and in between two parked vehicles. Generally, for any parking algorithm, prerequisite is proper path planning between the initial and final location

of a parking vehicle. It can be done in two ways: offline and online path planning. In offline mode, the role of path planning begins from a point where the vehicle has full knowledge of: its obstacle environment, its initial location, and its final target. The task of offline path planning simply connects the initial position to the final position, and then, the created path is followed by the vehicle. Path planning is carried out simultaneously in online mode when (a) moving toward the target and (b) perceiving the environment, including its modifications. For a static condition (or slowly changing), offline path planning is often used. However, online path planning must be used for dynamic environments, as the route must be adjusted according to environmental changes. This method is sometimes important for traveling under partially known environmental conditions, as the vehicle explores its parking area when traveling and has to adjust the path according to any new information.

The typical scenario for autonomous parallel parking with fifth-order polynomial path planning for forward and reverse parking is shown in Fig. 1a, b, respectively. The reason for addressing the fifth-order polynomial path is that it generates nonlinear paths taking care of non-holonomic constraints found in vehicle dynamics and found suitable for parallel parking. Any machine learning approach using soft computing technique generally takes care of the vehicles present state, i.e., postures, including location and orientation. To accomplish path tracking for parking, it can work in a free space environment, but practically, it has to take care of all the obstacles like objects like a curb, corner, and side or other parked vehicles or infrastructure sides. This fact points out the fact that just giving the only start and endpoints for parking may result in a collision of the vehicle for specific environmental conditions.

A typical scenario visualized in Fig. 1 demonstrates justification for the problem at hand (i.e., the requirement of environmental sensing) where a typical fifth-order path is shown for two different start points for a vehicle and the same endpoint inside the parking space. As shown, path I indicates that vehicle starts its parking from the parking space. When it tries to execute its parking without the help of any sensing mechanism, there may be chances that it will collide, as shown in the figure. At the



**Fig. 1** Typical scenario for **a** forward parallel parking, **b** reverse parallel parking

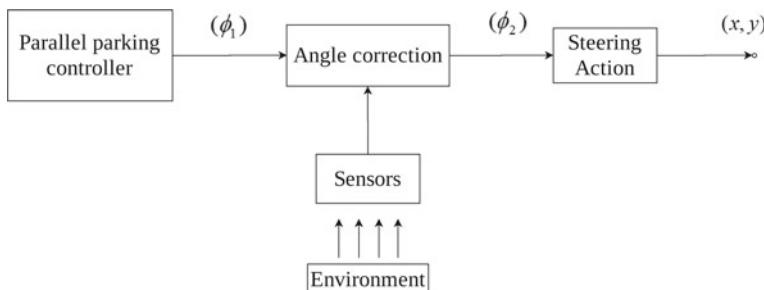
same time, a slightly different start point, e.g., path II as shown in Fig. 1, may give a collision-free path. These cases highlight a limitation of the path planning approach where typical start points will work well for a parking controller, which should not be the case for practical implementation.

To overcome this limitation, we propose an online path planning mechanism incorporating the vehicle's surroundings' general sensing element and feedback mechanism. This information can modify or correct the controlled output of the planned parking controller to lead the vehicle to a posture from where all conditions for parking are matched and make parallel parking feasible. This approach can eliminate the limitation of a selection of starting point can be used for any start point selection. In this paper, an online path planning approach, along with an ultrasonic sensor, is discussed. An angle correction module is integrated with a parking controller so that vehicle can change its motion path when the curb is present and parked itself without collision.

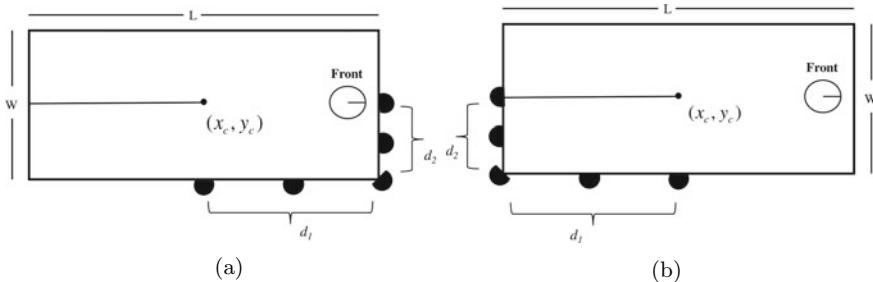
### 3 Sensor-Based Online Parking System

In this section, we describe the development of an individual sensor-based feedback module that can correct the steering angle of the vehicle computed by the parallel parking controller. A general block diagram of such a modified system is shown in Fig. 2.

Here, the core part of our system is a parallel parking controller, which is 2 input 1 output fuzzy system. It takes two distinct angle information, an angle with respect to target and an orientation angle as an input for the controller. The output of the parking controller is steering angle ( $\phi_1$ ). In our earlier work [16, 17], the detailed design and working of a parallel parking controller are already discussed. In addition to that controller, an angle correction module is introduced with an ultrasonic sensor model as shown in the block diagram. This angle correction module is a 3 input 1 output fuzzy control system whose inputs are steering angle ( $\phi_1$ ) computed by a parallel parking controller and two sensor information, as shown in Fig. 1. Its output is the corrected version of the steering angle ( $\phi_2$ ), which is given to the vehicle for steering action.



**Fig. 2** Block diagram of parallel parking with angle correction module



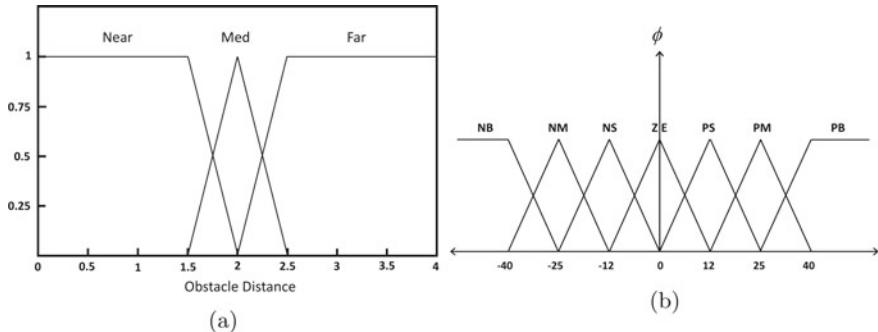
**Fig. 3** Sensor grouping for **a** forward parking, **b** reverse parking

The angle correction module used ultrasonic information to sense the map and make an intelligent decision for the vehicle for appropriate change in the motion path. To overcome the nature of the problem discussed in Fig. 1, the typical utilization of sensors that is assumed to be mounted on the perimeter of the vehicle is shown in Fig. 3. A total of nine ultrasonic sensors are presumed to be mounted for sensing parking side map. As shown in Fig. 3a, only five sensors are utilized in the forward way of parking, and similarly, another set of five sensors is being used in case of reverse parking as per Fig. 3b.

To use this sensor information as an input for angle correction fuzzy system, these ultrasonic sensors are grouped into two distances  $d_1$  and  $d_2$ . Here, the minimum distance is considered as a grouped distance  $d_1$  and  $d_2$ . To cater specific need by exploring the nature of parking, grouping can be modified as per requirement. The detailed algorithm for calculating and grouping sensor information is also discussed in our earlier work [16, 17].

These distances  $d_1$  and  $d_2$  are fuzzified into three membership functions near, medium, and far as shown in Fig. 4a, and the steering angle as an output of a parallel parking controller and angle correction module is having seven membership functions like negative big (NB), negative medium (NM), and zero (ZE). The shape and range of steering angle membership function are shown in Fig. 4b. Based on these membership functions, a total 63 rule base is generated for angle correction fuzzy module.

Few samples from a behavior-based rule base based on different conditions are given in Table 1. These rules are implemented for the angle correction module. It can be observed that whenever distance measures obstacles are at a far distance, it keeps the parking controller's output unchanged. Hence, the vehicle continues to move as per its parking path. At the same time, when any sensor detects any obstacle, say  $d_1$ , then depending upon the type of parking, it will correct the output of the parking controller, as shown as collision avoidance in Table 1. Such fuzzy laws are structured so that the car primarily changes its course of motion and drives smoothly to the target if no collision is likely. If sensor data shows that there are obstacles in the vicinity of the car, it must attempt to adjust its course in order to prevent collisions.



**Fig. 4** **a** Distance membership functions, **b** steering angle membership function

**Table 1** Fuzzy rule base for angle correction module

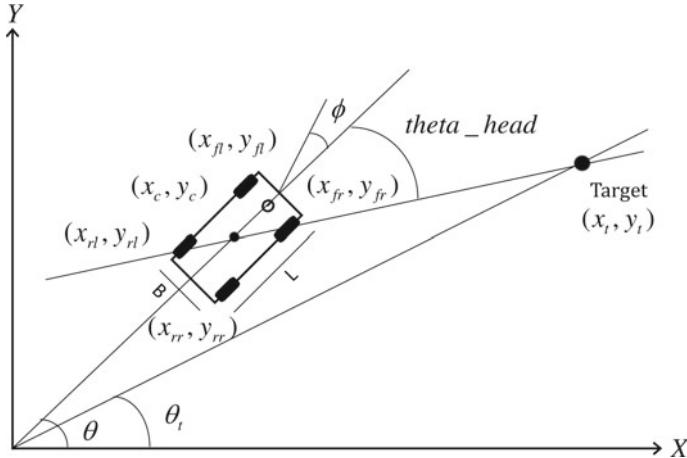
If				Then
Fuzzy behavior	Distance ( $d_1$ )	Distance ( $d_2$ )	Steering angle ( $\phi_1$ )	Steering angle ( $\phi_2$ )
Target steer	Far	Far	NS	NB
Target steer	Far	Far	ZE	ZE
Target steer	Far	Far	PS	PS
Collision avoidance	Near	Med, far	NB	NS
Collision avoidance	Near	Med, far	NS	PS
Collision avoidance	Near	Med, far	PS	NS
Collision avoidance	Near	Med, far	PB	PS

## 4 Simulation Results

To show the efficacy of our proposed angle correction module, simulation is provided in MATLAB software. A car-like mobile robot model used for our simulation is discussed in this section. Also, parallel parking like environment is generated that represents the challenge discussed earlier. An environment setup is thoroughly discussed in a later section.

### 4.1 CLMR Model

As shown in Fig. 5, a car-like mobile robot model (CLMR) for simulation is considered. It has a four-wheeled system that is compatible with actual vehicles. The



**Fig. 5** CLMR model

steering and the front side of the car are marked by the dotted circle between the front wheels. The control parameters for this model are steering angle and rear-wheel speeds. A car is limited to turning left and right on its front wheels, but they can stay parallel.

The kinematics equations of the CLMR model are as follows:

$$\theta_{\text{new}} = \theta_{\text{old}} + \dot{\theta} \times dt \quad (1)$$

$$x_{\text{new}} = x_{\text{old}} + v \times \cos(\theta_{\text{new}}) \times dt \quad (2)$$

$$y_{\text{new}} = y_{\text{old}} + v \times \sin(\theta_{\text{new}}) \times dt \quad (3)$$

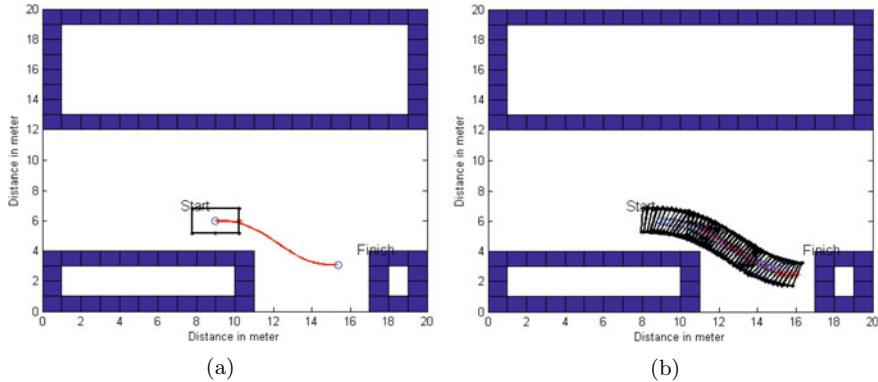
Equations (1), (2), and (3) are used to obtain a new position of the vehicle at each instance.

## 4.2 Environment Simulation

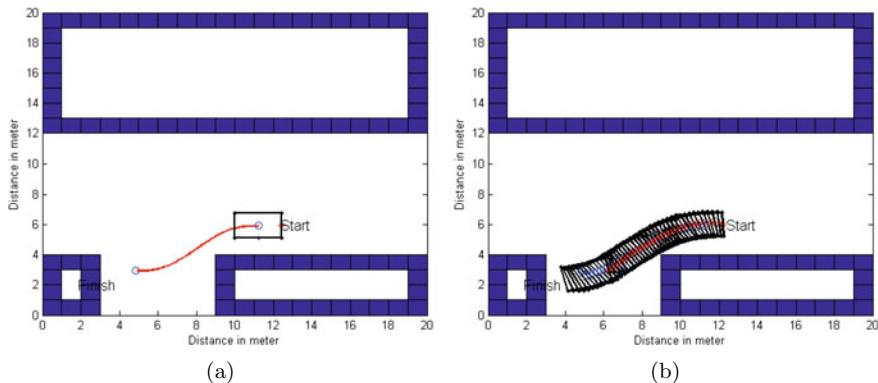
In this section, simulation results are provided to support the angle correction fuzzy system. A 20m by 20m MATLAB-based environment is created to similar to the scenario shown as per Fig. 1. A car-like mobile robot (CLMR) with Length = 2.4m and width = 1.6m is assumed. Total nine sensors are assumed and grouped into two

distances with minimum value into  $d_1$  and  $d_2$  as shown in Fig. 3. Single parking slot scenario is assumed, and parking space is taken enough so that CLMR able to park itself. The steering angle is assumed with respect to X-axis. The speed of CLMR is assumed constant, and the environment surface is taken flat. The defuzzification method used for all fuzzy systems is the centroid. Different case scenarios for forward and reverse parallel parking is discussed.

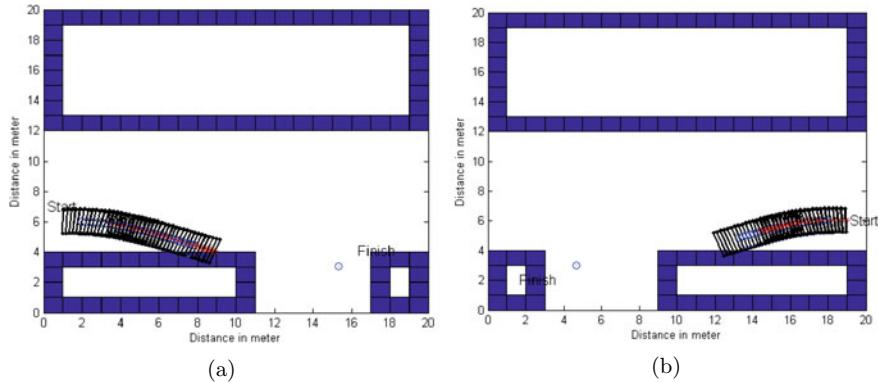
**Case 1: Feasible start point and success of typical parallel parking** Figure 6a shows the start and finish point for a vehicle to do a forward parallel park. Here, as the vehicle starts very near to parallel parking, it will able to park itself very easily without the use of an angle correction module. Figure 6b shows a continuous path taken by CLMR during the case described in Fig. 6a. A similar scenario for reverse parallel parking is shown in Fig. 7a, b.



**Fig. 6** **a** A case represents initial and final location of CLMR for forward parking, **b** a continuous run sequence of CLMR in forward parking



**Fig. 7** **a** A case represents initial and final location of CLMR for reverse parking, **b** a continuous run sequence of CLMR in reverse parking



**Fig. 8** Failed run sequence of CLMR with given start and finish point for **a** forward parking, **b** reverse parking

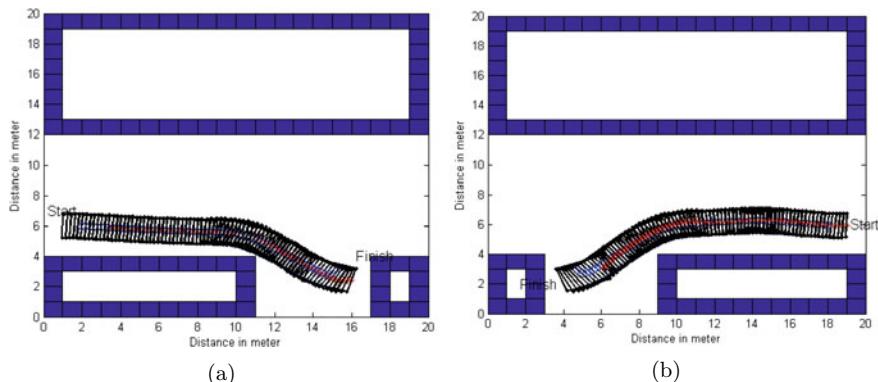
**Case 2: Failure of parallel parking (in absence of proposed angle correction module)** Figure 8 shows a different start point and the same final endpoint for forward and reverse parallel parking. Here, the initial start point for CLMR is far away from the parking space, and due to that, a path planned by the parking controller is also stretched compared to the previous case. It is clearly visible that if the traditional offline parking module (without the use of proposed module) is used for this case, it will collide with the side of an obstacle that can be a curb or other parked vehicle. Such failure continuous path for forward parking is also shown in Fig. 8.

**Case 3: Parking along with angle correction module** In this case, simulation results are given for the case, described in Fig. 8 with the angle correction module. Because here sensor information is incorporated to be used for angle correction module at any instance vehicle will be very near to collision boundary, angle correction module will correct its steering action temporarily as per the rules described in Table 1. When CLMR approaches toward parking and one of the sensors detect the parking space, angle correction module will allow parking to finish the rest of the task.

Figure 9a, b represents continuous sequence run of CLMR for forward and reverse parking with modified path using angle correction module as per the scenarios taken in Fig. 8a, b, respectively. It can be observed that using the angle correction module, CLMR does not deviate much from the finish point. So the chance of malfunction during path correction is very rare. It can be generalized for different dimensions of vehicles.

## 5 Conclusion

In this paper, a new fuzzy-based angle correction is introduced to the existing solutions with a standalone parking controller designed. Details of the idea and use of ultrasonic sensor information are also presented. Simulation results show that the



**Fig. 9** Continuous run sequence of CLMR with angle correction module for (a) forward parking (b) reverse parking

proposed work gives satisfactory results for forward as well as reverse parking scenarios. Results also demonstrate that introduction of correction of angle to CLMR is able to modify its path during runtime and results in collision avoidance. This makes the performance of parallel parking better, and in the future, one can work with optimization of the number of sensors to improve the performance.

## References

1. Lyon, D.: Parallel parking with curvature and nonholonomic constraints. In: IEEE Intelligent Vehicles Symposium, pp. 341–346 (1992). <https://doi.org/10.1109/IVS.1992.252283>
  2. Murray, R.M., Sastry, S.S.: Nonholonomic motion planning. Steering using sinusoids. *IEEE Trans. Automat. Contr.* **38**(5), 700–716 (1993). <https://doi.org/10.1109/9.277235>
  3. Liu, W., Li, Z., Li, L., Wang, F.Y.: Parking like a human: a direct trajectory planning solution. *IEEE Trans. Intell. Transp. Syst.* **18**(12), 3388–3397 (2017). <https://doi.org/10.1109/TITS.2017.2687047>
  4. Jing, W., Feng, D., Zhang, P., Zhang, S., Lin, S., Tang, B.: A multi-objective optimization-based path planning method for parallel parking of autonomous vehicle via nonlinear programming. In: 2018 15th International Conference Control Automation Robot Vision, ICARCV 2018, pp. 1665–1670 (2018). <https://doi.org/10.1109/ICARCV.2018.8581195>
  5. Zou, R., Wang, S., Wang, Z., Zhao, P., Zhou, P.: A Reverse Planning Method of Autonomous Parking Path, pp. 92–98 (2020). <https://doi.org/10.1109/acirs49895.2020.9162616>
  6. Laumond, J.P., Jacobs, P.E., Taix, M., Murray, R.M.: A motion planner for nonholonomic mobile robots. *IEEE Trans. Robot. Autom.* **10**(5), 577–593 (1994). <https://doi.org/10.1109/70.326564>
  7. Paromtchik, I.E., Laugier, C.: Autonomous parallel parking of a nonholonomic vehicle. In: Proceedings of Conference on Intelligent Vehicles, no. 33, pp. 13–18 (1996). <https://doi.org/10.1109/IVS.1996.566343>
  8. Paromtchik, I.E., Laugier, C.: Motion generation and control for parking an autonomous vehicle. In: Proceedings of IEEE International Conference Robotics Automation, vol. 4, pp. 3117–3122 (1996). <https://doi.org/10.1109/robot.1996.509186>

9. Gorinevsky, D., Kapitanovsky, A., Goldenberg, A.: Neural network architecture for trajectory generation and control of automated car parking. *IEEE Trans. Control Syst. Technol.* **4**(1), 50–56 (1996). <https://doi.org/10.1109/87.481766>
10. Müller, B., Deutscher, J., Groddeck, S.: Continuous curvature trajectory design and feedforward control for parking a car. *IEEE Trans. Control Syst. Technol.* **15**(3), 541–553 (2007). <https://doi.org/10.1109/TCST.2006.890289>
11. Kim, J.M., Il Lim, K., Kim, J.H.: Auto parking path planning system using modified Reeds-Shepp curve algorithm. In: 2014 11th International Conference Ubiquitous Robotics Ambient Intelligent. URAI 2014, no. Urai, pp. 311–315 (2014). <https://doi.org/10.1109/URAI.2014.7057441>
12. Vorobieva, H., Glaser, S., Minoiu-Enache, N., Mammar, S.: Automatic parallel parking in tiny spots: path planning and control. *IEEE Trans. Intell. Transp. Syst.* **16**(1), 396–410 (2015). <https://doi.org/10.1109/TITS.2014.2335054>
13. Du, X., Tan, K.K.: Autonomous reverse parking system based on robust path generation and improved sliding mode control. *IEEE Trans. Intell. Transp. Syst.* **16**(3), 1225–1237 (2015). <https://doi.org/10.1109/TITS.2014.2354423>
14. Chang, S.J., Li, T.H.S.: Design and implementation of fuzzy parallel-parking control for a car-type mobile robot. *J. Intell. Robot. Syst. Theory Appl.* **34**(2), 175–194 (2002). <https://doi.org/10.1023/A:1015664327686>
15. Li, T.-H.S., Chang, S.-H.: Autonomous fuzzy parking control of a car-like mobile robot. *IEEE Trans. Syst. Man, Cybern. Part A Syst. Hum.* **33**(4), 451–465 (2003). <https://doi.org/10.1109/TSMCA.2003.811766>
16. Nakrani, N., Joshi, M.: Fuzzy based autonomous parallel parking challenges in real time scenario. In: Advances in Intelligent Systems and Computing, pp. 789–802 (2016)
17. Nakrani, N., Joshi, M.: An intelligent fuzzy based hybrid approach for parallel parking in dynamic environment. *Procedia Comput. Sci.* **133**, 82–91 (2018). <https://doi.org/10.1016/j.procs.2018.07.011>

# Real-Time Proximity Sensing Module for Social Distancing and Disease Spread Tracking



Sreeja Rajesh, Varghese Paul, Abdul Adil Basheer, and Jibin Lukose

**Abstract** Low energy proximity sensing devices are being used in our daily life for various purposes. The concept of making a dedicated hardware module for measuring the distance arose from this concept, so as to provide reliable and accurate measurements for various applications. Analysing the need and severity of the present situation due to the spread of COVID-19, the proposed hardware and software architecture can be tuned for the efficient practice of social distancing. It also provides an effective measure to track the disease spread by the integration of a secure database. NTSA—a cryptographic algorithm that is specifically designed and developed to run on low energy microcontrollers can protect the identity of every user. Since the encryption is done by the embedded device, NTSA can ensure enhanced privacy protection compared to any algorithm run on the server. This also ensures the reliability of the collected data. The hardware module actively transmits and receives signals from similar hardware modules. The proximity or distance between the two modules is measured by analysing the signal strength received by each module. To achieve disease spread tracking the users can track their status or level of exposure on a scale of 4 and hence would provide a metric for having external interactions like first-hand contact with COVID-19 patients, secondary contact, tertiary contact, and so on.

## 1 Introduction

In the present age, rapid development is happening in the fields of augmented reality and virtual reality applications. Generally, the spatial geometry of the environment

---

S. Rajesh (✉)  
Bharathiar University, Coimbatore 641046, India

V. Paul  
CUSAT, Kochi, Kerala, India

A. A. Basheer · J. Lukose  
Beurokrat Business Management Solutions, Thrissur, Kerala, India

is generated by the computer by applying machine learning or deep learning algorithms to infer data from two-dimensional images. However, most of these algorithms give unreliable results. The signal strength measured in decibels (dB) is converted to metric units of length (cm) by calibration methods. For the purpose of effective social distancing, thresholds are set at 200 and 150 cm. If the distance between 2 modules is measured to be less than 200 cm and greater than 150 cm, an alert is triggered by the module. If the distance is less than 150 cm and remains within this area for over 10 seconds, the NTSA encrypted ID of the hardware modules are exchanged and stored in respective databases. The proposed NTSA algorithm is resistant to attacks like brute force attack, equivalent key attack and exhibited Avalanche property, a desired feature of a good security algorithm. The hardware module can be paired with any mobile device by means of Bluetooth and the supporting software application. The software application maintains the database of every interaction registered by the hardware module. Data is stored as a graph, which makes it easier to identify the level of exposure of the user to a COVID-19 patient. The entries have a lifetime of 14 days after which the data is discarded protecting the privacy as well as making sustainable use of resources. The proposed NTSA algorithm is a block symmetric encryption algorithm. As we are aware that cryptography is an art of hiding secret information from unauthorized users, it came into existence way back in 1900 BC and was found inscribed in the main chamber of a tomb in Egypt [1]. The advancement in technology and the drastic usage of the Internet led to many new cryptographic algorithms for secure transmission of information. Symmetric and Asymmetric cryptographic algorithms are the two-broad classification of cryptography [2]. Symmetric algorithm is faster, less complex when compared to an asymmetric encryption algorithm. This paper is structured as follows: Section 2 depicts the different literature that helped in shaping the proposed work. Section 3 discusses the methodology of the proposed system and the experiment results are discussed in Section 4. Section 5 concludes and gives light to the future scope of the system.

## 2 Literature Survey

With the drastic increase in the utilization of the Internet in our day-to-day life, the importance of secure transmission of information has become our priority. Various cryptographic algorithms were developed and the tiny encryption algorithm was most widely used for more than a decade for its simplicity and ease of implementation. But TEA [3] suffered from equivalent key attack. Even though its descendants XTEA [4] and XXTEA [5] could address the issues to some extent but resulted in more complexity than TEA. Also TEA had better performance than other symmetric encryption algorithms like DES [6], AES [7], IDEA [8], SEA [9], LEA [10], WAKE [11], HIGHT, MARS, SERPENT, RC5, Blowfish, 3DES to name a few. The proposed cryptographic algorithm NTSA is resistant to the attacks suffered by tiny encryption algorithms in addition to the improvement in terms of implementation and performance.

### 3 Methodology of the Proposed System

#### 3.1 Hardware

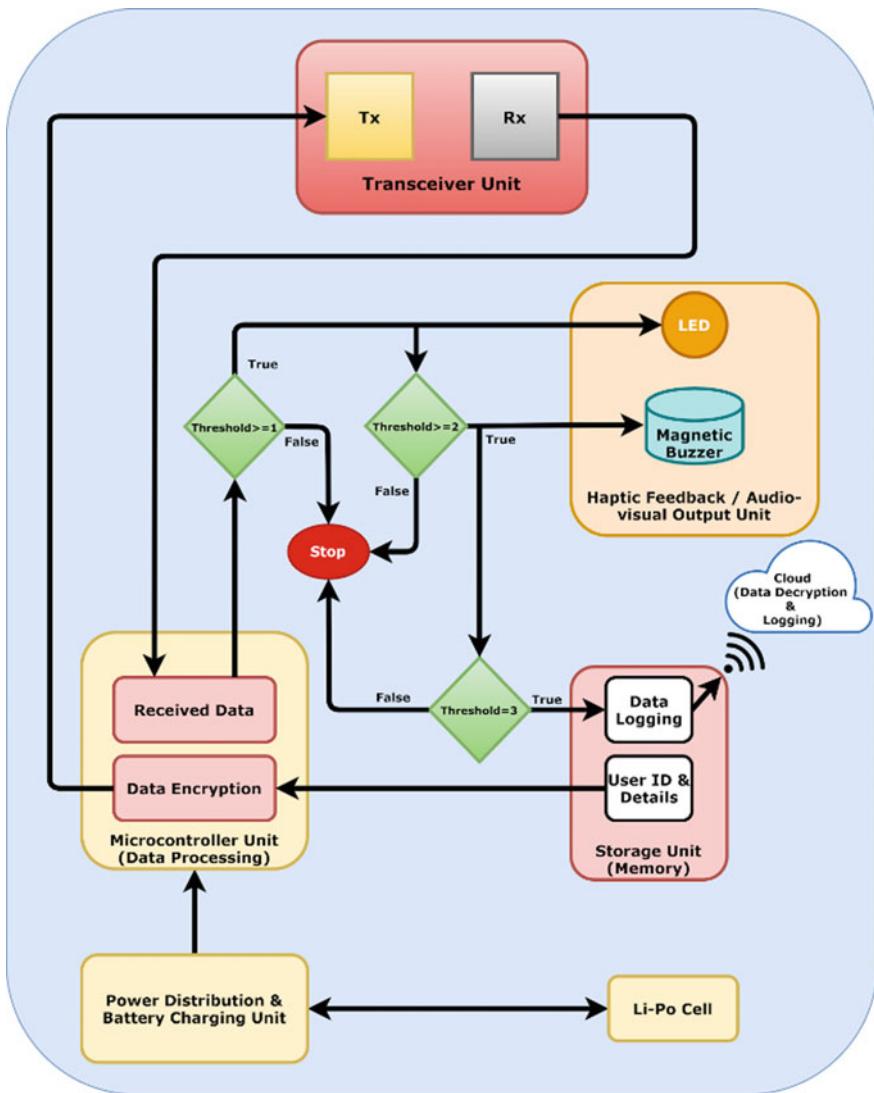
Real-time proximity sensing module, it is a compactly designed, low energy consuming hardware system equipped with a microcontroller unit powered by a Li-Po cell. The system holds an inbuilt storage/memory unit for data logging and a transceiver module for wireless data transmission and reception. The transceiver unit helps to broadcast the data and also scans for any other broadcasted data in the vicinity of the device. The architecture of the device-hardware is depicted in Fig. 1.

Figure 2 is the graphical representation of transmission of signal and thresholds at various perimeters. Say, ID1/User1 stays in the centre of a hall. ID2, ID3, ID4 and ID5 users enter into the same hall. The concentric coloured rings around each user represent the device signal range and the colorimetric representation is to convey and alert on how close each user is with each other, when in their signal scan range. Green coloured perimeter represents the safe distance and when moved closer towards the user—enters yellow, orange and red zones accordingly. When in the orange zone the users get notified regarding each other's presence in their perimeter and red zone marks that person has moved so close to the other person beyond the permissible social distance to be maintained. This triggers the devices of users within threshold distance, causing the device to generate an audio or visual output signal to catch the attention of the user; and the necessary details of the users gets logged into the memory and cloud database making record of the interaction. Later, if any of the users updates the status as 'infected with the disease', the graph is traversed to find and notify the users who made contact with the infected person.

It's the compact, low power consuming design of the device makes it very much suitable for the purpose of real-time proximity sensing and tracking, as the device could be a wearable gadget. The device runs on 2.7–5 V with a current consumption of a few milliAmps. The energy-efficient operating capability of both the data-processing unit and the transceiver unit makes this happen, to have a long run in a single charge.

#### 3.2 NTSA Algorithm Integrated to MCU

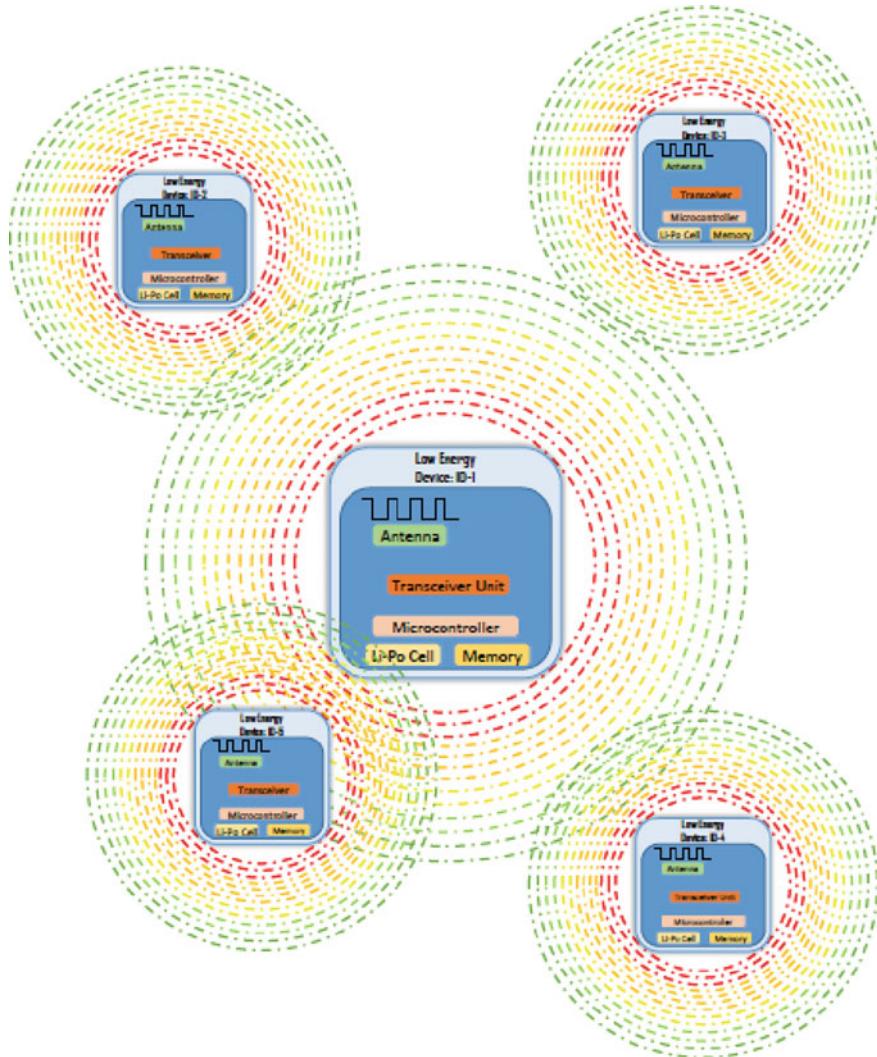
The proposed NTSA (Novel Tiny Symmetric encryption Algorithm) is a block symmetric encryption algorithm designed for secure transmission of information through insecure medium. NTSA has a very simple structure and can be implemented in both hardware and software. The inputs accepted by NTSA are 64-bit plaintext and 128-bit keys. It uses 32 cycles where 2 rounds make one cycle hence there are 64 rounds. The plaintext is split into two 32-bit blocks  $v_0$  and  $v_1$ . The key is partitioned into four 32-bit subkeys  $k_0$  through  $k_3$ . Throughout 64 rounds subkeys



**Fig. 1** Hardware architecture

$k_0$  and  $k_2$  are kept constant wherein  $k_1$  and  $k_3$  are recomputed in each cycle. The steps carried out while performing NTSA encryption are:

- Step 1: Splitting the plaintext.
- Step 2: Key partitioning.
- Step 3: Initialization of key constant  $k_c$ .
- Step 4: Computation of Key schedule constant  $k_{sc}$ .
- Step 5: Computation of 32-bit  $v_0$ .



**Fig. 2** Depiction of devices signal strength and proximity scan perimeter overlapping

Step 6: Recomputation of partial key  $k1$ .

Step 7: Recomputation of  $kc$ .

Step 8: Computation of 32-bit  $v1$ .

Step 9: Recomputation of partial key  $k3$ .

Step 10: Repetition of steps 5 to 9 for 32 cycles or 64 rounds.

Step 11: return  $k1$  and  $k3$ .

During step 1 of NTSA encryption process, the 64-bit plaintext is split into two 32-bit blocks  $v0$  and  $v1$ , respectively. Step 2 involves partitioning a 128-bit key into

4 subkeys  $k0$  through  $k3$  each of size 32-bit. There are two constants involved in the algorithm key constant  $kc$  and key schedule constant  $ksc$ . Key constant  $kc$  is initialized to 0 in step3. Step 4 involves computation of key schedule constant wherein  $ksc = (2^{31})/\Phi$ , where  $\Phi$  represents golden ratio whose value is 1.618033988749895. Steps from 5 until 9 are repeated for 32 cycles or 64 rounds. In step 5,  $v0$  is recomputed as shown in Eq. 1.

$$v0+ = ((v1 \ll 4) \text{ AND } k0) \text{ XOR } (v1 \text{ AND } kc) \text{ XOR } ((v1 \gg 5) \text{ AND } k1) \quad (1)$$

In step6, keeping  $k0$  constant, subkey  $k1$  is computed as.

$$k1+ = k0 \text{ XOR } \text{xtract}(v0)$$

where  $\text{xtract}()$  function returns value of array indexed by  $v0$ .

Step 7 involves computation of key constant  $kc = kc + ksc$  and Step 8 involves computation of 32-bit  $v1$  block as shown in Eq. 2

$$v1+ = ((v0 \ll 4) \text{ AND } k2) \text{ XOR } (v0 \text{ AND } kc) \text{ XOR } ((v0 \gg 5) \text{ AND } k3) \quad (2)$$

In step 9, keeping  $k2$  constant, subkey  $k3$  is computed as

$$k3+ = k2 \text{ XOR } \text{xtract}(v1)$$

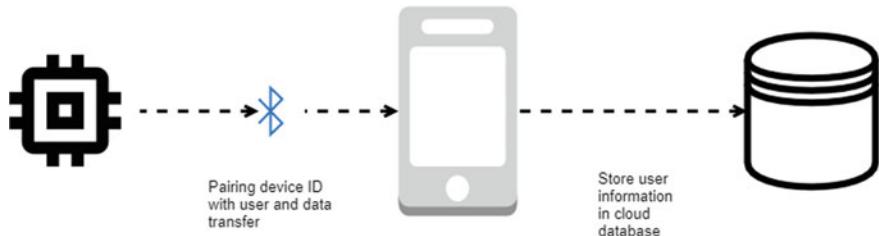
where  $\text{xtract}()$  function returns value of array indexed by  $v1$ .

The final value of  $k1$  and  $k3$  generated after the completion of 64 rounds will be used for decryption. The NTSA Decryption process is the reverse of encryption.

The NTSA encryption and decryption incurs very less time when compared to other symmetric encryption algorithms and due to the less memory utilization, it can be easily incorporated in hardware devices with limitations in storage.

### 3.3 Software Architecture

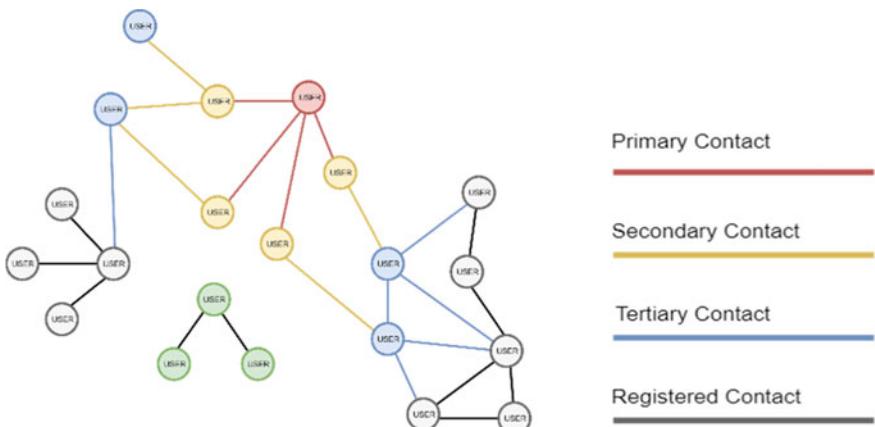
Initially, the hardware module has to be paired with a particular user. The user details along with the device ID are stored in a database. This process is achieved by means of a mobile application which uses the Bluetooth of the smartphone to pair with the proposed proximity sensing module. Once paired, the application prompts the user to enter the user details which will be validated for authenticity. After successful registration or pairing, the module will show indication to notify the user that the device has been successfully linked. Figure 3 shows device pairing and user registration. The real-time application of the device specifies that the device actively monitors as well as broadcasts a signal with the encrypted device ID. A protocol is established among the proximity devices in order to communicate or recognize the signal received from



**Fig. 3** Device pairing and user registration

a similar module. On detecting another proximity device, the firmware running on the microcontroller initiates the distance measurement.

Three functions are specified in the firmware at three different threshold levels based on the distance or proximity. The first 2 functions are triggered when the module crosses 200 and 150 cm radius boundaries, respectively. These two functions are handled at the client level, by providing visual, audio and haptic feedback in order to alert the user. When the proximity crosses a threshold of 100 cm and stays within the zone for over 30 seconds, the encrypted user IDs are exchanged and registered on individual storage memory. This remains in the local storage of the hardware until the device is synced with the smartphone and later it is erased from the local memory as depicted in Fig. 3. Upon syncing the device, the information related to the device interactions is uploaded to the cloud, where the secure information is decrypted and displayed at the client side by means of the graphical user interface of the smartphone application. This information is also logged in the cloud as a graph, for tracking the interactions as depicted in Fig. 4.



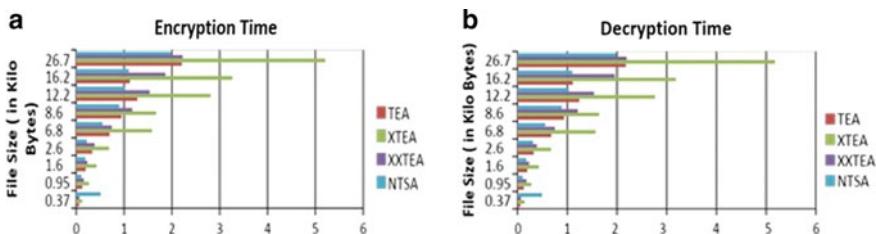
**Fig. 4** Data logged on the server stored in a graph

Every interaction is wiped from the cloud storage so as to reduce the consumption of resources, hence making sure sustainable use of storage resources which would in-turn cater to less energy consumption. For the purpose of disease tracking, a Breadth First Traversal (BFT) algorithm is implemented as the priority is given for immediate or primary contacts and the graph is traversed at a depth 5. The traversal is triggered when the status of a node changes status from negative to positive, hence alerting every node their level of exposure. Information about the patient is confidential and will not be shared with any other users. This path can also be analysed to study the behaviour patterns of the disease spread which will be very useful to take adequate measures or develop new technologies to control disease spread.

## 4 Experimental Results

Tiny Encryption Algorithm was used for many years to securely transmit information through the insecure medium. It was found very efficient in terms of performance, memory and resource utilization, etc. until it suffered from equivalent key attack. The proposed algorithm NTSA had all the features of TEA and additionally was resistant to all the attacks suffered by TEA. Also, NTSA was more efficient than XTEA and XXTEA which are the successors of TEA. NTSA exhibits good avalanche property which is a desirable feature for a good security algorithm wherein a small change in the input shows a significant change in the generated cipher. Various experiments were performed on TEA, XTEA, XXTEA and NTSA algorithms by providing varying plaintext or key, and the time taken to perform encryption and decryption was monitored. Figure 5a, b shows the encryption and decryption time of inputs (plaintext: varying, key: 128 bits). It was observed that the time taken for NTSA is less compared to other symmetric algorithms for performing encryption and decryption, respectively.

Further Fig. 6a, b depicts the encryption and decryption time taken for NTSA algorithm on 64-bit plaintext and 128-bit key.



**Fig. 5** **a** Comparison of encryption time of NTSA with TEA, XTEA and XXTEA for key size of 128 bits. **b** Comparison of decryption time of NTSA with TEA, XTEA and XXTEA for key size of 128 bits

```

a
ENCRYPTION
Plaintext : plaintext
Key String : cryptoalgo@ntsa1
Round 0 to 4 (sec:nanosec)= 0 : 5000
Round 4 to 8 (sec:nanosec)= 0 : 10000
Round 8 to 12 (sec:nanosec)= 0 : 9000
Round 12 to 16 (sec:nanosec)= 0 : 8000
Round 16 to 20 (sec:nanosec)= 0 : 9000
Round 20 to 24 (sec:nanosec)= 0 : 8000
Round 24 to 28 (sec:nanosec)= 0 : 8000
Round 28 to 32 (sec:nanosec)= 0 : 8000

Encrypted String:
??UL??

Time taken for Encryption(sec:nanosec)= 0 : 72000
Accumulated Time for Encryption(sec)= 0.000072 sec

Final text : ??UL??

b
DECRYPTION
Encrypted String :

??UL??

Key String : cryptoalgo@ntsa1
Round 0 to 4(sec:nanosec)= 0 : 6000
Round 4 to 8(sec:nanosec)= 0 : 10000
Round 8 to 12(sec:nanosec)= 0 : 8000
Round 12 to 16(sec:nanosec)= 0 : 9000
Round 16 to 20(sec:nanosec)= 0 : 8000
Round 20 to 24(sec:nanosec)= 0 : 9000
Round 24 to 28(sec:nanosec)= 0 : 8000
Round 28 to 32(sec:nanosec)= 0 : 9000

Decrypted String : plaintext
Time takenfor Decryption(sec:nanosec)= 0 : 73000
Accumulated Time for Decryption(sec)= 0.000073 sec

Final text : plaintext

```

**Fig. 6** **a** Encryption time of NTSA with plain text 64 bits and key 128 bits, **b** decryption time of NTSA with plain text 64 bits and key 128 bits

## 5 Conclusion

An effective measure to track the disease spread at a personal level hasn't been implemented yet. This may be due to the general stigma or the privacy issues faced by the patients. Our primary focus was to address this problem by securing the personal information with a novel cryptographic algorithm NTSA, which is designed to work at embedded level. The device is a low-cost, power-efficient gadget that will aid the users in maintaining effective social distancing and helps track the spread of any kind of contagious disease. All the database updation and data sharing will be through a secured cryptographic process to prevent breach on data privacy and enhance security. Proper measures to structure and collect data of the human interaction and disease spread have been implemented, so as to provide some more data for the research community to work with. The mentioned technology can have a wide variety of applications. However, our primary focus was to work for a socially relevant application.

## References

1. Stalling, W.: Text Book: Cryptography and Network Security, Principles and Practices (2006). Retrieved on 8 Dec 2006
2. Schneier, B.: Applied Cryptography, 2nd edn. Wiley, New York (1996)
3. Wheeler, D., Needham, R.: TEA, a tiny encryption algorithm. <https://www.cl.cam.ac.uk/ftp/papers/djw-rmn/djw-rmn-tea.html>; <https://www.cix.co.uk/~klockstone/tea.pdf>. Accessed 21 May 2007
4. Needham, R.M., Wheeler, D.J. (1997). TEA extensions. Technical Report, Computer Laboratory. Cambridge: University of Cambridge
5. Wheeler, D., Needham, R.: XXTEA: correction to XTEA. Technical Report, Computer Laboratory. University of Cambridge (1998)
6. Tang, H., Sun, Q.T., Yang, X., Long, K.: A Network coding and DES based dynamic encryption scheme for moving target defense. IEEE Access **6**, 26059–26068 (2018). <https://doi.org/10.1109/ACCESS.2018.2832854>

7. Banik, S., Bogdanov, A., Regazzoni, F.: Atomic-AES: a compact implementation of the aes encryption/decryption core. In: Dunkelman, O., Sanadhy, S.K. (eds) INDOCRYPT 2016, 10095. LNCS. Springer, Heidelberg, pp. 173–190 (2016). [https://doi.org/10.1007/978-3-319-49890-4\\_10](https://doi.org/10.1007/978-3-319-49890-4_10)
8. Hoffman, N.: A simplified IDEA algorithm. *Cryptologia* **31**(2), 143–151 (2007)
9. Standaert, F.X., Piret, G., Gershenfeld, N., Quisquater, J.J.: SEA: a scalable encryption algorithm for small embedded applications. In: Workshop on RFID and Light weight Crypto, Graz, Austria (2005)
10. Choi, J., Kim, Y.: An improved LEA block encryption algorithm to prevent side-channel attack in the IoT system. In: 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), Jeju, pp. 1–4 (2016). <https://doi.org/10.1109/APSIPA.2016.7820845>
11. Abdullah, D., et al.: Super-encryption cryptography with IDEA and WAKE algorithm. In: 1st International Conference on Green and Sustainable Computing (ICoGeS) 2017. J. Phys. Conf. Ser. **1019**, 012039 (2018)
12. Ramakrishna Murthy, M., Murthy, J.V.R., Prasad Reddy, P.V.G.D., et al.: Homogeneity separateness: a new validity measure for clustering problems. In: International Conference and Published the Proceedings in AISC and Computing. Springer (indexed by SCOPUS, ISI proceeding DBLP etc), vol. 248, pp. 1–10 (2014). ISBN 978-3-319-03106

# Automatic Depression Level Analysis Using Audiovisual Modality



Aishwarya Chordia, Mihir Kale, Mukta Mayee, Preksha Yadav,  
and Suhasini Itkar

**Abstract** Depression is a mental disorder which is quite common in today's modern world. Depression has a severe impact on a person's mood, feelings and health. For early detection and recognition of depression, use of the computer-vision field is found to be prominent. We have proposed an automated system using ML techniques to automate the process of depression analysis. The system is able to predict depression on Beck Depression Inventory (BDI-II) scale ranging from 0 to 63 by using audio-visual expressions. The video is segregated into two streams: visual and audio. In the visual stream, all the frames are extracted from the video and are preprocessed according to the pre-trained VGG-Face model. Further, the existing Feature Dynamic History Histogram (FDHH) algorithm is modified to capture all the patterns and applied on consecutive frames to extract dynamic features from the feature vector. Singular Value Decomposition (SVD) is applied for dimensionality reduction followed by regression to predict the depression level for visual stream. Mel-frequency cepstral coefficients (MFCCs), zero-crossing, and short-term energy features are extracted from the audio. Dynamic features are extracted by subtracting the components of the consecutive frames. The final feature vector is obtained by making short audio segments and regression is applied on each segment to predict the depression level of the audio stream. Finally, predictions from both the streams are combined by applying decision level fusion technique to predict the final depression level. This system has the best performance on depression level prediction on AVEC 2014 dataset, in comparison with all the existing methods.

## 1 Introduction

Depression is a mental disorder which is quite common in today's modern world [1]. Any mental or behavioural pattern which leaves a person in torment can be referred to as a mental or psychiatric disorder. The distinct causes of depression are biological causes, stress, social factors and lifestyle, influence of family, sociocultural factors and drug or alcohol addiction.

---

A. Chordia (✉) · M. Kale · M. Mayee · P. Yadav · S. Itkar  
PES Modern College of Engineering, Pune, Maharashtra, India

Recent study [2] indicates that for non-fatal health loss, depressive disorders are the single largest contributor. 322 million people in the world suffer from depression, out of which South-East Asia Region and Western Pacific Region contributes to almost half of them. According to Global Burden of Disease Study 2017 [3] depressive disorders were one of the leading causes for both male and female in 2017.

Although depression has severe effects on a person's life, it can be cured in several ways such as medication, physical therapy and psychotherapy. As the number of patients with depressive disorders increases, it puts an excessive load on doctors to diagnose the patients with accurate degrees of depression, therefore lack of long-term follow-up can be considered as a serious limitation for this diagnosis. In order to deal with this limitation, need for an automated and rigorous diagnosis is necessary. Even after diagnosis of depression, recognizing the need of treatment is important. According to [4], high-income countries (64.9%) have greater recognition as compared to upper-middle-income countries (52.2%) and low-/lower-middle-income countries (34.6%) have significantly lower recognition of need for treatment. They stated that only a few people with major depressive disorder (MDD) received minimally adequate treatment.

Recently, a lot of research has been going on for the study of automatic assessment of mental health. For early detection and recognition of depression, use of the computer-vision field is found to be prominent. Lately, to automate the assessment of depression, machine learning and deep learning methods are being used to learn human behaviour related to depression. Face is a vital part of the human body to express emotions. Research was done on facial movements and it was found that the face depicts accurate information about the emotional aspects [5]. A system was developed for expression recognition on the upper face based on FACS using Hidden Markov Models(HMM) [6]. Research states that not only the face, but also the body movements can be used in detection of depression. The process involved in the recognition includes relative orientation and radius of the body parts which was detected using the pictorial structured framework [7].

For automating the process of depression analysis, deep learning methods are popularly being used nowadays. The most common deep learning method used for visual modality is Convolutional Neural Network (CNN) [8]. CNN is considered to be a class of deep neural networks which is used to handle spatial data such as images. Applications of CNN are largest image classification dataset (ImageNet), computer vision, natural language processing (NLP), face recognition, etc. Pre-trained CNN is a network which is trained on large dataset for specific purposes. Example of pre-trained CNN model is VGG-Face which is used for recognition [9].

Audio/Visual Emotion Recognition Challenges (AVEC) is a competition in which they provide a dataset to participants for analyzing depression and emotion recognition. Participants provide solutions by using machine learning methods for automatic audio-visual emotion analysis. Till date, AVEC has conducted nine challenges ( AVEC 2013 [10], AVEC 2014 [11], AVEC 2015 [12], AVEC 2016 [13], AVEC 2017 [14], AVEC 2018 [15], AVEC 2019 [16]), AVEC 2019 being the recent one. The aim of this paper is to implement an artificial intelligent system that can achieve the

best performance on depression level prediction, in comparison with all the existing methods on the AVEC 2014 dataset. This system has many applications such that it can incorporate cognitive capability in robots to automatically recognize a human's mental state.

The main contributions of this paper are : (1) Deep features using VGG-Face and dynamic features using Feature Dynamic History Histogram (FDHH) algorithm are extracted from visual data followed by feature selection which are then given to regression for prediction. (2) Audio features using Mel-frequency cepstral coefficient (MFCC), zero-crossing, and short-term energy are extracted from audio data. From these audio features, dynamic audio features are extracted using the proposed method Audio Dynamic Feature Extraction (ADFE). Separate regression is applied on each segment to predict the BDI level. (3) Prediction level fusion is applied on both the modalities to predict the final BDI level.

## 2 Related Work

In the last 10 years, various methods and approaches for automatic depression assessment based on visual cues have been proposed. In the research field of machine vision, automatic recognition of facial features is becoming an intense area of interest in the field of computer vision. In order to recognize facial expressions from static face images, an automated system is developed by Pantic and Rothkrantz [17]. Contours of facial components like eyes and mouth are extracted and from the extracted contours, feature points of the contours are extracted. 32 AU (Action Units) which appear either in combination or alone are recognized. Behavioral changes can be seen distinctly in depressed patients. Changes in facial movements are proven effective in depression detection. A study was undertaken by Al-Gawwam et al. [18] to detect depression according to the eye blink feature. In the proposed system features such as eye blink per minute, blink amplitude and blink duration are extracted. These features are then classified to classify the subject to be depressed or non-depressed. The system was tested on datasets such as AVEC 2013, AVEC 2014. The above have achieved good accuracy on both the datasets. Various algorithms for visual feature extraction, dimensionality reduction and approaches of classification and regression are implemented. A quantitative analysis of these methods is done by Pampouchidou et al. [19], and accordingly, the results are shown. During the visual feature extraction, features are extracted from full face, AUs ( Action Units), facial landmarks mouth/eyes and heart rate. Feature landmarks are extracted. Variability of facial expressions is measured using AU. Features extracted individually from mouth and eyes contribute greatly. Blinking rate and pupil dilation are reported for the eye region. Out of various classification algorithms, SVM gave the best performance, whereas SVR (Support Vector Regression) gave the best accuracy.

A study was conducted by Dr. Venkataraman and Parameshwaran et al. [20] to detect depression among students using facial features. In this study, the frontal faces of students are captured and facial features are extracted from each frame while they

are answering different questionnaires. These facial features are classified into happy, contempt and disguised faces using SVM classifier. The face detection required for this study was done with the Viola-Jones face detection algorithm. To achieve this system was developed by Meng and Pears al. [21] to overcome the limitations of Motion History Image (MHI). They proposed a new method called Motion History Histogram (MHH) to capture dynamic information of the video. The classification was improved with a combination of MHH and MHI and further processed by a support vector machine (SVM) to give the final result. The proposed representation is improvisation of the previous method, the improvising is done by storing additional frequency information at every pixel. The study has a good performance over a large public human action database.

Audio can also be used for depression level analysis. Yang et al. [22] studied the effect of quantitative features of vocal prosody on depression severity. They concluded that change in depression severity divulged by combination of F0 and switching pauses. As vocal prosody is a powerful measure not just that but other vocal features are also useful. Mitra et al. [23] explored various features such as estimated articulatory trajectories during speech production, acoustic characteristics, acoustic-phonetic characteristics and prosodic features. Support vector regression, a Gaussian mixture model and decision trees are applied on the above features. They showed comparative results of effects of various features on depression analysis. According to them, Damped Oscillator Cepstral Coefficients showed better results.

For extracting handcrafted features a large amount of effort in terms of domain knowledge and labor is required which is quite time-consuming. Recently, deep learning techniques provide solutions to the above difficulties. For improving the method of automatic depression analysis using audio features He and Cao [24] proposed a method which combined deep learning and traditional methods. Deep Convolutional Neural Networks (DCNN) is implemented to extract deep learning features from spectrograms and raw speech waveforms. Median robust extended local binary patterns (MRELBP) are extracted manually from spectrograms. For enhancing the depression recognition performance, Joint fine-tuning layers are created to combine the handcrafted features and deep learning-based features. Lang He also stated that this method showed better performance on both AVEC 2014 and AVEC 2013 dataset of depression analysis.

Further research proved that systems consisting of both audio and visual features helped advance the computer vision field and one of the methods was proposed by Jan et al. [25] in which Motion History Histogram (MHH) was used along with the audio features. The dataset used in the system was AVEC 2014 which provides several audio features like spectral low-level- descriptors and MFCC 11-16. For visual processing a new method, MHH for 1D features was proposed based on a previous MHH for 2D feature. The combined features of audio and visual are tested on Partial Least Square regression and linear regression separately and combined to give the final depression level. The combined method achieves good results on the used dataset.

Deep learning techniques are also used for extracting visual features which are beneficial for recognizing depression level. Chao et al. [26] proposed a model for

depression analysis for both the modalities. They extracted facial features using CNN, face shape features using OpenCV Viola-Jones and audio features using YAAFE toolbox. SVD dimensionality reduction technique is applied on the individual features and then combined at a multimodal fusion layer. The multimodal feature sequence is then provided to LSTM-based neural network.

Jan et al. [27] proposed an automated artificial intelligent system, which uses two modalities audio and visual for depression level prediction. They used a VGG-Face pre-trained model for visual feature extraction. They proposed a method, Feature Dynamic History Histogram(FDHH) to extract dynamic features for visual. And for audio, they extracted Mel-frequency cepstral coefficients (MFCC) features. These features are then combined using feature level fusion techniques. They applied Principal Component Analysis(PCA) Dimensionality technique to reduce dimensions of fused feature vectors and performed Partial Least Square (PLS) and Linear regression (LR) to map it with BDI level. This system achieved good performance on the AVEC 2014 dataset.

Although this system is efficient, few drawbacks of this approach are, for visual processes, the patterns considered are not sufficient enough to make an accurate prediction. The results can be further improved if more such patterns are taken. In the above approach of the audio process, mean of the segments were taken which does not take into consideration the change in features for each segment. They extracted only MFCC features which are considered for the predictions, but incorporating other features can improve the results. The motive of this paper is to develop an automated system which can achieve better performance than all the existing ones on AVEC 2014 Dataset. Improvement will be made in both audio-visual streams as well as fusion and prediction techniques.

### 3 Framework

According to the study, it is observed that there is a significant change in human's facial and vocal expressions of a depressed person than that of a normal person. These human expressions can effectively contribute towards depression level prediction. A new method is developed which utilizes dynamic visual and vocal features of a person to automatically predict his depression level.

#### 3.1 Overview

For the visual process, the video is broken down into frames and each frame is preprocessed. Visual features are extracted using the deep learning technique for every frame. After normalization, FDHH algorithm is applied to get the patterns. Then the patterns are converted into one single vector to get the final feature vector for a single video sample. The dimensionality of the features is reduced using SVD

and linear regression is applied to get BDI value. For the audio process, audio data in which audio is separated from the video. Audio features are extracted followed by dynamic features. These features are converted into various segments and on each segment regression is applied. Final audio BDI value is obtained. Then both the streams predictions are combined by applying prediction level fusion to get the final BDI prediction.

### 3.2 Visual Module

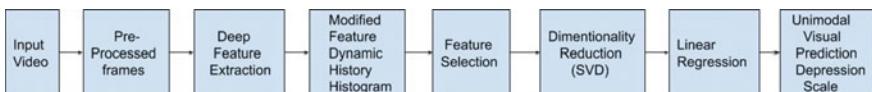
Figure 1 demonstrates the extraction of visual features using deep learning technique after which dynamic features are extracted and reduced using dimensionality reduction technique and given to regression to predict BDI level.

**Pre-processing** Pre-Processing of data is a vital technique that involves transforming raw data into an understandable format. In order to improve the performance of machine-learning algorithms, pre-processing offers solutions which help to accurately detect the outliers and missing values in all kinds of large datasets [28]. So necessary preprocessing steps are applied on the dataset.

**Deep Features Extraction** VGG-Face: We have utilized VGG-Face, a pre-trained model for feature extraction. VGG-Face is one of the most popular and widely used networks for face recognition in which 2.6 M facial images are trained for the application of face recognition. VGG-Face is more suited for depression analysis tasks as compared to other networks because here, we are dealing with facial images as opposed to objects from the ImageNet dataset. VGG-Face [9] contains a total of 36 layers. For facial feature extraction, we have used layer 32 as it gives the best results with 4096 dimensions.

**Modified FDHH algorithm** FDHH(Feature Dynamic History Histogram) has been used previously for dynamic feature detection [27]. A new special condition is added which was missed by the previous algorithm. As the algorithm counts the pattern only if ‘0’ is encountered, there is a condition such that if the count of  $CT \geq M$  and the sequence is completed the algorithm will move on to the next sequence without counting the last pattern. To overcome this limitation a new condition is added, that is if the component in the last frame is scanned to be ‘1’ then  $CT$  is checked, if  $CT \geq M$  then  $P_m$  is updated else the algorithm is skipped to the next sequence.

**Feature Selection** Feature selection is used to reduce the features. It can be done by removing irrelevant noisy and redundant features from the feature space and selecting



**Fig. 1** Architecture diagram

a subset of features with which best performance can be achieved. As proclaimed by Blum and Langley [29] Feature selection comprise of four main components which are:

- Starting point of feature selection
- Search procedure
- Evaluation function
- Criterion for stopping search.

In the proposed system, the patterns are extracted up to  $M = 10$ . Features are selected according to their single and combined performance on the data. From the experimental results, the subset of features consisting of features 2, 6, 9 and 10 is found to be the best subset to achieve the maximum accuracy on the system.

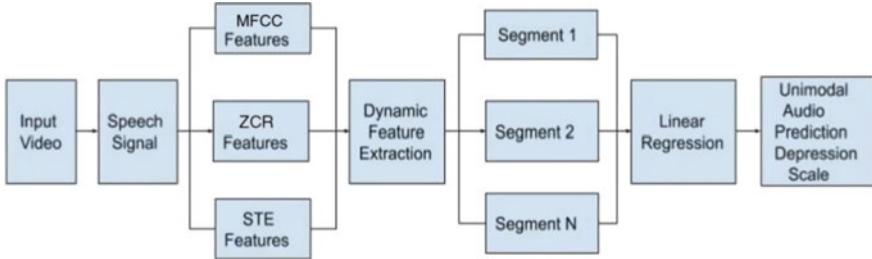
**Singular Value Decomposition** Dimensionality reduction helps to improve the performance of the system by reducing the features or representing the features in a lower dimensional space. Singular value decomposition is the technique that can be used to reduce the dimensionality in which higher dimensional data is represented into a lower dimensional space while retaining important information. SVD also has a feature of restoring the original matrix  $M$  which was decomposed into  $U$ ,  $S$  and  $V$ . Each of the singular values in diagonal matrix  $S$  can be used to understand the amount of variance. This is used in the proposed system to represent the final feature vector of the visual process in lower dimensional space and decorrelate the features. 99 dimensions are retained in the data.

**Regression** Linear regression finds its use in various practical applications because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters. We have performed linear regression on both the modalities, that is audio and visual and prediction level fusion is done to predict the final depression level.

### 3.3 *Audio Module*

Figure 2 demonstrates the process for audio data in which audio is separated from the video. Features such as Mel-frequency cepstral coefficients, zero-crossing, and short-term energy are extracted from the audio data. Dynamic features are extracted by taking difference of the 2 consecutive frames. These features are converted into various segments and on each segment regression is applied.

**Features Extraction** An individual's voice speaks a lot about their emotions, personality, mood, identity and other unique factors which are useful for voice recognition and other audio signal processing tasks. These audio features can also be useful in depression level analysis tasks [30]. For audio modules, audio is separated from the video provided by AVEC2014 dataset. According to the comparative study of several combined audio features and MFCCs in [27], their results show that MFCC is the



**Fig. 2** Architecture diagram

most dominant feature which is individually sufficient for recognizing depression. Our experiment shows that combining MFCCs with few more audio features such as zero-crossing rate and short-term energy will enhance the overall performance.

- MFCCs:

The most used feature in audio processing is MFCCs [31], Which are based on human ear scale and lower dimensions. These coefficients use a nonlinear frequency scale which is based on human auditory perception, that is Mel scale. MFCC usually provides 0-20 coefficients from which for our experiment we selected 16 coefficients.

- Zero-Crossing Rate:

When the sign of a mathematical function changes from positive to negative it is said to be a zero-crossing point. Zero-crossing rate is a count of the number of times the signal waveform crosses the amplitude line. ZCR is used to distinguish between audio classes such as noise, silence and speech [32]. For each analysis window, ZCR can be calculated using Eq. (1) :

$$\text{ZCR} = \frac{1}{2N} \sum_{n=0}^{N-1} |\text{sgn}[x_i(n)] - \text{sgn}[x_i(n-1)]| \quad (1)$$

where,  $N$  is total number of samples in a processing window,  $x(n)$  is the value of  $n$ th sample,  $\text{sgn}()$  is the sign function, i.e., Eq. (2)

$$\text{sgn}[x_i(n)] = \begin{cases} 1 & \text{if } x_i(n) \geq 0 \\ -1 & \text{if } x_i(n) \leq 0 \end{cases} \quad (2)$$

- Short Time Energy (STE):

Energy is the notable attribute according to human Auditory perception. Energy of speech signals vary over time in nature. It is also referred to as volume, loudness, intensity of speech signal [43]. The total energy is evaluated using Eq. (3):

$$\text{STE} = \sum_{n=-\infty}^{\infty} s^2(n) \quad (3)$$

**Dynamic Feature Extraction** While extracting above audio features frames are generated. These frames are with respect to time domain so considering adjacent frames will provide useful emotional characteristics which are changing over time. All these changes which are the dynamic features are beneficial for predicting depression level of the speaker. To enhance the results we calculated dynamic features using previously extracted features MFCC, ZCR, STE.

The Dynamic Feature Extraction process is as follows:

- Step 1: Read the extracted MFCC, ZCR and STE features of audio.
- Step 2: Combine these extracted features and form a final feature vector
- Step 3: Calculate the difference of all the features from the combined feature vector of Adjacent frames.
- Step 4: Perform normalization at different scales on each feature of above difference output vector.

**Segment-Wise Regression** The next step after dynamic feature extraction is applying regression to map the extracted features with given BDI-II scales. Dataset provided videos are of different length, so while converting feature vectors into same dimensions, we may lose certain important features. So instead of using full feature vector, we divide the vector into small segments containing a particular number of frames of each feature. We considered different numbers of frames per segment for evaluating results of regression. According to these results, we selected an optimal value 16 as the number of frames per segment. We generated a number of segments consisting of 16 frames of each feature according to length of videos. Every single segment is given as a sample to the linear regression for mapping it with respective BDI-II scale. All the predictions generated are then aggregated to generate a single BDI value for a video.

**Prediction Level Fusion** As a system output we require a single depression level to get this we used a technique for prediction level fusion. As we want to keep the efforts of both audio and visual modalities they are linearly fused together using weighted sum rule to aggregate their predictions. The Rule applied is as shown in Eq. (4)

$$S_{\text{fusion}} = w_1 s_1 + w_2 s_2 \quad (4)$$

where,  $w_1 w_2$  are weights for each modality,  $s_1$  are predictions from visual modality,  $s_2$  are predictions from audio modality

## 4 Experimental Results

### 4.1 AVEC 2014 Dataset

The approaches we have proposed are performed on the Audio/Visual Emotion Challenge (AVEC) 2014 dataset, a subset of audio-visual depressive language corpus (AViD-Corpus) which is used for depression sub-challenge [16]. AVEC2014 uses only 2 tasks for evaluation of depression which are referred as Freeform and Northwind tasks. The task includes 2 human-computer interaction tasks recorded by microphone and web-cam. Some subjects appear in more than one clip and the length of these clips is between 6 s to 4 min and 8 s. The participants are recorded between one to four times with a gap of two weeks between each recording. For both the tasks, the recorded videos are split into 3 partitions: training, development and test of 50 videos each, thus a dataset containing a total of 300 video clips. In our system, we merge the training and development set from both Freeform and Northwind data as one training set. The performance is measured for video clips from the test set. The BDI-II depression scale ranges from 0 to 63 where

- 0–10—normal mood
- 11–16—mild mood disturbance
- 17–20—mild depression
- 21–30—moderate depression
- 31–40—severe depression
- Over 40—extreme depression.

### 4.2 Experimental Setting

The experiment is tested on a dataset provided by AVEC challenge. The baselines are provided by [11]. The aim is to predict the BDI level by using the given dataset.

**Data Pre-processing** As the data is given to a pre-trained model, the pre-processing steps are necessary in order to ensure the optimal features. The features are extracted by inputting a single frame at a time to VGG-Face pre-trained network. The frames were resized to 227 \* 227 \* 3 before giving them as input. Also the mean image is subtracted. After following these steps, the frame is inputted to the network. In order to extract audio from video provided by AVEC2014, we used FFmpeg software suite. Librosa is a library used for extracting audio features that requires a wav file as input. Wav file is generated.

**Feature Extraction** The deep features for the visual process are extracted with the help of a pre-trained model which is provided by rcmail on GitHub. The library and the related information will be available on the specified GitHub repository [33]. For each frame in the video deep features are extracted. Total 4096 features are extracted from the 32nd layer of the VGG-Face network.

While extracting the dynamic features the value for M is taken to be  $M = 10$ . The threshold value is set to  $T = 0.00392$ . Feature selection is applied on the respective patterns. The number of patterns is chosen according to their individual performance on the system and analyzing experimental results. According to the output of feature selection, patterns 2, 4, 6, 9 are chosen for the further process. The dimensions for each pattern is  $1 * 4096$ , 4096 being the number of components therefore the total dimensions of the 4 patterns would be  $4 * 4096 = 16,384$  features per single video sample. As each subject has 2 videos (northwind and freeform) total 32,768 features are extracted for the final feature vector.

Librosa Library is the most preferred python package for audio analysis. It contains feature extraction as a submodule which includes low level feature extraction such as MFCC, ZCR, STE and many more methods. It also provides feature manipulation methods.

As the video provided by the AVEC2014 is recorded at a sample rate of 44,100 Hz and hop size of 512. So for every 1 s 86 frames are generated by using formula  $(\text{sec} * \text{sample rate})/\text{Hop size}$ , so for 1 s  $1 * 44,100/512 = 86$  frames. For every feature total frames are generated depending upon video length. As we extracted 3 different features that are 16 MFCCs, ZCR, and STE, so total dimension of feature vector for 1 video is  $18 * \text{frames}$ . As mentioned earlier dataset has two tasks freeform and northwind so the combined feature vector will be of  $18 * (\text{frames of Freeform} + \text{frames of Northwind})$  dimensions.

For dynamic feature extraction, every consecutive frame is subtracted for each feature, So the resultant feature vector will be of  $18 * (\text{frames of Freefrom} + \text{frames of Northwind} - 1)$  dimensions. To eliminate the differences in numerical value in each feature of the final feature vector of differences, it is then normalized.

Before applying regression we generated segments containing 16 frames per feature. So the segment dimension will be  $18 * 16$  (i.e., 16 MFCCs, ZCR and STE). Every individual segment is considered as a sample for regression. All segments of that video are mapped with the same BDI-II scale. For obtaining the best result the weights are optimized by observing the performance of audio and visual modalities.

### **4.3 Performance Comparison**

For measuring accuracy of a system, two most common metrics used are Mean Absolute Error and Root Mean Square Error.

Starting with best performing Uni-modal audio, Table 1 demonstrates the performance of dynamic features (DF) of Mel-frequency-cepstrum coefficients features (MFCC), zero-crossing rate (ZCR), and short-term energy (STE) followed by segment-wise linear regression (LR) to map the feature vector with provided BDI-II scale. Results show that audio modules RMSE and MAE are 9.9 and 9.7 on the development set and 9.3 and 9.0 on the test set, respectively.

Table 2 depicts the performance of visual modality by applying modified FDHH algorithm on deep features extracted from VGG-Face (denoted by VGG-F) and

**Table 1** Performance measure of audio stream

Method	Develop		Test	
	RMSE	MAE	RMSE	MAE
DF (MFCC + ZCR + STE) + LR	9.9	9.7	9.3	9.0

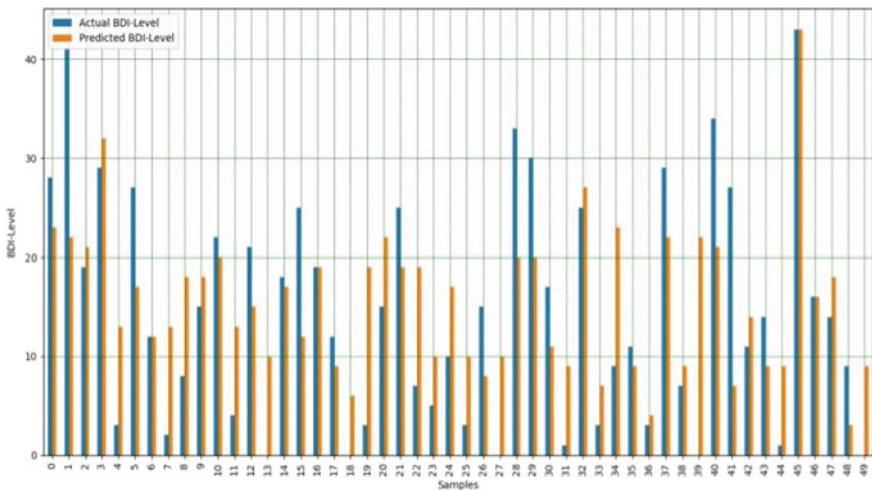
**Table 2** Performance measure of visual stream

Method	Develop		Test	
	RMSE	MAE	RMSE	MAE
FDHH(VGG-Face) + LR	9.22	7.14	9.4	7.7

**Table 3** Performance measure of combined model (Audio + Visual)

Method	Develop		Test	
	RMSE	MAE	RMSE	MAE
Visual + Audio	9.11	7.42	8.88	7.18

features are selected based on their individual performance, RMSE of individual pattern were used for evaluation and the best performing patterns were selected and then linear regression (LR) was applied on it to map it with BDI-II scale.

**Fig. 3** Comparison graph

**Table 4** System performance using audio and visual modality

Method	System Testdata	
	RMSE	MAE
System (Visual + Audio)	7.7	5.2

**Table 5** Performance comparison against other approaches of audio stream

Method	Develop		Test	
	RMSE	MAE	RMSE	MAE
Baseline [11]	11.52	8.93	12.567	12.567
Jan [27]	10.92	8.86	10.28	8.07
Mitra et al. [23]	7.71	6.10	11.10	8.83
Jan et al. [25]	10.69	8.92	11.30	9.10
Chao et al. [26]	11.16	8.94	10.61	8.70
Our method	9.9	9.3	9.7	9

These two models audio and visual are then combined using prediction level fusion technique, i.e., weighted sum rule. The results for the same are shown in Table 3.

Figure 3 illustrates the system performance by comparing the predicted depression level from combined audiovisual modality and the actual depression level values provided by the dataset.

Table 4 shows the performance of the proposed standalone system which was trained using 90% of the AVEC4014 Dataset (i.e., training set, development set and some of test set). The system is tested on the remaining 10% of the dataset and the results are remarkable in terms of RMSE and MAE both.

Table 5 shows performance comparisons of performance of audio modality with other paper's performance on only audio modality including baseline. We can conclude by the comparison that our proposed audio modality gives better results in terms of RMSE for both development and test set.

## 5 Conclusion

In this paper, an automatic system is proposed to predict the depression level in human beings according to their visual and vocal features extracted from the video recordings. The system consists of two modalities. In visual processing, deep features are extracted from the facial expressions. Modified FDHH algorithm is applied to extract the dynamic features. SVD is used to reduce the dimensionality of the dynamic features, and finally, regression is applied to predict the final BDI level. In audio processing MFCC, zero-crossing and energy features are extracted and dynamic

features are obtained from them. The extracted dynamic features are divided into smaller segments which are further given for regression separately to predict the BDI level. Output for all the segments is combined to give a single final predicted value for the audio stream. Predicted values from both the streams are combined by decision level fusion and single value for each subject is predicted. The results for the multimodal approach were better than most of the existing systems. The results of audio approach were remarkable in comparison with the existing unimodal approaches.

A multimodal system is developed which gives better performance on the development set than the baseline result and the previous state of art result. There are certain limitations of this system which can be improved to further increase the performance of the system. The basic limitation of the system is, the BDI level given by the dataset is the response to a certain questionnaire and therefore the scale can certainly be related to the questions itself and may not be the present depression level of the patient.

The performance of the system can be further improved if the VGG-Face network is re-trained by adding more data. Other dimensionality reduction techniques can be considered to improve the results.

## References

1. Marcus, M., et al.: Depression: A global public health concern (2012)
2. Depression and Other Common Mental Disorders (2017)
3. Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. vol. 392 (2018)
4. Thornicroft, G. et al.: Undertreatment of people with major depressive disorder in 21 countries. Br. J. Psychiatr. (2017)
5. Girard, J.M., et al.: Social risk and depression: Evidence from manual and automatic facial expression analysis. In: IEEE International Conference on Automatic Face and Gesture Recognition (2013)
6. Lien, J.J., et al.: Automated facial expression recognition based on FACS action unit. IEEE (1998)
7. Kaletsch, M., et al.: Major depressive disorder alters perception of emotional body movements. Frontiersin (2014)
8. Al-Zawi, S., Albawii, S., Mohammed, T.A.: Understanding of a convolutional neural network. ICET (2017)
9. Zisserman, A., Parkhi, O.M., Vedaldi, A.: Deep face recognition. Visual Geometry Group Department of Engineering Science University of Oxford (2015)
10. Valstar, K.M., Schuller, B.: AVEC 2013—The continuous audio/visual emotion And depression recognition challenge. ACM (2013)
11. Valstar, M., et al.: AVEC 2014—3D dimensional affect and depression recognition challenge. ACM (2014)
12. Ringeval, F., et al.: AVEC 2015—The 5th International Audio/Visual Emotion Challenge and Workshop. ACM (2015)
13. Valstar, M., et al.: AVEC 2016—Depression, Mood, and Emotion Recognition Workshop and Challenge. ACM (2016)

14. Valstar, M., et al.: AVEC 2017—Real-life Depression, and Affect Recognition Workshop and Challenge. ACM (2017)
15. Valstar, M., et al.: AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition. ACM (2018)
16. Ringeval, F., et al.: AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. ACM (2019)
17. Pantic, M., Rothkrantz, L.J.M.: Facial Action Recognition for Facial Expression Analysis From Static Face Images. IEEE (2004)
18. Al-Gawwam, S., Benaaissa, M.: Depression Detection From Eye Blink Features IEEE (2018)
19. Pampouchidou, A., et al.: Automatic assessment of depression based on visual cues: a systematic review. IEEE (2017)
20. Venkataraman, D., Parameshwaran, N.S.: Extraction of facial features for depression detection among students. Int. J. Pure Appl. Math. **118**(7) (2018)
21. Meng, H., Pears, N.: Descriptive temporal template features for visual motion recognition. Pattern Recogn. Lett. (2019)
22. Fairbairn, C., Yang, Y., Cohn, J.F.: Detecting depression severity from vocal prosody. IEEE Trans. Affect. Comput. **4**(2) (2013)
23. Mitra, V., et al.: The SRI AVEC-2014 Evaluation System. ACM (2014)
24. He, L., Cao, C.: Automated depression analysis using convolutional neural networks from speech. Elsevier (2018)
25. Jan, A., et al.: Automatic depression scale prediction using facial expression dynamics and regression. IEEE (2014)
26. Tao, J., Chao, L., et al.: Multi task sequence learning for depression scale prediction from video. IEEE (2015)
27. Jan, A., et al.: Artificial intelligent system for automatic depression level analysis through visual and vocal expressions. IEEE (2017)
28. Saleem, A., et al.: Pre-Processing Methods of Data Mining, 7th International Conference on Utility and Cloud Computing. IEEE and ACM (2014)
29. Langley, P., Blum, A.L.: Selection of relevant features and examples in machine learning. Elsevier (1997)
30. Alfàs, F., Socoró, J.C., Sevillano, X.: A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds. Appl. Sci. MDPI (2016)
31. Ranjan, R., Thakur, A.: Analysis of feature extraction techniques for speech recognition system. Int. J. Innov. Technol. Explor. Eng. (IJITEE) (2019)
32. Bachu, R.G., et al.: Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In: Conference ASEE Regional Conference (2008)
33. Keras-Vggface: <https://github.com/rcmalli/keras-vggface> (2017)

# A Notification Alert System with Heartbeat and Temperature Sensors for Abnormal Health Conditions



V. Sireesha, M. S. V. Sashi Kumar, S. Vinay Kumar,  
and R. M. Shiva Krishna

**Abstract** Generally, we go to a hospital for two reasons, that is, for a regular check-up for aged people and the other is during health issues. For some cases, we need an Ambulance or any means of transport to reach the hospital when the condition becomes critical. Unfortunately, in such cases, we don't have any predictive systems which keep track of the health conditions of the people and notifies when the situation becomes abnormal. Therefore, in our paper, we have addressed this issue with the help of a system which keeps track of our health and raises an alert notification to our relatives and the doctors of the nearby hospitals or maybe a family doctor who knows our condition well, in case of any abnormal health condition. The proposed system consists of device trackers and sensors to keep track of data and uses a messaging service to send alert notification. This work can further be improvised to suit the current pandemic situation (COVID—19) which can be used as a tracker for tracing the COVID patients. We can place all the required sensors like temperature sensor, heartbeat sensor, pressure sensor, wearable sensors.

## 1 Introductions

A wireless sensor network is a large node of small wireless sensor nodes. There are small batteries, limited power, and a limited microprocessor at the sensor nodes. These sensor nodes primarily aim to collect and transfer the collected information from all sensors to a base station.

---

V. Sireesha (✉) · M. S. V. Sashi Kumar · S. Vinay Kumar · R. M. Shiva Krishna  
Department of CSE, Vasavi College of Engineering, Hyderabad, Telangana, India  
e-mail: [v.sireesha@staff.vce.ac.in](mailto:v.sireesha@staff.vce.ac.in)

M. S. V. Sashi Kumar  
e-mail: [m.sashikumar@staff.vce.ac.in](mailto:m.sashikumar@staff.vce.ac.in)

S. Vinay Kumar  
e-mail: [s.vinaykumar@staff.vce.ac.in](mailto:s.vinaykumar@staff.vce.ac.in)

If we consider the present situation, the person who has any health issues, needs to get admitted in the hospital and get treated till it gets cured. If the case is not too severe, they will follow the medical prescription prescribed by the doctors. But there are few health issues which persists for a longer time like Blood Pressure, Diabetes, Thyroid issues and others. So for these kinds of problems, the treatment should be taken regularly, and the seriousness of the case may go to any extent. The patient condition cannot be judged at any time. Such patients should be admitted in the hospital for a longer time and get treated, which may not be feasible for them economically as private hospitals charge higher costs for those treatments. Therefore, if the person is at home, and once we connect the sensors to him, his or her health can be monitored minute to minute. Once the condition becomes critical, instead of calling for emergency services like Ambulance, the notification would be sent to the nearby doctor along with the location of the patient so that the doctors can send the Ambulance to the site.

This paper mainly focuses on tracking user data and sending an alert notification to concerned people to alert about the condition of the patient. The message sent includes the values of the heartbeat and temperature of the person. So the patients need not be in the hospital all the time. The regular check-up and monitoring can be done at home, which is an economically feasible solution from the patients' point of view. Also during this pandemic this system is beneficial to most of the people and it may become the new norm for treatments. The paper is further organized as follows. Literature survey included in Sect. 2; Proposed system included in Sect. 3. Implementation details included in Sect. 4 and Results and Discussion in Sect. 5. Finally, the Conclusion and future scope included in Sect. 6.

## 2 Literature Survey

The rapid increase in the emergence of new technologies have paved the way for the introduction of many smart notification systems for the convenience of the users. The authors in [1–3], have proposed an intelligent alert system in case of accidents or emergencies. In [4], the authors have developed a smart alert system for waste management and issuing an alert when the bin is full. The authors in [5–8] have proposed smart notification systems for different health care applications. These smart alert system approaches have motivated us to think in the direction of the proposed method.

## 3 Proposed System

Sending alert notification in case of emergency is the key concern in this system. The registered users can only opt for this service. The sensors used in this design are

1. LM35 Temperature Sensor
2. Heartbeat sensor
3. Breadboard
4. 10 K Ohm Resistors
5. 9 V battery (power supply)
6. Esp8266 NodeMCU with in-built Wi-Fi module.

The technologies required in developing this system are Arduino IDE, Web Application (000webhost.in), and SQL Database.

### ***3.1 Launching of System***

The heartbeat and the temperature sensors aforementioned are connected to the human body, and the other terminals are connected to the breadboard along with the resistors. The breadboard is, in turn, connected to the power supply from the battery. The temperature sensor records the body temperature and displays it on the User Interface. The heartbeat sensor also tracks the heartbeat of the patient and shows in the UI.

The Arduino IDE has code which collects the readings from sensors and sends it to the web page. The web page has the threshold value calculated based on the medical history given by the user during sign up. Therefore, the computed value is taken as the threshold value. Now the current heartbeat values are compared to the threshold, and a grace value of a random integer is taken. If the value crosses the total reading, then the counter starts to increment continuously. Once it reaches the threshold, then the messaging service is triggered, and an alert is sent to the registered contact numbers.

This system is useful for people of all the ages instead of focusing on a specific age group. The sensors data is read and compared with the threshold, and a counter is maintained. Once the counter reaches a particular count, the alerts would be sent because we consider any case to be abnormal or critical if it exists, for a considerable amount of time. For example, when a person runs or walk for a few seconds, their heartbeat increases. This condition doesn't mean the person is suffering from a heart attack. So a counter is used to handle these kinds of exceptions.

### ***3.2 Algorithm Description***

**Step 1:** Initially, the setup is made to connect the NodeMCU module to Wi-Fi so that it can send the readings to the web page where the user is signed in.

**Step 2:** Import packages such as ESP8266HTTPClient.h, ESP8266Wifi.h. PulseSensorPlayground.h packages to capture the heartbeat and temperature from the sensors.

**Step 3:** The pulse readings are captured and compared against the threshold value, and the counter is incremented based on the value.

**Step 4:** Once the counter reaches the value of 20, SMS service is triggered, and the alert is sent to the concerned authorities and family members.

**Step 5:** Once the alert is sent, the counter resets itself and starts re-recording the values, and the alerts would be sent continuously till the patient is in contact with those sensors.

### 3.3 Flow Chart

The proposed system, as shown in Fig. 1, has been developed as there is no kind of system currently available so that the patient can be treated at home and can be taken to hospital only in case of emergency. The present systems suggest the patients to get admitted into a hospital over a random reading of these vitals, which could very well be a false positive.

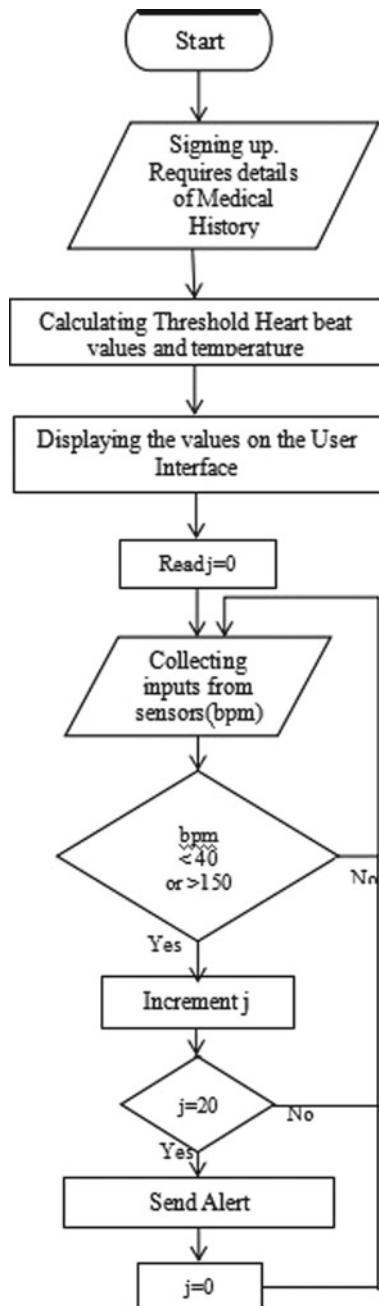
This work can also be considered as an instant notification system as the alert goes immediately, and the patient needs not respond or call anyone. The proposed setup has a limitation of sending a limited number of messages. It can be replaced with any paid services like Twilio, Vonage, and TextLocal to be able to send messages from time to time. Within the time interval, if the messages reach the count, they stop sending messages. This is handled as the counter resets itself and starts sending again after the time interval. This counter value is based on the service provider. For a free service, the cost is less, but if users opt for a premium service and pay monthly charges, the messages count would be larger considerably.

## 4 Implementation

There is a web application hosted on the server where the user who wants to use this server needs to register first. There is a login portal and sign up page where the user needs to give necessary details like name, age, gender and medical history. So once the user logs in, he or she can see the details (heartbeat, temperature) on the User Interface in the website. Based on the details provided, the system predicts the functionalities of the body parts of the patient. It foretells the heartbeat and temperature of the person based on medical history. These defaults are considered as a threshold value and beyond that and below values are cases of abnormalities.

There are sensors connected to the human body, and they always track the heartbeat and temperature. The platforms used to build this system is Arduino IDE [9] which is used to collect data from sensors, 000webhost platform to host the website for the users to login to use this system and IFTTT [10] services to send messages.

**Fig. 1** Flow chart of the proposed system



IFTTT is a specific service which is open source and can be opted once you sign up. It is a kind of platform where there are many services available of which we can opt based on our requirement. IFTTT stands for if-this-then-that. It is a triggering service which is triggered based on the counter, which counts the abnormal heartbeat values. IFTTT has many services available like messaging service, calling service, email service, reminder service, and many more.

The service is triggered using an URL which is provisioned the subscribed users. For an account, a limited number of messages can be sent, and after a specific period, the service starts automatically. It resets itself after a particular period. If it crosses the limit of the messages, the service stops itself, but the counts are considerably large.

## ***4.1 Setup of NodeMCU***

This section deals with connecting nodemcu module to wifi by providing username and password.

```
#include <ESP8266WiFi.h>
void setupWifi()
{
    WiFi.mode(WIFI_STA);
    status = WiFi.begin(ssid, password);
    Serial.print("Attempting to connect to SSID: ");
    Serial.println(ssid);
    // Wait for connection
    while (WiFi.status() != WL_CONNECTED) {
        delay(1000);
        Serial.print(".");
    }
    Serial.println("Connected to wifi");
}
```

## ***4.2 Pulse Monitoring***

This section deals with using in-built libraries in Arduino IDE to get the heart beat values. The inbuilt function `getBeatsPerMinute()` reads the value from sensor.

```
#include <PulseSensorPlayground.h>
int bpm=0;
PulseSensorPlaygroundpulseSensor;
void loop() {
    if (WiFi.status() != WL_CONNECTED)
    {
        setupWifi();
    }
    if (pulseSensor.sawNewSample()) {

        if (--samplesUntilReport == (byte) 0) {
            samplesUntilReport = SAMPLES_PER_SERIAL_SAMPLE;
            // pulseSensor.outputSample();
            bpm=pulseSensor.getBeatsPerMinute();
            if (pulseSensor.sawStartOfBeat()) {
                bpm= pulseSensor.getBeatsPerMinute();
            }
        }
    }
}
```

### 4.3 Evaluation

This section deals with comparing the heartbeat with threshold values and increment the counter. Once the counter reaches a specific value, the messaging service is triggered through get\_http2 function.

```
int j=0;
if(bpm>150 || bpm<40)
{
    j++;
    if(j==20)
    {
        get_http2(String(bpm));
        j=0;
    }
}
```

#### 4.4 Messaging Service

This section of code deals with invoking the messaging service and send the values to the server.

```
#include <ESP8266HTTPClient.h>
int get_http2(String state)
{
    HTTPClient http;
    int ret = 0;
    Serial.print("[HTTP] begin...\n");
    // conFig.ifttt server and url should be HTTP only..not
    https!!! (http://)
    http.begin("http://heart-beat-project.000webhostapp.co
m/updateHeart.php?id=2"+ "status=" + state); //HTTP

    http.begin("http://maker.ifttt.com/trigger/HeartAttack
    /with/key/h0CbltIVx2s_b9L4pBIEq96OyiKLpQSAZTIi8dMNaMF?val
    ue1=" + state); //HTTP

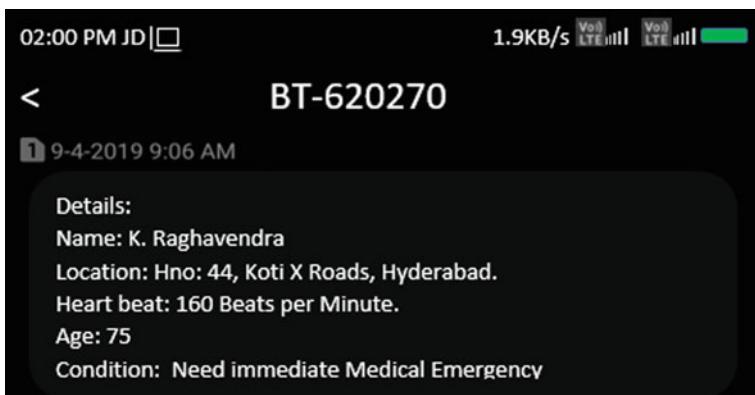
    // start connection and send HTTP header
    int httpCode = http.GET();
    // httpCode will be negative on error
    if(httpCode > 0) {
        // HTTP header has been send and Server response header
        has been handled
        Serial.printf("[HTTP] GET code: %d\n", httpCode);
        if(httpCode == HTTP_CODE_OK) {
            String payload = http.getString();
            Serial.println(payload);
        }
    } else {
        ret = -1;
        Serial.printf("[HTTP] GET failed, error: %s\n",
        http.errorToString(httpCode).c_str());
        delay(500); // wait for half sec before retry again
    }

    http.end();
    return ret;
}
```

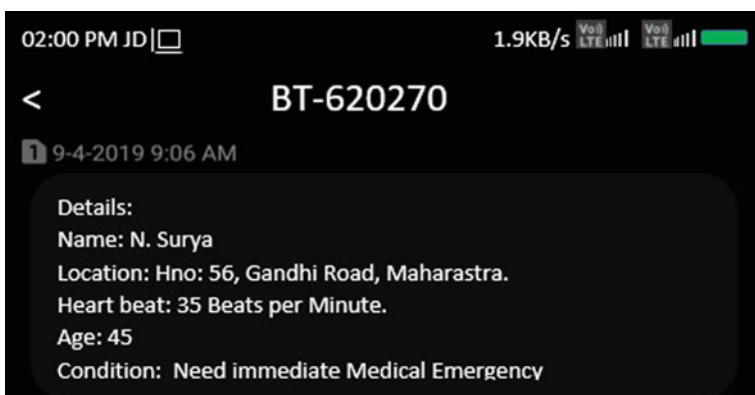
## 5 Results and Discussion

The alerts would be sent in the form of a message to the family members and doctors, as shown in Figs. 2 and 3. The heartbeat values change instantly from time to time, and thereby the alerts also have different values.

The messaging service name varies every time it sends an alert. Here the service rendering message is received from BT-620270.



**Fig. 2** Notification alert-1



**Fig. 3** Notification alert -2

## 6 Conclusion and Future Scope

The proposed system explained in this paper is helpful for the people who are suffering from genetic diseases and the problems which don't have a permanent cure but only control. It is also useful mostly for aged people who can't visit the hospitals frequently for regular check-ups. Also, as the entire system would be present in the home itself, the patient can self-monitor his health conditions and react accordingly.

The instruments present in the hospital are not affordable for everyone, and this system is very economical and beneficial. More additions can be made to this system where we can set the SMS service based on the periods for the patients to take medicines, including exercises related to their problems.

## References

1. Sharma, S., Sebastian, S.: IoT based car accident detection and notification algorithm for general road accidents. *Int. J. Electr. Comput. Eng.* **9**(5), 4020–4026 (2019). ISSN: 2088-8708, <https://doi.org/10.11591/ijece.v9i5.pp4020-4026>
2. Soujanya, K.L.S., Rajasekhar Gutta, S.S.: Accident alert system with IoT and mobile application. *Int. J. Recent Technol. Eng.* **7**(5), S2 (2019). ISSN: 2277-3878
3. Pattikonda, M.R., Reddy Pottipati, S., Senduru, S.: Automated emergency rescue alert system. *Int. J. Recent Technol. Eng.* **8**(2), S11 (2019). ISSN: 2277-3878
4. Hisham Che Soh, Z., Azeer Al-Hami Husa, M., Afzal Che Abdullah, S., Affandi Shafie, M.: Smart waste collection monitoring and alert system via IoT. In: 2019 IEEE 9<sup>th</sup> Symposium on Computer Applications & Industrial Electronics (ISCAIE), (pp. 50–54). Malaysia (2019) <https://doi.org/10.1109/ISCAIE.2019.8743746>
5. Valliappan, S., Mohan, B.P.R., Kumar, S.R.: Design of low-cost, wearable remote health monitoring and alert system for elderly heart patients. In: 2017 International Conference on IoT and Application (ICIOT) (pp. 1–7), Nagapattinam (2017). <https://doi.org/10.1109/ICIOTA.2017.8073612>
6. Ameta, D., Mudaliar, K., Patel, P.: Medication reminder and healthcare—an android application. *Int. J. Manag. Public Sector Inf. Commun. Technol. (IJMPICT)* **6**(2) (2015)
7. Cook, D.A., Enders, F., Caraballo, P.J., Nishimura, R.A., Lloyd, F.J.: An automated clinical alert system for newly-diagnosed atrial fibrillation. *PLoS One* **10**(4), e0122153 (2015). Published 2015 Apr 7. <https://doi.org/10.1371/journal.pone.0122153>
8. Mauney, J., Furlough, C., Barnes, J.: Developing a better clinical alert system in EHRs. **4**(1), 29–36, Article firstonline: August 13 2015; Issue published: (2015)
9. [https://en.wikipedia.org/wiki/Arduino\\_IDE](https://en.wikipedia.org/wiki/Arduino_IDE)
10. <https://en.wikipedia.org/wiki/IFTTT>

# Recommender System for Resolving the Cold Start Challenges Using Classification



Chandrima Roy, Siddharth Swarup Rautray, and Manjusha Pandey

**Abstract** Recommender Systems are quite popular and very useful for predictions of various products to consumers by providing recommendations. It deals with the definite type of items and produces the recommendations that are personalized to deliver effective and valuable suggestions to the consumer. The cold-start problem is one of the challenges in recommender systems. The cold-start situation arrives when products added to the collection have either no experiences, or very little. This causes a challenge for collaborative filtering algorithms mainly because they rely on the interactions of the item to make recommendations. In general, it is much harder to ask new user about their personal information (users don't want to answer too many questions). But it is easier to ask a lot of information about a new item (people who add it are interested in filling in this information to make their products recommended to the customers). The proposed system delivers the recommendation of new items to existing users with high consistency and precision.

## 1 Introduction

Cold start means the system doesn't have sufficient knowledge to make recommendations for a new user, or a new item [18]. Any Recommendation system will face cold start problems with the introduction of new client, product, or software. A new product registry or an introduction of new products makes it impossible to introduce an item to a customer as less information is available. The collective filtering cannot efficiently make recommendations for new user and new object event. Collaborative

---

C. Roy (✉) · S. S. Rautray · M. Pandey

School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT) Deemed To Be University, Bhubaneswar, Orissa, India

S. S. Rautray

e-mail: [siddharthfcs@kiit.ac.in](mailto:siddharthfcs@kiit.ac.in)

M. Pandey

e-mail: [manjushafcs@kiit.ac.in](mailto:manjushafcs@kiit.ac.in)

filtering is suffering from the “cold start problem”. It is difficult to recommend item to new users or suggest a new item to any user. This drawback can be approached in a different way: A system can recommend random items to novel users or new items to any random users (random approach). It can also recommend popular items to new users or new items to existing users (highest expectation approach), recommending a collection of different items to new users or a new item to a set of different users (exploratory approach) or, eventually, using a non-collaborative approach [1]. Any rule-based similarity requirements can be established for a new item. For example, to estimate a new listing, Airbnb used the average of three geographically closest listings of the same kind and price range.

Recommended systems should not exploit the data that means overfitting past user-item preference data in order to avoid getting trapped in the small neighborhood. Training data set should not be fully collided by past recommendations. YouTube contains videos embedded for training at other websites [2]. The videos viewed outside of YouTube’s platform are not from the recommendation framework, and can potentially generate new content. Injecting randomness into the framework can also be considered that means the system can make random recommendations).

Simple guidelines may be added to the System to expand the variety of recommendations. Methods are also available from multi-armed bandits. Uber eats applied the upper confidence bound to improve the variety of suggested restaurants/dishes. The upper confidence bound uses the upper bound of the estimated success rate. The confidence interval is large when a new item enters the system without any previous information, and thus the upper bound is high. As the item gets more views, the estimate would be more precise and closer to its actual value. The content-based recommendation method [3] is the solution to this problem since it does not rely on the ranking of products. Another way is to identify visitors as browsers that are only there to search items. Two basic types of cold start problem are:

## 1.1 New User Cold Start Problem

The term cold start means that the program does not have sufficient information to suggest a new user or a new object. With the addition of new user, an object or a program, any Recommendation program will face cold start problem [4]. Limited number of information is available which makes it problematic to recommend an item to a user. The issue of cold start is linked to the sparseness of information (i.e. for users and items) available in the algorithm recommendation.

The theory is that people with a common culture would most possibly have similar preferences. A model will be implemented from the training data. The goal is to find a neighborhood where the neighbors are users that belong to the same group as the group predicted by the model [5]. After this step, Similarity index will be calculated that combines similarity from the neighbors. Finally, the similarity measure and neighbors’ ratings are combined to get predictions.

## 1.2 New Item Cold Start Problem

The issue with the item cold-start will cause the new item to miss the opportunity to be recommended and stay “cold” all the time. The proposed model would like to suggest new products to potentially interested users—for which no interests have been expressed so far [6].

The method employs interrelationships mining. It can derive functionality based on a comparison of various attribute values. One of the main concepts of interrelation mining is to reflect similar characteristics based on a comparison of new attributes called interrelated attributes between values of different attributes [7]. First, most similar item will be chosen to include a new item’s neighborhood, and neighborhood rating information will be used to estimate the new item’s rating value. Lastly, highest rating product will be recommended to a target customer.

## 2 Proposed Work

This research will try to mitigate the New item cold start problem issue. For a case study, this paper will analyze the efficiency and predictive ability of a model; trained and evaluated on data gathered from homes in Boston, Massachusetts suburbs trained model on this data could then be used as a good fit to make various predictions about a newly enlisted house on that particular suburb, specially its market price. This model will be beneficial for users who are in real estate business, so that they can make use of such knowledge in their daily work cycle. Some websites like housing.com, MagicBricks or 99 acres can use this technique when some new house or flat enters the system with no ratings. This framework will help to recommend the new houses to the existing user.

### 2.1 Dataset Description

The UCI Machine Learning Repository is the source of the data set for this project. The data for the Boston housing was obtained in 1978 and the 506 entries reflects demographic information on 14 features for houses from different neighborhoods in Boston, Massachusetts.

### 2.2 Data Exploration

The findings are provided by a brief survey of the Boston housing data. Since the proposed system will create a working model which can predict the value of houses,

**Table 1** MEDV statistics calculation

Maximum	Minimum	Mean	Median	Standard deviation
\$1,024,800.00	\$105,000.00	\$454,342.94	\$438,900.00	\$165,171.13

we will need to split the dataset into features and the target component. The features, ‘PTRATIO’, ‘RM’, and ‘LSTAT’ provide us quantitative data point details. The target component, ‘MEDV’, would be the one we are trying to forecast. These are stored in respective features and prices.

where,

MEDV: Median value of houses occupied by the owner in \$1000s

LSTAT: Percentage of lower-income group of the population

PTRATIO: Student–teacher ratio of the neighborhood

RM: Average number of rooms per house.

### 2.3 Statistics Calculation

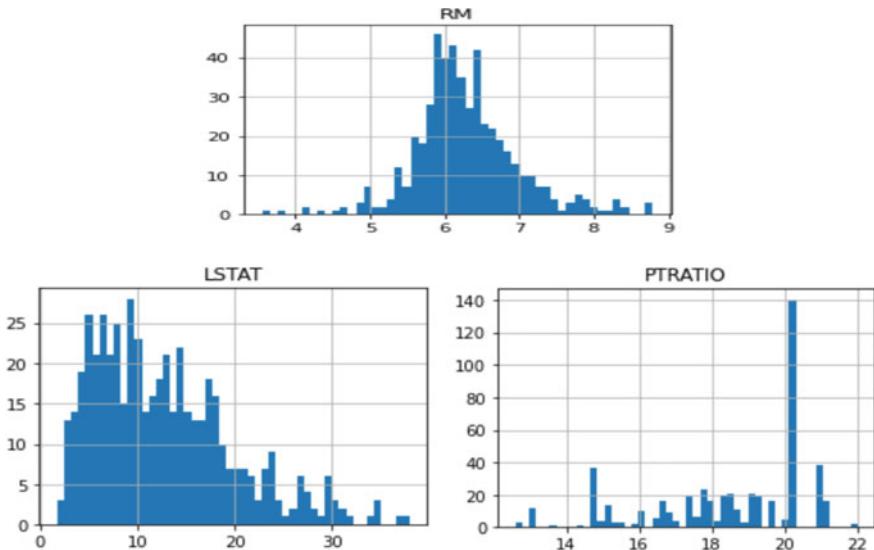
Maximum, minimum, mean, median, and standard deviation of ‘MEDV’ has been calculated and stored in prices (Table 1).

### 2.4 Feature Observation

I have taken three features from this dataset such as ‘RM’, ‘LSTAT’, and ‘PTRATIO’ for every single neighborhood. Figure 1 shows how frequently each distinct value occurs in a data set.

### 2.5 Train-Test Splitting

The method consists of taking a dataset and separating it into two subsets. The first subset is used for model fitting and is called the training dataset. The second subset will not be used to train the system; instead, the model is provided with the input element of the dataset, then predictions are made and compared with the expected values [6]. This second dataset is called test dataset. The goal is to estimate the model’s output on new item. In this case study, I have taken 80% of the data for training purposes and the remaining 20% for testing the model. Rows in train set is 404 whereas rows in test set: 102 train/test split provides a high variance estimate since it can significantly change the accuracy of testing. It can change depending



**Fig. 1** Histogram of RM, PTRATIO, and LSTAT

on observation in the test. Therefore, to solve this problem this research used k-fold cross-validation [8].

## 2.6 Correlation Coefficient

The effectiveness of linear connection between two or more variables is quantified by the coefficient of correlation. The correlation coefficient often takes a value of  $-1$  to  $1$ , with  $1$  or  $-1$  suggesting perfect correlation (in this case, all points will lie in a straight line).  $1$  Represents that the two variables shift in equilibrium. They move up and down together and the correlation is fine.  $-1$  Implies the two variables are in absolute contrast. One keeps increasing and the other decrease, in a completely negative way. If any two variables aren't correlated then the correlation value would be  $0$  [9].

### 2.6.1 Equation

To measure the Pearson product-moment correlation, the covariance of the two variables have to calculate first, then standard deviation for each variable is calculated. The correlation coefficient is measured by dividing covariance of two variables by the product of standard deviations of the two variables [10].

**Table 2** Correlation coefficient with MEDV

MEDV	1.000000
RM	0.680857
B	0.361761
ZN	0.339741
DIS	0.240451
CHAS	0.205066
AGE	0.364596
RAD	0.374693
CRIM	0.393715
NOX	0.422873
TAX	0.456657
INDUS	0.473516
PTRATIO	0.493534
LSTAT	0.740494

$$\rho_{xy} = \frac{\text{Cov}(x, y)}{\sigma_x \sigma_y}$$

where

$\rho_{xy}$  = Pearson product-moment correlation coefficient

$\text{Cov}(x, y)$  = covariance of variables  $x$  and  $y$

$\sigma_x$  = standard deviation of  $x$

$\sigma_y$  = standard deviation of  $y$ .

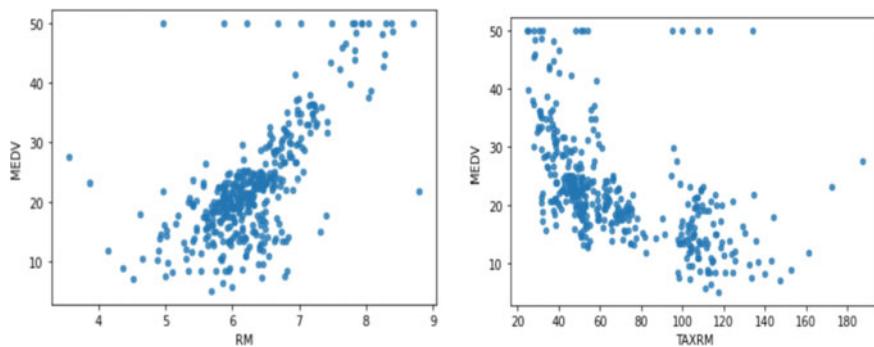
Standard deviation is a metric of data dispersal from its average [11]. Covariance measure, how two variables evolve together, but its magnitude is infinite, so the analysis is difficult [12]. The normalized version of the statistics can be determined by dividing covariance by the dividing of the two standard deviations. This is the Coefficient of Correlation.

From Table 2 we can see that RM has the highest correlation with MEDV whereas PTRATIO and LSTAT has the lowest correlation with MEDV respectively. Figure 2 Represent the scatter plot of RM and TAXRM with MEDV.

Prices rise as RM's value rises linearly. There are few outliers and it appears that the data is limited to 50. With a rise in TAXRM the prices begin to decrease. Although it doesn't appear to exactly follow a linear pattern.

### 3 Result Analysis

Our aim is to find the neighborhood of each new entry of house in the list. This research will use C4.5 Algorithm to find the similar neighbors. This algorithm aims

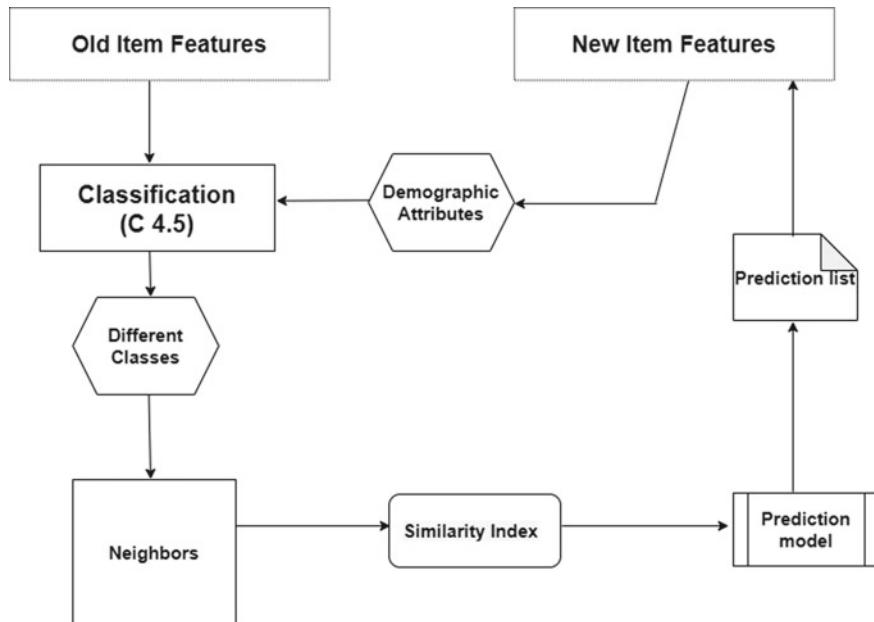


**Fig. 2** Scatter plot of RM and TAXRM with MEDV

to match the category of the new houses and category of the old houses whose price already has been evaluated. The outcome of this algorithm is the set of neighbors.

C4.5 algorithm is a classification algorithm that can be used to produce a decision based on a certain data set represented in Fig. 3.

There are different phases in the decision-making process with C4.5 algorithm, namely:



**Fig. 3** Proposed model for mitigating cold start item issue

1. The training data can be obtained from the data that has not been Classified into certain groups.
2. Determining a tree's root by calculating the highest Gain value for every attribute, or from the index of the

Lowest entropy value.

$$\text{Entropy}(s) = \sum_{i=1}^c -p_i \log_2 p_i$$

$C$ : Total Number of entities

$P_i$ : Positive/Negative entity of given dataset

3. Calculate the Information Gain

$$\text{Gain}(S, A) = \text{Entropy}(S) - \sum_{v \in \text{values}(A)} \frac{S_v}{|S|} \text{Entropy}(S_v)$$

C4.5 has benefits over other Decision Tree systems. Inherently the algorithm uses Single Pass Pruning Mechanism to prevent overfitting. It can function for discrete, as well as continuous data. C4.5 can manage incomplete data issue very well.

### 3.1 Evaluation Metrics

Root mean squared error (RMSE) is a quadratic scoring method which calculates the average error magnitude. It is the square root of the square differences average between prediction and real observation.

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^n (P_i - O_i)^2}{n}}$$

Final RMSE while testing the model is 2.9382073149269416.

### 3.2 Comparison with Different Models

I also performed Linear Regression and Random Forest Regression on this dataset. Table 3 represent C4.5 algorithm performed best in this dataset.

**Table 3** Mean and SD for different models

Model	Mean	Standard deviation
C 4.5	3.494650261111624	0.762041223886678
Random forest regression	3.945666155289214	0.757829838438667
Linear regression	4.221894675406022	0.7520304927151625

## 4 Conclusion

This research is based on similarity factor that overcomes Conventional Collaborative Filtering disadvantages. Collaborative Filtering fails when a new item, in this case, mew building, with no ratings is added to the system. This paper presents a framework to mitigate the new item cold Start Issue. This research proposed to implement a method which takes the demographic data and based on similarity measure techniques discover items with similar characteristics with the new item. Item with a similar features and common characteristics can be recommended to users who have similar likings. So, with addition of new items in the system, the old user will get the recommendation.

## References

1. Ramzan, B., et al.: An intelligent data analysis for recommendation systems using machine learning. *Scientific Programming* (2019)
2. Fu, M., et al.: A novel deep learning-based collaborative filtering model for recommendation system. *IEEE Trans. Cybernet.* **49**(3), 1084–1096 (2018)
3. Zhang, Y.: GroRec: a group-centric intelligent recommendation system integrating social, mobile and big data technologies. *IEEE Trans. Serv. Comput.* **9**(5), 786–795 (2016)
4. Shu, J., et al.: A content-based recommendation algorithm for learning resources. *Multimed. Syst.* **24**(2), 163–173 (2018). <https://doi.org/10.1007/s00530-017-0539-8>
5. Rohit, et al.: Proposed approach for book recommendation based on user k-NN. *Adv. Comput. Computat. Sci.* **554**, 543–558 (2017). [https://doi.org/10.1007/978-981-10-3773-3\\_53](https://doi.org/10.1007/978-981-10-3773-3_53)
6. Wei, J., et al.: Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Syst. Appl.* **69**, 29–39 (2017)
7. Chen, J., et al.: Attentive collaborative filtering: multimedia recommendation with item-and component-level attention. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (2017)
8. Correlation Coefficient, AKHILESH GANTI, <https://www.investopedia.com/terms/c/correlationcoefficient> (2020)
9. Lam, X.N., et al.: Addressing cold-start problem in recommendation systems. In: Proceedings of the 2nd International Conference on Ubiquitous Information Management and Communication (2008)
10. Vartak, M., et al.: A meta-learning perspective on cold-start recommendations for items. In: Advances in Neural Information Processing Systems (2017)
11. Roy, C., Rautaray, S.S., Pandey, M.: Big data optimization techniques: a survey. *Int. J. Inf. Eng. Electron. Bus.* **10**(4) (2018)

12. Roy, C., Barua, K., Agarwal, S., Pandey, M., Rautaray, S.S.: Horizontal scaling enhancement for optimized big data processing. In: Abraham, A., Dutta, P., Mandal, J., Bhattacharya, A., Dutta, S. (eds) Emerging Technologies in Data Mining and Information Security. Advances in Intelligent Systems and Computing, vol. 755. Springer, Singapore (2019)

# A Skyline Based Technique for Web Service Selection



Yamini Barge, Lalit Purohit, and Soma Saha

**Abstract** Service-Oriented Architecture (SOA) proves to be the primary reason behind the advancements of web platforms. Therefore, an enormous quantity of web services with similar functionalities has evolved over the web. However, wide-ranging web services which possess Quality of Service (QoS) with varying dimensionality may be inefficient in terms of performance due to web service selection strategy with traditional service selection methods. Skyline technique yields to improve the efficiency of web service selection. Skyline is based on a domination theory in which only those services can survive which are better than the other services available in the set. The concept of Web Service Selection using Skyline mainly tends to satisfy the user experience and user requirements. The participation of QoS is considered to be a base criterion for the selection of the most favorable services. The presented paper represents the work in the direction of Web Service Selection for composite web services using the Skyline technique.

## 1 Introduction

The advantages in SOA have accelerated the focus on the core functionalities associated with web services to provide more convenient options. Convenient options tend to facilitate better web services to the user in the context of desired factors satisfaction that leads the better user experience and satisfaction of the user requirements while utilizing the web service. Web services are propagators between the software entities that evolved in a full or partially different environment. At present, almost ten thousand web services are available over the web. Advancement and technological revolution has standardized web services and created a competition among the service providers to provide better web services to the users according to their demands. In a nutshell, service selection becomes a vogue in research area and grabs up user satisfaction.

---

Y. Barge (✉) · L. Purohit · S. Saha

Shri Govindram Seksaria Institute of Technology and Science, Indore 452003, India

To fulfill the requirements and offer a value-added service, many times web services need to be composed. The composition of services is directly associated with web service selection because selecting the set of services and to compose into a single web service is a crucial task. Thus to pick the most eligible set of these services, QoS is considered as base criteria. Every web service is associated with some quality parameters. Quality of Service is taken as the base criterion of service for web service selection. The main research focuses on selecting a web service for composite services using the Skyline technique. The selection of web services on the basis of QoS parameters can be cast as a multi-objective problem. Consequently, Skyline technology is opted to solve the different tradeoffs easily [1]. Skyline assists to apply a dominance check over a set of web services and abandon those services from the pursuit of being better which are not better in terms of selection criteria. Section 2 provides a detailed review of related works. In Sect. 3, the background details and terminologies are discussed. Performed experiments are elaborated in Sect. 4 and the obtained results are explained in Sect. 5. The conclusion is delivered in Sect. 6.

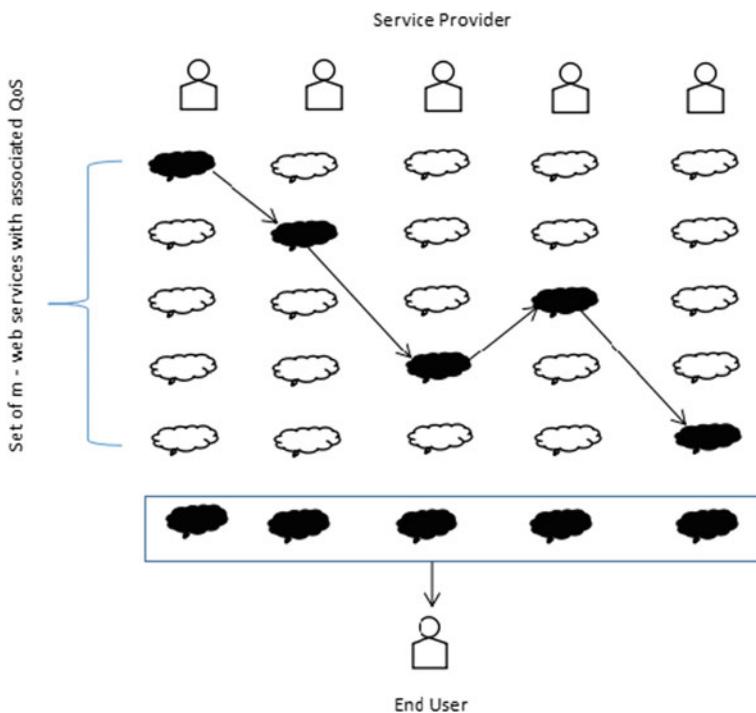
## 2 Related Work

Web service selection has pulled plenty of focus from researchers. Apart from Skyline, many other techniques have been involved in the web service selection. Approaches formulated based on the concepts of linear programming intended to find the optimal solution for QoS-aware web services. A global optimization approach based on a linear program to find the best service component was introduced [2, 3]. Heuristic algorithm based on linear programming is also employed to identify optimum services for making runtime decisions [4]. Although linear programming proved better to small datasets, but could not able to do justice with large datasets and has to suffer from poor stability and efficiency. QoS levels were extracted with the help of a greedy approach [5]. For the most preferential QoS, this approach lacks potential dependencies and correlations due to handling each QoS dimensions independently. The service selection problem was modeled as a combinatorial problem and employed a multi-objective optimization technique to meet the optimal solutions among the available services. The experimentation involved the use of the NSGA-II algorithm for achieving the desired results [6]. The problem of web service selection can be solved efficiently by applying classification before selection [7]. Thus to overcome the problem of handling large datasets with better efficiency, Skyline was introduced [8]. Skyline is very good at handling large datasets. Extensive work has been done in the direction of proving efficiency of Skyline technique. Along with Skyline, some well-known problems are also associated with skyline queries like NN-query, Top-N problems, contour problem, convex hulls, and multidimensional indexing [9]. Many problems have been addressed using the skyline technique for optimizing service selection such as handling frequent requests through lottery scheduling [10]. Multicriteria decision-making approach was adopted for effective

selection of skyline services using Best-worst, TOPSIS, and PROMETHEE method [11]. Skyline with k-means is also used to prefilter the explored web services [12]. The authors experiment Block Nested Loop (BNL) algorithm which is a Skyline algorithm for the service selection. The advantages of the BNL algorithm have been discussed along with the best quality attribute parameters.

### 3 Background

1. **Web Service:** Web Service establishes communication between the software entities that are developed in their respective environment. Web services use four standard pillars to establish communication with other services, named; SOAP, XML, UDDL, and WSDL.
2. **Composite Service:** A composite service is a combination of multiple atomic services. The functionality of the entire composite services is fulfilled by the completion of functionalities of all its constituent atomic services. The Abstract Web Service (AWS) explains the task functionality which is responsible for



**Fig. 1** Web service selection for web service composition

service composition [13]. Figure 1 depicts the role of service composition in Web Service Selection.

3. **Skyline:** Skyline is based on the concept of dominance in which the services are applied a dominant check so that the dominated solutions are pruned to reduce the unimportant services from the race. Computing Skyline services is similar to determine the maximal vector problem in computational geometry [14]. Non-dominant and dominant are two base pillars of the Skyline technique. **Dominance:** Consider a service class  $S$ , and services  $x, y \in S$ , characterized by a set of  $Q$ , QoS attributes.  $x$  dominates  $y$ , denoted as  $x \prec y$ , if  $x$  is as good or better than  $y$  in all parameters in  $Q$  and better in at least one parameter in  $Q$ , i.e.  $\forall k \in [1, |Q|]: q_k(x) \leq q_k(y)$  and  $\exists k \in [1, |Q|]: q_k(x) < q_k(y)$  [6].

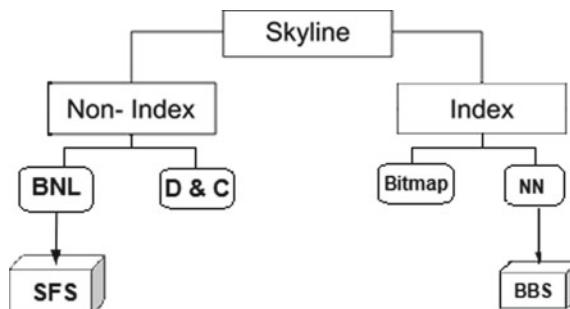
**Skyline Services:** Let the service class  $S$  represents Skyline services. These Skyline services are represented as SS and include the services in  $S$  which are not dominated through another service, i.e.,  $SS = \{x \in S | \neg \exists y \in S: y \prec x\}$  [8]. There exists multiple solutions in the dimension space. These solutions are compared with each other through a dominance check to find non-dominant solutions which are better among the available solutions based on considered parameters. Skyline can be defined as “The solution is said to be Skyline Solution if all the dimensions of the solution have better values than other solutions or have a strictly better value at least in one dimension” [6]. Skyline can be considered as the pruning process to reduce the underrated solutions which are not important. Skyline concept is used to provide better web services to extend the good performance in all considered parameters. Skyline does not consider user preferences [15]. Skyline of a given set Ds of the point is that any set of evaluation criteria that arises from user’s preferences can be modeled in the form of the monotone scoring function  $f: Di \rightarrow R$ , like L1 norm  $f(x, y) = |x - y|$  or Euclidian form [14]. The goodness or badness of any solution is checked through the Skyline operator. The categorization of Skyline algorithms is based on Index-based and Non-Indexed based approaches in which other categories have been covered.

1. **Index-based:** Index-based Skyline algorithms demand the access to certain unique data structure like R-tree and B-tree. The following are the algorithms based on Index-based approach.
2. **Bitmap:** Bitmap is a progressive algorithm. It does not need to scan the complete dataset to return results. Bitmap algorithm is based on the bitmap structure that encodes all the required information to examine those data points which belong to Skyline family.
3. **Nearest Neighbor (NN):** The algorithm uses R\*-tree data structure. This algorithm cuts off the redundant dominant checks and eliminates massive non-skyline points. The algorithm starts with searching for the nearest neighbor from the origin point defined.
4. **Branch and Bound Skyline (BBS):** It is an improved version of NN algorithm. BBS overcomes the recursive searching effort which was the main problem of NN. BBS requires a single search of R\*-tree. In the R\*-tree, the data points are

arranged in such a way that each internal R\*-tree node consists of maximum tree nodes and each leaf can also have maximum three entries.

5. **Non-Index-based:** It is more generic approach than Index-based. This approach does not require any special access to a data structure like R\*-tree. The approach does not require any pre-processing on the underlying dataset. Algorithms based on Non- Index approach that have been used in particular for Web Service Selection and described are as follows:
6. **Block Nested Loop (BNL):** The algorithm reads the input data and each point is retrieved and compared against the points in the buffer. BNL proves to be efficient and results in better performance if the resulted size of Skyline is small. The algorithm terminates in single iteration when the best case fits into the window.
7. **Short Filter Skyline (SFS):** The data points are checked on the basis of their score, arranged in ascending order and kept in-memory buffer. This algorithm ensures a reduction in the pairwise comparison between the points. Exhaustive search is performed on existing skyline points for the dominance test.
8. **Divide and Conquer (D&C):** The D&C algorithm recursively divides the input dataset into m partitions  $P_1 \dots P_m$  (m-way partitioning), to fit values in the main memory. The partition boundaries are determined by computing the q-quintiles of the dataset which results in the division of the dataset into  $q - 1$  equal subsets. Figure 2 represents the Skyline algorithms particularly employed in the web service selection. Algorithms belong to the particular category are mentioned in the square box and the arrows depict the extended variation of the algorithms.
1. **BNL Algorithm:** BNL is certainly like the Naïve-Nested Loop algorithm. The algorithm repeatedly reads the set of tuples and eliminates points by finding other points in the dataset that dominate them. BNL allocates a buffer (window) in the main memory that contains several points to sequentially track the dominance between them [14]. The algorithm reads the input data and each point is retrieved and compared against the points in the buffer. In the first run of the algorithm, no point will exist in the buffer so it is trivial to insert the first point in the buffer. For the next runs, if the point retrieved is dominated by at least one point in the buffer there is no need to continue the comparison with the other points

**Fig. 2** Skyline algorithms for web service selection



that may already exist in it and the point is discarded. Otherwise, if the point is incomparable or dominates one or more points in the buffer, those points that are dominated are removed from the buffer and the new point is inserted.

### **Algorithm: BNL Algorithm**

**Step 1:** Use the normalized values of the dataset

**Step 2:** Checks for dominating values.

**Step 3:** If value  $y$  is dominated by others then, eliminate  $y$  from buffer.

**Step 4:** If value  $x$  is dominating other values then, keep  $x$  in the buffer.

**Step 5:** If the value is incomparable then it is inserted in the memory buffer.

**Step 6:** Check for the dominance status.

**Step 7:** If status =  $-1$ , found dominant; remove the value from Skyline.

**Step 8:** If status =  $1$ , dominating other values. Add to the Skyline service.

**Step 9:** Return the indices of the Skyline points.

**Step 10:** Print Skyline services.

**Normalization:** It is a pre-processing stage. This step requires data points associated with the particular attribute can have the values that lie in the range of [0, 1]. Normalization tends to manipulate the data points which can be used for further calculations or observations. Min\_Max normalization is used over the data points for further calculations. Linear transformation which is performed on the original data can be depicted by the following Eq. (1)

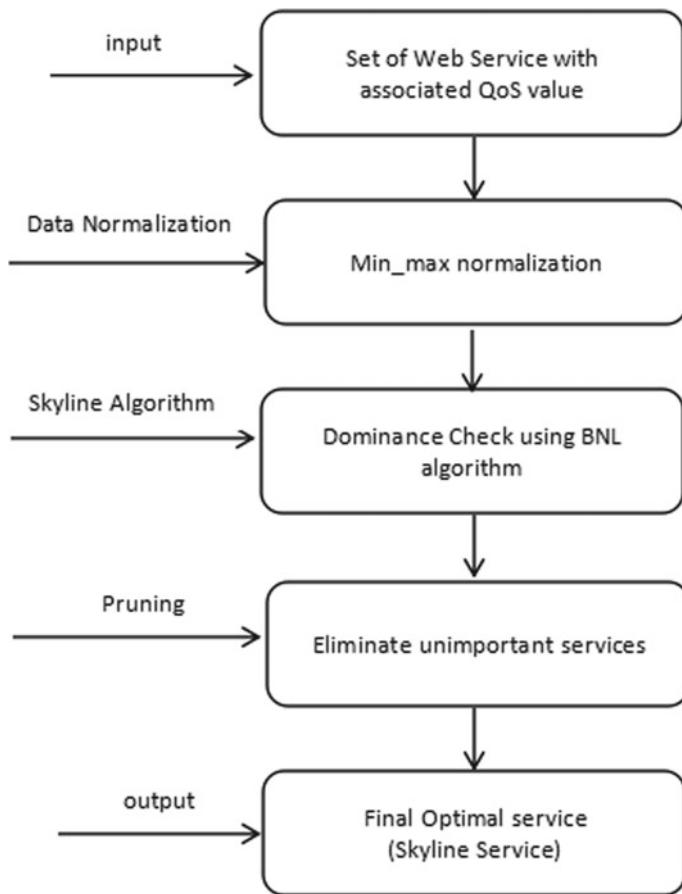
$$v' = \frac{v - \text{max}(A)}{\text{max}(A) - \text{min}(A)} (\text{new\_max}(A) - \text{new\_min}(A) + \text{new\_min}(A)) \quad (1)$$

where,  $v'$  is the current value,  $v$  is the previous value,  $\text{new\_max}(A) = \text{max}$  value of range,  $\text{new\_min}(A) = \text{min}$  value of range.

Figure 3 depicts basic flow of Skyline for Web Service Selection at different levels.

## 4 Experiment

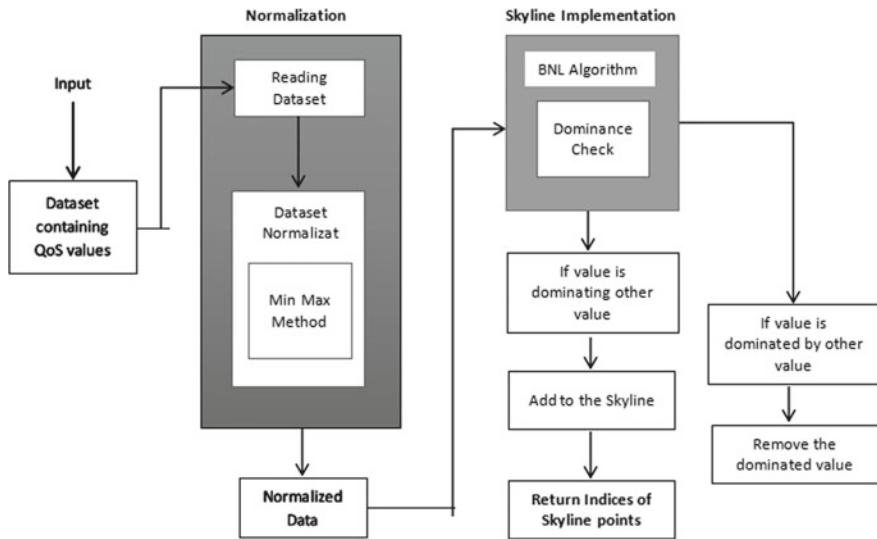
In this paper, authors have performed the BNL algorithm for the Web Service Selection which is a Non-Indexed based Skyline algorithm. BNL algorithm is more generic than other algorithms. The experiment has been performed using Java language. The proposed experiment is conducted on PC with Intel (R) Core (TM) i3 CPU @ 2.00 GHz and 4 GB RAM memory. The dataset that has been used for the experiment is contained of 365 real web services with measurement of nine QoS parameters per services and additional attributes. The dataset is recovered as a text file. The dataset also contained name of services, the URL of its WSDL file. Figure 4 represent the implementation diagram for the proposed work.



**Fig. 3** Algorithmic flow chart

## 5 Results and Discussion

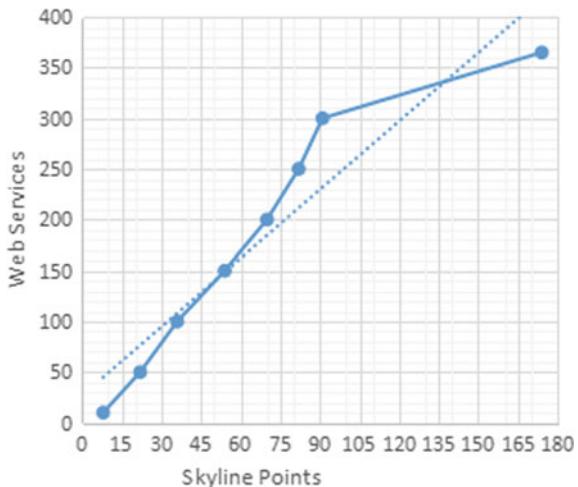
The experimental result shows that 174 skyline web services are obtained out of 365 web services. The computational time of the resultant services is recorded as 471  $\mu$ s on an average. According to the further experiment on QoS, Attribute Selection methods are applied on the datasets. Further experiments are performed on the datasets to deeply analyze the result patterns and to check the relevancy of the attributes. Attribute selection methods have been used to extract the ranking of the specified attributes. By using Classifier AttributeEval and Ranker method, Latency is determined as the best QoS parameter while Response time as worst. Similarly, as a result of using Correlation AttributeEval and Ranker method, Response Time is shown as the best parameter while Reliability as worst. Further, by using CFSub-setEval and Best first method of feature selection, three QoS parameters are selected.



**Fig. 4** Implementation diagram

to be best among 8 parameters and they are Throughput, Successability, and Reliability. Using these three parameters as the base QoS, results were totally different. Experimentation with these three QoS parameters results in less number of Skyline services. It indicates that number of attributes plays an important role in generation of Skyline services. Besides this, there is one more attentive point need to be focused on is that even if the number of attributes remain same in quantity but the number of services are increased or decreased also affects the Skyline results. Figure 5 shows the

**Fig. 5** Skyline growth with number of web services



**Table 1** Computational efficiency

S. No	Number of web services	Number of QoS	Execution time ( $\mu$ s)
1	365	8	459
2	365	3	396
3	2507	8	1728
4	2507	3	1309

**Table 2** Runtime comparison of skyline

S. No.	No. of services	Runtime of selection of web services (in $\mu$ s)		
		Dynamic programming [16]	Pisinger's [16]	Skyline
1	10	1395	116	48
2	100	35,668	304	242
3	1000	7 s 67,617	3240	1046
4	2000	–	–	1817

graph plotted for number of web services. Different quantity of web services generates different figures of Skyline points. It can be observed that larger the number of services, larger the number of Skyline points will generate. The growth of Skyline points increased exponentially with larger number of web services.

QoS parameters also make an observable difference in the results. Relevancy of QoS attributes has an impact on the generated Skyline results. This is because the Skyline uses the concept of Pareto Accumulation. Pareto Accumulation tends to give equal importance to all the QoS parameters. This means that all QoS parameters are treated equally while the selection of web services. Due to which a large number of Skyline web services are produced as output. This reason is enough for Skyline to grow exponentially with a large number of attributes. The computational efficiency of the BNL algorithm is depicted in Table 1

The runtime performance of Skyline is compared with similar traditional approaches like Dynamic Programming and Pisinger's algorithm [16]. The depiction of Table 2 clearly points out the performance difference of Skyline and other similar approaches. Skyline performs better in terms of computing better services as compared to other techniques. The performance of Skyline has been calculated in  $\mu$ s.

## 6 Conclusion

As Skyline offers the best services from the available collection of web services, Quality of Service has a vital role in the selection aspect of these Skyline services. It is observed that in the mentioned work, efforts have been made to improve the

web service selection criteria through Skyline by employing various hybrid or mixed approaches. It is concluded that the user-defined constraints are still needed to be on an efficient track to deal with user preferences and satisfy the Quality of user experience. However, Skyline becomes a little vulnerable when it comes to deal with more constraints conditions. Further, most of the traditional techniques including linear programming, Greedy approach, Genetic Algorithm suffer from insufficient capabilities to handle large datasets. Skyline with its potential capability of handling large datasets has solved much of the Service Selection dilemma. According to our work the Skyline services exponentially grow with the increase in the number of services this is because Skyline uses Pareto accumulation due to which the attributes are considered to be equally important and as a result of dataset comes out with large Skyline points.

## References

1. Borzsony, S., Kossmann, D., Stocker, K.: The Skyline operator proceedings. In: International Conference on Data Engineering (pp. 421–430) (2001). <https://doi.org/10.1109/ICDE.2001.914855>
2. Zeng, L., Benatallah, B., Dumas, M., Kalagnanam, J., Sheng, Q.: Quality driven web services composition (2003). <https://doi.org/10.1145/775152.775211>
3. Sun, Q., Wang, S., Zou, H., Yang, F.: QoS-aware web service selection with the skyline (2010). <https://doi.org/10.1109/ICBNMT.2010.5705226>.
4. Ardagna, D., Pernici, B.: Adaptive service composition in flexible processes. IEEE Trans Softw Eng **33**(6), 369–384 (2007). <https://doi.org/10.1109/TSE.2007.1011>
5. Torkashvan, M., Haghghi, H.: A greedy approach for service composition. In: 2012 6th International Symposium on Telecommunications, 929–935, IST 2012 (2012). <https://doi.org/10.1109/ISTEL.2012.6483119>
6. Claro, D.B., Albers, P., Hao, J.: Selecting web services for optimal composition. *SDWP@ICWS* (2005)
7. Purohit, L., Kumar, S.: Web services in the internet of things and smart cities: a case study on classification techniques. IEEE Consum. Electron. Mag. **8**(2), 39–43, (2019). <https://doi.org/10.1109/MCE.2018.2880808>
8. Alrifai, M., Skoutas, D., Risse, T.: Selecting skyline services for QoS-based web service composition. In: Proceedings of the 19th International Conference on World Wide Web, WWW (2010). ‘10. <https://doi.org/10.1145/1772690.1772693>
9. Mourad, F., Asaidi, H., Bellouki, M.: Comparative study of skyline algorithms for selecting web services based on QoS. Procedia Comput. Sci. **127**, 408–415 (2018). <https://doi.org/10.1016/j.procs.2018.01.138>
10. Wang, Y., Song, Y., Liang, M.: A skyline-based efficient web service selection method supporting frequent requests, 328–333 (2016). <https://doi.org/10.1109/CSCWD.2016.7566009>
11. Serrai, W., Abdelli, A., Mokdad, L., Hammal, Y.: An efficient approach for Web service selection, 167–172 (2016). <https://doi.org/10.1109/ISCC.2016.7543734>
12. Purohit, L., Kumar, S.: Clustering based approach for web service selection using skyline computation. In: IEEE International Conference on Web Services (ICWS), (pp. 260–264). Milan, Italy (2019). <https://doi.org/10.1109/ICWS.2019.00052>
13. Kumar, S., Purohit, L.: Exploring K-means clustering and skyline for web service selection, 603–607 (2016). <https://doi.org/10.1109/ICIINFS.2016.8263010>
14. Kalyvas, Tzouramanis, T.: A Survey of Skyline Query Processing (2017)

15. Purohit, L., Kumar, S.: Replaceability Based Web Service Selection Approach, 113–122 (2019). <https://doi.org/10.1109/HiPC.2019.00024>
16. Yu, T., Lin, K.-J.: Service selection algorithms for web services with end-to-end QoS constraints. *Inf. Syst. e-Bus. Manage.* **3**, 129–136 (2004). <https://doi.org/10.1109/ICECT.2004.1319726>

# Novel Trust Model to Enhance Availability in Private Cloud



Vijay Kumar Damera, A. Nagesh, and M. Nagaratna

**Abstract** The benefits of cloud computing for companies are numerous. However, among all the advantages, high availability in the cloud is one of the factors that draws the attention of most of cloud users. High availability, as the name suggests, ensuring that IT resources are available at all times. For this, it is necessary to implement processes to detect single points of failure in your system and reduce the chances of their occurrences through strategies such as redundancy and/or replication. Widely accepted solution for enhancing availability in cloud is replication. However, deciding on minimum number of replicas and where to place replica is a major research concern in cloud. Hence, where to keep replicas in the system to satisfy availability requirement is a matter of concern. To address above issue, this paper proposes a novel trust model which helps in selecting appropriate node for replica placement.

## 1 Introduction

Cloud computing technology appeared relatively recently and became very popular among companies of all types and sizes. This rapid expansion is not surprising when we stop to analyze all the benefits that this technology can offer. One of its flagships, without a doubt, is high availability in the cloud [1, 2].

High availability is ultimately the holy grail of the cloud. It incorporates the idea of accessing services, tools ,and data anywhere and anytime and is the foundation of modern companies. Availability is also related to reliability: a service that works 24/7, but is constantly unstable is not considered good. To have true high availability

---

V. K. Damera (✉)  
JNTU Hyderabad, Hyderabad, Telangana, India

A. Nagesh  
Department of CSE, MGIT Hyderabad, Hyderabad, Telangana, India

M. Nagaratna  
Department of CSE, JNTUCEH Hyderabad, Hyderabad, Telangana, India

in the cloud, your services need to not only be always active, but also have 99.2% availability [3–5].

In general, high availability means eliminating all points of failure and creating redundancy in processes and equipment. Thus, if a server fails for any reason, there is another one to continue the operation. The change is made instantly (failover) and the user does not even notice the change [6, 7].

The cloud computing service creates redundancy in the different layers of operation. The data center, for example, has two power supply systems and two different forms of cooling to ensure that everything works correctly in any situation. Not to mention the fact that there are at least two telephone operators to guarantee communication between the data center and traffic collection points such as ix.br for example.

At the hardware layer, there are redundant servers available to take over the operation in the event of a hardware failure. Thus, if one presents a problem, the other is immediately activated to continue operations. Likewise, high network availability is guaranteed with redundant switch and edge routers [8].

In the third layer, that of IaaS, where the servers are virtualized, it is also possible to implement high availability. The load balancer manages the load and directs it to the application servers that are active. Thus, it is possible to perform maintenance on the virtual servers, stopping one server at a time, without interrupting the operation of the application [9].

In short, high availability guarantees the operation of a service 365 d a year, regardless of the number of people who are connected or an incident in the company. Therefore, you can continue with business activities even if the workspace suffers from a flood or fire [10].

Trustworthiness, as a basic requirement of Internet-based application systems, faces many new challenges in new computing environments and application models. How to effectively manage trust in the Internet environment to adapt to the needs of the development of computing environment and application mode has become a hot issue. Under the background of a variety of trust problems in cloud computing environment, in this paper, an attempt is made to present a trust evaluation models to enhance availability in private cloud.

## ***1.1 Availability in Cloud Computing***

Availability refers to the proportion of the time that the software system normally works in a given time to the total time, usually measured as a percentage. In traditional data centers, factors that affect service availability include abnormal server downtime, service attacks, operating system crashes, software crashes, power outages, and network interruptions. Data center administrators need to use redundancy and disaster backup to ensure service availability. However, the introduction of these redundant or disaster backup systems has brought new problems, such as the problem of copy consistency caused by redundant backup, higher procurement, and manage-

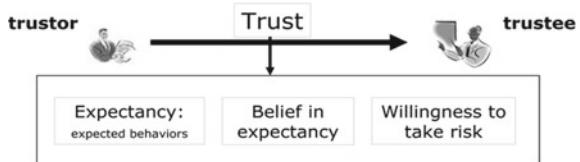
ment overhead. The cloud environment can minimize the impact of unavailability of resources on business systems and create a highly available computing environment. In cloud computing, providing uptime guarantees and service-level agreements has become a standard requirement for most cloud computing providers. Most of these cloud computing platforms claim to provide 99.999% availability. But in fact, the existing cloud computing environment has also experienced availability problems. When a physical failure occurs, the server hardware shuts down in a short time, and recovery from the backup state often takes longer. A tiny cloud computing failure may lead to a chain reaction of software failures, causing a software service that depends on cloud computing to be interrupted for hours, dozens of hours, or even days. This means that the availability of the overall cloud computing environment may be able to reach 99.999%, but the availability of individual services or applications that users are concerned about cannot reach 99.999%. In order to provide truly high-availability services, cloud computing providers are studying common failure analysis and prediction models. Based on the research of these models, cloud computing service providers hope to be able to anticipate possible availability problems and avoid these failures or reduce the losses caused by failures by preparing copies in advance, resolving failures in advance and notifying users.

## 2 Trust

### 2.1 *Definition of Trust*

As a basic factor in human society, trust plays a decisive role in social organization. For example, at a traffic intersection, we always believe that cars in the other direction will follow the instructions of the signal lights to make our decisions and travel smoothly. It is precisely because of the importance of trust that trust research has received attention in various fields, including psychology, sociology, philosophy, etc. [9, 11, 12]. With the development of the times, it has been integrated into business management, economic theory, engineering, computer science, and other application fields. Knowledge. Due to the complexity and multi-faceted nature of trust, there is currently no precise and widely accepted definition of trust in academia and industry, which is often understood as an intuitive concept. There are various definitions around trust. Trust in the Oxford Dictionary is defined as “a belief in the reliability, authenticity, ability, and strength of someone or something”ž. Hwang et al. [8] define trust as “based on the prediction of a certain behavior of the believer, the relying party is willing to accept the risk of believing the other party, regardless of whether it can monitor or control the trusted party”. This definition emphasizes the risk of trust, indicating that trust is essentially an assessment of the risk of acceptance. As shown in Fig. 1, Huang [13] et al. define “trust is a state of mind, which includes three aspects: (1) expectation: the service that the trustee wants to obtain from the trusted person, and (2) the belief: Based on the judgment of the ability and will of

**Fig. 1** Composition of trust elements



the trusted person, the trustee believes that the expectation is correct, (3) **the risk willingness:** the trustee is willing to bear the possible failure of the belief”.

## 2.2 Nature of Trust

The dynamics of trust are the biggest challenge of trust evaluation and trustworthiness prediction. It is determined by the natural attributes of the entities in the trust relationship. This section summarizes the nature of trust as follows:

1. **Subjective Uncertainty:** refers to the fact that the trustee cannot clearly judge the dynamic change of the trustee as the context and time change. The trust can only be evaluated according to the previous interaction history; trust is credit. The party has a subjective judgment on the recipient, and different entities will have different criteria. Even for the same trusted party, the same context, the same time period and the same behavior, the difference of the creditors, the given quantitative judgment is likely to be different.
2. **Context-Dependent:** The specific state of trust is closely related to the context. It is meaningless to discuss the trust issue from the specific context.
3. **Time Asynchrony:** It means that the evaluation result of trust relationship between entities has time asynchrony. The solution to the problem is to average the time slot; trust will decay with time, and the most direct performance is: The longer the trust evaluation, the worse its persuasiveness.
4. **Multi-Objectivity:** Trust is often associated with multiple attributes of the trusted party and is influenced by multiple attributes. It is a concept of multi-attribute interaction. Taking online shopping as an example, customer evaluation of the seller may include evaluations of the quality, price, service attitude, and speed of the delivery.

## 2.3 Principles of Trust

To guarantee a maximum level of confidence certain principles must be followed:

1. Trust transitivity, if entity A trusts entity B and entity B trusts entity C, then we can come to the conclusion that A can trust entity C by referring to the trust of entity B;

2. Trust is a function of perception of risk, it represents a belief in a person for his correct actions. Thus, trust must also assess the uncertainty that the other party is acting properly and incorporate the associated risks;
3. Trust is determined by time, it is built over time based on past experiences;
4. The trust can be measured, it is measurable by a numerical value, generally in the interval  $[0 - 1]$ ;
5. Formal and social tools are necessary for the evolution of trust, trust can be modeled according to various formal models.

## 2.4 *Trust in Cloud Computing*

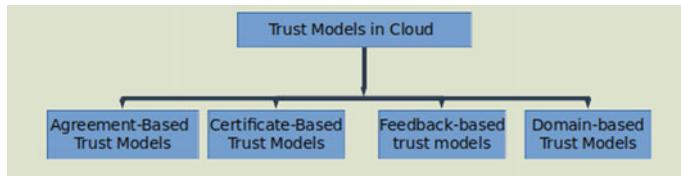
Trust is initially an idea that rises in humanism, and ideas in human science are frequently obscure. Trusted computing guides the idea of trust to the field of software engineering. Trusted Computing Group (TCG) characterizes trust as: a substance that can generally accomplish the ideal objective in the normal way, at that point the element is trustable. That is, trust underlines the desire for the element's conduct, while likewise focusing on the security and dependability of the system [14–16].

Trust is basically derived from a twofold relationship. One can express this relationship in one among the following forms: one-to-one, many-to-many, one-to-many, and many-to-one. Different mechanisms to evaluate trust are: direct trust, recommended trust, multi-level recommendation trust, and hybrid trust [15, 16].

In majority of e-commerce applications, trust is achieved through a belief that the opposite party is reliable and is able accomplish all its promises. Smooth online transactions will be achieved when the participating entities trust each other. This shows the significance of trust for online transactions. In cloud environment, trust value for entities will be evaluated using either of two mechanisms: direct or indirect trust. Direct trust is evaluated based on past experiences between entities where as indirect trust is by the recommendation of other entities.

## 3 Trust Evaluation Models for Cloud

Cloud computing is an emerging computing model with the advantages of large scale, high reliability, and extremely low cost. At present, there have been a large number of public cloud service providers at home and abroad, such as Google's GFS (GoogleFileSystem), IBM's blue cloud computing platform. On the one hand, different cloud service providers may provide some of the same basic services and also provide some unique services; on the other hand, the service quality of different cloud computing vendors also differs greatly. With the further popularization and development of cloud computing, users will face an increasingly important issue called How to choose a service provider that best suits users' needs from many cloud service providers? In order to solve this problem, trust evaluation models for cloud



**Fig. 2** Trust models in cloud

computing must be used. The credibility of the service quality of the business is evaluated. The trust evaluation in the existing cloud computing environment can be used for service quality transactions [1, 3], secure storage [4], resource allocation [5, 6], access control [8], cloud environment security [8, 9], and other aspects. Service quality is an important factor affecting the development of cloud computing.

Before opting any cloud service consumer always depends on trust evaluation model to select the services of a CSP and outsources its confidential sensitive data to the cloud environment. Trust is highly subjective and context-sensitive. Due to this nature, the service selection from a cloud provider becomes most challenging task. Further trust value may change with time based on the experiences of user with provider. The trust level may also vary with feedback from other cloud users getting cloud services from the same provider [17–19]. Organizations basically evaluate trustworthiness of different CSPs whenever they wants to migrate its sensitive data on to the cloud environment. The trust level of that CSP is evaluated using a trust model. A trust model can be defined as “a coded implementation that relies on concepts of trust in order to assign a trust value for a CSP, based on which the interactions with that specific cloud provider are restricted and controlled [11, 20, 21]. Categorization of different trust models is presented in [12, 22–25].

Based on the evaluation criteria, trust evaluation models are categorized into four different models as shown in Fig. 2 namely: agreement-based, certificate-based, feedback-based, and domain-based.

## 4 Proposed Trust Model

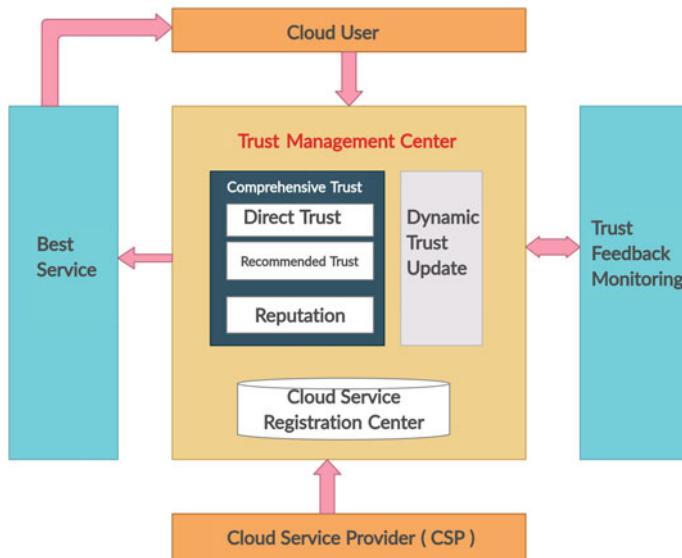
Cloud data center administrators need to use redundancy and disaster backup to ensure service availability in cloud. Selecting appropriate node in data center for replication is a challenging task. To solve this, a trust model is proposed. Based on the trust value, a node will get selected for replication which ultimately enhances availability in cloud.

## 4.1 Methodology

For this work, we opted for simulation method for evaluating the trust value of data center node using Eclipse IDE. This work is done in two phases: In the first phase design of algorithm for trust model and in the second phase simulating the model using CloudSim. The simulation is conducted for selected trust issues like confidentiality, availability, and security. For performing simulations, we have used open-source cloud environment modeling and simulation tool called CloudSim.

## 4.2 Model Framework

The trust model used for evaluating trust score for this work consists of five modules namely: Cloud Service Provider (CSP), Cloud User as Service Requestor, Cloud Services Trust Management Center (CSTMC), Dynamic Trust Update Mechanism, and Cloud Service Registration Center as shown in Fig. 3.



**Fig. 3** Trust evaluation model

### 4.3 Trust Evaluation Process

The entire trust evaluation process is shown in Fig. 4. The evaluation process considers the direct trust, recommendation trust, and reputation trust. The direct trust indicates the historical data related to direct interactions between user and the CSP node. In absence of direct interactions, the evaluation process makes use of recommended trust. The reputation trust represents the feedback collected from cloud users regarding service quality of various service attributes provided by CSP node.

### 4.4 Trust Evaluation Algorithm

The process of calculating trust value using the proposed model is shown in Fig. 5.

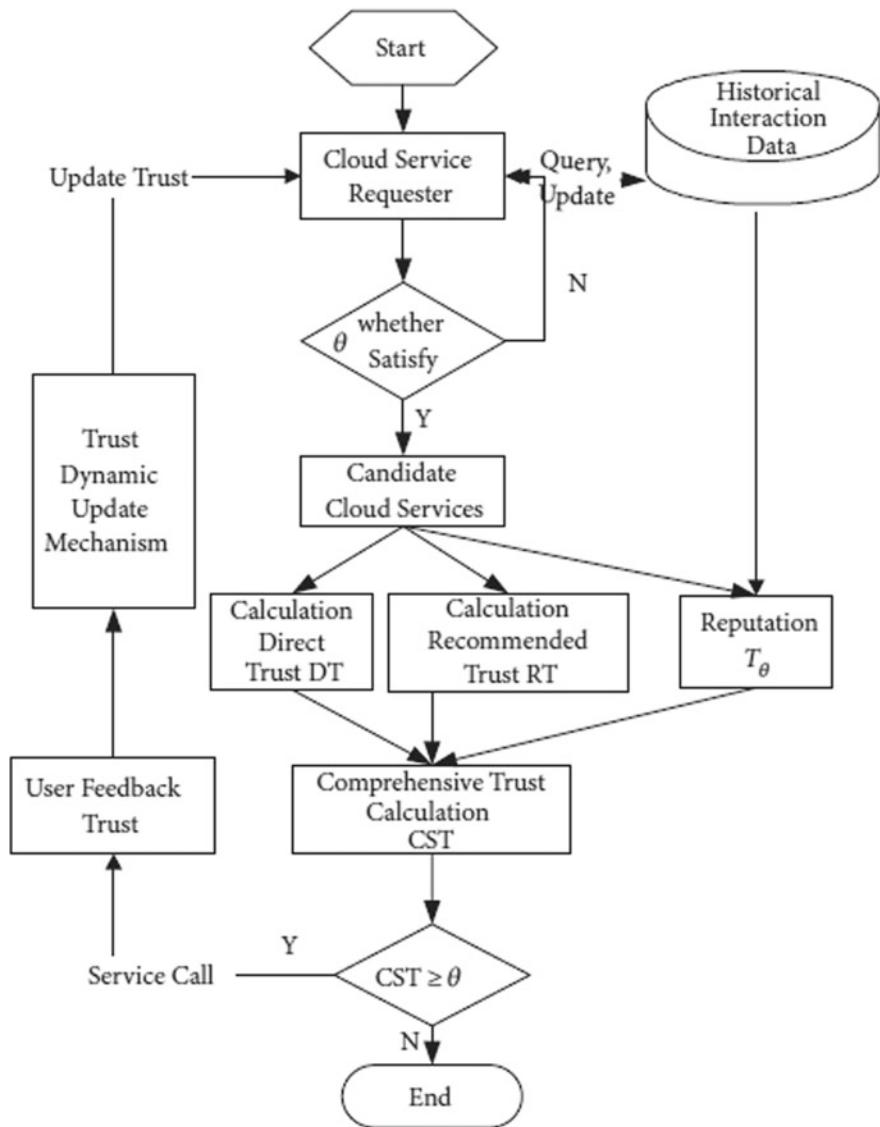
## 5 Results and Discussions

The simulation work is presented to show which node has to be selected for replication in the given data center. This model is effective in implementing trust issues for cloud computing environment. This simulation work is conducted on an Intel system with the following configuration: 1.60 GHz with 2 GB of RAM running a Java version 8.0.2 and JDK 1.8.

The proposed algorithm was run on simulated cloud computing environment using CloudSim. The virtual environment is created with two data centers, a broker and a user with series of requests. For each trust issue, a separate experiment is conducted. For each simulation, we took a data center consisting of hosts varying in the range of 100–1000. And each host with the following configuration: single-core CPU with processing capacity of 1000 MIPS, 512 MB RAM and 1 GB storage capacity. Scheduling policy used for VMs was time-shared. Users of cloud are modeled in such a way to request creation of 5 VMs. Each VM with the following configuration and constraints: single-core CPU, 1GB storage, and 512 MB physical memory. The application unit is designed in such way it consists of 5 task units. Each task unit requiring 1000 MIPS.

Simulation results are shown in Figs. 6, 7 and 8.

Figure 6 shows the comparison of trust evaluation of confidentiality when implemented the proposed algorithm under different trust evaluating classes like domain-based, certificate-based, feedback-based, and agreement-based trust models using workload traces in homogeneous environment. The  $x$ -axis is being used to represent the number of cloudlets and the  $y$ -axis is being used to represent confidentiality. The simulation results show that the proposed model gives maximum confidentiality when compared with other trust models.



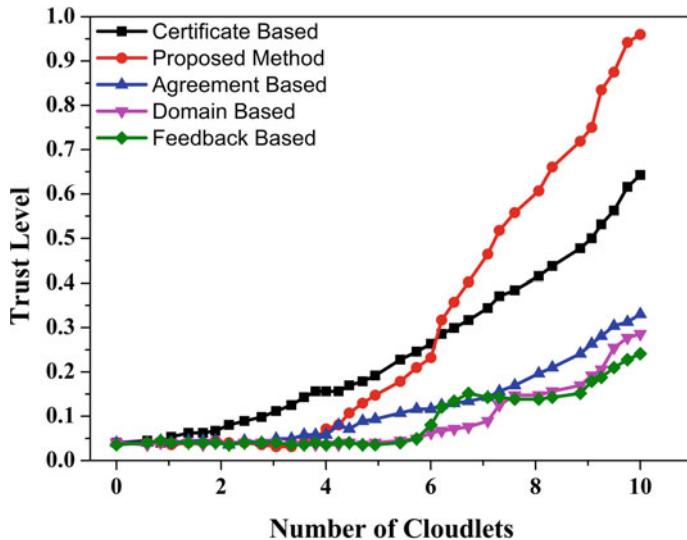
**Fig. 4** Trust evaluation process

```

Input: Requester  $u_r$ , Provider CS
Output: CST ( $u_r$ ,CS,Time).
D=Find( $u_r$ ,CS,Time);
Public Void CSTrust_Eval()
D=Find( $u_r$ ,CS,Time);
C=Count(CSr,CSp,Time); // History interaction records are inquired by cloud trust
management center
If C≥1;
Calculate  $W_j^*$  =  $W_j W'_j / \sum_{j=1}^m W_j W'_j$ 
Calculate  $\phi(i) = T(i)P(i)$ 
Calculate  $DT^{t_i} = \sum_{i=1}^n E(Q)W_j^{*T} \phi(i)$ ;
else
DT=CST0 //CST0 is a trust value that is published by cloud service
Then
Calculate  $Sim_{u_r, u_i} = \gamma_{u_r, u_i} / \sum_1^m \gamma_{u_r, u_i}$ 
 $RT = (1/m) \sum_{i=1}^m Sim_{u_r, u_i} \times T_{u_i} \times DT_{CS_k, u_i}$ 
Then
Calculate the comprehensive trust in  $t_i$ 
 $CST_{u_r, CS_k}^{t_i} = \alpha * DT_{u_r, CS_k}^{t_i} + \beta * RT_{u_r, CS_k}^{t_i} + \chi T_\theta^{t_i} // \alpha=0.7, \beta=0.15, \chi=0.15$ 
If CST ≥ θ Send service request to CS, then user feedback trust
else
Return request failed

```

**Fig. 5** Proposed trust model



**Fig. 6** Comparison of confidentiality

Figure 7 shows the comparison of trust evaluation of detection of malicious behavior when implemented the proposed model along with the algorithm under different classifications of trust evaluation models namely certificate-based, agreement-based, feedback-based, and domain-based. Horizontal line indicates number of cloudlets and vertical line indicates malicious behavior. The simulation result shows that the proposed model provides better behavior than other models.

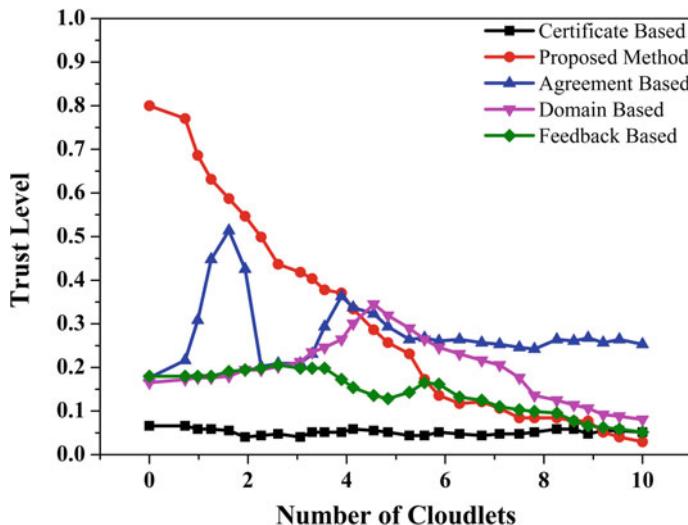


Fig. 7 Comparison of malicious behavior

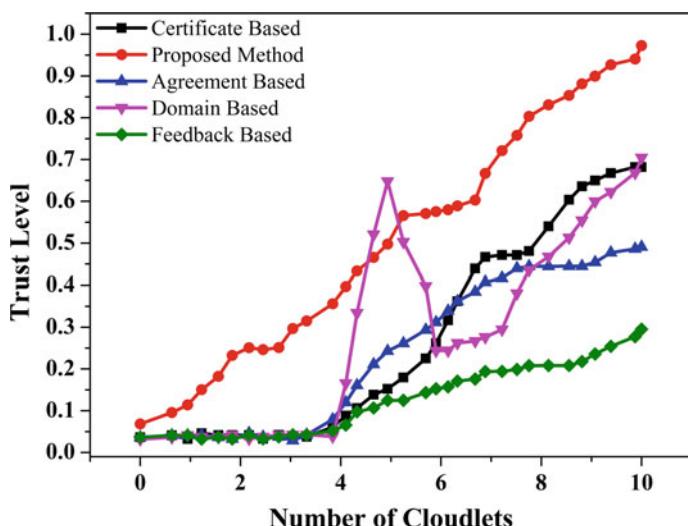


Fig. 8 Comparison of availability

Figure 8 shows the comparison of trust evaluation of data availability when implemented the proposed algorithm under different trust evaluating classes like domain-based , certificate-based, feedback-based, and agreement-based trust models. In the figure, *x*-axis is represented with number of cloudlets and the *y*-axis is represented with availability. Simulation shows that if the numbers of cloudlets are less, domain-based models give enhanced availability. When a number of cloudlets are more, the proposed model gives better availability.

## 6 Conclusions

In this paper, a detailed classification of trust evaluation models for cloud namely: agreement-based, certificate-based, feedback- based, and domain-based are given. This paper presented a trust evaluation model for selection of better data center node for replica placement which ultimately helps in providing enhanced availability in cloud. The proposed trust model is based on combining direct trust, recommended trust, and reputation together to form a comprehensive trust, which gives accurate overall trust value of the data center node. The simulation results of proposed model have shown that the node selected for replication under this model provides enhanced availability. Further a detailed comparison of proposed model with existing trust models under parameters like security, confidentiality, and availability is given.

## References

1. Ardagna, D., et al.: Cloud and multi-cloud computing: current challenges and future applications. In: 7th IEEE International Workshop on Principles of Engineering Service-Oriented and Cloud Systems, Piscataway, Page: 1–2 (2015)
2. Wang, Y., et al.: Trust and reputation model in peer-to-peer networks. In: 3rd IEEE International Conference on Peer-to-Peer Computing, Piscataway, pp. 150–157 (2003)
3. Tang, W., et al.: Z. Research on a fuzzy logic-based subjective trust management model. *J. Comput. Res. Dev.*, 1654–1659 (2005)
4. Liu, Y., et al.: The Research of Dynamic Trust Evaluation in Mobile AdHoc Networks. Anhui University (2012)
5. Zhang, Y., et al.: Research of Trust Relationship in Multi-domain Access Control. South-Central University for Nationalities (2010)
6. Dellarocas, C., et al.: Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior. In: The 2nd ACM Conference on Electronic Commerce, pp. 150–157. New York (2000)
7. Zhou, Q., et al.: Defense system model based on trust for cloud computing. *J. Comput. Appl.* 1531–1535 (2000)
8. Hwang, K., Kulkarni, S., et al.: Cloud security with virtualized defense and reputation-based trust management. In: IEEE 8Th International Conference on Dependable, Autonomic and Secure Computing, Piscataway, pp. 717–727 (2009)
9. Alhamad, M., et al.: Conceptual SLA framework for cloud computing. In: the 4th IEEE International Conference on Digital Ecosystems and Technologies, pp. 606–610. Piscataway (2010)

10. Alhamad, M., et al.: SLA-based trust model for cloud computing. In: 13th IEEE International Conference on Network-Based Information Systems, pp. 321–324, Piscataway (2010)
11. Noor, T.H., et al.: Cloud Armor: A Platform for Credibility-Based Trust Management of Cloud Services, pp. 2509–2511 (2013)
12. Asadullah, A., Oyefolahan, I.O., Bawazir, M.A.: Factors influencing users' willingness to use cloud computing services: an empirical study. **769**, 227–236 (2019)
13. Huang, et al.: Trust mechanisms for cloud computing. *J. Cloud Comput. Adv. Syst. Appl.* **2**(1), 1–14 (2013)
14. Liu, X., et al.: Performance analysis of cloud computing services considering resources sharing among virtual machines. *Int. J. Super Comput.* 357–374 (2014)
15. Celesti, A., et al.: Adding long-term availability, obfuscation, and encryption to multi-cloud storage systems. *J. Netw. Comput. Appl.* 208–218 (2016)
16. An, K., et al.: 'A cloud middleware for assuring performance and high availability of soft real-time applications. *J. Syst. Archit.* 757–769 (2014)
17. Tjang, C., et al.: Research on evaluation of SaaS SP service quality based on SLA. *J. Comput. Eng.* 31–36 (2013)
18. Dantas, J., et al.: Eucalyptus-based private clouds: availability modeling and comparison to the cost of a public cloud, pp. 1130–1140 (2017)
19. Damera, V.K., et al.: Trust issues in cloud computing. *Int. J. Recent Technol. Eng.* **8**(4), 8303–8308 (2019) ISSN: 2277-3878
20. Cho, J., et al.: A Survey on Trust Modeling. *ACM Comput. Surv.* **48**(2), 1–40 (2015)
21. Lansing, J., Sunyaev, A.: Trust in cloud computing: conceptual typology and trust-building antecedents. *Data Base Adv. Inf. Syst.* **47**(2), 58–96 (2016)
22. Tang, M., Dai, X., Liu J., Chen, J.: Towards a trust evaluation middleware for cloud service selection. *Futur. Gener. Comput. Syst.* **74**, 302–312 (2017)
23. Zhang, P.Y., et al.: Security and trust issues in Fog computing: a survey, *Futur. Gener. Comput. Syst.* **88**, 16–27 (2018)
24. Lins, S., et al.: Trust is good control is better: creating secure clouds by continuous auditing. *IEEE Trans. Cloud Comput.* **6**(3), 890–903 (2018)
25. Damera, V.K., et al.: Trust evaluation models for cloud computing. *Int. J. Sci. Technol. Res.* **9**(2), 1964–1971 (2020)

# Feature Impact on Sentiment Extraction of TEnglish Code-Mixed Movie Tweets



S. Padmaja, M. Nikitha, Sasidhar Bandu, and S. Sameen Fatima

**Abstract** Sentiment extraction is a natural language processing task dealing with the detection and classification of sentiments in various monolingual and bilingual texts. In this context, the automation of extracting sentiments from social media text is one of the pertinent areas of research as there is an enormous noisy multilingual content. This work focuses on extracting sentiments for code-mixed Telugu–English (TEnglish) bilingual Roman script movie tweets extracted using Twitter API. Initially, every tweet in the dataset was annotated with the source language of all the words present and also the sentiment expressed in the code-mixed tweet. The annotated data was automated for sentiment extraction through machine learning-based approach. Sentiment classification was accomplished with features like character N-grams, emoticons, repetitive characters, intensifiers, and negation words using support vector machine classifier with radial basis function as it performs efficiently in high-dimensional feature vectors. The study was to focus on identifying the type of feature which has more impact in capturing sentiments. The results show that character N-grams, emoticons, and negation words are the features that affect the accuracy most.

## 1 Introduction

Sentiment extraction is a part of natural language processing that investigates the automatic deduction for opinion mining of textual data [1]. The ongoing development of Web-based life made people utilize multiple languages in social networking. Yet there are different tasks conducted on code-mixed texts, the task of sentiment

---

S. Padmaja · M. Nikitha (✉)  
Keshav Memorial Institute of Technology, Hyderabad, India  
e-mail: [padmaja@kmit.in](mailto:padmaja@kmit.in)

S. Bandu  
Prince Sattam Bin Abdul Aziz University, Al-Kharj, Saudi Arabia

S. Sameen Fatima  
Anurag University, Hyderabad, India

extraction, particularly, has been rarely explored for multilingual code-mixed texts [2].

It is thus normal for the speakers in multilingual societies to mix and switch codes according to certain personal and social conditions of the communication they are involved in. The process of code alternation is called code mixing and code switching. Code switching describes the switch of languages at a level of either block of speech, sentences, or words. Further code mixing describes at word level mixing of two languages [3]. Speakers of more than one language tend to mix their language during communication. In the context of a single conversation, code switching refers to alternating between two or more languages or language varieties [4].

Code mixing is utilizing one language in another language, the mixing of at least two or more languages or language categories in content [5]. The below example (e.g. 1) is a mix of two languages, namely Telugu transliterations and English. This frequently happens when the utilization of two languages or two cultures cannot be separated from the components of one language well and frequent overlap between the two systems. Code mixing can happen with bilingual or multilingual community society. Its importance of meaning hidden in the the language cannot be clearly separated [2].

**eg 1:** '#RaviTeja Sir you are great. #Touchchesichudu movie chala bagundhi. Chala ante chala bagundhi.... Movie ante ila undali...'

The remaining paper is as discussed. In Sect. 2, previous works on sentiment analysis of code-mixed text have been discussed. Section 3 focuses on the building of corpus and the process of its annotation. Section 4 elaborates on the extraction of different key features from annotated data. Section 5 describes the implementation of ML technique using SVM classifier and the impact of different key features when eliminated while extracting sentiments in TEnglish code-mixed data. Section 6 concludes with future work.

## 2 Related Work

Sentiment extraction became difficult as social media text contains informal text, bilingual or multilingual text of different scripts, or within a script, so sentiment extraction is a major task in processing the text for future applications. There are few authors who worked on these lines.

Ghosh et al. in [6] had worked on Facebook posts and utilized altered information from ICON—2015 to automatically extract positive and negative opinions for the English–Hindi and English–Bengali code-mixed information. To perform this task, he used machine learning approaches. But the use of arbitrary emoticons was not taken care of. This work also did not deal with handling negation in data.

Sarkar et al. in [7] presented their work as a part of the shared task at the ICON 2017 challenge. Hindi–English and Bengali–English code-mixed online networking text were labelled with positive, negative, and neutral classes using ML technique. They built a model by training multinomial Naive Bayes classifier only with n-gram

and sentiwordnet as features. But they did not handle machine learning algorithms like SVM which is one of the most efficient ML algorithms for such complex data.

Jhanwar et al. [8] proposed an ensemble of only character tri-grams-based long short-term memory model and word n-grams-based multinomial Naive Bayes model to identify the sentiments of Hindi-English code-mixed data with an accuracy of 70.8% and F1 score proved to be 0.661. But our work is more focused on the impact of various features in code-mixed data.

### 3 Corpus Creation and Annotation

#### 3.1 Corpus Creation

The proposed work was focused on TEnglish code-mixed tweeter data on movie reviews which was online posted in the last 3 years. 54,109 tweeter data was scrapped using the Twitter Python API.<sup>1</sup>

Semi-automated extensive pre-processing was carried out to remove all the noisy tweets. All those tweets which were without movie-related keywords like movie, casting, lyrics, entertainment, celebrity, actor, actress wrote either in pure Telugu or in pure English language were removed considering them as noisy, thus keeping only the code-mixed movie tweets. As a result, out of 54,109, only 22,085 TEnglish code-mixed movie tweets were left for further processing.

#### 3.2 Data Annotation

Following two phases were used to annotate the text:

**Language Annotation:** For each word, three kinds of tags, namely ‘other’, ‘en’, and ‘te’, were assigned to its source language by bilingual speakers.

The tag ‘en’ was assigned to English vocabulary words such as ‘amazing’, the tag ‘te’ was assigned to Telugu vocabulary word such as ‘Aavesham’ (anger), and the tag ‘other’ was given to symbols, acronyms, punctuations, emoticons, named entities, and URLs as shown in e.g. 2.

**eg 2:** ‘Aavesham||te movie||en is||en really||en amazing||en :)||other

**Sentiment Label:** Each tweet is labelled with its respective sentiment labels referenced in Table 1. The data is manually annotated by two bilingual authors with prescribed guidelines.

**Annotation guidelines** For annotation, we adopted the approach taken by Mohammad [9], and each sentence was annotated according to the following schema:

---

<sup>1</sup><https://pypi.org/project/twitterscraper/0.2.7>.

**Table 1** Labels used for sentiment

Sentiment polarity	Class label used
Positive class	1
Negative class	-1
Neutral class	0

- **Positive class:** There is an implicit or explicit clue within the text suggesting that the speaker is in a positive class, i.e. relaxed, forgiving, happy, and admiring.
- **Negative class:** There is an implicit or explicit clue within the text suggesting that the speaker is in a negative class, i.e. anxious, violent, sad, and angry.
- **Neutral class:** There is an implicit or clue in the text suggesting that the speaker is experiencing both negative and positive feelings of neither of them which is labelled as neutral: Comparing two movies or there is no implicit or explicit indicator of the speaker's emotional state. For example, asking for like or subscription or questions about the release date or movie dialogue. This is neither positive nor negative, and hence, the statement can be labelled as neutral class.

### 3.3 *Inter Annotator Agreement*

To identify sentiment in the tweets, two human annotators who have linguistic background and proficiency in both Telugu and English have carried out the annotation of dataset. In order to validate the quality of the annotation, the inter annotator agreement was calculated between the two annotation sets each of 22,085 code-mixed tweets using Cohen's Kappa coefficient. It was found that the agreement is 0.801 which is significantly high. This implies that the presented schema and quality of the annotation are productive.

## 4 Machine Learning-Based Approach

**Feature Extraction** To train our supervised machine learning model, the work has been carried out with the following feature vectors.

**Character N-grams:** They are language independent and very efficient in text classification. These are also additionally helpful in situations when the text suffers from errors such as incorrect spellings. In the code-mixed language, groups of characters can help in capturing semantic meaning where there is an informal use of words. Character n-grams has been used as one of the features with n varying from 1 to 5.

**Emoticons:** Emoticons in the form of symbols represent textual portrayal of a writer's emotion. For example, ':o' and ':(' express sadness, and ':)' and ';' express happiness. A list of western emoticons were used from Wikipedia.<sup>2</sup>

**Repetitive Characters:** To stress upon a particular emotion, users on social media often repeat some characters in a word. For example, 'lol' (abbreviates laughing out loud) can be written as 'lool', 'loooool'. 'Happy' can be written as 'happyyy' and 'haaaappy'. Storage of such words in which a particular character is repeated more than two times in a row is used them as one of the features.

**Intensifiers:** Intensifiers are used by users to lay out emphasis on sentiments. For example, 'Movie chaala disappointing ga undi because the casting was too bad.', Translation : 'Movie was very disappointing because the casting was too bad.' Here, 'chaala' and 'too' are used to emphasize on the movie to be bad. English intensifiers' list was extracted from Wikipedia.<sup>3</sup> For creating the list of Telugu intensifiers, English intensifiers were transliterated to Telugu. The list of Telugu intensifiers were incorporated which were found in the corpus. The number of such intensifiers was counted as feature in a tweet.

**Negation Words:** Negation words were selected to address variance from the desired sentiment caused by negated phrases like 'not sad' or 'not happy'. For example, though the tweet 'Emaina chepu, I did not like the movie!' has a like unigram, it should be classified as a sad tweet. Thus, to encounter this problem, negation was defined as a separate feature. Christopher Pott's list of English negation words was used for this purpose.<sup>4</sup> Telugu negation words were selected manually from the corpus. The count of feature was used as the number of negations in a tweet.

**Feature Experiment** In order to determine the effect of each feature on classification, several experiments were performed by eliminating one feature at a time. In all the experiments, tenfold cross-validation was carried out. Experiments were performed using SVM classifier with radial basis function as they are efficient in case of high-dimensional feature vectors. For training the system classifier, Scikit-learn was used.

## 5 Results

22,085 TEEnglish code-mixed movie tweets were tested and evaluated using precision, recall, F-measure, and accuracy.

Further, experiments were performed after eliminating one feature at a time and using SVM classifier with radial basis function are as shown in Table 2. The results show that the accuracy of features like character N-grams, emoticons and negation words affected the most. An accuracy of 77.63% was able to achieve using the character N-grams, emoticons, repetitive characters, intensifiers, and negation words as features trained with SVM classifier.

---

<sup>2</sup>[https://en.wikipedia.org/wiki/List\\_of\\_emoticons](https://en.wikipedia.org/wiki/List_of_emoticons).

<sup>3</sup><https://en.wikipedia.org/wiki/Intensifier>.

<sup>4</sup><http://sentiment.christopherpotts.net/lingstruc.html>.

**Table 2** Impact of each feature when eliminated while extracting sentiments from TEnglish code-mixed data

Feature eliminated	Precision	Recall	F-measure	Accuracy (%)
None	0.771	0.716	0.726	<b>77.63</b>
Character N-grams	0.711	0.721	0.720	<b>71.84</b>
Emoticons	0.719	0.763	0.727	<b>72.22</b>
Repetitive characters	0.701	0.711	0.721	77.01
Intensifiers	0.714	0.712	0.701	76.71
Negation words	0.729	0.731	0.737	<b>73.15</b>
Character N-grams, emoticons	0.681	0.641	0.630	<b>69.84</b>
Character N-grams, repetitive characters	0.699	0.701	0.720	71.02
Character N-grams, intensifiers	0.704	0.701	0.701	71.01
Character N-grams, negation words	0.609	0.631	0.601	<b>69.15</b>
Repetitive characters, emoticons	0.709	0.703	0.710	71.22
Repetitive characters, intensifiers	0.729	0.769	0.772	71.22
Repetitive characters, negation words	0.705	0.712	0.703	71.09
Emoticons, intensifiers	0.706	0.693	0.692	71.07
Emoticons, negation words	0.619	0.663	0.627	<b>69.22</b>
Intensifiers, negation words	0.704	0.705	0.733	71.21
Character N-grams, emoticons, repetitive characters	0.612	0.625	0.638	69.06
Character N-grams, emoticons, negation words	0.603	0.612	0.622	<b>68.53</b>
Character N-grams, emoticons, intensifiers	0.608	0.615	0.634	69.01
Character N-grams, repetitive characters, intensifiers	0.631	0.638	0.651	69.89
Character N-grams, repetitive characters, negation words	0.611	0.621	0.653	69.01
Character N-grams, intensifiers, negation words	0.603	0.629	0.641	69.01
Emoticons, repetitive characters, intensifiers	0.637	0.622	0.652	69.71
Emoticons, repetitive characters, negation words	0.632	0.618	0.629	69.08
Emoticons, intensifiers, negation words	0.621	0.628	0.625	69.07
Repetitive characters, intensifiers, negation words	0.651	0.661	0.682	69.78

## 6 Conclusion and Future Work

In this work, a machine learning approach was used to build a model using TEEnglish code-mixed annotated data. Sentiment classification was accomplished with key features like character N-grams, emoticons, repetitive characters, intensifiers and negation words using support vector machine classifier with radial basis function as it performs efficiently in high-dimensional feature vectors. The study has been focused on identifying the type of feature which has more impact in capturing sentiments. The results show that with an accuracy of 77.63 %, character N-grams, emoticons, and negation words are the features which affect the accuracy most.

This work can be extended further by adding more features to enhance performance of the sentiment extraction model. It can further be extended to improve and refine the techniques to resolve ambiguous words related to movie domain and to sarcasm. Domain-specific sentiment lexicons are to be created pertaining to movie domain.

## References

1. Padmaja, S., Sameen Fatima, S.: Opinion mining and sentiment analysis-an assessment of peoples belief: a survey. *Int. J. Ad Hoc Sens. Ubiquit. Comput.* **4**(1), 21 (2013)
2. Padmaja, S., Bandu, S., Sameen Fatima, S.: Text processing of Telugu–English code mixed languages. In: International Conference on E-Business and Telecommunications, pp. 147–155. Springer (2019)
3. El-Saghir, K.: Code-Switching in Sociolinguistic Studies: Review and Analysis. *LIN 5770–Sociolinguistics*, 1–7 (2010)
4. Muysken, P.C.: Code-switching and grammatical theory (1995)
5. Ranjan, P., et al.: A comparative study on code-mixed data of Indian social media vs formal text. In: 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), pp. 608–611. IEEE (2016)
6. Ghosh, S., Ghosh, S., Das, D.: Sentiment Identification in Code-Mixed Social Media Text In: arXiv preprint [arXiv:1707.01184](https://arxiv.org/abs/1707.01184) (2017)
7. Sarkar, K.: JUKS@ SAILCodeMixed-2017: Sentiment Analysis for Indian Code Mixed Social Media Texts (2018). In: arXiv preprint [arXiv: 1802.05737](https://arxiv.org/abs/1802.05737) (2018)
8. Gopal Jhanwar, M., Das, A.: An Ensemble Model for Sentiment Analysis of Hindi-English Code-Mixed Data (2018). In: arXiv preprint [arXiv:1806.04450](https://arxiv.org/abs/1806.04450)
9. Mohammad, S.: A practical guide to sentiment annotation: challenges and solutions. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 174–179 (2016)

# Linear and Ensembling Regression Based Health Cost Insurance Prediction Using Machine Learning



M. Shyamala Devi, P. Swathi, M. Purushotham Reddy, V. Deepak Varma, A. Praveen Kumar Reddy, Saranya Vivekanandan, and Priyanka Moorthy

**Abstract** The health insurance is an important big eye-openers during the emergency need during accidents and disease pandemic situations. Many of the people will lag to hit financially and to bear the operational expenses during treatment. the need for health insurance changes from youth to old age depending on your lifestyle and genetics. Due to the change in lifestyle and diseases, the health insurance is much needed for each individual. Since it is uncertain that a medical emergency can attack anyone, anytime that impact the person so emotionally and financially. With all this background, this paper attempts to predict the Health cost insurance based on the accessible parameters like age, sex, region, Smoking, Body Mass Index, Children with the following contributions. Firstly, the Health Cost Insurance dataset is extracted from UCI machine repository and the data is preprocessed along with exploratory data analysis. Secondly, the anova test is applied to verify the features with Probability of F-Statistic  $PR(>F) < 0.05$  that highly influence the Target. Thirdly, the raw dataset and the feature scaled dataset is applied to all the Linear Regression models and the performance is analyzed. Fourth, the raw dataset and the feature scaled dataset is applied to all the Ensembling Regression models and the performance is analyzed through intercept, MAE, MSE, R2Score, and EVS. Anova Test Reults shows that the variable ‘region’ does not influence the target as the F-statistic value is 0.14. Experimental results show that polynomial regression is achieving 88% of R2Score before and after feature scaling. The Random Forest regression is achieving 86% of R2Score before and after feature scaling.

---

M. Shyamala Devi (✉) · P. Swathi · M. Purushotham Reddy · V. Deepak Varma ·  
A. Praveen Kumar Reddy

Computer Science & Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science  
and Technology, Chennai, Tamil Nadu, India  
e-mail: [shyamaladev@veltech.edu.in](mailto:shyamaladev@veltech.edu.in)

S. Vivekanandan  
Infologia Technologies, Chennai, Tamil Nadu, India

P. Moorthy  
RedBlackTree, Chennai, Tamil Nadu, India

## 1 Literature Review and Shortcomings

Predictive machine learning models [1] were used to forecast the expenditures, especially for high-cost, high-need (HCHN) patients. It demonstrates temporal correlation and predicts future health care expenditures using machine learning and observes that variables with limited predictive accuracy, population-level models that offer limited information at patient-level [2, 3]. The model demonstrates the transformational power of machine learning and artificial intelligence in care management, which would allow healthcare payers and providers to introduce care management programs [4, 5]. Interpretable regression method based on evidence Regression model to predict the cost for insurance [6, 7]. Machine learning models identify and predict potential high-cost patients and explore the key variables of the forecasting model, by comparing differences in the predictive performance of variable sets [8]. An exploratory data analysis was performed on the claims data set [9]. The EDA study aimed at understanding the data properties, inspecting qualitative features, and discovering new patterns and associations in the data through summarization and visualization [10]. The risk in building the automated health risk prediction model have the inability to directly ascertain clinical phenotypes, potential utility in disease risk prediction, when combined with data-driven machine learning [11].

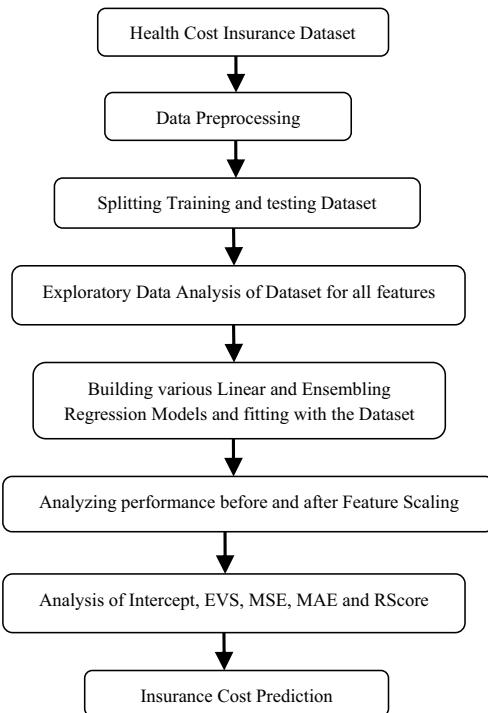
## 2 Overall Architecture

### 2.1 Dataset Preparation

The Health Insurance Cost Prediction dataset is used from UCI Machine Repository with independent variable such as age, sex, region, Smoking, Body Mass Index, Children, and dependent variable as Charges. The overall workflow is shown in Fig. 1. The contributions of this paper is given below.

1. Firstly, the Health Cost Insurance dataset is preprocessed along with exploratory data analysis.
2. Secondly, the anova test is applied to verify the features with Probability of F-Statistic  $PR(>F) < 0.05$  that highly influence the Target.
3. Thirdly, the raw dataset and the feature scaled dataset is applied to all the Linear Regression models and the performance is analyzed.
4. Fourth, the raw dataset and the feature scaled dataset is applied to all the Ensembling Regression models and the performance is analyzed through intercept, MAE, MSE, R2Score, and EVS.

**Fig. 1** Overall workflow of this paper



### 3 Feature Analysis

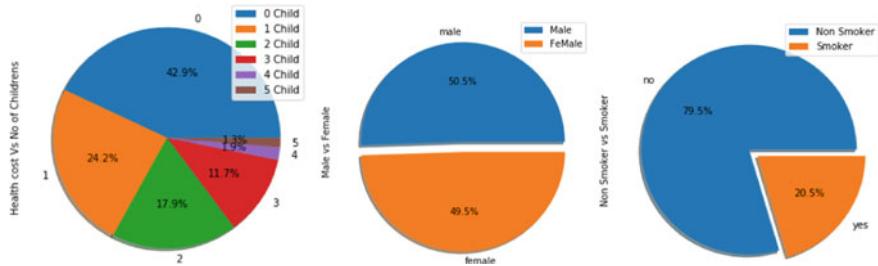
#### 3.1 Anova Test Analysis

Anova test is applied to dataset features and results shows that “Region” have value of  $PR(>F) > 0.05$  as 0.14 and does not contribute to target as in Table 1.

Anova test is used to analyze the features of the dataset by comparing both the null and alternate hypothesis. If the  $P$  value associated with the Fstatistic is less than 0.05, then the existence of that feature highly influence the target.

**Table 1** Anova test analysis with the dataset features

Features	Sum_sq	df	F	PR(>F)
Age	1.753019E + 10	1	131.174013	4.88669E-29
Sex	6.4359021E + 8	1	4.3999702	0.036133
bmi	7.7133911E + 9	1	54.709308	2.459086E-13
Children	9.0659991E + 8	1	6.206037	0.012852
Smoker	1.2151999E + 11	1	2177.6148	8.271436E-28
Region	3.0555127E + 8	1	2.084934	0.148993



**Fig. 2** Distribution of health cost with children, sex, and smoking

### 3.2 Dataset Exploratory Analysis

The extracted dataset is analyzed in order to extract the relationship of each of the independent variables with respect to the dependent variable health cost charges and the distribution of the health cost with respect to children, sex, and smoking are shown in Fig. 2. The distribution of charges with the dataset independent variables is shown in Figs. 3 and 4.

## 4 Results and Discussion

### 4.1 Implementation Setup

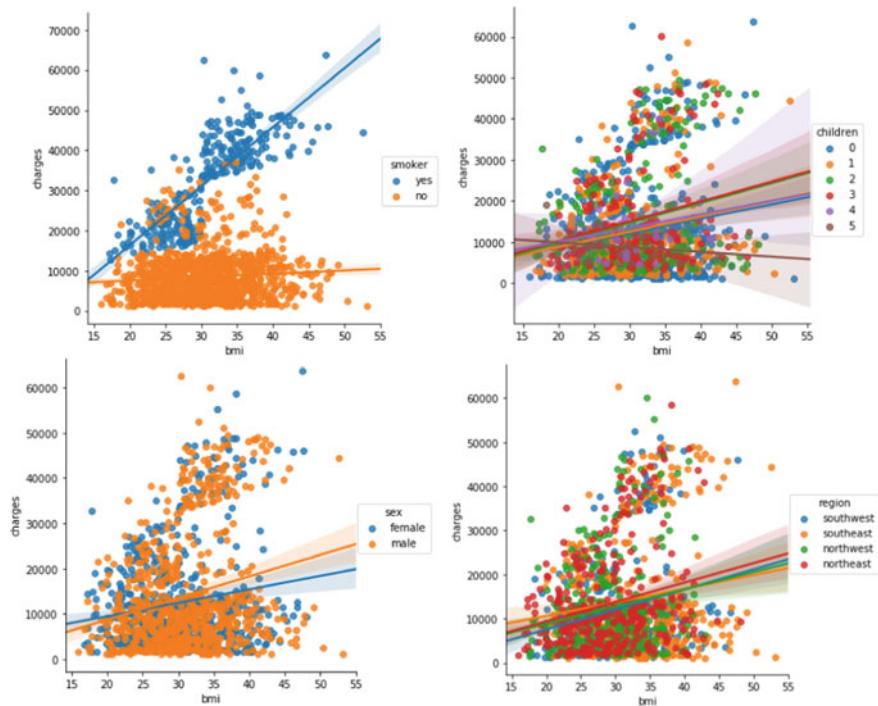
The Health Insurance Cost Prediction dataset is used from UCI Machine Repository with independent variable as age, sex, region, Smoking, Body Mass Index, Children, and dependent variable as Charges. The Anaconda IDE is used to implement python in Spyder editor for predicting the health cost. Dataset is fitted with linear regression models and Training VS testing data is shown in Fig. 5.

The dataset is fitted with ensembling regression models and analysis of Training and testing data is in Fig. 6.

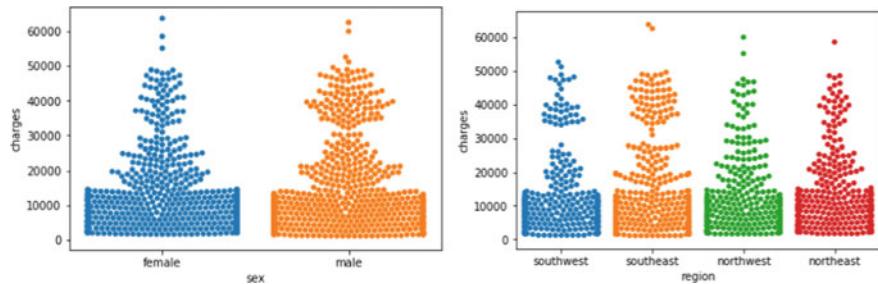
### 4.2 Performance Analysis

The performance analysis of the ensembling models before and after feature scaling is done with EVS, MAE, MSE, and R2Score and is shown in Figs. 7 and 8.

The performance analysis of the Linear models before and after feature scaling is done with EVS, MAE, MSE, and R2Score and is shown in Figs. 9 and 10.



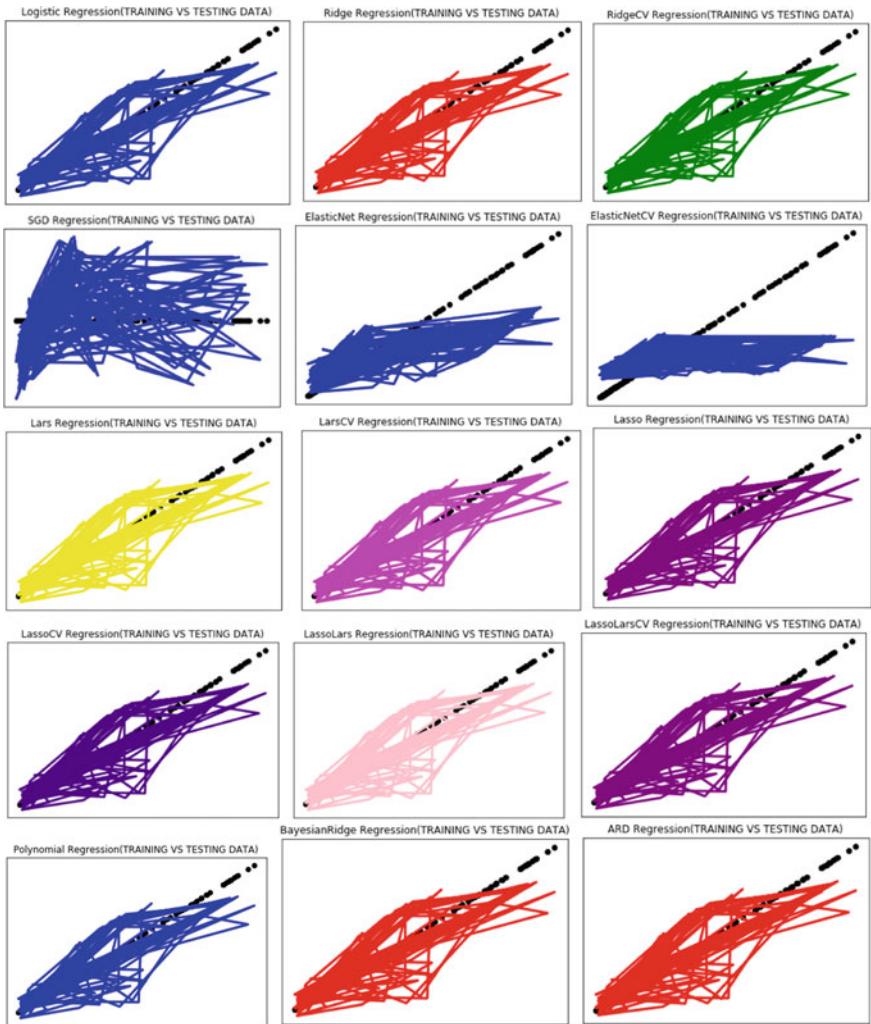
**Fig. 3** Relationship of charges and bmi based on [Top] (left) smoker (right) children. [Bottom] (left) sex (right) region



**Fig. 4** Relationship of charges with (left) sex (right) region

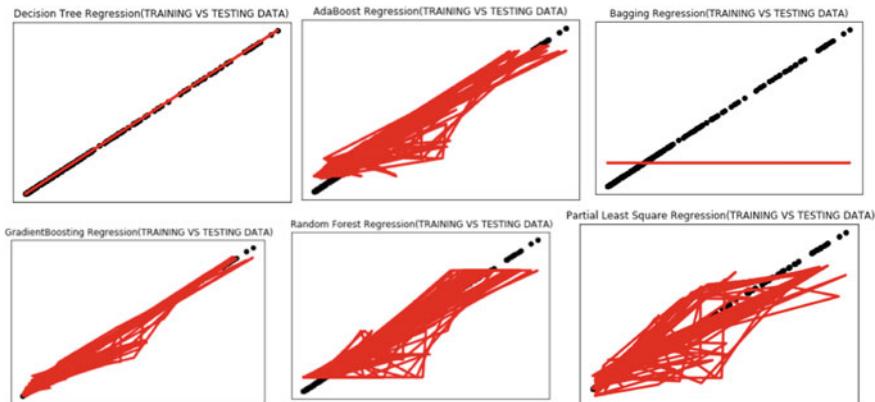
## 5 Conclusion

This paper attempts to explore the prediction of health cost insurance by applying to linear and ensembling regression models before and after feature scaling. Anova test is applied to dataset and results show “Region” have F-statistic value of PR(>F)



**Fig. 5** Training data versus testing data prediction analysis for linear regression models

>0.05 as 0.14 and does not contribute to target. The dataset is applied to linear regression models like Linear, Ridge, RidgeCV, ElasticNet, ElasticNetCV, Lars, LarsCV, Lasso, LassoCV, LassoLars, LassoLarsCV, LassoLarsIC, BayesianRidge, Polynomial, Linear Support vector, and NuSupport Vector Regression models before and after feature scaling. The dataset is applied to ensembling regression models like Decision Tree, Extra Tree, Ada Boost, Bagging, Gradient Boosting, Random Forest, and Partial Least Square Regression models before and after feature scaling. Experimental results show that polynomial regression is achieving 88% of R2Score before



**Fig. 6** Training data versus testing data prediction analysis for ensembling regressions

dfEnsemblemetrics\_BeforeFeatureScaling - DataFrame

Index	Ensemble Regression	EVS	MSE	MAE	R2SCORE
0	Decision Tree Regression	0.728109	4.38834e+07	3096.31	0.724229
1	Extra Tree Regression	0.803519	3.12805e+07	2545.62	0.803428
2	AdaBoost Regression	0.88717	2.20901e+07	3799.42	0.861182
3	Bagging Regression	0.000314844	1.75445e+08	8614.84	-0.102525
4	GradientBoosting Regression	0.897442	1.63711e+07	2366.4	0.897121
5	RandomForest Regression	0.866283	2.1279e+07	3120.8	0.866279
6	Partial Least Square Regression	0.798315	3.2162e+07	3938.26	0.797889

**Fig. 7** Performance analysis of ensembling regression models before feature scaling

dfEnsemblemetrics\_AfterFeatureScaling - DataFrame

Index	Ensemble Regression	EVS	MSE	MAE	R2SCORE
0	Decision Tree Regression	0.726236	4.42424e+07	3131.85	0.721976
1	Extra Tree Regression	0.803519	3.12805e+07	2545.62	0.803428
2	AdaBoost Regression	0.88717	2.20901e+07	3799.42	0.861182
3	Bagging Regression	0.00764808	1.74406e+08	8566.43	-0.0959983
4	GradientBoosting Regression	0.897569	1.63488e+07	2360.1	0.897261
5	RandomForest Regression	0.866283	2.1279e+07	3120.8	0.866279
6	Partial Least Square Regression	0.798315	3.2162e+07	3938.26	0.797889

**Fig. 8** Performance analysis of ensembling regression models after feature scaling

dflogmetrics\_BeforeFeatureScaling - DataFrame

Index	Regression	intercept	EVS	MSE	MAE	R2SCORE
0	Linear Regression	-11901.1	0.798281	3.21658e+07	3939.78	0.797864
1	Ridge Regression	-11865	0.797941	3.22198e+07	3952.48	0.797525
2	RidgeCV Regression	-11897.5	0.798249	3.2171e+07	3941.06	0.797832
3	ElasticNet Regression	-6755.28	0.428275	9.10448e+07	7281.37	0.427859
4	ElasticNetCV Regression	-11901.1	0.798281	3.21658e+07	3939.78	0.797864
5	Lars Regression	-11871.4	0.798218	3.21757e+07	3940.39	0.797802
6	LarsCV Regression	-11899.7	0.798257	3.21696e+07	3940.54	0.797841
7	Lasso Regression	-11761.4	0.797354	3.23138e+07	3965.81	0.796935
8	LassoCV Regression	-11638.6	0.797827	3.22381e+07	3942.11	0.79741
9	Lassolars Regression	-11871.4	0.798218	3.21757e+07	3940.39	0.797802
10	LassoLarsCV Regression	-11821.9	0.79811	3.21925e+07	3941.41	0.797697
11	LassolarsIC Regression	-11888.7	0.798168	3.21837e+07	3944.14	0.797752
12	BayesianRidge	-11757	0.797953	3.22191e+07	3937.87	0.79753
13	Polynomial Regression	-5120.3	0.881167	1.89574e+07	2822.37	0.880868
14	Linear Support Vector Regression	-17.57	0.120973	1.74333e+08	6838.62	-0.0955403
15	Nu Support Vector Regression	21235.38	6.80875e-05	2.18229e+08	13464.7	-0.371389

Fig. 9 Performance analysis of linear regression models before feature scaling

dflogmetrics\_AfterFeatureScaling - DataFrame

Index	Regression	intercept	EVS	MSE	MAE	R2SCORE
0	Linear Regression	13201.2	0.798281	3.21658e+07	3939.78	0.797864
1	Ridge Regression	13201.2	0.798217	3.21761e+07	3941.59	0.7978
2	RidgeCV Regression	13201.2	0.798217	3.21761e+07	3941.59	0.7978
3	ElasticNet Regression	13201.2	0.692204	4.90531e+07	5118.97	0.691742
4	ElasticNetCV Regression	13201.2	0.798281	3.21658e+07	3939.78	0.797864
5	Lars Regression	13201.2	0.798218	3.21757e+07	3940.39	0.797802
6	LarsCV Regression	13201.2	0.798265	3.21683e+07	3939.93	0.797849
7	Lasso Regression	13201.2	0.798132	3.2189e+07	3941.2	0.797719
8	LassoCV Regression	13201.2	0.797827	3.22381e+07	3942.11	0.79741
9	Lassolars Regression	13201.2	0.798218	3.21757e+07	3940.39	0.797802
10	LassoLarsCV Regression	13201.2	0.79811	3.21925e+07	3941.41	0.797697
11	LassolarsIC Regression	13201.2	0.798167	3.21841e+07	3942.97	0.79775
12	BayesianRidge	13201.2	0.797953	3.22191e+07	3937.87	0.79753
13	Polynomial Regression	-1.70458e+13	0.881167	1.89575e+07	2822.54	0.880868
14	Linear Support Vector Regression	-17.57	0	3.14802e+08	12476.9	-0.978271
15	Nu Support Vector Regression	21235.38	0.00161318	2.17909e+08	13455.8	-0.369379

Fig. 10 Performance analysis of ensembling regression models after feature scaling

and after feature scaling. The Random Forest regression is achieving 86% of R2Score before and after feature scaling.

## References

1. Yang, C., Delcher, C., Shenkman, E., et al.: Machine learning approaches for predicting high cost high need patient expenditures in health care. *Bio. Med. EngOnLine* **17**, 131 (2018)
2. Maisog, J., Li, W., Xu, Y., Hurley, B., Shah, H., Lemberg, R., Borden, T., Bandeian, S., Schline, M., Cross, R., Spiro, A., Michael, R., Gutfraind, A.: Using massive health insurance claims data to predict very high-cost claimants: a machine learning approach (2019)
3. Sethi, P., Jain, M.A.: Comparative feature selection approach for the prediction of healthcare coverage. *Commun. Comput. Inf. Sci.* **54**, 392–403 (2010)
4. Panay, B., Baloian, N., Pino, J., Peñafiel, S., Sanson, H., Bersano-Méndez, N.: Feature selection for health care costs prediction using weighted evidential regression. *Sensors* **20** (2020)
5. Luo, L., Li, J., Lian, S.: Using machine learning approaches to predict high-cost chronic obstructive pulmonary disease patients in China. *Health Inf. J.* **26**(3), 1577–1598 (2020)
6. Xie, Y., Schreier, G., Chang, D., Neubauer, S., Liu, Y., Lovell, N.: Predicting days in hospital using health insurance claims. *IEEE J. Biomed. Health Inf.* (2015)
7. Park, J.H., Cho, H.E., Kim, J.H.: Machine learning prediction of incidence of Alzheimer's disease using large-scale administrative health data. *npj Digit. Med.* **3**, 46 (2020)
8. Dhibe, N., Ghazzai, H., Besbes, H., Massoud, Y.: A secure AI-driven architecture for automated insurance systems: fraud detection and risk measurement. *IEEE Access* **8**, 58546–58558 (2020)
9. Blough, D.K., Ramsey, S.D.: Using generalized linear models to assess medical care costs. *Health Serv. Outcomes Res. Method.* **1**, 185–202 (2000)
10. Lysaght, T., Lim, H.Y., Xafis, V., et al.: AI-Assisted decision-making in healthcare. *ABR* **11**, 299–314 (2019)
11. Boodhun, N., Jayabalan, M.: Risk prediction in life insurance industry using supervised learning algorithms. *Complex Intell. Syst.* **4**, 145–154 (2018)

# An Adaptive Correlation Clustering-Based Recommender System for the Long-Tail Items



Soanpet Sree Lakshmi, T. AdiLakshmi, and Bakshi Abhinith

**Abstract** This is the study of long-tail problem of a recommender systems. The long tail has few ratings and is therefore difficult to use in recommendation systems. The approach presented in this paper separates the items according to their popularities. The head items are recommended based on their popularities. Long-tail items are correlation clustered according to their similarities. This method is applied on a subset of real-life dataset, and accuracy of rating prediction is studied. The diversity of recommendations is also ensured by including popular items from head part and also niche items from the long-tail part.

## 1 Introduction

There are various recommender systems like news recommender, movie recommender, music recommender to name a few. Such recommendations tend to improve user experience. They also help the organizations to improve their businesses through targeted marketing [1]. The various issues of recommendation systems are addressed in [2–4]. Many recommendation systems are ratings based, and most of them often give less importance to unpopular items. These items have lesser ratings compared to the popular items. However, in most real-life datasets, the number of popular items is far less compared to these niche items in the long tail [5]. The long-tail items have great potentiality to improve businesses, enabling personalization [6]. Various approaches are proposed in the literature to improve recommendations of the items in the long tail [7–9]. This work is based on adaptive clustering method proposed in [9]. The approach divides the items based on their popularity into head and tail sets. The head items are recommended based on conventional each item. Method. The items in the tail are then clustered based on their popularities. The author of [9] proposed that such a method reduces the computation overhead of head part and improves performance in the tail part of the dataset.

---

S. S. Lakshmi (✉) · T. AdiLakshmi · B. Abhinith  
Vasavi College of Engineering, Hyderabad 500031, India

We propose an adaptive recommender approach focussing on correlation clustering of the items in the long tail while keeping the recommendation in the head part simple.

The rest of this paper is organized into five sections:

Section 2 explains related work.

Section 3 presents the experimental settings.

Section 4 describes the adaptive correlation clustering-based recommendation.

Section 5 presents the experimental results, and finally, concluding remarks are described in Sect. 6. The distribution of movie ratings for considered dataset is included in the Appendix.

## 2 Related Work

### 2.1 Adaptive Clustering

The clustering method proposed in [9] showed that it outperformed many conventional recommendation systems by employing different recommendation approaches in head and tail part. Hence, the approach proposed is chosen for recommendation for long-tail items.

### 2.2 Correlation Based Clustering

The clustering technique used in proposed method is based on a similar approach proposed by Bohm et al. [10]. The authors propose a correlated connected clustering to improve clustering accuracy. In their paper, density-based clusters were generated, and the next point in the neighbourhood is included in the cluster only if it has desired correlation similarity with core point.

### 2.3 Proposed Adaptive Correlation Clustering

The proposed solution is recommendation system based on adaptive correlation clustering technique. This paper implements the adaptive approach where popular genre-based recommendation is applied for the head part of the data, and correlation clustering-based recommendation is chosen for the tail part. By applying correlation clustering to the items in the tail part, we propose to fine-tune the recommendations. The items can thus be suitably grouped, thereby improving the quality of recommendation.

### 3 Experimental Settings

The method proposed is applied to a subset of popular dataset: Movielens [11]. The movielens dataset contains movie-related tags information and the user movie ratings. It originally contained 20,000,263 ratings and 465,564 tag applications across 27,278 movies. The user demographics information is not included in the given dataset. The ratings are given on a scale of 1 to 5 from 138,493 users. The tag genome is a data structure that contains tag relevance scores for movies. The tag genome encodes how strongly movies exhibit particular properties represented by tags in genome scores.

From this data, a subset of first 1000 users was considered. The movies rated by these 1000 users were extracted. The movie was considered only if it had a minimum of 25 ratings. The total number of movies rated by these users all together is 1490.

Splitting point ( $\alpha$ ) for head and tail distribution was chosen to be 100. Distribution of ratings for the 1000 users—1490 movies subset of movielens is shown in the Appendix.

#### 3.1 *The Derived Variables*

##### User related

1. User\_fav\_genre: the favourite genre of the user based on ratings.
2. User\_genre\_avg: the average rating given by user for the movies genre wise.
3. User\_fav\_cluster: cluster of movies with more than average ratings given by user.

##### Movie Related

1. Movie\_avg\_rating: the average ratings value given by the users.
2. Movie\_count: number of ratings for a movie.
3. Movie\_genre: the genre to which the movie associates itself to is calculated.

### 4 The Proposed Method

#### 4.1 *Method to Predict Rating Value for a Given UserId and MovieId (Fig. 1)*

#### 4.2 *o Generate Recommendation List for a Given UserId (Fig. 2)*

The recommendation list generation list is based on:

**(i)Adaptive\_CorrClus (userId, movieId)****Step1:**

**Calculate user -related and Movie related derived variables** User\_fav\_genre, User\_genre\_avg, Movie\_avg\_rating, Movie\_count, Movie\_genre

**Step 2:** depending on movie\_count , a criterion ( $\alpha$ ) is chosen to split the movies data into 2 parts : head, tail . these are used for adaptive recommendation.

- (a) If the movie\_count>( $\alpha$ ), go to step 3
- (b) else if movie\_count<=( $\alpha$ ) go to step 4.

**Step 3:** Predict rating based on User\_genre\_avg calculated for input value of userId and movieId.

**Step 4 :** predict rating based on clustered movies rating.

- (a) the movies are clustered based on correlation connectedness
- (b) find the cluster to which the movieId belongs to.
- (c) Predict the rating based on the average of user\_genre\_ratings for movies of same genre in this cluster.

**Step 4:** calculate error between actual rating and predicted rating .

**Step 5:** Calculate MAE and RMSE

**Fig. 1** Method to predict rating value for a given userId and movieId

**(ii) recommend\_movies(userId)**

**Step1:** Calculate user -related and Movie related derived variables- User\_fav\_genre, User\_fav\_cluster, User\_genre\_avg, Movie\_avg\_rating, Movie\_count, Movie\_genre

**Step2:** compute the User\_fav\_genre and User\_fav\_cluster

**Step3:** depending on movie\_counts, a criterion ( $\alpha$ ) is chosen to split the movies data into 2 parts: head part, tail part.

**Step4:** generate movie\_clusters based on correlation clustering.

**Step 4:** The recommendation list is generated for userId with 10 movies. the list consists of two sublists of 5 movies each.

- a) sublist\_h composed of five popular movies from the head part based on user\_fav\_genre
- b) sublist\_t is composed of movies chosen from user\_fav\_cluster

**Step 5:** display the list with sublist\_h and sublist\_t.

**Fig. 2** To generate recommendation list for a given userId

## 1. Head part recommendation:

The recommendation of movies in head part is straight forward. The movies in the head part are sorted in the descending order of ratings. The fav\_genre of user is extracted, and the top five movies of the users' favourite genre are chosen for this sublist\_h, and the recommendation ensures that the movies already seen by the user will not be included in this sub list.

## 2. Tail part recommendation:

The movies are clustered based on correlation connectedness. The clusters thus formed are used for the recommendation of movies in the long tail. Each movie is described in 1128 tags. These movies were correlation clustered based on their corresponding tags value.

The movies thus extracted were all assigned to the appropriate cluster.

When a user id is provided as input, the user\_fav\_genre is extracted based on the users rating patterns. The top five ratings values are considered for recommendation of genre that is most liked by a user. From this list, the maximum—preferred genre is extracted and assigned to the user as the User\_fav\_genre. For the given userId, the movies from the User\_fav\_cluster are extracted and corresponding clusterId is identified.

Five popular movies are extracted randomly from this cluster, and the other sub-list is generated. The recommendation ensures that the movies already seen by the user will not be included in both the sub-lists.

## 5 Experimental Results

### Sample Outputs:

#### 5.1 Rating Prediction for 105 Users (Userids 751–855)

105 users were chosen for testing prediction of rating. These users watched 5017 movies altogether. The MAE and RMSE are calculated for these 5017 predictions (Table 1).

#### 5.2 Rating Prediction for 20 Movieids

20 MovieIds were chosen for testing prediction of rating. These movies have 740 ratings altogether. The MAE and RMSE are calculated for these predictions (Table 2).

#### 5.3 Recommendation List

Recommendation list is generated for two users; sample list for user userId 751 and 994 is generated. Ten movies are recommended for each user (Fig. 3).

## 6 Conclusion

In this paper, we employed the adaptive correlated clustering technique, and the MAE and RMSE were calculated. In calculation of predicted ratings, the next highest rate

**Table 1** Output\_screen\_snip-Calculated MAE and RMSE for user 751-855

	A	B	C	D	E	F	G
1							
2	uid	movid	predict rat	actual rat	error	MAE	RMSE
3	751	10	4	2	2	0.644	0.887
4	751	11	4	4	0		
5	751	16	4.5	5	0.5		
6	751	21	4	5	1		
7	751	47	4	4	0		
8	751	50	4.5	4	0.5		
9	751	110	4	5	1		
10	751	150	3.5	3	0.5		
11	751	266	4	3	1		
12	751	300	4	4	0		
13	751	318	4.5	4	0.5		
14	751	350	4	3	1		
15	751	356	4	4	0		
16	751	357	4	4	0		
		<b>MAE</b>		<b>0.644</b>			
		<b>RMSE</b>		<b>0.887</b>			

value is considered. Further, the items in the recommendation list were based on combination of popular items and niche items. Thus, diversity in recommendation is ensured. However, the performance of recommender system improves further if user demographics are considered.

**Table 2** Output\_screen\_snip for movie specific recommendation 20 sample movies

uid	movid	predic f rat	Actual rat	error	MAE	RMSE
9	4369	3.063	4	0.94	0.706459	0.90677
24	4369	3.185	3	0.19		
96	4369	3.164	3	0.16		
116	4369	2.579	1	1.58		
131	4369	2.623	0.5	2.12		
156	4369	3.361	4	0.64		
206	4369	3.328	3.5	0.17		
247	4369	3.19	3	0.19		
258	4369	3.172	4	0.83		
294	4369	3.172	3	0.17		
313	4369	3.034	2.5	0.53		
337	4369	3.488	3.5	0.01		
367	4369	3.05	2.5	0.55		
369	4369	3.36	3.5	0.14		
383	4369	2.632	3	0.37		
388	4369	2.599	1	1.6		
422	4369	3.033	5	1.97		
430	4369	3.329	4.5	1.17		
462	4369	3.285	3.5	0.22		
469	4369	2.666	0.5	2.17		
489	4369	3.104	2.5	0.6		
520	4369	3.585	5	1.41		

**MAE 0.707****RMSE 0.907.**

**Fig. 3** Sample recommendation lists

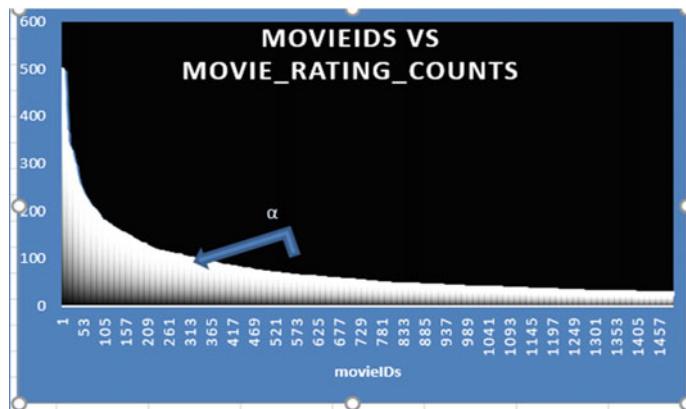
```
In [47]: rec(751)
Six Days Seven Nights (1998)
Licence to Kill (1989)
Scream 3 (2000)
Little Nicky (2000)
Arachnophobia (1990)
In July (Im Juli) (2000)
Four Adventures of Reinette and Mirabelle (1987)
American: The Bill Hicks Story (2009)
Animals are Beautiful People (1974)
Bad Medicine (1985)

In [48]: rec(994)
Threesome (1994)
Under Siege 2: Dark Territory (1995)
Hard Target (1993)
Diamonds Are Forever (1971)
Godzilla (1998)
Era of Vampires, The (2003)
Inception (2010)
Waist Deep (2006)
Strul (1988)
The Raid 2: Berandal (2014)

In [49]:
```

## Appendix

Distribution of ratings for the 1000 users—1490 movies subset of movielens:



## References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state -of-the art and possible extensions IEEE Trans. Knowl. Data Eng. **17**(6), 734749 (2005)
2. Sree Lakshmi, S., Adi Lakshmi, T.: The survey of recommender systems. Int. J. Eng. Trends Technol. (IJETT) Special Issue (2017)
3. Ramakrishna Murty, M., Murthy, J.V.R., Prasad Reddy, P.V.G.D., Sapathy, S.C.: Performance of teaching learning based optimization algorithm with various teaching factor values for solving optimization problems. In: International Conference FICTA-13 at Bhuvaneswar, Springer AISC (indexed by SCOPUS, ISI proceeding DBLP etc), vol. 247, pp. 207216 (2013). ISBN 978-3-319-02931-3
4. Ramakrishna Murty, M., Murthy, J.V.R., Prasad Reddy, P.V.G.D., Satapathy, S.: Homogeneity separateness: a new validity measure for clustering problems. In: International Conference and Published the Proceedings in AISC and Computing, pp. 110. Springer, (indexed by SCOPUS, ISI proceeding DBLP etc) vol. 248 (2014). ISBN 978-3-319-03106
5. Cremonesi, P., Garzotto, F., Pagano, R., Quadrana, M.: Recommending without Short Head, www 14 companion, April 711, 2014
6. Anderson, C.: The Long Tail. Hyperion Press (2006)
7. Yin, H., Cui, B., Li, J., Yao, J., Chen, C.: Challenging the long tail recommendation. In: Proceedings of the VLDB Endowment, vol. 5(9)
8. Park, Y.J., Tuzhilin, A.: The long tail of the recommender systems and how to leverage it. In: Proceedings of ACM Conference Recommender Systems, pp. 1118 (2008)
9. Park, Y.J.: The adaptive clustering method for the long tail problem of recommender systems. IEEE Trans. Knowl. Data Eng. **25**(8) (2013)
10. Bohm, C., Kailing, K., Kroger, P., Zimek, A.: Computing clusters of correlation connected objects. In: Proceedings of ACM International Conference on Management of Data (SIGMOD), pp. 455466. Paris, France
11. <https://les.grouplens.org/datasets/movielens/ml-20m-README.html>
12. [https://les.grouplens.org/papers/tag\\_genome.pdf](https://les.grouplens.org/papers/tag_genome.pdf)

# Plant Leaf Identification Using HOG and Random Forest Regressor



Jyotisagar Bal, Manas Kumar Rath, and Prasanta Kumar Swain

**Abstract** According to Indian state of Forest Report 2019, the total forest in India is 712,249 km<sup>2</sup> which is 22% of total geographical area of India. India is rich in biodiversity which includes more than 40,000 species of plants. It is very difficult to distinguish between different plant species. Botanist can identify these plants using the characteristic of leaf but the process is very difficult and time taking. So, in this paper, we proposed a method to recognize plants on the basis of its leaf pattern. Here, Histogram of Oriented Gradients (HOG) and Random Forest Regressor techniques are used for plant leaf image classification and recognition. Experimental results show that the proposed model achieves accuracy up to 99% in leaf recognition.

## 1 Introduction

Plants play a vital role in sustaining life on Earth. Ecosystem plants are the primary producer as they can only convert solar energy to chemical energy through the process of photosynthesis. Plants also play an important role in curbing pollution and release fresh oxygen to atmosphere. Plants contain medicinal properties which help us to cure many diseases. AYUSH organization of India has identified over 8000 herbal remedies. World health organization (WHO) in its survey told that 80% of people worldwide rely on herbal medicine for their primary healthcare needs. Plants have also economic significance as it produces wood, timber, flower, fruit, etc. Many flowering plants are used as a source of raw material for perfume manufacturing industry. Horticulture is also helpful in creating employment in India. Tribal people depend upon minor forest products for their livelihood.

India is rich in its biodiversity. India has four biodiversity hotspots, i.e., Eastern Himalaya, Western Himalayas, Western Ghats and Andaman and Nicobar Island.

---

J. Bal · P. K. Swain (✉)  
North Orissa University, Baripada, India

M. K. Rath  
KIIT Deemed University, Bhubaneswar, India

India is home for more than 40,000 species of plant including a variety of endemics. Therefore, it is important to identify different plant species to know their medicinal and other economic benefits so that appropriate steps could be taken for their preservation. But unfortunately, due to limited research and time taking process, small number of plants have been identified. Although plant species may resemble morphologically, each plant has a distinct leaf shape. Therefore, proper study of leaf pattern can pave a way for easy and fast identification of plant species. Hence, we used HOG and Random Forest Regressor [1–3] to analyze leaf shape and identify different plant species.

Remaining of the paper is organized as follows. Section 2 represents literature review. Overview of image identification method is given in Sect. 3. Dataset creation and experiment is done in Sect. 4. Performance study is presented in Sect. 5, and Sect. 6 concludes the paper.

## 2 Literature Review

After a deep literature review of papers based on wood identification, human identification, plant leaf identification using several methods like HOG, CNN, visual system and algorithms like SVM, Linear SVM, GoogleNet, Random Forest Regressor, we followed the following three papers as they have higher relevance to the proposed work.

Ecosystem consists of hundred numbers of trees and it is very difficult to distinguish between them. While working with plants and leaves it is believed that information can be obtained from plant leaf images [4] are sufficient to identify different plant species. For classification of images, better accuracy, faster execution, a proper feature descriptor, and suitable classifier are very much essential. So, HOG as feature descriptor and algorithm of image feature extraction by Navneet Dalal and Bill Triggs is applied on the ICL database to make the model [5]. HOG feature extraction is done on the basis of calculating well-normalized HOG in the detection window. The features are extracted based on these steps:

1. Counting histograms of oriented gradient cell.
2. Segment images into cell.
3. Normalization of histograms on overlapping blocks.

The dataset used here is ICL public database, in which all the images are taken by cameras or scanners in a white background under various illumination conditions. All the images are of uniform size and colorful. The database includes 200 species of plant and a total of 6000 images. A series of experiment is done by changing the cell size and block size of image, increasing and decreasing the number of orientation bins, changing the block stride size and found that it affects the final result in a great way.

Identifying humans in an image is also a challenging task. To detect a human in an image, first it needs to separate him from the background which can contain a

thousand number of data and information. Human beings can also have a wide range of poses for which we need a powerful and advanced model to identify. Navneet Dalal and Bill Triggs made a model using HOG as feature descriptor and Linear SVM algorithm for classification purpose [6].

The MIT pedestrian dataset containing 509 training and 200 testing images and “INRIA” dataset containing 1805 of  $64 \times 128$  images of humans cropped from several photos is used for experimentation. In the preprocessing step, image is normalized into gamma and color, and the number of gradients is calculated. Then the contrast of overlapping spatial blocks is normalized to reduce the noise. After that, HOG feature is collected over detection window, and then linear SVM algorithm is applied to classify persons from other objects. This experimental result shows that normalized HOG features similar to SIFT descriptors in a dense overlapping grid and gives very good result. Machine learning is a growing technique and advanced hardware with big data made this more practical. Convolutional Neural Network (CNN), which is more often used for deep learning is applied to check its robustness over plant leaves [7].

CNN model works more likely visual systems in humans. When a human sees an object, the edges of the object are detected on the basis of light intensity difference and then this information is transferred to lateral geniculate nucleus (LGN). The LGN neurons compress the entire shape around the corner and send it to primary visual cortex ( $v1$ ). It identifies corners, contour and direction and difference between the objects, and the result is sent to secondary visual cortex ( $v2$ ). Then  $v2$  neuron identifies overall shape and sends the result to visual cortex ( $v3$ ).  $V3$  neuron identifies the color of whole object and overall shape and color of the object is re-identified in lateral optical cortex (LOC). CNN has several simple computing units of artificial neuron, and every unit is connected to each other through weight connectors. These weights are calculated of the given input to produce the desired output. Here, CNN and GoogleNet are used together for training and testing of the dataset.

### 3 Overview of Image Identification Method

In this paper, we have implemented HOG and Random Forest Regressor to develop the image identification system. It mainly consists of three phases as shown in Fig. 1.

1. Preprocessing of the sample image.



**Fig. 1** An overview of image identification system with HOG and Random forest

2. Feature extraction [8] using HOG method.
3. Classification [9] of the plant species using Random Forest Regressor.

In preprocessing step, the leaf images are resized and set to a ratio of 1:2 of width and height for further processing. At the second step, the HOG is used to extract features from the leaves. In the last step, Random Forest is used to classify the species of the plant.

### ***3.1 Histogram of Oriented Gradients (HOG)***

Mostly for computer vision and object detection, feature extraction from a captured or live image is done after defining HOG feature description. From Fig. 2, it is clearly visible that the image has the information like object shape, color, edges and background. But, Fig. 3 has only limited information; still, it is easy to identify because it has only the essential information that we want to recognize that object. This is the main task of a feature descriptor [10]. For this purpose, HOG is used in our research so that it not only identifies the object but also reduces the complexity and processing time.

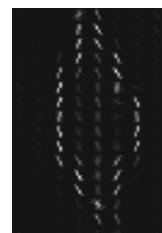
#### **3.1.1 Working Principle of HOG**

HOG scans the image bit by bit, and it mainly focuses on the shape or structure of the image. It can also give the edge direction by mapping the orientation and gradient of the image. The whole image is divided into small regions, and for every region,

**Fig. 2** Ballot paper



**Fig. 3** Histogram of ballot paper



the orientation and gradients are calculated. Then HOG is created for each block separately. The histogram is created using orientations and gradients of each pixel value for this it is named as HOG.

### 3.1.2 Steps to Calculate HOG

The image given below is of (6000 \* 8000) size. Let's calculate HOG for this image:

(i) **Preprocess the data:**

While working on images with machine learning, preprocessing [11] of the data is very much important. In the first step, we have to balance the ratio to 1:2 between width and height that means balancing to the size of (64 × 128). Such size is due to the division of the image block wise into (8 × 8) or (16 × 16) for feature extraction. Image with a particular size will make our calculations very simple. Figure 4 is the sample of a china leaf from our proposed dataset and Fig. 5 is the resized image (64 × 128) of china leaf.

(ii) **Calculating the gradients (direction x and y):**

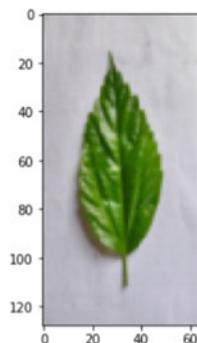
Subsequent step involves the calculation of gradient value of each image pixel. Let us take a table and calculate the gradients.

Let's, calculate the gradient for the colored pixel. The gradient is calculated in X direction by subtracting the left pixel value from the right pixel intensity value and in the same way the gradient is calculated in Y direction by subtracting the below

**Fig. 4** A china leaf



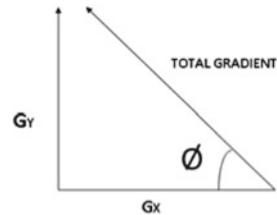
**Fig. 5** Resized china leaf



**Fig. 6** Pixel values of small patch of an image

50	70	81	93	23
79	85	128	125	48
25	156	170	178	93
33	59	112	81	110
37	135	127	89	75

**Fig. 7** Pythagoras formula



pixel value from the above pixel intensity value. The result is shown in Figs. 6 and 7.

Change in  $X$  direction ( $G_x$ ) =  $178 - 156 = 22$ .

Change in  $Y$  direction ( $G_y$ ) =  $128 - 112 = 16$ .

Now, we have two resultant matrices, first one stores the  $X$  direction gradients and the other stores the same in  $Y$  direction. After that similar process is iterated for each image pixel. The next step is to calculate the orientation and magnitude with the help of gradients that we have calculated in the previous step. Here Pythagoras theorem will be used to calculate the orientation.

Previously, we got that  $G_x = 22$  and  $G_y = 16$ . Applying Pythagoras theorem, the total gradient magnitude is calculated as:

$$\text{Total gradient magnitude} = \sqrt{(G_x)^2 + (G_y)^2}$$

$$\text{Total gradient magnitude} = \sqrt{(22)^2 + (16)^2} = 27.2$$

Further, pixel orientation can be calculated by taking the tangent for the angles:

$$\tan(\theta) = G_y / G_x$$

So, the angle value would be,

$$\theta = \text{atan}(G_y / G_x) = 72^\circ$$

Now, the histogram can be created using the orientations and gradients. With the generated frequency table (as shown in Fig. 8), HISTOGRAM with angle is created on X-axis and the Y-axis frequencies. After HOG is carried out to decrease the noise of the image.

50	70	81	93	23
79	85	128	125	48
25	156	170	178	93
33	59	112	81	110
37	135	127	89	75
<b>FREQUENCY</b>				
<b>ANGLE</b>				
	1	2	3	..
	70	71	72	73
	74	....	177	178
			179	180

**Fig. 8** Pixel and frequency table for HOG calculation

### 3.2 Random Forest Algorithm with Python and Sci-Kit-Learn Library

Random forest algorithm undergoes supervised machine learning based on ensemble learning. In ensemble learning, different types of algorithms can be joined or same algorithms joined multiple times to build a more powerful prediction model. In random forest algorithm, multiple decision trees combined multiple times resulting in a forest of trees that is why it is known as “Random Forest”.

#### How Random Forest works

Firstly, it picks N number of random records from the given dataset, and then it builds a decision tree. Then the numbers of trees are chosen according to required model and size of the dataset. Then the above steps are repeated again. If the user adds a new data and regression problem [12, 13] occurs then each tree in the forest predicts a value for the output. Then the final value is calculated by taking the average of all values.

## 4 Dataset and Experiment

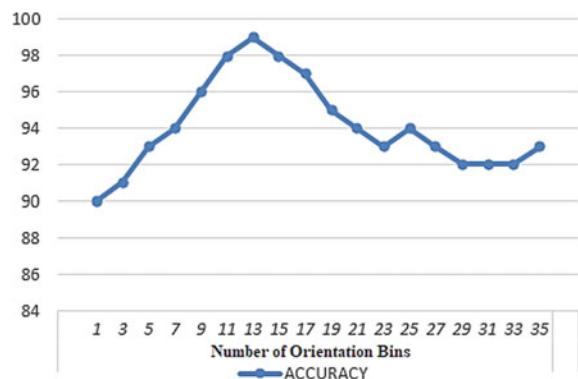
The dataset containing leaves of 11 different plant species is collected (shown in Fig. 9) from rural areas of Odisha which includes 1100 leaf images. All these images are taken by mobile camera under several lighting conditions. After leaves are plucked from plant it placed on a white piece of paper for background. Only upper part of the leaf images is taken. All the images are of uniform size. Some damaged leaves are also taken in the dataset for more complexity and to make a better and powerful model.

Throughout the experiment, it is seen that several factors like number of orientation bins, number of trees in Random Forest and image size affect the final result as well as computing time also as shown in Figs. 10 and 11.

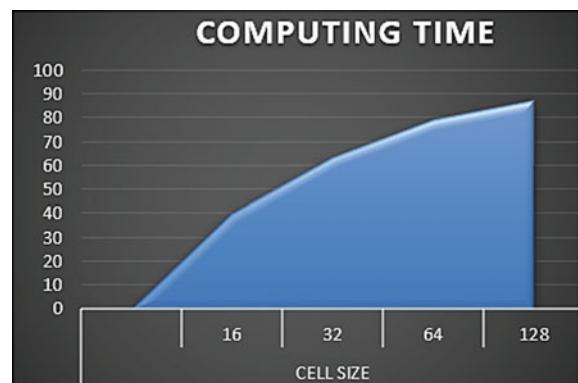


**Fig. 9** Several types of leaf samples from dataset

**Fig. 10** Effect of no. of orientation bins



**Fig. 11** Effect of block size and cell size



## 5 Performance Study

See Table 1.

## 6 Conclusion

In this paper, we have proposed a technique to identify plant leaves based on their color and structure using HOG and Random Forest Regressor. A dataset of 1100 different types of leaf images has been created. The experimental result shows that our model performs very well with 99% accuracy when all the variables are set in

**Table 1** Comparative study of the proposed method with different methods used in the literature

Paper name	Algorithms	Accuracy	Dataset	Pros and Cons
Plant leaf recognition using histogram of oriented gradients [5]	HOG	85–97%	ICL dataset	<p>Pros:</p> <ul style="list-style-type: none"> <li>• Independent of algorithm</li> <li>• Great result</li> </ul> <p>Cons:</p> <ul style="list-style-type: none"> <li>• Third party data</li> <li>• Images captured with white background</li> <li>• Noisy data affecting the accuracy</li> </ul>
HOG for human detection [6]	HOG and linear SVM	84–89%	MIT pedestrian test dataset	<p>Pros:</p> <ul style="list-style-type: none"> <li>• Work tested on several large and complicated dataset</li> <li>• Works fine with a great result</li> </ul> <p>Cons:</p> <ul style="list-style-type: none"> <li>• Needs a powerful processor with more RAM for computation</li> </ul>
Plant leaf recognition using a convolution neural network [7]	CNNVisual system and Google Net	94% with good data and 90% with noisy data	Flavia dataset	<p>Pros:</p> <ul style="list-style-type: none"> <li>• Achieves accuracy even when 30% of the images are destroyed</li> </ul> <p>Cons:</p> <ul style="list-style-type: none"> <li>• Computation level is high</li> <li>• Generalization is difficult due to dependency on specific data</li> </ul>

(continued)

**Table 1** (continued)

Paper name	Algorithms	Accuracy	Dataset	Pros and Cons
<i>Our work:</i> Plant leaf identification using HOG and Random Forest Regressor	HOG and Random Forest Regressor	99%	Own dataset	<p>Pros:</p> <ul style="list-style-type: none"> <li>• Great accuracy</li> <li>• Faster than other models</li> <li>• Own dataset, model predicts well even for damaged leaves and noisy data</li> <li>• Real-world implication is possible</li> </ul>

proper values. Further, we are working on how to identify leaves attached to plant without plucking it and also without any white background.

## Reference

1. Tsolakidis, D.G., Kosmopoulos, D.I., Papadourakis, G.: Plant leaf recognition using Zernike moments and histogram of oriented gradients. In: Hellenic Conference on Artificial Intelligence, pp. 406–417. Springer, Cham (2014)
2. Liu, J., Lv, F., Di, P.: Identification of sunflower leaf diseases based on random forest algorithm. In: International Conference on Intelligent Computing, Automation and Systems (ICICAS), pp. 459–463. IEEE (2019)
3. Chaki, J., Parekh, R.: Plant leaf recognition using shape based features and neural network classifiers. Int. J. Adv. Comput. Sci. Appl. (IJACSA) **2**(10) (2011)
4. Raut, S.P., Bhattacharya, A.S.: Plant recognition system based on leaf image. In: 2nd International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1579–1581. IEEE (2018)
5. Xia, Q., Zhu, H.D., Gan, Y., Shang, L.: Plant leaf recognition using histograms of oriented gradients. In: International Conference on Intelligent Computing, pp. 369–374. Springer, Cham (2014)
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1, pp. 886–893. IEEE (2005)
7. Jeon, W.S., Rhee, S.Y.: Plant leaf recognition using a convolution neural network. Int. J. Fuzzy Logic Intell. Syst. **17**(1), 26–34 (2017)
8. El Massi, I., Es-Saady, Y., El Yassa, M., Mammass, D., Benazoun, A.: Automatic recognition of the damages and symptoms on plant leaves using parallel combination of two classifiers. In: 13th International Conference on Computer Graphics, Imaging and Visualization (CGIV), pp. 131–136. IEEE (2016)
9. Lukic, M., Tuba, E., Tuba, M.: Leaf recognition algorithm using support vector machine with Hu moments and local binary patterns. In: 15th International Symposium on Applied Machine Intelligence and Informatics (SAMI), pp. 485–490. IEEE (2017)
10. Es-saady, Y., El Massi, I., El Yassa, M., Mammass, D., Benazoun, A.: Automatic recognition of plant leaves diseases based on serial combination of two SVM classifiers. In: International Conference on Electrical and Information Technologies (ICEIT), pp. 561–566. IEEE (2016)

11. Ali, R., Hardie, R., Essa, A.: A leaf recognition approach to plant classification using machine learning. In: IEEE National Aerospace and Electronics Conference (NAECON), pp. 431–434. IEEE (2018)
12. Srivastava, V., Khunteta, A.: Comparative analysis of leaf classification and recognition by different SVM classifiers. In: International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 626–631. IEEE (2018)
13. Elhariri, E., El-Bendary, N., Hassani, A.E.: Plant classification system based on leaf features. In: 9th International Conference on Computer Engineering & Systems (ICCES), pp. 271–276. IEEE (2014)

# Deep Learning Based Facial Feature Detection for Ethnicity Recognition



Sujitha Juliet Devaraj, R. Catherine Joy, I. Santhosh, and I. C. Kevin

**Abstract** In the study of ethnical group recognition, facial feature discovery is one of the essential tasks. Specially, with the recent development of deep learning methods, significant evolution has been made in computer-based facial recognition. In this paper, we proposed a deep learning model for ethnicity recognition, considering the facial features. At first, we gathered the dataset for ethnical group which includes Southeast Asian, East Asian, Black, White, Latin, Indian and Middle Eastern. Also, we have collected real-time face images to recognize the ethnicity group. Every ethnicity group has its own characteristics and facial features. Therefore, we considered the particular ethnical groups and intended to find the similar categorization. Since human facial information will vary based on the geographical location, we have collected data from various regions to achieve better results for our experiment. In this analysis, deep leaning-based ethnicity identification has been exploited for various details in human faces. This study is proposed to understand and analyze the relation between the ethnicity and ethnical groups, by considering the images with different intensity and registered range. We trained 85,000 images and managed to give a precise result. The research work on ethnicity recognition with facial features helps in the study of facial features evolution in anthropology (the study of human societies and cultures and their development).

---

S. J. Devaraj (✉) · R. Catherine Joy · I. Santhosh · I. C. Kevin  
Karunya Institute of Technology and Sciences, Coimbatore, India  
e-mail: [sujitha@karunya.edu](mailto:sujitha@karunya.edu)

R. Catherine Joy  
e-mail: [catherinejoy@karunya.edu](mailto:catherinejoy@karunya.edu)

I. Santhosh  
e-mail: [santhoshi@karunya.edu.in](mailto:santhoshi@karunya.edu.in)

## 1 Introduction

In recent years, facial feature-based analysis of race, nation, and ethnical groups has become a popular field of research in face recognition society [1]. Especially, with a swift development in human globalization, facial identification and recognition methods have great demand in providing security to public, monitoring the country borders and regulating the movement of people across the border, and also in customs check.

Generally, the factors like gene, environment and society will greatly influence the facial features of a person. Out of which, gene is one important factor which has very important role in ethnical groups. It is hardly unique and it's very hard to analyze it from other ethnical groups. The only possible way is by understanding the various gene systems and analyzing the similarities between them. This analysis is significant to find the similarity between facial features for different ethnicities.

In the study of ethnicity recognition, the group of people is distinguished based on their language, genes, culture and their geographical location [2]. Race and ethnicity terms are strongly correlated although they have dissimilarity. Few examples for ethnical groups are American Indian, African American, Black American, Indian, Asian, Native Hawaiian, White America and Pacific Islander.

Face recognition system detects, recognizes and verifies a person from the given digital image. Using this technology, various ranges of ethnical groups can be easily studied and the person's ethnicity also can be recognized. In general, ethnicity is identified through facial features such as face texture, geometrical alignment, eyes, hair and physical structures. This research work will be helpful to Anthropologist who is mainly engaged in the study of human societies, cultures and their development.

In recent years, there has been a remarkable investigation carried out on deep learning-based algorithms for Computer vision applications. These algorithms are effectively used for feature extraction, image classification, and image recognition purposes. A major breakthrough in deep learning models has been achieved by Convolutional Neural Network (CNN). CNN is presently considered as the preference of neural networks for classification of images since it detects patterns in tiny parts of an image, for instance, a small curve of an eyebrow. It extracts more detailed and higher-level information from the given image progressively. In this work, we have proposed CNN-based model to extract facial features from the ethnical group consists of Asian, Black, Indian and Latin faces and to recognize ethnicity.

## 2 Related Works

In the recent past, the field of ethnicity recognition has gained extensive consideration and significant development. Researchers have proposed remarkable techniques, which can perform viably and accomplish a better accuracy. All the methods are commonly designed with two important parts: Feature extraction, and Classification.

Wang et al. [3] proposed a face feature analysis method for ethnicity identification. They built an ethical face dataset with Chinese Uyghur, Korean and Tibetan. They proved that if the feature is dependent on entire face image, the sparse approach is not suitable for ethncial class recognition. They considered three ethncial classes and analyzed the individuality of each class to predict the common classification for the three classes. In this recognition method, the comprehensive facial features are end up being incapable for ethnicity examination. The proposed “T” region for ethnic characteristic description via data mining technique was investigated and concluded that this method is effective to recognize for ethnicity, however, not for face recognition.

Very recently, Gao et al. [4] have proposed a recognition model for Chinese ethnic class using transfer learning from deep convolution network. The model achieved an accuracy of 80.5% with better generalization performance. They also investigated the feasibility of the model for ethnicity recognition and proved its effectiveness. The technique proposed by Achkar et al. [5] extracted geometric features from images, including the nostril width, nose tip, lips width, brow width and coloration with the aid of the Viola-Jones method, followed by the execution of artificial neural networks. The system is executed through 3 steps: face detection, segmentation and ethnicity categorization. The final classified output facilitates to conclude the race of the individual.

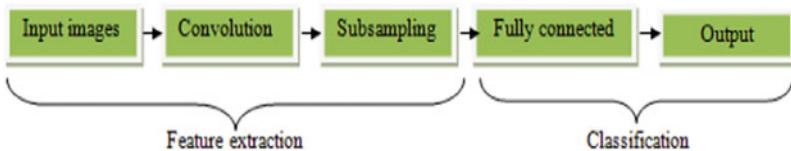
Wang et al. [6] proposed a solution for ethnicity recognition using deep convolution neural networks. In this model, the feature extraction and classification are done at the same time. This recognition model was analyzed in different scenarios such as the categorization of white and black, Chinese and Non-Chinese, and classification of Han, Uyghurs and Non-Chinese people. Experimental results verified the usefulness of the model.

Hui Lin et al. [7] have proposed recognition system that includes Gabor filtering, AdaBoost learning and SVM classifier for face-processing application. Kernel Class-dependent Feature Analysis method proposed by Xie et al. [8] handles the ethnicity recognition on huge database with face images. It deals with the periorbital areas rather than the whole face district and achieves better accuracy for Caucasian, Asian and African American ethnicity group. Liu et al. [9] have given a review on Cross-ethnicity Face Anti-spoofing Recognition Challenge based on the CASIA-SURF CeFA dataset and presented the overview of the challenge with the design methodology and experimentation analysis.

To sum up, despite the fact that these techniques have increased a high recognition rate for ethnicity groups, there is still a room for improvement. Therefore, this paper proposes deep learning-based model for ethnicity recognition by analyzing the facial features.

### 3 Proposed Methodology

CNN is presently considered as the preference of neural networks for classification of images since they detect patterns in tiny parts of an image, for instance, a small curve



**Fig. 1** Basic CNN architecture

of an eyebrow. CNN consists of convolution layer, subsampling and fully connected layers as revealed in Fig. 1. CNN understands any two-dimensional images as input, providing an exclusive gain for image recognition applications.

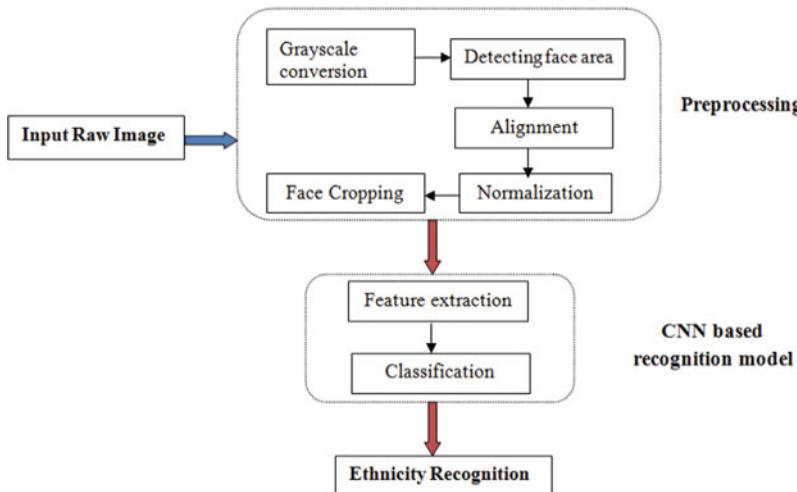
Figure 1 illustrates a basic CNN model where it receives 2D images as input and convolves with a number of modifiable convolution kernels to produce the subsequent feature map function. This function forms the Convolutional layer. Further, the mapping function is down-sampled to eliminate the size of the data and generates the Subsampling layer. In general, the pooling size of  $2 \times 2$  is preferred. The same process is continued in subsequent convolution and subsampling layers.

Following the extraction of adequate feature variables, the image pixels in two dimensions are regressed toward one-dimensional information and geared up for neural network classifier. In general, Softmax regression function is favored as the ultimate multiple classifier for practical applications. Later, the top feature map layer is separated into a huge number of narrow regions and convolved with intelligent kernels. Afterward, the convolutions are handled with activation functions and the novel attribute maps are generated.

The subsampling layer consists of two basic and commonly used methods such as mean-pooling and max-pooling. The final fully connected layer calculates the class scores on the whole image. The majority of the deep learning methods [10] have tailored CNN for face emotion identification. The CNN algorithm is composed of a number of basic essentials, which fulfills the purpose of facts entry, training and testing specifications. Finally by evaluating and comparing the experimental investigations of training records by the testing labels, the recognition accuracy is calculated. Figure 2 shows the proposed block diagram for CNN-based ethnicity recognition with facial features.

In the proposed work, we have considered *FairFace* dataset, a huge dataset which contains 85 K train images and 14 K test images. The different ethничal groups are Southeast Asian, East Asian, Black, White, Latino-Hispanic, Indian and Middle Eastern. To avoid confusion, both East Asian and Southeast Asian races are combined into a single Asian race. The input images are observed based on the file names and the image pixels are stored as a column. Each image has 3 channels and of the size  $224 \times 224 \times 3$ . The steps that have been followed in the proposed ethnicity recognition are: Preprocessing of the input images, detection of facial points from the given input images, extraction of facial features, classification and finally, recognition of features for ethnicity identification.

At first, the input face image is preprocessed with gray-scale conversion, face detection, face alignment, normalization and then cropping steps. After selecting



**Fig. 2** Proposed block diagram for CNN-based ethnicity recognition

the facial points from the given input image, the facial distances between head to chin, the width of head, eye centers, nose tip, and mouth corners are calculated for face alignment. Later, they are normalized and cropped to the size of  $64 \times 64$  by excluding the background. Now the CNN model which contains three convolutional layer and two fully connected layers and Softmax layer is trained with cropped image to envisage the ethnicity group.

## 4 Results and Discussion

In the proposed work, FairFace dataset and UTKFace dataset are used to effectively classify the individuals according to their ethnicity. FairFace is a large-scale dataset with 108,501 face images, in which 85 K images are considered for training and 15 K images for testing. In order to have fair race composition, there are seven ethnical groups such as Southeast Asian, East Asian, Black, White, Latino-Hispanic, Indian and Middle Eastern [11]. The dataset is tagged as race, gender, and age groups. UTKFace dataset is again a face dataset with age group from 1 to 116 years. It consists of 20,000 images with observations of gender, age and ethnicity.

In addition, real-time facial images of Indian ethnical group, captured by ourselves, are also used. Figure 3 shows the arbitrary samples taken from FairFace and UTKFace datasets.

The model is trained with the above-mentioned datasets. After preprocessing the input image, the CNN model is used to predicate the ethnicity class. The average accuracy achieved by the proposed method is 88%, and thus we declare the robustness of the proposed model

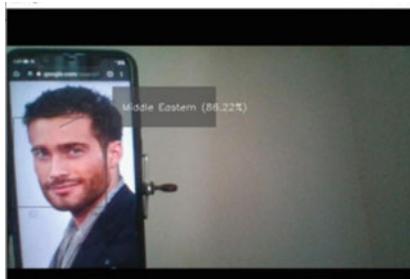


**Fig. 3** Face samples from **a** FairFace and **b** UTKFace datasets

The experiments are conducted on a laptop with i7 processor and NVIDIA GeForce GTX 1650 UHD graphics, and 32 GB memory. The model is implemented in Keras framework, an open-source neural network library written in Python and the proposed method takes around 3.12 ms to process one face image. Figure 4 shows sample screenshots of ethnicity recognition for Middle Eastern, Black, White, Asian ethnicity group and finally the real-time ethnicity recognition for Indian face.

## 5 Conclusion

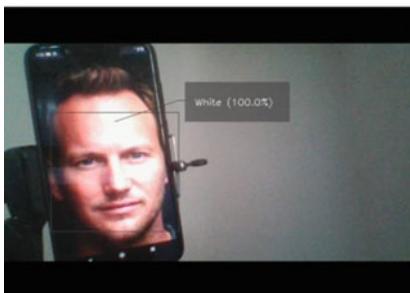
This paper mainly focuses on extracting the facial features for ethnicity recognition using CNN-based recognition method. The proposed model is experimented with different ethnical group dataset which consists of Southeast Asian, East Asian, Black, White, Latin, Indian and Middle Eastern. Also, we have tested our model with real-time face images to recognize the ethnicity group. The average accuracy achieved by the proposed method declares the robustness of the model. In future, this algorithm may be helpful for detecting the missing children, search investigations and refugee crisis.



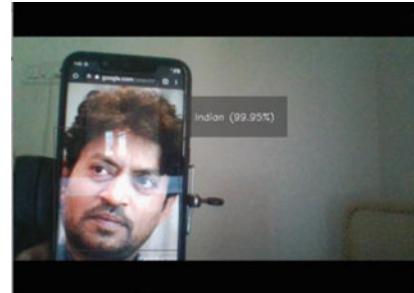
a) Middle eastern ethnicity



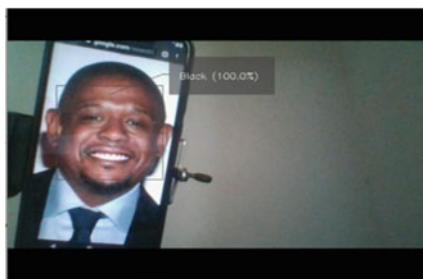
b) Asian ethnicity



c) White ethnicity



d) Indian ethnicity



e) Black ethnicity



f) Real time ethnicity recognition

**Fig. 4 a–f** Sample screenshots of ethnicity recognition for Middle Eastern, Black, White, Asian and the real-time ethnicity recognition for Indian face

## References

1. Chenlei, L., Zhongke, W., Dan, Z., Xingce, W., Mingquan, Z.: 3D Nose shape net for human gender and ethnicity classification. *Pattern Recogn. Lett.* **126**, 51–57 (2019)
2. Ding, H., Huang, D., Wang Y., Chen, L.: Facial ethnicity classification based on boosted local texture and shape descriptions. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Shanghai, pp. 1–6 (2013)
3. Wang, C., Zhang, Q., Liu, W., Liu, Y., Miao, L.: Facial feature discovery for ethnicity recognition. *WIREs Data Mining Knowl. Discov.* **9**, e1278, **10**(5), 1–17 (2019)
4. Gao, S., Zeng, C., Bai, M., Shu, K.: Facial ethnicity recognition based on transfer learning from deep convolutional networks. In: 2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), Shenzhen, China, 310–314 (2020)
5. Achkar, R., Haidar, G., Assal, M., Habchy, D., Ashi D., Maylaa, T.: Ethnicity recognition system using back propagation algorithm of an MLP. In: 2019 Fourth international conference on advances in computational tools for engineering applications (ACTEA), Beirut, Lebanon, 1–5 (2019)
6. Wang, W., Feixiang H., Zhao, O.: Facial ethnicity classification with deep convolutional neural networks, biometric recognition. *Biometr. Recogn.* 176–185 (2016)
7. Lin, H., Zhang, L.: A new automatic recognition system of gender, age and ethnicity. In: 6th World Congress on Intelligent Control and Automation, vol. 2, 9988–9991 (2006)
8. Xie, Y., Luu, K., Savvides, M.: A robust approach to facial ethnicity classification on large scale face databases. In: 2012 IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS), Arlington, VA, pp. 143–149 (2012)
9. Ajian L., Li, X., Wan, J., Escalera S.: Cross-ethnicity face anti-spoofing recognition challenge: a review. *IET Res. J.* 1–12 (2020)
10. Siyue, X., Haifeng, H., Yongbo, W.: Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition. *Pattern Recogn.* **92**, 177–191 (2019)
11. Kimmo, K., Jungseock, J.: FairFace: face attribute dataset for balanced race, gender, and age. *Comput. Vis. Pattern Recogn.* 1–11 (2019)

# Scanning Array Antenna Radiation Pattern Design Containing Asymmetric Null Steering Based on L-ASBO



Anitha Suresh, C. Puttamatappa, and Manoj Kumar Singh

**Abstract** The involvement of heuristic computational design for an antenna array in the field of electromagnetic application has improved the quality and cost of the solution significantly. In this paper, a learning form of adaptive social behavior optimization has been presented to obtain the radiation pattern of the scanning array antenna which carried the high gain for the main lobe and satisfied the constrained of null steering in the different directions by controlling the phases of current in the array elements. The existing form of ASBO does not use the relevant information of the solution which has been explored by the previous population in the first stage. In the proposed form of modification, the best-explored information in past has been utilized by the current population under a probabilistic environment. The proposed modification has improved the quality of the solution as well as the rate of convergence. The obtained performances of radiation patterns through the proposed method has been compared with results obtained through particle swarm optimization available in the literature.

## 1 Introduction

An antenna array is considered one of the finest ways to detect the signal originated from various directions. The purpose of array synthesis is to configure the array elements geometrically and electrically to obtain the desired radiation pattern. Many important characteristics can be associated with radiation patterns like reducing the level of sidelobe without compromise in the gain of the main lobe, nullify the effect of interference, and jamming appeared from a certain direction. There are appreciable benefits like coverage, signal quality, and enhancement of spectrum efficiency that can be observed with the use of antenna arrays in various applications as a requirement

---

A. Suresh · C. Puttamatappa

Department of Electronics and Communication, Dayananda Sagar University, Bangalore, India

M. K. Singh (✉)

Manuro Tech Research Pvt. Ltd, Bangalore 560097, India

e-mail: [mksingh@manuroresearch.com](mailto:mksingh@manuroresearch.com)

or cost-effectiveness or both by providing the steerable beams and maintaining higher directivity. It is well known that in any wireless communication system, the quality of system performance is heavily decided by the design quality of the antenna. For most of the long-distance coverage applications, the characteristics of radiation pattern need should carry high directivity and small beam width along with satisfying the constraints of low sidelobes and nulls have to appear in the interferential directions. The desired objectives of radiation patterns of an array antenna have been achieved by controlling the various structural and electrical parameters values. In a broad way, synthesis of desired array pattern can be considered as an optimization problem and various possibilities of structural and electrical optimization exist like geometric optimization approach in terms of defining the proper spacing between elements, electrical characteristics optimization approach by providing the proper amplitude and phase of the current in each element. Practically, the needed pattern characteristic depends upon applications. In many cases, there is a need for high gain in the main lobe and minimum level of side lobes while in some other cases there is a demand for interference reduction from unwanted directions by placing the nulls.

The uses of null steering to minimize the interferences are very common in the application areas of radar and sonar. In comparison with geometrical parameters, controlling electrical parameters like amplitude and phase of the current in the array elements are more effective in the placement of nulls. There is a possibility exist of controlling the amplitude and phase in combination or individually and have their advantages and limitations. Null steering problems can be solved more efficiently by considering the problem as a nonlinear optimization model consisting of the optimization parameters like excitation amplitudes, phases, and/or array element positions. The cost of the phase shifter and variable attenuators provides the constraint in the design; hence to make the solution cost-effective to control the nulls, the research community prefers to focus on either phase only or amplitude only. In this paper, null synthesis of the phased array antenna has been achieved by controlling phase one. There is considerable complexity that exists in the development of an optimal solution of having high directivity and multiple null steering with phase adjustment; however, the proposed solution L-ASBO has handled this challenge efficiently.

## 2 Literature Survey

The standard form of ASBO has been discussed in [1]. Based on PSO, the synthesis of various patterns has been presented in [2]. To obtain the desired radiation pattern through an array, antenna in the past several different approaches have been applied. The conformal phased array antenna radiation pattern has been analyzed in [3] using field vector synthesis. In [4], real-coded GA has been applied to optimize the geometrical (space between elements) as well as electrical parameters (weight of elements current amplitude). To diagnose the array elements of PAA, reflected signal from radiator-free space junction has been applied in [5]. RGA [6] has also been used to

obtain the optimal weight value of the array element to reduce the level of sidelobe for uniformly spaced array elements geometry. There were various ways that PSO has been used in [7, 8] to obtain the desired radiation pattern. For a large phased array, a fixed beamwidth has been maintained in [7] when the direction of the main lobe is away from the direction of the broadside. For different configurations of hexagonal array [8] in which each element is excited with the same current value, PSO has been applied to achieve a directive beam, and side lobe level also has minimized. Taguchi's method has been used in [9] to obtain the reduction in SLL for linear array antenna. For a linear array antenna, the null placement and SLL reduction using ant colony optimization have been discussed in [10]. Microstrip technology-based hardware implementation has been used in [11] to develop the array antenna which carried the benefit of low SLL value. The design of the circular array antenna using the firefly algorithm [12] and using symbiotic organisms search (SOS) algorithm [13] has been discussed. A deterministic approach has been applied in [14] to design an a-periodic antenna array.

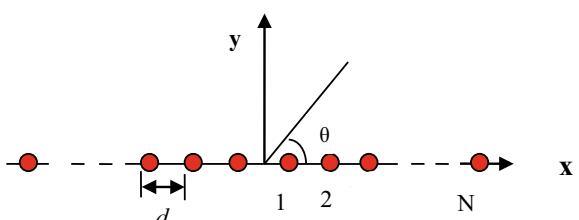
### 3 Linear Array Antenna

In this work, a linear array antenna has been considered a set of  $2N$  elements where each element is represented as an infinitesimal dipole. All elements have been arranged along the X-axis and maintained the uniform distance from their neighbor elements on either side as shown in Fig. 1. The total far-field of such array antenna is defined as the multiplicative result of the field generated by one element positioned at the center and the array factor.

Assuming any two consecutive elements separated by distance is ' $d$ ' and the center position of the considered linear array formation as a reference point, the array factor (AF) in normalized form for such a linear array can be obtained by Eq. (1) and assumption has made about constant amplitude value and odd symmetry of phase shifts.

$$AF_{\text{norm}}(\theta) = \frac{1}{N} \sum_{n=1}^N \cos[(n - 0.5)\psi + \beta_n] \quad (1)$$

**Fig. 1** Geometry structure of linear array antenna with  $2N$  elements



where  $\psi = \frac{2\pi}{\lambda} d \sin(\theta) = kd \sin(\theta)$ ,  $N$  is the total number of array element one side of reference point;  $\beta_n$  is the ‘ $n$ th’ array element phase shift weight, and  $\theta$  is the desired direction angle.

To obtain the desired radiation pattern, the objective function has been developed by combining the different normalized array factors in an additive or subtractive manner depends upon their needs. The objective function has to optimize by obtaining the optimal values of array antenna parameters.

In this work, considering the phase of elements excitation as variable parameters, the number of asymmetrical nulls and placement of the major lobe in any direction have obtained by a modified form of meta-heuristic ASBO called learning ASBO (L-ASBO) algorithm.

## 4 Objective Function Definition

In this work, objectives of maximize the gain over the desired direction and minimize the level of interference over specified directions have been considered simultaneously and formulated as a maximization problem as given in Eq. 6.

$$F = |AF_d|^2 - \sum_{i=1}^p |AF_{N_i}|^2 \times \sum_{i=1}^p w_i \times pv \quad (2)$$

where first term  $AF_d$  is the array factor in the direction of the major lobe while the second term carried the sum of all array factors where nulls have to place. Weighted penalty ( $pv$ ) has been assigned according to the difference between the obtained and expected level of nulls ( $w_i$ ).

## 5 Proposed Method

Based on the social evolutionary process, in [1] a heuristic algorithm called adaptive social behavior optimization (ASBO) has been proposed. It has been observed that there is faster change occurred in the social entities through social evolution instead of genetic evolution; hence, same can help in designing the more powerful mathematical model for the optimization algorithms. Considering the abstract model of society where a number of inspiring factors like the leader, neighbors’ and self-motivation control the change of an individual in a dynamic manner, ASBO has also taken the same factors in the improvement of individual fitness. To capture the dynamic nature of the influence of influencing parameters, in ASBO there are two populations exist, one exists as the solution population while the other ones carried the influence depth. The benefit of phenotype representation of solution provides the easing in the implementation. The criteria to be the neighbors of an individual are based upon

the next higher fitness while the present population best has considered as leader-member. The best performance obtained in the past by an individual becomes a self-motivation factor for that individual in the current iteration.

Equations (3) and (4) define the change in  $i$ th dimension of the individual solution where  $C_g$ ,  $C_s$ ,  $C_n$  are positive constant.  $R_g$ ,  $R_s$ ,  $R_n$  are the random numbers generated through  $U[0 \ 1]$ .  $G_b$ ,  $S_b$ , and  $N_c$  are the positions of the leader, self-best, and neighbor center.

$$\begin{aligned}\Delta X(i+1) = [ & C_g R_g (G_{bi} - X_i) \\ & + C_s R_s (S_{bi} - X_i) \\ & + C_n R_n (N_{ci} - X_i) ]\end{aligned}\quad (3)$$

$$X(i+1) = X_i + \Delta X(i+1) \quad (4)$$

The dynamic behavior of influencing factors have achieved through a self-adaptive mutation strategy of progress constant values as given by Eqs. 5 and 6

$$p'_i(j) = p_i(j) + \sigma_i(j) N(0, 1) \quad (5)$$

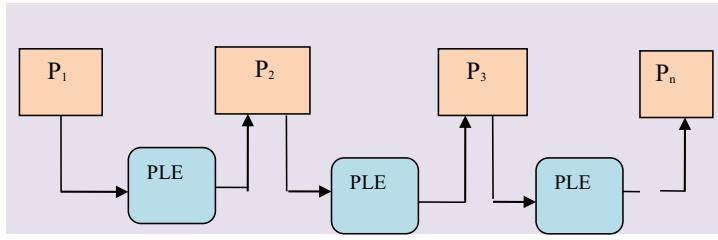
$$\begin{aligned}\sigma'_i(j) = \sigma_i(j) e^{\left(\tau' N(0,1) + \tau N_j(0,1)\right)}, \\ \forall j \in \{1, 2, 3\}\end{aligned}\quad (6)$$

where  $p_i(j)$ ,  $p'_i(j)$ ,  $\sigma_i(j)$ ,  $\sigma'_i(j)$  denote the  $j$ th component of the vectors  $p_i$ ,  $p'_i$ ,  $\sigma_i$ ,  $\sigma'_i$ , respectively, and  $N(0,1)$  a is Gaussian distributed random number and will be same for a solution while  $N_j(0,1)$  is a new random number generated for  $j$ <sup>th</sup> attribute under a solution using Gaussian distribution.  $\tau'$  and  $\tau$  are problem dimension depend constants.

Mimicking the development behavior of human social development, ASBO has evolved the solution under two stages. In the 1<sup>st</sup> stage, the number of independent random populations evolved, and in the 2<sup>nd</sup> stage, a new population formed by some of the fittest members available in the different populations at the end of the 1<sup>st</sup> stage.

## 5.1 L-ASBO: Learning ASBO

In the existing form of ASBO, in the first stage, each population evolved independently and later they combined to form a pool from where the fittest members were selected to form the population for the second stage. The independence of each population can cause to reduce the speed of convergence as well as may cause the end of the result which can improve further. Under the 1<sup>st</sup> stage, when the first population has completed its evolution, it has a certain level of exploration of the solution domain,



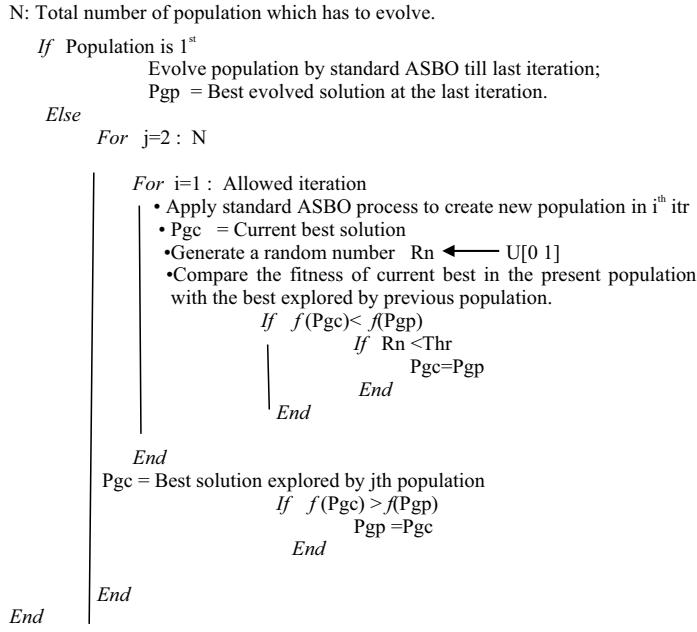
**Fig. 2** Process flow in L-ASBO

while when the second population started the exploration, previously explored information by the first population does not use. Similarly, the third population does not use the information explored by the 1st and 2nd population. This process continued for all the population. In other words, issue with the existing model of ASBO 1<sup>st</sup> stage is not having any inspirational factor to improve the quality of search from previously completed evolution of the population under different trials.

There may be very good possibilities that the previous population explored information can help to guide the next population, particularly when the current best solution has lesser fitness in comparison with the past population explored solution fitness. But if considered blindly, the past explored solution by the previous solution can behave like a strong attractor particularly in the beginning which may cause loss of diversity immediately. Hence, a probabilistic model has been applied in which rather than continuous interference, an occasional interference of the previous population explored information has been utilized by the current population. The process has shown in Fig. 2 while algorithmic steps have shown in Fig. 3. In Fig. 2,  $P_i$  is the populations while PLE is the probabilistic learning environment as given in Fig. 3.

## 6 Experimental Results

To synthesizes the different patterns and understanding the quality behavior of the proposed algorithm L-ASBO, in the simulation three different cases of desired pattern (PS1, PS2, and PS3) synthesis have been considered as given in Table 1. Pattern PS1 has one null while pattern PS2 has two nulls on the same side of the major lobe while nulls for pattern PS3 have distributed on either side of the major lobe. As desired pattern characteristics have variability in their complexity, there were a different number of array elements have considered and were 10, 20, and 50 for PS1, PS2, and PS3 correspondingly. Each element assigns a phase value; hence, the dimension of the problem varied accordingly. To obtain the optimal phase of elements in all the three different cases, L-ASBO has uses the population size as 100. There was 10 independent population that have been considered in the 1st phase of L-ASBO, and each population has evolved up to the 100 iterations. For the 2nd phase, 100 best members have been selected from the different populations evolved in the 1st stage

**Fig. 3** Algorithmic steps for proposed L-ASBO**Table 1** Desired pattern characteristics

Pattern synthesis	Desired direction	Null direction	Elements number
PS1	-60°	30°	20
PS2	-60°	[-20°, 40°]	40
PS3	-20°	[-40°, 1°, 30°, 60°]	100

and allowed to evolve for the 1000 iterations. The search region of solution bounded in  $[-\pi, \pi]$  radians and any violation forced to randomly bound within limit range. The fitness value has estimated by Eq. 2 where the penalty factor has considered as a high value of 1000. The whole simulation setup has been developed in the MATLAB environment. To observe the variability in performances with different trials, 20 independent trials have given over each problem, and obtained performances have shown in Tables 2, 3, and 4. The statistical performances over 20 trials have given in terms of mean and standard deviation along with confidence intervals. It is clear that for all the cases L-ASBO performed very well and always more than -50 dB gain has achieved for the position of null. For the pattern problem PS2, the obtained convergence characteristic has shown in Fig. 4. The desired pattern PS3 is very

**Table 2** Obtained pattern characteristics by L-ASBO over PS1

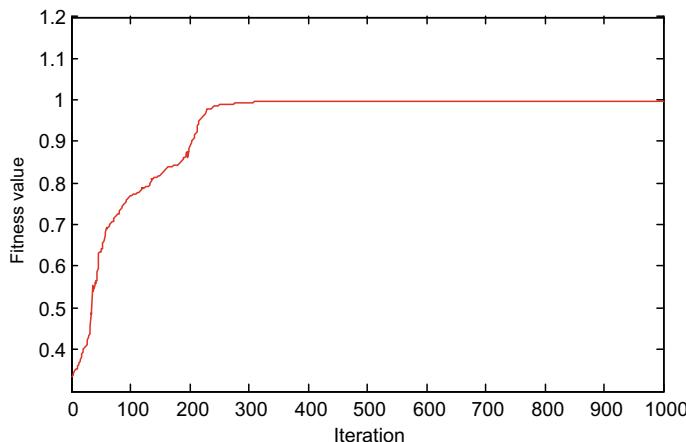
	-60°	30°
Mean (dB)	-0.0210	-53. 22
Std.Dev (dB)	0.0000	0.0000
CI (95%) (dB)	$\mp 0.0002 * 10^{-6}$	$\mp 0.5637 * 10^{-6}$
CI (99%) (dB)	$\mp 0.0002 * 10^{-6}$	$\mp 0.7437 * 10^{-6}$

**Table 3** Obtained pattern characteristics by L-ASBO over PS2

	-60°	-20°	40°
Mean (dB)	-0.0082	-58. 83	-77. 57
Std. Dev (dB)	0.0001	0.0358	0.0761
CI (95%) (dB)	$\mp 0.0000$	$\mp 0.0143$	$\mp 0.0305$
CI (99%) (dB)	$\mp 0.0000$	$\mp 0.0201$	$\mp 0.0404$

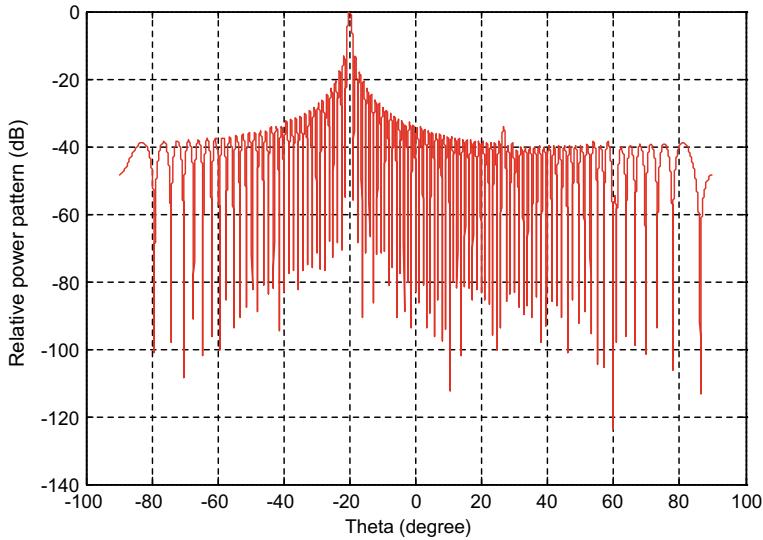
**Table 4** Obtained pattern characteristics by L-ASBO over PS3

	-20°	-40°	1°	30°	60°
Mean (dB)	-0.0007	-81.878	-82.024	-68.451	-59.23
Std.Dev (dB)	0.0000	1.183	0.1454	0.566	0.0723
CI (95%) (dB)	$\mp 0.0000$	$\mp 0.7209$	$\mp 0.0849$	$\mp 0.3281$	$\mp 0.0471$
CI (99%)(dB)	$\mp 0.0000$	$\mp 0.9203$	$\mp 0.1202$	$\mp 0.4642$	$\mp 0.0624$



**Fig. 4** L-ASBO convergence in a trial over pattern synthesis of PS2

difficult to achieve but it can observe from Table 4 that L-ASBO performed very well and the obtained radiation pattern has shown in Fig. 5. In [2] over the pattern, PS1



**Fig. 5** Radiation pattern for pattern synthesis PS3 by L-ASBO

**Table 5** Performance comparison for PS1

PS1	-60°	30°
VAT [2]	-0.0347 dB	-57.1472 dB
L-ASBO	-0.0223 dB	-51.7636 dB

**Table 6** Performance comparison for PS2

PS2	-60°	-20°	40°
VAT [2] (dB)	-0.4975	-58.0635	-70.5191
L-ASBO (dB)	-0.0093	-54.0090	-80.0149

and PS2 in PSO have been applied and comparative performances against L-ASBO have shown in Table 5 and Table 6 correspondingly.

It was observed that over simple pattern synthesis PS1, PSO performance is somehow acceptable but as the complexity increased in PS2, performances degraded. Unlike PSO, L-ASBO has shown consistency in performances over different problem complexity.

## 7 Conclusion

Based on the previous exploration by the previous population, a modified version of ASBO has been presented. The past explored information learned by the current

population in the probabilistic environment. By controlling the phase of array elements, asymmetrical patterns containing various nulls have been synthesized. The proposed solution has delivered a better and efficient radiation pattern, what it was obtained with particle swarm optimization. The proposed method of solution is having the capability to deliver a global solution with faster convergence.

## Reference

1. Singh, M.K.: A new optimization method based on adaptive social behavior: ASBO. *AISC* 174, pp. 823–831. Springer (2012). [https://doi.org/10.1007/978-81-322-0740-5\\_98](https://doi.org/10.1007/978-81-322-0740-5_98)
2. Zuniga, V., Erdogan, A.T., Arslan, T.: Adaptive radiation pattern optimization for antenna arrays by phase perturbations using particle swarm optimization. In: NASA/ESA Conference on Adaptive Hardware and Systems 2010
3. Zeng, G., Li, S., Wei, Z.: Research on conformal phased array antenna pattern synthesis. In: Proceedings of the 2012 International Conference on Information Technology and Software Engineering, pp. 13–21
4. Goswami, B., Mandal, D.: A genetic algorithm for the level control of nulls and side lobes in linear antenna arrays. *J. King Saud Univ. Comput. Inf. Sci.* **25**(2), 117–126 (2013)
5. Manichev, A.O., Balagurovskii, V.A.: Methods for diagnostics of an array antenna by the signals reflected from radiator-free space junctions in the presence of mutual coupling of array elements. *J. Commun. Technol. Electron.* **58**(4), 307–317 (2013)
6. Zhang Z., Li T., Yuan F., Yin L. (2014) Synthesis of Linear Antenna Array Using Genetic Algorithm to Control Side Lobe Level. In: Wong W.E., Zhu T. (eds) Computer Engineering and Networking. Lecture Notes in Electrical Engineering, vol 277. Springer, Cham
7. Yang, S.-H., Kiang, J.-F.: Adjustment of beam width and side-lobe level of large phased-arrays using particle swarm optimization technique. *IEEE Trans Antennas Propag.* **62**(1), (2014)
8. Bera, R., Mandal, D., Kar, R., Ghoshal, S.P.: Application of particle swarm optimization technique in hexagonal and concentric hexagonal antenna array for side lobe level reduction. In: Mandal, D., Kar, R., Das, S., Panigrahi, B. (eds.) Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 343. Springer (2015)
9. Kumar, R., Mohanty, S.K., Mangaraj, B.B.: Side lobe level and null control design of linear antenna array using Taguchi's method. In: Jain, L., Patnaik, S., Ichalkarane, N. (eds.) Intelligent Computing, Communication and Devices. Advances in Intelligent Systems and Computing, vol 309. Springer (2015)
10. Saxena, P., Kothari, A.: Ant lion optimization algorithm to control side lobe level and null depths in linear antenna arrays. *AEU Int. J. Electron. Commun.* **70**(9), 1339–1349 (2016)
11. Hussein, A.H., Metaw'e, M.A., Abdullah, H.H.: Hardware implementation of antenna array system for maximum SLL reduction. *Eng. Sci. Technol. Int. J.* **20**(3), 965–972 (2017)
12. Singh, G.N., Rattan, M., Patterh, M.S.: A non-uniform circular antenna array failure correction using firefly algorithm. *Wirel. Pers. Commun.* (1), (2017)
13. Dib, N.: Design of planar concentric circular antenna arrays with reduced side lobe level using symbiotic organisms search. *Neural Comput. Appl.* **30**(12), 3859–3868 (2018)
14. Caratelli, D., Toso, G.: Deterministic constrained synthesis technique for conformal aperiodic linear antenna arrays—Part I: theory. *IEEE Trans. Antennas Propag.* **67**(9), 5951–5961 (2019)

# A Modified Novel Signal Flow Graph and Memory-Based Radix-8 FFT Processor Design



A. Anitha, B. Triveni, Pinninti Kishore, and Makkenna Madhavi Latha

**Abstract** The Fourier Transform is a noteworthy calculation in digital signal processing region. An enhanced signal flow graph is required for FFT processor structures, whose sources of information and outputs are in normal order. In this paper, a memory-based conflict-free FFT processor design by utilizing a modified radix-8 DIF (Decimation-In-Frequency) signal flow graph is proposed. The number of iterations is decreased, and therefore, the execution time is low when compared with memory-based radix-4 FFT. The architecture is implemented on Xilinx Virtex-6 FPGA platform. The proposed FFT processor with radix- 8 has reported lower processing time when compared with radix-4 FFT processor.

## 1 Introduction

An important algorithm within the area of signal and image processing is Fast Fourier Transform (FFT). Two categories of architectures are available to implement FFT processors. The first one uses pipelined architecture [1–4] with independent butterfly unit and memory elements in each section and, the later one is the memory-based architecture [5–9], but this will operate with identical memory and butterfly unit in each section. Unique memory-based radix-2 DIF FFT architecture [7] reported reduced computation complexity with respect to number of clock cycles and improved utilization of the Processing Element (PE) for a Real Fast Fourier Transform (RFFT) as results of bit reversal order of the output. Architecture for a 4096-point radix-4 memory-based FFT using DSP slices [10] is aimed to improve the utilization of DSP Slices and reduce the utilization of distributed logic with a

---

A. Anitha (✉) · B. Triveni  
V. R. Siddhartha Engineering College, Vijayawada, India

P. Kishore  
V. N. R. Vignan Jyothi Institute of Engineering and Technology, Hyderabad, India

M. Madhavi Latha  
J. N. T. University, Hyderabad, India

conflict-free strategy. A low power Split Radix FFT(SRFFT) [8] processor using radix-2 Butterfly units is introduced which reported 20%, lower power consumption compared to other FFT algorithm with address conflicts and output is in bit reversal order as a demerit. An efficient memory organization for memory-based FFT architectures [11, 12] is demonstrated using an 8192-point radix-2 Cached Memory (CM) FFT processor. CM can lead faster operation and low dissipation, which consumes 10.1–29% less area and 9.6–67.9% less power.

Majority of the literature concentrated on radix-2 and radix-4 memory-based FFT architectures. This work considers a memory-based architecture to implement a FFT processor using a modified radix-8 signal graph. A memory conflict strategy that aligns with the proposed flow graph is employed. The rest of the document is organized as follows. In Sect. 2 the modified radix-4 graph of signal flow is discussed. Section 3 discusses the proposed modified Signal Flow Graph (SFG) and memory based radix-8 FFT processor design. Finally in Sect. 4, the results are presented and followed by Sect. 5 concludes the discussion.

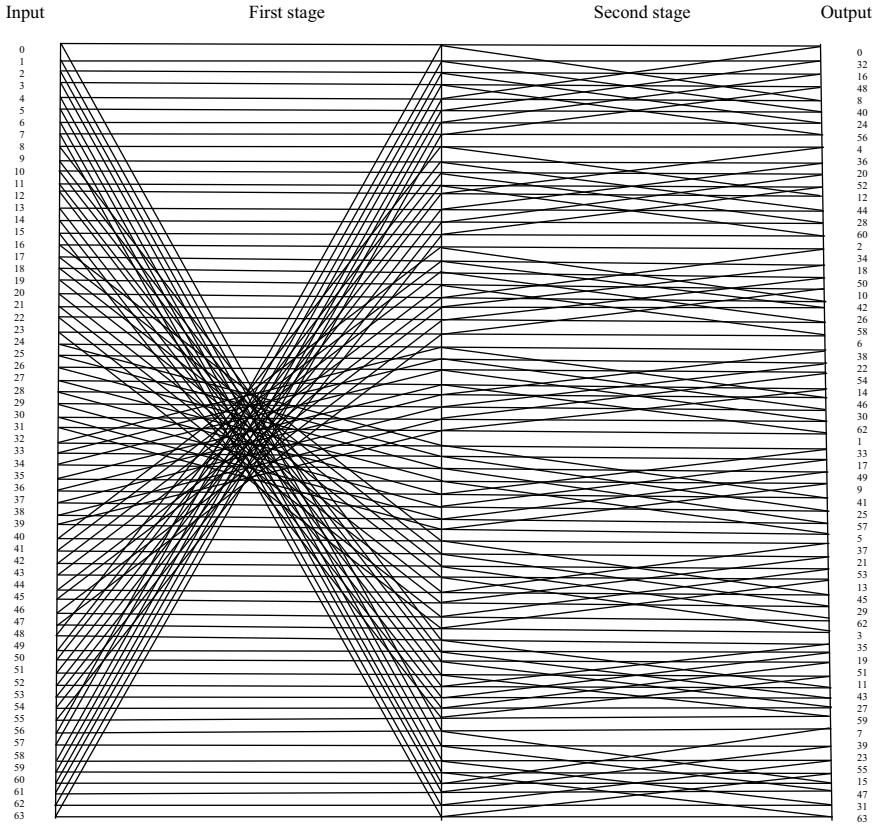
## 2 Modified Signal Flow Graph of Radix 8 FFT

Figure 1 shows conventional SFG for a 64-point radix-8 DIF FFT. The input and output samples are in regular and bit reversal order respectively. Therefore, in order to achieve a regular output sample order, a bit reversal operation must be performed which adds to additional computation time.

A modified SFG with input and output samples in regular order for a 64-point radix-8 FFT is presented in Fig. 2. To achieve output in regular order, the output node and twiddle factors are adjusted accordingly. The butterfly operation performed in both conventional SFG and modified radix-8 SFG as shown in Fig. 3a. The corresponding architectural design of radix-8 computation unit is shown in Fig. 3b. Position of the input of the butterfly unit in each stage of the modified SFG is identical to the conventional one. The conventional one is different from the position of the result of the butterfly unit in modified SFG. The modified SFG results are stored in the consecutive locations of  $n$ ,  $n + (N/8)$ ,  $n + (2*N/8)$ ,  $n + (3*N/8)$ ,  $n + (4*N/8)$ ,  $n + (5*N/8)$ ,  $n + (6*N/8)$ ,  $n + (7*N/8)$  where ‘ $N$ ’ represents the number of points of FFT ‘ $n$ ’ represents the butterfly unit executed.

## 3 Memory-Based FFT Architecture

Single butterfly and memory units are used in each stage of memory-based FFT architecture. Hence, it utilizes less area and power consumption comparative to other conventional architectures. The FFT architecture, shown in Fig. 4, contains the radix-8 butterfly computation unit, two number of single port memories each consisting of 8 banks, interface multiplexer network, data multiplexer networks, ROM used for

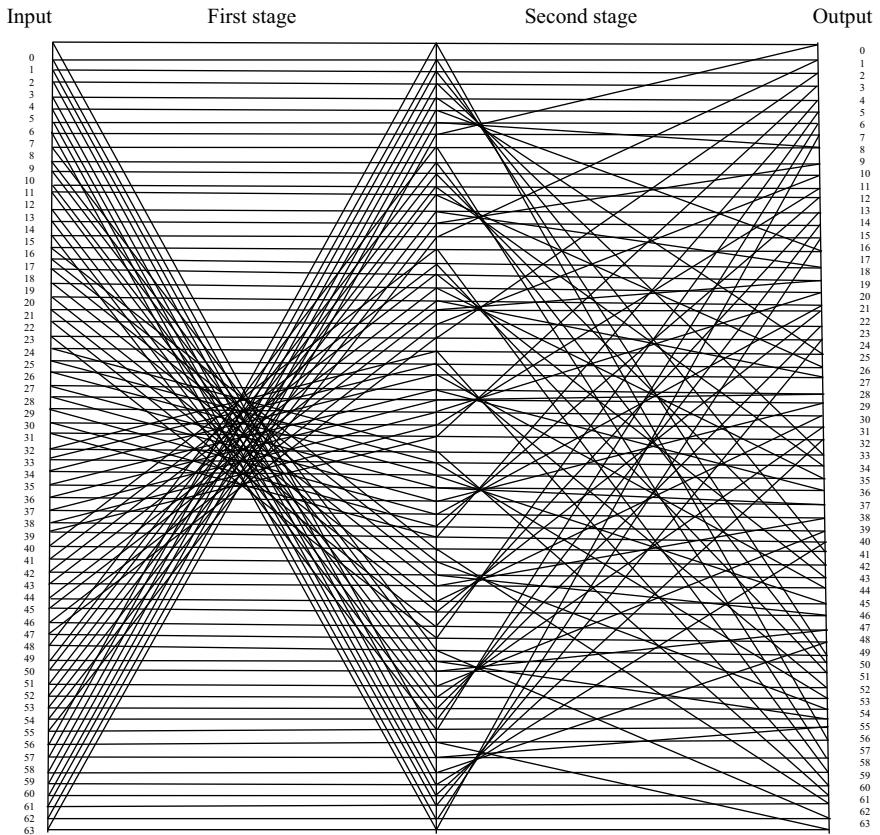


**Fig. 1** Conventional radix-8 signal flow graph for a 64-point FFT

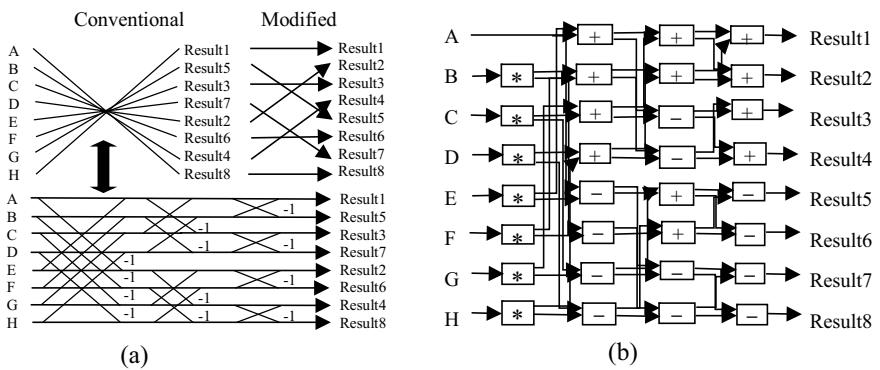
storing the twiddle factors and two control units used to control different states of memory banks and each module in FFT processor respectively. To store the input data and/or intermediate/final results of the butterfly unit will be done the memories. These memories work in ping-pang mode among each stage and to avoid the address conflicts in memories three multiplexers are used.

### 3.1 Conflict-Free Address Strategy

Each memory is organized as number of banks depending on the radix. Data is stored in the memory within the specified addresses corresponding to the modified SFG. The bit reversal operation is not necessary as results are directly obtained in the stage 2. To have efficient memory organization for input and output in regular order, conflict-free addressing strategy should be used.

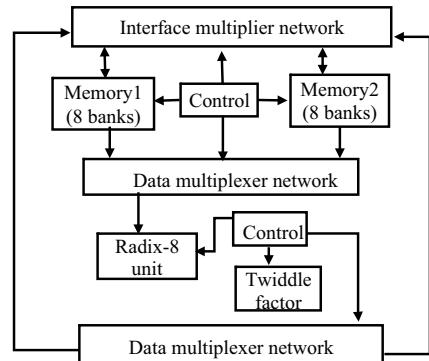


**Fig. 2** Modified radix-8 signal flow graph for a 64 point FFT



**Fig. 3** **a** Radix-8 butterfly computation, **b** architecture of radix-8 computation unit

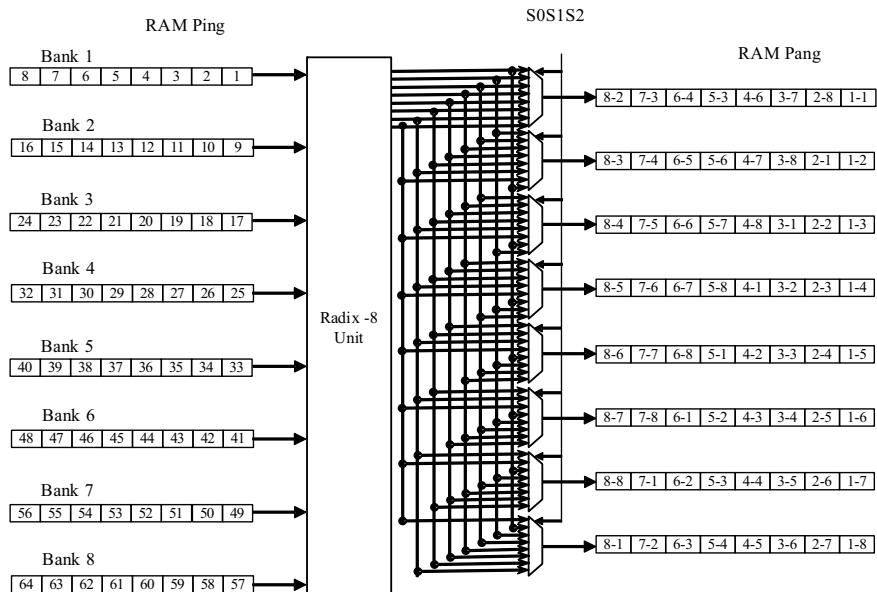
**Fig. 4** Architecture of the memory based FFT



### 3.1.1 The First Stage of FFT

Stage I of 64-point radix -8 FFT is shown in Fig. 5, where, the input samples are stored in sequential order in eight banks of RAM Ping. The data in each bank is stored in incremental addresses of banks. For example, data 13 is stored in bank2 at address 5, data 26 is stored in bank4 at address 2, and so on. After the computation, the corresponding output results are generated in RAM Pang.

In order to avoid memory conflict, Table 1 represents the pattern for storing the results of butterfly unit in every stage of 64-point FFT processor. First column of



**Fig. 5** Stage I of 64-point radix-8 FFT

**Table 1** Conflict-free memory addressing strategy for radix-8 FFT

Address (cycles)	Selection	Bank 1	Bank 2	Bank 3	Bank 4	Bank 5	Bank 6	Bank 7	Bank 8
000 (1)	0	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8
001 (2)	1	2_8	2_1	2_2	2_3	2_4	2_5	2_6	2_7
001 (3)	2	3_7	3_8	3_1	3_2	3_3	3_4	3_5	3_6
011 (4)	3	4_6	4_7	4_8	4_1	4_2	4_3	4_4	4_5
100 (5)	4	5_5	5_6	5_7	5_8	5_1	5_2	5_3	5_4
101 (6)	5	6_4	6_5	6_6	6_7	6_8	6_1	6_2	6_3
110 (7)	6	7_3	7_4	7_5	7_6	7_7	7_8	7_1	7_2
111 (8)	7	8_2	8_3	8_4	8_5	8_6	8_7	8_8	8_1

Table 1 indicates the address of the butterfly computation, second column indicate the address selection and remaining columns expressed as  $x\_y$ , indicate the results of radix-8 butterfly computations.  $x$  represents which iteration of butterfly computation and  $y$  represents which results of the butterfly computation.

In Table 1, the eight-butterfly computation results are 1\_1, 1\_2, 1\_3, 1\_4, 1\_5, 1\_6, 1\_7, and 1\_8 are produced during the first cycle, and the initial result 1\_1 from the first butterfly computation are stored in bank 1 first address. The multiplexer selects one of the eight results from the radix-8 butterfly unit.

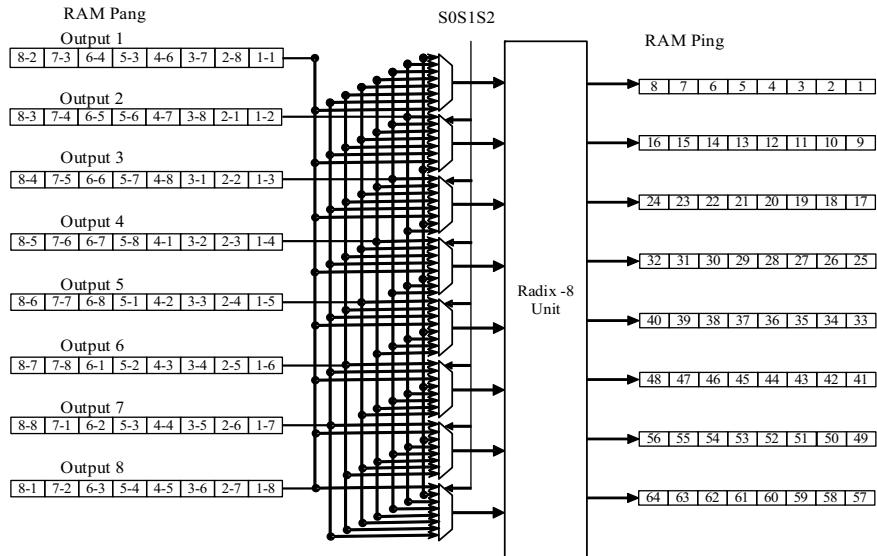
### 3.1.2 Second Stage of FFT

The last/second stage of 64-point FFT processor is shown in Fig. 6. In this stage, the output results are in the natural order because of the modified SFG. The computation time for this processor is reduced and iteration time decreased.

## 4 Results and Comparison

A 4096-point radix-8 FFT is designed to verify the performance, and resource utilization of the proposed architecture. The design is simulated using Verilog HDL for Virtex 6 XC6VCX75T FPGA device. The theoretical comparison of hardware resources and processing time is presented in Table 2.

It can be noticed that resources have increased with increase in radix while the total processing time has decreased by a factor of 3. Table 3 shows the timing report of radix-8 FFT and existed method, the computation time required for the radix-8 FFT is 14.020 us, and existed method is 73.519 us. The total processing time of proposed method is less compared to existed method. The macro utilization is summarized in Table 4. It is observed that macro utilization is higher for radix-8 due to increase in logic operations per iteration. Similar increase is reported in device utilization



**Fig. 6** Stage II of 64-point radix-8 FFT

**Table 2** Performance estimation of radix-4 and radix-8 processor

	Radix-8(proposed)	Radix-4 [13]
Parallel process	Single	Single
Memory size	$2 N$	$2 N$
Memory bank	8	4
Iteration	$(\log_8^N)$	$(\log_2^N)/2 - 1$
Cycles per iteration	$N/8$	$N/4$
Total processing time	$N(\log_8^N)/8$	$N(\log_2^N)/8 - N/4$
Complex adders	24	8
Complex multiplier	7	3

**Table 3** Performance estimation of radix-8 and radix-4 processor

Memory-based FFT Processor	Radix-8 (proposed)	Radix-4 [13]
Delay per cycle	6.846 ns	11.966 ns
Total computation time	14.020 us	73.519 us

of radix 8 FFT memory-based processor over radix-4 FFT processor. The device utilization summary for Virtex 6 XC6VCX75T FPGA device is reported in Table. 5.

**Table 4** Macro utilization summary

	Radix-8 (proposed)	Radix-4
RAM'S	2	4
Multipliers	22	12
Adders/Subtractors	93	40
Registers	162	99
Multiplexers	65	4

**Table 5** Device utilization

Logic utilization	Radix-8 (proposed)	Radix-4
No. of registers	1330	556
No. of LUT's	1279	1795
No. of fully used LUT'S-FF pair	537	295
No. of bonded IOB's	152	65
No. of BUFGS	1330	556

## 5 Conclusion

A novel modified signal flow graph and memory-based Radix-8 FFT processor design has been presented in this work, which implemented with 4096-point. The experimental results show the proposed design achieves reduced iterations and lower computation time for realization of FFT processor that is 80.7% less compared to existing processor design. Our proposed method also achieves lower address conflicts but more hardware utilization. The design is mostly used in communications and signal processing for lower computation time processors.

## References

1. Garrido, M., Huang, S.J., Chen, S.G., Gustafsson, O.: The Serial Commutator FFT. *IEEE Trans. Circ. Syst. II Exp. Briefs* **63**(10), 974–978 (2016)
2. Yin, X.B., Yu, F., Ma, Z.G.: Resource-efficient pipelined architectures for radix-2 real-valued FFT with real data paths. *IEEE Trans. Circ. Syst. II Exp. Briefs* **63**(8), 803–807 (2016)
3. Liu, J.Q., Xing, Q.J., Yin, X.B., Mao, X.B., Yu, F.: Pipelined architecture for a radix-2 fast Walsh-Hadamard-Fourier transform algorithm. *IEEE Trans. Circ. Syst. II Exp. Briefs* **62**(11), 1083–1087 (2015)
4. Hazarika, J., et al.: Energy efficient VLSI architecture of real-valued serial pipelined FFT. *IET Comput. Digital Tech* **13**(6), 461–469 (2019)
5. Xing, Q.J., Ma, Z.G., Xu, Y.K.: A novel conflict-free parallel memory access scheme for FFT processors. *IEEE Trans. Circ. Syst. II Exp. Briefs* **64**(11), 1347–1351 (2017)
6. Mao, X.B., Ma, Z.G., Yu, F., Xing, Q.J.: A continuous-flow memory-based architecture for real-valued FFT. *IEEE Trans. Circ. Syst. II Exp. Briefs* **64**(11), 1352–1356 (2017)
7. Ma, Z.G., Yin, X.B., Yu, F.: A novel memory-based FFT architecture for real-valued signals based on a radix-2 decimation-in-frequency algorithm. *IEEE Trans. Circ. Syst. II Exp. Briefs* **62**(9), 876–880 (2015)

8. Qian, Z., Margala, M.: Low-power split-radix FFT processors using radix-2 Butterfly units. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **24**(9), 3008–3012 (2016)
9. Tian, Y.H., Hei, Y., Liu, Z.Z., Di, Z.X., Shen, Q., Yu, Z.H.: A memory-based FFT processor using modified signal flow graph with novel conflict-free address schemes. *IEICE Electron. Express* **14**(15), 1–11 (2017)
10. Garrido, M., Sanchez, M.A., Vallejo, M.L.L., Grajal, J.: A 4096-point radix-4 memory-based FFT using DSP slices. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **25**(1), 375–379 (2017)
11. Xia, K.F., Wu, B., Xiong, T., Ye, T.C.: A memory-based FFT processor design with generalized efficient conflict-free address schemes. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **25**(6), 1919–1929 (2017)
12. Luo, H.F., Liu, Y.J., Shieh, M.D.: Efficient memory-addressing algorithms for FFT processor design. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **23**(10), 2162–2172, 2015
13. Tian, Y., Hei, Y., Liu, Z., Shen, Q., Di, Z., Chen, T.: A modified signal flow graph and corresponding conflict-free strategy for memory-based FFT processor design. *IEEE Trans. Circ. Syst. II Express Briefs* **66**(1), (2019)

# Vouch augmented Program Courses Recommendation System for E-Learning



**K. B. V. Rama Narasimham, C. V. P. R. Prasad, J. Jyothirmai,  
and M. Raghava**

**Abstract** In the realm of remote working and online living, it is imperative for educational institutions to steer their teaching learning process toward virtual classrooms and online evaluations. However, the existing learning management systems offer online video conference tools and facilitate the stakeholders with recommendations based on hard coded rules. As a result, the participating students also are failing to consolidate the theoretical concepts and started shying away from the critical core courses which are highly an unwanted situation for the next generation knowledge building. This paper aspires to propose developing a cognizant system which continually collects the multi-modal data from diversified sources, integrate them with the emotions of the students, intuitions of the torch bearers of various fields and evolve authentic recommendations to the students. The proposed vouch augmented learning management system, which we refer as vLMS, offers a framework that does a deep poll in the background on various data sources, derives the semantic relations into cognizance and offers hard as well as soft recommendations for a student to rediscover himself. Finally, the article presents the detailed architecture, the suitable soft-computing models and the technology stack support for implementing the vLMS framework.

---

K. B. V. Rama Narasimham  
Career Point University, Aalniya, India

C. V. P. R. Prasad  
Malla Reddy Engineering College for Women, Hyderabad, India

J. Jyothirmai  
VNRVJET, Secunderabad, India  
e-mail: [jyothirmai\\_j@vnrvjet.in](mailto:jyothirmai_j@vnrvjet.in)

M. Raghava (✉)  
CVR College of Engineering, Hyderabad, India  
e-mail: [raghava.m@cvr.ac.in](mailto:raghava.m@cvr.ac.in)

## 1 Introduction

A sprawling undergraduate college campus offers well-crafted curriculum, necessary classroom and computing infrastructure and recreation facilities for students. Students from various backgrounds, learning habits, career ambitions, once poured into new academic environment, tend to take time to adapt to the learning environment and better utilize the facilities. The parents also show lot of concern on performance of their wards in the competitive environment. However, neither the students nor the parents share their ambitions and concerns with the college due to the absence of single point of contact. Many a time, these interactions could not extract the inhibitions of the students and anxiety of the parents leading to administrative vagaries. On the other hand, the electronic learning, E-Learning [1], conditioned by Internet, is aspired to unleash the education systems in delivering the content to the remote learners in virtual mode [2]. The facilitator presents the contents or demonstrates the program models either synchronously or asynchronously [3] through networks and smart devices. This paradigm shift has prompted academicians to evolve novel methodologies in content dissemination and sharing in various formats. However, the students community with varied societal positions and skills set tend to take time to embrace the emerging teaching-learning models. queryPlease check the clarity of the sentence 'The latest Web 5.0 [4] standard which can cater to emotions of the...'. The latest Web 5.0 [4] standard which can cater to emotions of the students involved with E-Learning through ultra-intelligent electronic agents. More interestingly, these agents can place the anonymity of the user into a question and can capture the emotions of individuals and turn the Web platform as emotive Web. Andre [5] analyzed the significant role played by the micro loops in flipped class-based learning. Many authors modeled teaching-learning process as an optimization problem [6] and exploited the merits of swarm intelligence to get better outcomes. The recent National Educational Policy-2020 (NEP-2020) [7] approved by the Government of India fosters the skill and code-based education that is going to become the game changer in the academics of India. The NEP-2020 emphasizes the importance of mother tongue as the first language during primary education, and the pedagogy must make the learning environment enjoyable which will essentially expand the mental horizon of the students. Implementation of this policy is possible only through integrating the technology ecosystem with the learning environments. However, the exaggeration of agent-based Web intelligence can end up as a hype-cycle and doom to fail if adapted in its native form for E-learning [8]. Please check the clarity of the sentence 'In this direction, this research article proposes a new methodology and framework that hand holds the...'. In this direction, this research article proposes a new methodology and framework that hand holds the students and help them realizing their potentials. Along with that, it also presents the future prospects of its implementation. Owing to its complexity, we restrict our discussions to circuit branches of engineering program at undergraduate level offered in India. Further works gradually aspire to accommodate the other disciplines of education. The following sections are organized into the pitfalls of the existing LMS platforms and the features of the proposed comprehen-

sive vLMS model. Finally, it also explores the agility gained by the model from the developments in the entire ecosystem.

## 2 Shortcomings of the Existing LMS

The conventional teaching-learning(T-L) systems are purely class room-based and facilitate the students to learn the content in class rooms and laboratories. The pitfalls of the conventional teaching-learning process and its relevance in the online learning scenario are detailed below.

- The success stories of the classroom-based teaching process rooted in the proven practices: eye-to-eye contact, chalk-and-talk, flipped models, learning while doing, etc. However, these practices are conspicuous for their absence in online learning scenarios.
- The existing T-L process evaluation models utilize the measures either absolute or derivable from the data collected from the theory and practical exams, typically, spread over a period of 16 weeks. However, students are also encouraged to participate in group events that foster their dynamics, and the outcomes are not accounted for in any course outcome evaluations.
- Due to the virtual nature, the concentration levels of student greatly get affected in online scenario, and hence, there is a need to remodel the content delivery process, consolidation of learning outcomes to ensure the success of the online learning model.
- The online T-L process casts a bug-bear to the teachers as they are evaluated in the open house. It is imperative for the teachers to equip themselves with new content delivery skills and display new balanced reasoning skills which are absent in the LMS.
- Perhaps, this new online realm is an unwanted situation for the ambivalent teachers who doubt the safe delivery of the contents as a myth, and further, the loss of personal relations might lead to isolated and self-centered individuals which are really unwantedqueryPlease check the clarity of the sentence ‘Off late the revolutions in the Web Technologies stack driven by Web 5.0 standard compelled researchers to...’.

Off late the revolutions in the Web Technologies stack driven by Web 5.0 standard compelled researchers to support the LMS with lot of functional capabilities. Veselina and Snejana [9] attempted to develop an intelligent system in Web searching, document management and organizational control in the form of webOS. Knowing which category a student–learner belongs to can help the instructor to resolve what information to discuss and how to deliver the content. Dunn and Kennedy [10] assessed the delivery of the technology in enhancing the learning curve of the students. However, these models serve as recommendation systems that process the data collected by the model and offer advisory rules to the learners. Also, they do not evaluate the

efficacy of the suggested rule and its impact on the learner. Hence, this article introduces vouch augmented course recommendation system which learns in a reinforced manner and assures the student and the teacher with the expected outcome.

### 3 Objectives of the Proposed Research

The proposed research work aspires to address all the concerns specified above by leveraging the latest technologies and artificial Intelligence models that are capable of extracting the latent semantics and establish effective logistics with the following expected outcomes:

- To develop an intelligent cyber-physical model which enhances the capabilities of self-learning in the students.
- To increase the contents complexity dynamically that adjust to the learning curve of the aspirant learner who goes beyond the expectations.
- To qualitatively model the interaction scenarios between the host and the participant for realizing the program outcomes.
- To quantify the level of focus a learner is maintaining in a time series.
- To design and develop a technology fabric to offer wide variety of use cases for all the actors. and finally,
- To evolve an agent-based model that retrospects and reinforces its recommendations on the time line and offers a symbiotic ecosystem for both the learners and patrons. Hence, the LMS with self-evaluation feature is referred to as vouching-based LMS (vLMS) which is coined for the first time in the literature of recommendation systems. The self-accountable and self-reformable features are realized by exploiting the technology platform and to extract suitable metrics (standards) and measures.

#### 3.1 *To Quantify the Level of Focus (LoF) for a Learner in a Time Series*

Here, the student activity can be framed as supervised learning [2] which focuses on the learner's active [5] participation in learning the content or the completion of the course within the time-bound. The model will also resolve if the participant was not that active and sends an alert notification to the learner through various channel to complete the course. For example, if  $x$  is the learner who got registered with the LMS application ( $y$ ) and got registered for a particular course for about 40 h, here the course can be completed in two days or he may complete the course in 10 days also by spending at 4 h per day in learning the content. Here, the machine (supervises) records the entry of the learner, the log-in and exit activities and analyzes using time series models.

### ***3.2 To Enhance the Capabilities of Self-learning in the Students***

This is a difficult task for the machine to make the learner to be self-learning [11]. As of now, all the learners might have been learning the content in the form of spoon-feeding, when it comes to self-learning, the learner will face a difficult task as he/she needs to focus on the content and needs to solve the problems on their own. Here, the machine learning algorithm manifested by virtual reality can enable the learner to solve the obtained problems. For that, the system has to be designed in such way, a sample or the similar problem has to be solved step by step in a diagrammatic way, by seeing the process, the learner can understand how to solve the problem, and in the next step, the learner by himself can able to solve the problem which enables self-learning.

### ***3.3 To Fiddle the Complexity of the Course Contents with the Learning Curve of the Aspirant Learner***

This is one of the interesting points the proposed machine learning algorithms focus upon and add a value to the system. This can be done based on the active participation of the learner if the learner has registered for a particular course through LMS, if he is completing each and every module, supervised machine learning algorithm will play a vital role, and here, the supervised machine learning algorithm would like to assess the participant how far he/she has learned the content by putting some random questions dynamically in the form of the quiz so as to get the information about the participant how far he/she following the content through LMS. By this, if he/she answered the questions properly, then also, we can consider he/she is the active participant, i.e., testing the participant from easy to complex mode, as the participant moving ahead in learning the course.

### ***3.4 To Qualitatively Model the Interaction Scenarios Between the Host and the Participant for Realizing the Program Outcomes***

The learner should be able to understand the content and the outcomes clearly, in addition to that, the learning can be considered as a dynamic process in which the learner actively “constructs” new knowledge as he or she is engaged and immersed in a learning activity, and along the other side, there should be a keen focus on two factors, i.e., behaviorism and cognitive psychology, where these things can be evaluated by using different machine learning algorithms for the better results.

### ***3.5 To Design and Develop a Technology Fabric to Offer Wide Variety of Use Cases for All the Actors***

A necessary design should be made where all kinds of interfaces are developed where a layman can also understand and have some kind of interaction between the learner and the instructor, where the learner requirements could be fulfilled.

### ***3.6 To Evolve a Vouching-Based Online Teaching Learning Technology Platform with Suitable Metrics and Measures***

This objective clearly focuses on how the assessment of the student is taking place, rather than the traditional assessment like physical evaluation of the student in the form of marks [7]. This feature of the system assesses the level of understanding of the student and the additional efforts required by the student to meet the essential rubrics. Along sides, the vLMS also evaluates the strength and outcome of the recommendation and reinforces itself by consuming additional input from the system.

## **4 Proposed Agent-Based Reinforced Learning Model (vLMS)**

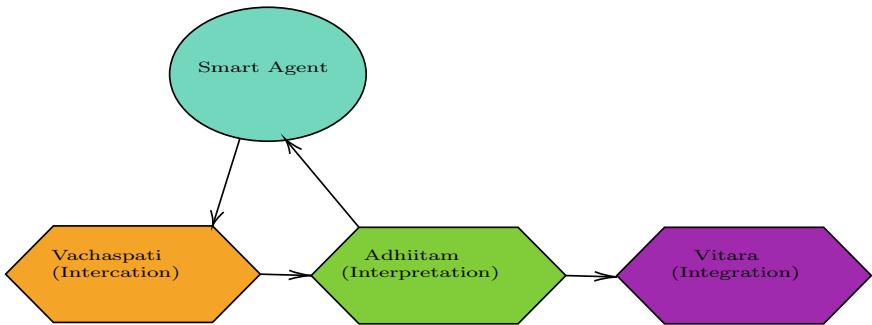
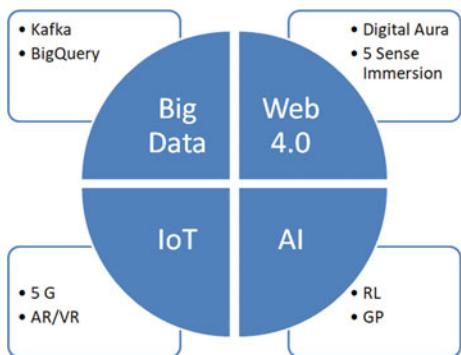
The vLMS offers a comprehensive solution to all the stakeholders of the system and is composed of two functional modules, namely SPSS and smart vouching agent. It is proposed to foster the outcome-based learning model of the students with a joyful learning environment on the campus by developing smart student profiling system (SSPS). The vouching module observes and analyzes the recommendations of SPSS and offers a reinforced learning capability. This research work is proposed to augment the existing student monitoring and mentoring process adopted by an academic institute with smart technologies. The fundamental idea behind this research is to evolve a smart mentoring system which relieves the students from intangible stress caused by the various factors such as strict submission deadlines, nearing last dates, changing industry trends, peer-pressure, temporary setbacks and other distractions. Also, it pushes the event-based messages to the parents and generates the summary sheets useful for many accreditation agencies. The unique feature of SSPS is to serve as an oracle to parents and students in quashing all misgivings.

The proposed SSPS model is a distributed computing model which utilizes cutting edge technologies to process data collected from various sources and formats, offers a analytics prism and fills the semantic gap between the ambitions of the students and capabilities of the system. The sources include the day-to-day class room attendance, assignment submission due dates, course registration deadlines, regular classwork reading hours, experiences and observations in lab sessions, the

summary of mentoring reports, performance in internals of theory and lab examinations and outlook of experts on fast changing industry trends. The value derivation from such a data is accomplished by various mathematical and statistical models available in predictive and prescriptive data analytics libraries. SSPS also serves as a recommendation system on professional electives and open electives to the students. Along with the data available in the documents generated, visual data from IP-enabled cameras is collected seamlessly without hampering the privacy of the students. This data throws light on behavioral dynamics and pedagogy effectual aspects of the students on the campus. It is imperative to install an IoT-based visual sensor networks to deploy the service model of this functional requirement. The openCV libraries and Python API available in the public domain can be utilized. SSPS offers a interactive chat box as a frontend to all the stakeholders and dispenses the most relevant and artificial intelligence-driven information flakes, measures and metrics. The salient features of this interactive model are to offer personalized prescriptive analytics model, to raise the future spirits of the students and to repose confidence in the parents on the prospects of the students, by employing natural language processing, applied machine learning and deep learning models. Yet another usecase of SSPS is to generate summary sheets and attainment measures. Overall the SSPS helps ingest management of all the data resources and present visual analytics to all the stakeholders. The vouching system which is a reinforced learning-based module constantly observes the recommendations of the SPSS and compares the actual outcome with the expected outcome and suggests new recommendations. We propose to employ a genetic programming-based methodology to solve the inference problem of the vouching agent.

## 5 Technology Stack to Implement the vLMS

Figure 1 gives the schematic diagram of the proposed vLMS along with the building blocks. The agent-based vLMS can primarily augment the mission of an academic institute to realize its vision statement. It is intended to collect data from various sources and interprets it to derive actionable intelligence. It also generates new learning scenarios for the students and also accelerates the decision agility of a teacher. Further, it offers nudge factors to the students to elevate his spirits. Vitara module in Fig. 1 offers a data integration pipeline that ingests the data into the Adhiitam, the cognitive engine of the system. The legacy databases, teaching learning process outcomes (TLP), flat files (mentoring sheets, survey forms), API (social media channels) serve as the sources of the data. This module is responsible for data customization and integration of interpretations available in the data sources. The Adhiitam module offers a data analytics prism and is capable of solving multi-subjective optimization problems and suggests some actionable items. Vachaspathi module offers the interfaces to users to have interactions with vLMS leading to conversational experiences for all the end users. It dispenses some bonafied recommendations for trivial queries and extracts the latent semantics while handling complex and involved questions.

**Fig. 1** vLMS**Fig. 2** Technology stack

It also serves appraisals and recommendations to the end user. The composition of Vachaspati, Adhiitam and Vitara implements the SSPS, and the smart agent which offers the vouchers through genetic programming and deep learning models turns SSPS into vLMS. The smart system agent is capable of implementing the sequence models and gains insights of the teaching learning events.

Figure 2 presents the technology support to implement the vLMS along with the building blocks. The Web 5.0, AI, IoT and big data ecosystems play a vital role in implementing the vLMS. Web 5.0 communication mechanisms, Digital Aura and the entertainment API 5 senses can be used to offer better user experiences and immersion features. The AR/VR libraries help to implement visualization and presentations capabilities. The big data and AI-based tools process the time series data, and the NLP packages bridge the semantic gap between the student and the system. Overall the vLMS improves the operational efficiency of T-L process by evolving new set of standard operational procedures. The AR and VR technology can be used in taking the e-Learning or vouching learning management system (vLMS) to another level.

## 6 Conclusions

This article discussed the imperatives of new online learning models, and a self-accountable vouching-based LMS is proposed. The technological feasibilities are elaborated along with the architecture of vLMS. The future work aspires to implement the model and evaluate it on industry workbench data sets. This research work is going to offer a tech conclave for implementing NEP-2020.

## References

1. Sunita, A., Lobo, M.J.R.: A framework for recommendation of courses in e-learning system (2011)
2. Hwang, G-J., Lai, C-L., Wang, S.-Y.: Seamless flipped learning: a mobile technology-604 enhanced flipped classroom with effective learning strategies. *J. Comput. Educ.* **2**(4), 449–473 (2015)
3. Castillo, G., Millan, E.: Discovering Student Preferences in E-Learning (2007)
4. Benito-Osorio, D., Peris-Ortiz, M., Armengot, C.R., et al.: Web 5.0: the future of emotional competences in higher education. *Glob. Bus. Perspect.* **1**, 274–287 (2013)
5. Mastmeyer, A.: Quantitative and Qualitative Evaluation of Transforming to Flipped-Classroom from Instruction Teaching using Micro Feedback (Version 0.5), Aug 2020
6. Suresh, S.C., Naik, A.: Modified teaching–learning-based optimization algorithm for global numerical optimization—A comparative study. *Swarm Evol Comput* (2014)
7. NEP: [www.mhrd.gov.in/sites/upload\\_files/mhrd/files/NEP\\_Final\\_English\\_0.pdf](http://www.mhrd.gov.in/sites/upload_files/mhrd/files/NEP_Final_English_0.pdf) (2020)
8. Romn, P.E., Torres, E.O., Hermndez, R.L.: Virtual learning environments as a continuous assessment tool in university students, Chap. 12, pp. 229–252. IGI Global (2019)
9. Veselina, N., Snejana, D.: Intelligent e-Learning with New Web Technologies. ICVL (2015)
10. Dunn, T., Kennedy, M.: Technology enhanced learning in higher education; motivations, engagement and academic achievement. *Comput. Educ.* **137**, 104–113 (2019)
11. Zaiane, O.R., Luo, J.: Towards evaluating learners' behaviour in a web-based distance learning environment (2001)

# Heart Disease Prediction Using Extended KNN (E-KNN)



R. Sateesh Kumar and S. Sameen Fatima

**Abstract** The WHO estimates that deaths due to heart disease are the number one cause worldwide, accounting for around 30% annually taking an estimated 1.5 crores who die due to this disease. In this study, an extension of KNN algorithm known as E-KNN is used and compares with the results of different machine learning methods such as K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Classification and Regression Trees (CART) (Kumar and Thomas in Int J Recent Technol Eng (IJRTE) 9(1) [1]) in the prediction of heart disease. To improve the efficiency of the proposed system, the most important features are selected using chi-square test. The performance and efficiency of the algorithms are evaluated and compared on the basis of accuracy, recall, precision, and F1 score. The results of the proposed algorithm were more accurate with lesser attributes than all attributes. The performance of E-KNN by using 11 attributes has an accuracy value of 90.10%. It is followed by SVM with 89% accuracy.

## 1 Introduction

WHO estimated that there are 17 million people die of heart diseases (CVD) every year. These death rates can be reduced with early diagnosis and informed to the patients. Common cardiovascular diseases include coronary heart disease, cardiomyopathy, hypertensive heart disease, heart failure, etc. Smoking, diabetes, lack of physical activity, hypertension, high cholesterol diet are some of the common causes of heart diseases.

---

R. Sateesh Kumar (✉)

Department of Computer Science and Engineering, Vasavi College of Engineering, Hyderabad, Telangana, India

e-mail: [sateeshramatenki@staff.vce.ac.in](mailto:sateeshramatenki@staff.vce.ac.in)

S. Sameen Fatima

Osmania University, Hyderabad, Telangana, India

Research leading to the sciences of cardiovascular diseases using machine learning and data mining has been an endeavor encompassing optimum treatment plans, expedited and disease prognosis, identification of risk factors. Numerous CVD surveys have been conducted with the primary dataset as Cleveland dataset. Recommending the parameters from this dataset proposes a predictive system to apply logistic regression model, random forests, SVM, and KNN to obtain predictions from several learners which are in turn used in meta models. The results of each ensemble modeling are then compared on the basis of several evaluation parameters.

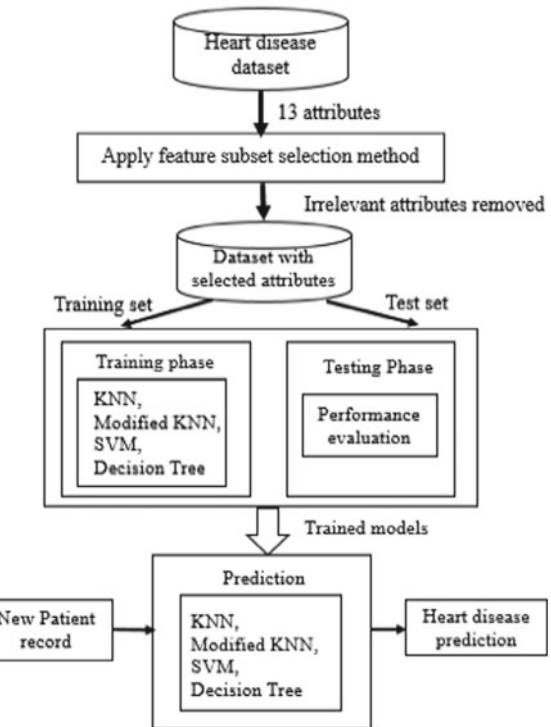
In this paper, we are proposing extension of KNN algorithm, namely, Extended K-Nearest Neighbor algorithm (E-KNN). The results of this algorithm compared with K-Nearest Neighbor algorithm, Decision Tree algorithm, and Support Vector Machine algorithm. The specified data set is used to train these algorithms. With enough data and time, it is good to all the attributes or features including the irrelevant features to train the classification algorithm. The problems of using the irrelevant features are that it leads to lesser accurate results and it may also lead to overfitting. Feature subset selection is a dimensionality reduction method that is used to select the most relevant attributes from the dataset [1]. In this paper, the Chi-square test is used as the feature subset selection method. The selected features are fed to the algorithms as input.

## 2 Literature Review

Revathi et al. [2], provides an analysis of various data mining methods for prediction of heart disease. Decision tree, Back-propagation algorithm, and Nave Bayes are used. The system used 14 parameters including blood pressure, chest pain, cholesterol, and heart rate to enhance the system accuracy. A comparison of the performance of the three algorithms is done. Study shows that the neural network performs best to predict heart disease and has the accuracy of 100%. It has outperformed the other two algorithms.

Shinde et al. [3], using the data mining methods introduced a system for predicting heart disease. K-means clustering algorithm and Nave Bayes were used in this system. For prediction, a combination of both techniques has been used. The K-means clustering algorithm is used to enhance overall system efficiency. It is used to group the different attributes present in the dataset and the prediction is done by the Nave Bayes algorithm. The system used 10 attributes. Compared with the other algorithms, the system produces better results.

Sateesh et al. [1], used ensemble learning method to predict the heart disease. They used KNN, SVM, CART algorithms and designed an ensemble using these algorithms.

**Fig. 1** Approach

### 3 Approach

The methodology used.

Figure 1 shows the approach used in this paper. The study uses dataset which has 300 records of patients. The original dataset contains thirteen attributes. From the dataset, the required attributes using the statistical measure (Chi-square test) are identified.

### 4 Dataset

Table 1 describes the data set used in this study. It is an openly available dataset and can be found at the UCI Machine Learning Repository. The original data set consists of 76 features. We used 13 features from them. The features are described in the following table. The disease attribute is the class label which specifies whether the patient is suffering from heart disease or not.

**Table 1** Dataset description [1]

Feature No.	Name of feature	Feature description
1.	age_years	Age of the patient
2.	Gender	1 Male 0 Female
3.	cpain_type	Chest pain {0, 1, 2, 3}
4.	rest_bp	This feature represents the value of the blood sugar in rest
5.	serum_chol	Measure the serum cholesterol
6.	fb_sugar	Measure fasting blood sugar {0 if fb_sugar < 120, 1 if fb_sugar > 120}
7.	rest_ecg	ECG in resting {0, 1, 2}
8.	max_heartrate	maximum heart rate
9.	ex_angina	It takes {0, 1}
10.	ex_ST_depression	ST depression during exercise
11.	peak_slope	Takes the value s{0, 1, 2}
12.	Number of vessels	Indicates number of vessels blocked {0, 1, 2, 3}
13.	defect_type	Heart defect type {1, 2, 3}
14.	disease	Identification of heart disease 1 Suffering from disease 0 No disease

#### 4.1 Chi-Squared Value

Sometimes our dataset may contain both relevant and irrelevant features. It is very important to find and discard the unimportant features that do not contribute to the outcome to be predicted by the system. There are various methods to identify the relevant features from our input attributes. The methods to select the relevant features are called Feature Selection Methods. Chi-squared test applies the statistical method to identify the relation between the input variable and the target variable.

The following are the steps to determine the Chi-squared value.

Create a contingency table for two attributes. The values present in the table are called as observed values (O).

Calculate the expected frequency (E) value for each of the cell present in the table.

$$E = (\text{row total} * \text{column total}) / \text{overall sum}$$

Calculate the chi-squared value

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

Arrange the chi-squared value in the descending order.

Select the top-ranked attributes from the above list.

## 4.2 *Extended K-Nearest Neighbor*

The KNN classifier uses polling to assign the class label to the new data item. This leads to the missclassification. In order to overcome this problem, the E-KNN is implemented. In KNN, a new data item belongs to certain class C1 which is surrounded by classes C2 data points. The new data point is misclassified as class C2 because of the majority of class. To avoid this problem in E-KNN, the selection of top k neighbors is varied. To reduce the class C2 neighbors, it skips 3 Neighbors and select the next 3 Neighbors, i.e.,  $3 + k$  Neighbors and also assign the weight to the each neighbor based on the distance from the new data item. Class label to the new data item is selected based on the  $k + 3$  neighbors and their weights.

E-KNN Algorithm steps

1. Divide the data set in training and testing
2. Choose the K values based on the requirements
3. For every new data item compute the following:
  - 3.1 Find the Distance from the new data item to all the data items in the training set using distance measure
  - 3.2 Sort the distances in descending order and identify the K nearest data points
  - 3.3 Assign the weights to all the K nearest neighbors (Weight  $w = 1/d$  where  $d$  is the distance from the new data point to its neighbor)
  - 3.4 To obtain the top k rows from the above-sorted distance array, calculate the following two values start =  $k / 2$ , end =  $k + \text{start}$
  - 3.5 The top k rows are obtained by adding  $k = \text{sorted distance}(0: \text{start}) + \text{sorted distance}(k-1: \text{end})$  and based on the weights
  - 3.6 The most frequent class from the sorted rows and their weights
4. Stop

## 4.3 *K-Nearest Neighbor*

1. Divide the data set in training and testing
2. Choose the K values based on the requirements
3. For every new data item compute the following:
  - 3.1 Find the Distance from the new data item to all the data items in the training set using distance measure

- 3.2 Sort the distances in descending order and identify the K nearest data points
- 3.3 The most frequent class from the sorted rows and their weights
4. Stop

## 5 Results and Discussion

This section provides the results of implementing the various above-discussed algorithms on the heart disease dataset. Table 2 indicates the top 11 attributes ranked from the Chi-square test.

Comparison of algorithms according to the number of attributes used show in Tables 3 and 4 (Figs. 2, 3 and 4; Table 5).

**Table 2** Top 11 attributes selected by chi-squared test

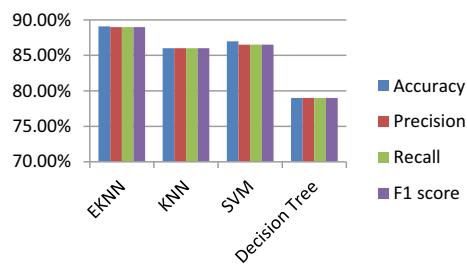
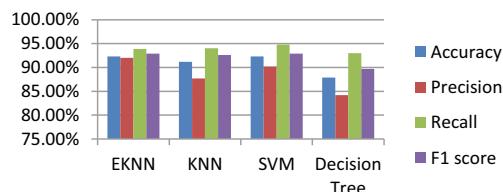
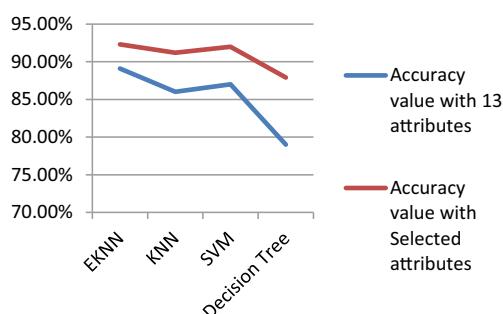
Attribute	Rank
A1	max_heartrate
A2	ex_ST_depression
A3	num_vessels
A4	cpain_type
A5	ex_angina
A6	serum_chol
A7	Age
A8	rest_bp
A9	peak_slope
A10	Gender
A11	defect_type

**Table 3** The comparison of algorithm result with 13 attributes

	EKNN (%)	KNN	SVM	Decision Tree
Accuracy	89.10	86	87	79
Precision	89	86	86.5	79
Recall	89	86	86.5	79
F1 score	89	86	86.5	79

**Table 4** The comparison of algorithm result with selected attributes

	EKNN (%)	KNN (%)	SVM (%)	Decision tree (%)
Accuracy	92.31	91.21	92.31	87.91
Precision	92	87.72	90.20	84.21
Recall	93.88	94.04	94.83	93
F1 score	92.93%	92.59%	92.93%	89.72%

**Fig. 2** Shows the results with 13 attributes**Fig. 3** Shows the results with selected attributes**Fig. 4** Shows the accuracy of all algorithms**Table 5** Comparison of accuracy according to the number of attributes used

Algorithm	Accuracy value with	
	13 attributes (%)	Selected attributes (%)
EKNN	89.1	92.31
KNN	86	91.21
SVM	87	92
Decision tree	79	87.91

## 6 Conclusion

The Chi-squared test was used as a feature selection method. This helps in identifying the suitable attributes for this study. In this paper, we used E-KNN, KNN, SVM, and Decision Tree algorithms. The overall objective of this project was to predict more accurately the presence of heart disease. The performance of four data mining classification techniques, namely, Extended KNN (E-KNN), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Decision tree algorithm (CART) compared. The performances of the models were evaluated using the standard metrics of accuracy, precision recall, and F1-score. There is also a decrease in the misclassification rate in E-KNN when compared to KNN.

The overall work showed that E-KNN predicts heart disease more accurately compared to base classifiers. Finally, it is concluded that the algorithms efficiently predict heart disease with less number of attributes.

**Acknowledgements** Authors would like to express gratitude to Vasavi College of Engineering (A), Hyderabad for their constant supervision as well as for providing necessary information regarding this research and also their support in completing this endeavor.

## References

1. Sateesh Kumar, R., Thomas, A.: Heart disease prediction using ensemble learning method. Int. J. Recent Technol. Eng. (IJRTE) **9**(1) (2020) ISSN: 2277-3878
2. Jabbar, M.A., Deekshatulu, B.L., Chandra, P.: Classification of heart disease using K-Nearest Neighbor and genetic algorithm. Procedia Technol. **10**, 8594 (2013)
3. Kalaiselvi, C.: Diagnosing of heart diseases using average k-nearest neighbor algorithm of data mining. In: 3rd International Conference on Computing for Sustainable Global Development (INDIACoM), New Delhi, pp. 30993103 (2016).

# Prediction Analysis of Diabetes Using Machine Learning



Srikanth Bethu, G. Charles Babu, B. Sankara Babu, and V. Anusha

**Abstract** Prescient frameworks are the frameworks that are wont to foresee some result based on some example, acknowledgment. Diabetes illness discovery is that the technique by which a patient's determination is performed based on indications examined, which may cause trouble while foreseeing infection influence. For instance, fever itself could be a manifestation of the numerous scatters that do not tell the human services proficient what precisely the sickness is. Since the outcomes or feelings fluctuate from one doctor to an alternate, there is a necessity to help a restorative doctor, which will have comparative assessment positively side effects and clutters. It may finish by breaking down the data created by medicinal information or therapeutic records. In this way, applying the AI calculations to foresee diabetes ought to be completed.

## 1 Introduction

The human body needs significance for approval. The sugars are separated into glucose, which is the significant centrality hotspot for human body cells. Insulin is dependent upon to convey the glucose into body cells—the blood glucose given insulin and glucagon hormones made by the pancreas. Insulin hormones produced by the beta cells of the islets of Langerhans and glucagon hormones are passed on by the alpha cells of the islands of Langerhans in the pancreas. Exactly when the blood glucose grows, beta cells are invigorated, and insulin given to the blood. Insulin empowers blood glucose to get into the cells, and this glucose utilized for vitality. In this way, blood glucose kept in a tight range.

Inherent Diabetes [1] happens in humans because of the natural flaws of insulin release, cystic fibrosis-related Diabetes, and large fragments of glucocorticoids brief

---

S. Bethu (✉) · B. Sankara Babu · V. Anusha

Department of Computer Science and Engineering, GRIET, Hyderabad 500090, India

G. Charles Babu

Department of Computer Science and Engineering, MREC, Hyderabad 500100, India

steroid diabetes. As a result, concerning the human, our bodies glitch as indicated by produce insulin and requires the person between an impersonation of supplement insulin and raise an insulin siphon [2, 3]. This class is once recently shown as much permanency “Insulin-Dependent Diabetes Mellitus.” The second classification about DM is perceived to be specific “Type II DM” along these lines a result as respects insulin encounter, a circumstance of any cells are ineffectual of agreement with endeavor insulin properly, incidentally combined all in all with an outright insulin deficiency. At last, “gestational diabetes” happens when considered ladies without a before. The prior finding of Diabetes, the danger of the intricacies can be evaded. Diabetic patients endure different infections, and it influences different pieces of various organs. Subsequently, successful measures must be taken to anticipate the sickness at the most punctual and control.

These days data mining [2] devices and methods are generally utilized in pretty much every field like social insurance frameworks, promoting, climate determining, E-business, retails, and so on. The medicinal services system is one of the new developing exploration territories where information mining methods and apparatuses can be adequately connected. Our therapeutic services frameworks are wealthy in data. However, they are deficient in learning, so there is a considerable need for having strategies and devices for removing the data from the sizeable informational collection with the goal that restorative analysis should be possible.

Numerous natural frameworks [4] are on a fundamental level nonlinear, and their parameters restrictively reliant. Various necessary physical structures are straight, and their settings are free. Achievement in AI is not ensured continuously. Similarly, as with any strategy, an excellent comprehension of the issue, and a valuation for the confinements of the information are significant. On the off chance that an AI examination appropriately planned, the students effectively actualized and the outcomes vigorously approved, at that point, one more often than not, has a decent possibility at progress. Whether the information is of low quality, the outcome will be of low quality (trash in = trash out).

## 2 Literature Survey

Evaluation of Glycaemic Control, Glucose Variability and Hypoglycaemia on Long-Term Continuous Subcutaneous Infusion vs. Multiple Daily Injections: Observational Study in Pregnancies With Pre-Existing Type 1 Diabetes, Aleksandra Jotic, Diabetes Therapy, 11, pages 845–858 (2020)—This paper clarifies the assessment of the adequacy of long haul persistent subcutaneous insulin imburement (CSII) contrasted and numerous everyday insulin (MDI) infusions for glycaemic control and fluctuation, hypoglycaemic scenes and maternal/neonatal results in pregnant ladies with previous sort 1 diabetes (pT1D).

Impact of Simultaneous Versus Sequential Initiation of Basal Insulin and Glucagon-like Peptide-1 Receptor Agonists on HbA1c in Type 2 Diabetes: A Retrospective Observational Study, Vivian Fonseca, Diabetes Therapy, 11, pages 995–1005

(2020)—This paper determines When and how to strengthen treatment in patients with type 2 diabetes (T2D) not accomplishing glycated hemoglobin (HbA1c) focuses with oral antidiabetic drugs (OADs) in clinical practice stays a matter of clinical inclination. This pilot study was directed utilizing the review observational information from such patients to assess the effect on HbA1c of three treatment groupings: synchronous inception of basal insulin (BI) and a glucagonlike peptide-1 receptor agonist (GLP-1 RA; Cohort 1); BI followed by GLP-1 RA commencement inside a 90-day time span (Cohort 2); or BI followed by GLP-1 RA inception past 90 days (Cohort 3).

Foresee The beginning of Diabetes Disease Using Artificial Neural Network (ANN) by Manaswini Pradhan, Dr. Ranjit Kumar Sahu—This paper speaks to 8.8% of the absolute ladies grown-up populace of the 18 years old or more in 2003, and this is almost a two overlap increment from 1995 (4.7%). Ladies of minority racial and ethnic gatherings have the most noteworthy predominance rates with two to multiple times the prices of the white populace. With the expanded development of minority populaces, the number of ladies in these gatherings who are analyzed will increment essentially in the coming years.

(2018) Prediction of diabetes using classification algorithms. Procedia computer science, by Deepti Sisodia and Dilip Singh Sisodia—This investigation work bases on pregnant women encountering diabetes. In this work, Naive Bayes, SVM, and Decision Tree AI request counts are used and surveyed on the PIDD dataset to find the figure of diabetes in a patient. The preliminary execution of all three estimations is pondered on various measures and achieved extraordinary precision.

In this work, WEKA gadget is used for playing out the preliminary. The principal purpose of this examination is the desire for the patient impacted by diabetes using the WEKA mechanical assembly by using the therapeutic database PIDD.

(2019) Comprehensive Review of Artificial Neural Network Applications to Pattern Recognition—The assessment outfits per users with an all the more precise understanding of the current and new example in ANN models that effectively addresses Pattern Recognition troubles to investigate focus and subjects. Likewise, the sweeping review reveals the different zones of the achievement of ANN models and their application to PR. In this paper, we have used ANN models to generate results and compared the remaining Machine Learning algorithms with it.

### 3 Methodology

#### 3.1 *Proposed Method*

In this paper, we talked about how the Artificial Intelligence [5] method helps foresee diabetes and its significance in social insurance applications. Proposed a human services framework utilizing brilliant dress for sustainable wellbeing checking [6]. I had altogether examined the various structures. I accomplished the best outcomes

for cost minimization on tree and primary way cases for different frameworks. Here we use utilize A.I. methodologies to recognize the prevalent parts causing diabetes in individuals. In the beginning, factors that are accepted to be colossal like age, B.M.I., High Cholesterol, Hyperthyroid, Hypertension, Age, and Skin Thickness are considered. Among these, the most basic ones provoking diabetes are perceived. Characteristics of each essential factor are analyzed in diabetic and non-diabetic individuals provoking learning disclosure of particularly essential explanations behind diabetes with everything taken into account. The entire educational gathering is moreover subject to portrayal using four A.I. computations [7], and a close examination of the strategies is similarly grasped.

### 3.2 Algorithms

In Supervised learning, the structure must “adjust” inductively a limit called target work, which is an outpouring of a model depicting the data. The objective limit used to envision the estimation of a variable [8], called a subordinate variable or yield variable, from a lot of elements, called self-sufficient parts or information components or characteristics or features.

The Modules are divided based on various steps involved in the determination of best algorithms are Feature Selection and Performance Evaluation.

- (a) Feature Selection [9]: AI AI and measurements, include choice, otherwise called variable determination, trait choice, is the procedure of choosing a subset of applicable highlights for use in model development. There are numerous strategies for highlight determination; we are utilizing univariate include choice technique in this structure. There are two various primary methodologies in the component determination process.

The first is to make an independent appraisal, given general qualities of information. Strategies have a place with this methodology called channel techniques because the list of capabilities is sifted through before model development. The last calculation will be utilized at last to construct a prescient model. Strategies in this class are called wrapper techniques, which wraps the entire element determination process.

- (b) Performance Evaluation [10]: Exhibitions of all classifiers are assessed by various estimation factors as precision (ACC), affectability (SE), particularity (SP), positive prescient worth (PPV), negative prescient worth (NPV), and so forth. These estimation variables are determined by utilizing genuine positive (TP), genuine negative (TN), false positive (FP), and false negative (FN).

Accuracy, it is the extent of the entirety of the genuine positive and genuine negative against absolute number of populace. It very well may be communicated numerically as pursues:

$$ACC(\%) = \left( \frac{TP + TN}{TP + FN + FP + TN} \right) * 100 \quad (1)$$

Sensitivity, it is the extent of the positive condition against the anticipated condition is sure. It very well may be communicated scientifically as pursues

$$SE(\%) = \left( \frac{TP}{TP + FN} \right) * 100 \quad (2)$$

Specificity, it is the extent of the negative condition against the anticipated condition is negative. It tends to be communicated numerically as pursues

$$SP(\%) = \left( \frac{FP}{FP + TN} \right) * 100 \quad (3)$$

Positive prescient worth, the positive prescient worth is the extent of the anticipated positive condition against the genuine condition is sure. It very well may be communicated scientifically as pursues

$$PPV(\%) = \left( \frac{TP}{TP + FP} \right) * 100 \quad (4)$$

Negative prescient worth, it is the extent of the anticipated negative condition against the genuine condition is negative. It very well may be communicated mathematically as pursues

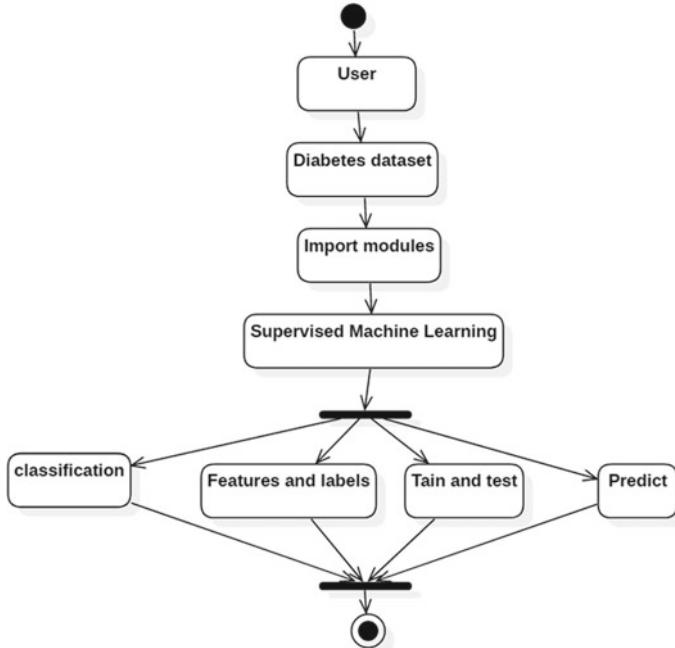
$$NPV(\%) = \left( \frac{TN}{FN + TN} \right) * 100 \quad (5)$$

Genuine constructive (TP): Those Sick individuals who are accurately analyzed as debilitated. False constructive (FP): The Healthy individuals who are mistakenly recognized as debilitated. Genuine adverse (TN): The Healthy individuals who are accurately recognized as sound. False pessimistic (FN): The Sick individuals who are inaccurately distinguished as solid.

## 4 Implementation

### 4.1 Technologies Used

Operating System was Windows, Technology was Python, Libraries are pandas, numpy IDE was Jupyter notebook, Google Colabatory are used for experimental environmental setup.



**Fig. 1** Functional diagram

In Fig. 1, supervised machine learning algorithms like KNN, Logistic regression, SVM and ANN are compared to identify classification and feature selection. Figure 1 also shows the functional process of results prediction using machine learning algorithms.

## 4.2 Logistic Regression

Logistic Regression [11] is another methodology gotten by AI from the field of bits of knowledge. It is the go-to methodology for the twofold request (issues with two class regards). The determined limit, in like manner called the sigmoid limit, was made by examiners to depict properties of people improvement in nature, rising quickly and augmenting at the passing on the point of confinement of the earth. It's an S-encircled bend that can take any affirmed respected number and guide it into an inspiration some spot in the extent of 0 and 1, at any rate never precisely at those cutoff centers.

$$X = \left( \frac{1}{1 + e^{-\text{VALUE}}} \right) Y = e^{\left( \frac{b_0 + b_1 * x}{1 + e^{b_0 + b_1 * x}} \right)} \quad (6)$$

From Eq. (6) Calculated relapse utilizes a condition as the portrayal, particularly like straight relapse. Information esteems ( $x$ ) are joined directly utilizing loads or coefficient esteems to anticipate a yield esteem ( $y$ ). Where  $y$  is the predicted yield,  $b_0$  is the propensity or catch term and  $b_1$  is the coefficient for the single information respect ( $x$ ). Each fragment in your information has a related b coefficient that must be grabbed from your arranging information. The veritable delineation of the model that you would store in memory or in a record is the coefficients in the condition.

### 4.3 Support Vector Machine

In SVM [12], we take the yield of the straight capacity, and if that yield is more noteworthy than 1, we distinguish it with one class, and if the return is—1, we recognize it with another type. Since the limit esteems are changed to 1 and—1 in SVM, we acquire this fortification scope of values( $[-1, 1]$ ), which goes about as edge. In the SVM calculation in Eq. (7), we are hoping to amplify the edge between the information focuses and the hyperplane. The misfortune work that boosts the edge is pivot misfortune.

$$\begin{aligned} C(x, y, f(x)) &= \begin{cases} 0 & \text{if } y \neq f(x), \\ 1 - y & \text{if } y = f(x) \geq 1 \end{cases} \quad \text{else} \\ C(x, y, f(x)) &= (1 - y * f(x)) \end{aligned} \quad (7)$$

Pivot misfortune (work on left can be spoken to as a capacity on the right). The expense is 0 if the anticipated worth and the genuine worth are of a similar sign. In the event that they are not, we at that point figure the misfortune esteem. We likewise include a regularization parameter for the cost capacity. The goal of the regularization parameter is to adjust the edge augmentation and misfortune. In the wake of including the regularization parameter, the cost capacities look as beneath in Eq. (8).

$$\min_w \tau \|w\|^2 + \sum_{i=1}^n (1 - Y_i \langle X_i, w \rangle) \quad (8)$$

Since we have the misfortune work, we take incomplete subsidiaries as for the loads to discover the angles. Utilizing the slopes, we can refresh our loads in Eq. (9).

$$\begin{aligned} \frac{\delta}{\delta W_k} \lambda |w|_2^2 &= 2\lambda W_k \\ \frac{\delta}{\delta W_k} \left( 1 - Y_i \langle X_i, W \rangle \right) &= \begin{cases} 0 & \text{if } Y_i \langle X_i, W \rangle \geq 1 \\ 1 - Y_i X_{ik} & \text{else} \end{cases} \end{aligned} \quad (9)$$

Table 1 sample data are used for prediction of accuracy scores and selecting the best accurate model. We used Jupyter notebook to work on this dataset. We first go with importing the necessary libraries and import our dataset to Jupyter notebook.

## 5 Results and Discussion

In the proposed system, we utilized machine calculations with Artificial Neural Network, K-Nearest Neighbor, Support Vector Machine, and Logistic Regression [8] for diabetes expectation. The analyses performed on the information dataset (Diabetes Dataset) given the proposed method. From Table 2, it examined that Artificial Neural Network appears the greatest exactness. So the Artificial Neural Network AI classifier can anticipate the odds of diabetes with more accuracy when contrasted with other classifiers. Better test precision of 80.5% gotten alongside other measurable execution parameters for the Diabetes forecast model. Table 2 gives algorithm results.

Figure 2, shows the variables spotting after applying the data measures on algorithms. The main variables like Glucose, Insulin, BMI are considered to identify sickness. The above plotting values are ranged from minimum rate 0% to maximum rate 100%. Figure 2, also shows the outcome values as 0 to define mellitus in blue color as negative and green color value as 1 as positive. The outcome values are used to identify diabetes mellitus and insulin, so that risk factor is also identified for better treatment. The graph is generated separately for all the modules Glucose, Insulin, BMI. The individual comparison of all the results suggests the patient to identify side effects and clutters in the body. Visualizing all attributes given as below.

## 6 Conclusion

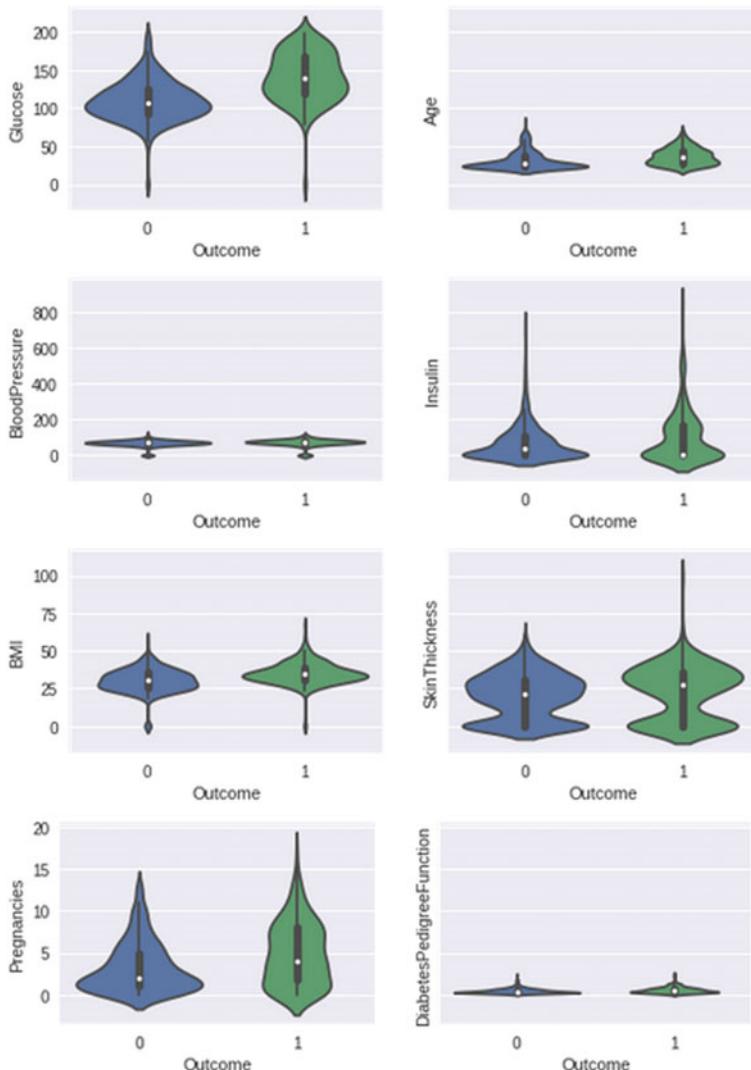
The diabetes forecast framework was created utilizing four information mining order demonstrating systems. These models are prepared and approved against a test dataset. Every one of the four models can concentrate designs in light of the anticipated states. The best model to foresee understanding with diabetes gives off an impression of being ANN trailed by KNN and Logistic relapse. The following restriction was that we did not straightforwardly gauge drug adherence. At last, our information was, for the most part, dependent on patient data. Nonetheless, this investigation represents a potential utilization of the information mining strategy. In the medicinal field, exactness in expectation of the ailments is the most significant factor. In the examination of information mining systems, the ANN classifier gives 80% of the most noteworthy exactness utilizing the Jupyter scratchpad instrument. Future works may address crossover order models utilizing KNN with different methods of AI.

**Table 1** Data set design description using pandas

S. No.	Pregnancies	Glucose	Blood pressure	Skin thickness	Insulin	BMI	Diabetes pedigree function	Age	Outcome
Count	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000	767.000000
Mean	3.847458	121.598641	72.431749	29.113796	156.938543	32.443489	0.471742	33.203390	0.348110
Std	3.371117	30.359639	12.113731	8.547994	88.900636	6.882979	0.331524	11.721879	0.476682
Min	0.00000	44.00000	24.0000	7.0000	14.00000	18.20000	0.07800	21.0000	0.00000
25%	1.00000	99.50000	64.0000	25.00000	121.0000	27.50000	0.243500	24.0000	0.00000
50%	3.00000	117.0000	72.0000	28.00000	130.827879	32.0000	0.371000	29.0000	0.00000
75%	6.00000	141.0000	80.0000	32.631285	206.846154	36.60000	0.626500	41.0000	1.00000
Max	17.00000	199.0000	122.0000	63.00000	846.0000	67.10000	2.242000	81.0000	1.00000

**Table 2** Algorithm comparison based on given data

S. No.	Algorithm	Accuracy
1	ANN	80.519480
2	KNN	77.922077
3	Logistic regression	77.656384
4	SVM	64.767885



**Fig. 2** Dataset attributes visualization

## References

1. Osarech, A., Shadgar, B.: A computer-aided diagnosis system for breast cancer. *Int. J. Comput. Sci. Issues* **8**(2) (2011)
2. Krawczyk, B., Galar, M., Jelen, L., Herrera F.: Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Article in *Appl. Soft Comput.*, Elsevier B.V., pp. 1–14 (2016)
3. Vijayan, V., Ravikumar, A.: Study of data mining algorithms for prediction and diagnosis of diabetes Mellitus. *Int. J. Comput. Appl.* **95**(17) (2014) (0975-8887)
4. Huang, J., Ling, C.X.: Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. Knowl. Data Eng.* **17**(3), 299–310 (2005)
5. Devi, M.R., Maria Shyla, J.: Analysis of various data mining techniques to predict diabetes Mellitus. *Int. J. Appl. Eng. Res.* **11**(1), 727–730 (2016)
6. Kaur, G., Chhabra, A.: Improved J48 classification algorithm for the prediction of diabetes. *Int. J. Comput. Appl.* **98**(22). (0975-8887) (2014)
7. Wang, H., Yoon, S.W.: Breast cancer prediction using data mining method. In: IEEE Conference paper (2015)
8. Fonseca, V.: Impact of simultaneous versus sequential initiation of basal insulin and glucagon-like peptide-1 receptor agonists on HbA1c in Type 2 diabetes: a retrospective observational study. *Diab. Ther.* **11**, 995–1005 (2020)
9. Pradhan, M., Sahu, R.K.: Foresee The beginning of Diabetes Disease Using Artificial Neural Network (ANN) (2011)
10. Lakshmi, K.R., Premkumar, S.: Utilization of data mining techniques for prediction of diabetes disease survivability. *Int. J. Sci. Eng. Res.* **4**(6) (2013)
11. Wajid, S.K., Hussain, A., Huang, K., Bonilla, W.: Local energy-based shape histogram feature extraction technique for breast cancer diagnosis (2015)
12. Bagdi, R., Patil, P.: Diagnosis of diabetes using OLAP and data mining integration. *Int. J. Comput. Sci. Commun. Netw.* **2**(3), 314–322.

# Enhanced Goodput and Energy-Efficient Geo-Opportunistic Routing Protocol for Underwater Wireless Sensor Networks



V. Baranidharan, B. Moulieshwaran, V. Karthik, R. Sanjay,  
and V. Thangabalaji

**Abstract** The acoustic transmission in the aquatic environment is inherent for communications in underwater wireless sensor networks (UWSN). These acoustic waves based underwater wireless sensor networks are having many challenges due to their large propagation delay, energy consumption is also comparatively high and the available bandwidth is also very less. The major disadvantage of this UWSN is high latency due to its channel impairments. In this paper, the geo-opportunistic routing protocols are proposed for maximizing the goodput and to improve the residual energy range. In these routing protocols, we proposed the new metric called EEFL to ensure at least one forwarder of  $F_i$  to receive the packets successfully without loss. The proposed routing protocols formulates the OR problem based on the non-linearization model. In order to solve these issues, we proposed a multi-step heuristic algorithm that is implemented and composed in each and every node for calculating the forwarding set determination and prioritization. The proposed routing protocols are evaluated by analyzing its performance metrics and it shows that this proposed scheme is outperforming than the existing geo-opportunistic routing protocols in terms of energy and energy costs.

---

V. Baranidharan (✉) · B. Moulieshwaran · V. Karthik · R. Sanjay · V. Thangabalaji  
Department of Electronics and Communication Engineering, Bannari Amman Institute of  
Technology, Sathy, India

B. Moulieshwaran  
e-mail: [moulieshwaran.ec18@bitsathy.ac.in](mailto:moulieshwaran.ec18@bitsathy.ac.in)

V. Karthik  
e-mail: [karthik.ec18@bitsathy.ac.in](mailto:karthik.ec18@bitsathy.ac.in)

R. Sanjay  
e-mail: [sanjayr.ec18@bitsathy.ac.in](mailto:sanjayr.ec18@bitsathy.ac.in)

V. Thangabalaji  
e-mail: [thangabalaji.ec18@bitsathy.ac.in](mailto:thangabalaji.ec18@bitsathy.ac.in)

## 1 Introduction

Underwater Wireless Sensor Networks are encompassing various number of applications are formed for long-term gradients obtained from the seawater or underwater scenarios. There are a large number of sensornodes are employed over the oceanic environments and terrestrial areas in the underwater scenario [1]. Knowledge should be gathered from all the counterparts of the sensors, terrestrial networks (ie. Classical Wireless sensor Networks), those sensor networks are cannot be directly used in the field of Underwater acoustic environment based research because the transmission media and the type of signals used are not supports in underwater environment. In all other networks, there is more number of critical applications in different environments such as pollution monitoring. Submarines detection, weather forecasting, etc. the underwater channels are not reliable because of the path loss, noise, data rate is very low, multipath and high probability for occurring the error in the received signals [2].

A good, novel, and effective design of energy-efficient routing protocol will enable the effective monitoring of the underwater environment and route the data packets effectively in UWSN. Enhanced Goodput and Energy Efficient based Geo-Opportunistic Routing protocol (UWWOR) is one of the promising approach to provide eth reliable data communication to route the data packets in underwater environments. These OR protocols select the suitable neighbors as forwarding sets and enables them to route the packets to its neighbors. This will improve the one-hop reliability by more than employing even a single node and this will be delivered the data to the respective sink node in the monitoring stations. The major issues in this OR protocols are,

1. Forwarding set selection—To choose the proper forwarding node from its neighbor node based on its distance.
2. Relay Prioritization—Based on its distance, the next forwarder node has a high prioritization in the forwarder set.

The performance of this proposes high delay and energy cost-sensitive OR routing protocol to address these two issues. In this proposed work, the Geo-opportunistic routing schemes are used to maximize the good-put of the effective data received and it also satisfies the end to end latency related issues and its requirements for the routing protocols [3, 4]. The number of deadlines in the deployed sensor nodes are also minimized the energy consumption of sensor nodes. This modified UWOR is focused on the delay and Energy cost in terms of energy consumption. EEFi metric was introduced in this method. This metric will select the appropriate sensor node in the data forwarder set to choose the best neighbors and it also ensures the packets will be forwarded to at least one of its neighboring node. This routing protocol will have proposed the heuristic routing algorithm to compose the forwarding of the set neighbor effective nodes determination and data packet forwarding and assigning the various prioritization. The results of the performance metrics are shows that this proposed scheme outperforms more effective than the existing systems in terms of the network good-put and energy costs.

The research articles are organized as follows, Sect. 2 gives the details about the UWSN routing protocols by reviewing the related papers. Next, in Sect. 3 explains the mathematical modeling of the optimized objective function and the heuristic solution of the correlated algorithms is discussed. The performance metrics are evaluated in Sect. 4 and comparison with other existing routing protocols is discussed. The paper is concluded with future scope in Sect. 5.

## 2 Related Work

There are many different routing algorithms and schemes are designed to deal with the unique characteristics of the Underwater Wireless Sensor Networks including very high propagation delay in the received signals, it has a very low data rate (up to 100 kHz), and high error probability. Some of the researches are discussed here,

VBF routing protocol was proposed from the geographic information based routing approach for UWSN. It uses the geographic routing algorithm; this algorithm will do not requires any type of the state information about all other deployed sensors. This routing involves and allowed to deploy the very limited number of the sensor nodes. This routing has more energy consumption of the sensor nodes for transmission based on its geographic position [5]. In an energy-efficient layer-based routing protocol [6, 7] is proposed to find the position of the relay and intermediate nodes and this will also improve the transmission energy of the sensor network, the maximum probability of effective and successful data transmission of information packets, and it will improvise the energy consumption over the deployed region. The relay nodes are used for the route the data packets from the source node to its destination node deployed in the UWSN. The major disadvantages of this routing protocol will increase the end-to-end delay.

The Depth Based Routing protocol (DBR) [8] is proposed to continuing the challenge in UWSN for obtaining the location information of underwater sensor nodes. The DBR routing protocol is requiring only the local depth information of the sensor nodes. This information can be easily obtained at the time of the localization process (i.e. Beaconing). The sonobuoy is floated over the sea surface in the ocean. The sink or monitoring stations are drifted over the surface of the underwater or oceanic environments and the sender node (ie. The node in which the event occurs) will select the next hop forwarder by using greedily method based on the lowest pressure among the neighbors. The modified algorithm is beside the process of this modified algorithm is to find and select the forwarding nodes in the UWSN. The IDBR algorithm proposed can be easily computed and obtains the location information of the neighbors with the calculated depth of the deployed sensors. The receiving node or sink or monitoring stations node are always a sonobuoy which is get drifted over the environment surface, and after the calculation of depth, the sensor information is transmitted to a sender node greedily to select the forwarder node by calculating the lowest pressure values among the neighbors in the sensor networks. This proposed mechanism will

develop to predefined size of the data packet and redundancy or duplication of the packets are also reduced in the underwater communications [9, 10].

In this paper, the author employs the pressure information from the deployed sensors in order to develop the anycast data transmission based routing algorithm. This algorithm is called as a Hydro cast routing algorithm [11]. This routing protocol is based on the dead-end data recovery method which is widely used to enhance the data packets delivery rate by avoiding the retransmissions of the data under the unreliable, complex, constraints over the underwater communication environments. Even though, this routing protocol will get effects the coordination delay with end-to-end latency. In this MPT (Multi-path Control Transmission) routing protocol scheme which will combine all the power control strategies based on the multipath routing protocols [12, 13]. The sensor nodes on the multi-path to its destination to adjust the transmission power to adjust its transmission and reception power for data packets aggregation in center nodes or sink. The major disadvantage of these routing protocols is that having high end-to-end latency for the delay-sensitive applications. This will focus to adjust its power in the transmission power in order to minimize the end to end delay. This paper exploits the Geo-opportunistic routing for the UWSN for the QoS provisioning because of its end-to-end reliability, delay and power consumption constraints with Underwater wireless sensor networks. The routing optimization problem is formulated which is changed into a multi-objective multi-constraint optimization problem that is addressed in the EGOR. These protocols will consider the coordination of the delay among the forwarder to reduce the end-to-end delay and guaranteed the packet delivery ratio.

### 3 Geo-Opportunistic Routing Protocol Design

Consider the network model which consists of the various sensors consisted nodes that are placed randomly over the entire deployment region of the underwater environment. The sensor nodes are represented as  $S_N = \{S_1, S_2, \dots, S_N\}$  having the  $N$  represents the total number of sensor node deployed in the underwater environments. Whenever the sensor nodes are receiving the data packets, it will compute the forwarding the data packets to its sensor nodes and priority values of the relay nodes  $F_i = \{f_{i1}, f_{i2}, \dots, f_{iN}\}$ , where  $f_{ij}$  is always represents the data forwarder node of the  $j$ th level of the highest priority. So, the  $T_d$  is denoted the total time remaining before the data packet deadline at the destination and the allowable time. (ie. The ratio of data packets good-put is increased). This routing protocol is first used to determine its relay priorities from the given set of forwarders. Here, a cluster-based forwarding set selection algorithm is used to choose the best neighbor in its consideration set [14, 15]. This will include an iterative prioritization scheme for priority-based data transmission schemes. In this last step, the next hop sensor node data forwarding set algorithm is always chosen with the best hop among the deployed various sensor nodes in one-hop by hop reliability from these data aggregated clusters or centered node from actual forwarding set for its neighboring nodes.

### 3.1 Packet Forwarding Prioritization

The geo-opportunistic routing protocol EELS ( $F_i$ ) is defined as an objective function of the relay prioritization  $F_i$  value and adjacent sensor nodes prioritization values. These relay priorities are maximizing this metric. The data forwarding set of sensor nodes  $F_i$  values are always given at the time of simulation, the terms of the performance metric values are do not change among the candidate nodes [16, 17]. In order to achieve the goal, the packet forwarding the priority, to maximize the  $P_{di} \rightarrow F_i$ . The metric values can be effectively simplified by ignoring the constant values in the objective functions in terms of the error probability value occurrences of the  $j$ th forwarder relays as shown below,

$$EELi = \sum_{j=0}^{|F|-1} \left( EELi + d(i, j) + \sum_{j=0}^{|F|} (d(i, j) + f^{k-1}) * p[fwd = f^j] \right) \quad (1)$$

According to the relay or sensor nodes priority, this algorithm is always possess highly probability values for the  $p[fwd = f]$  is always larger than the  $p[fwd = f_j^{j+1}]$ . The  $EEF_F$  of the higher order prioritized more suitable data forwarders node which has a great impact value on  $EEL^- (F_i)$ . the higher relays or sensor nodes the priority values should be assigned randomly or based on the distance between the data forwarding nodes in which has the smaller value of the total energy cost and time for consumed over the propagation delay to the pre-forwarder and the value the  $EEL^+$  having the algorithms in the UWMGOR relay prioritization procedure.

### 3.2 Forwarding Set Determination Algorithm

The proposed routing protocols are optimized by the objective function problem, the objective function is always maximized by the one-hop reliability. The forwarding set can be improved for this routing protocol for this performance based on adding the sensor nodes. This forwarding the data packets to its neighbor nodes through the relay nodes to prevent the packet duplication. For finding the forwarding data set, the optimization-based objective problem in a comprising of NP value variation over the optimized objective problem. The clique and the most important problem is the special case of the optimized problem in which  $EEF_l$  of the sensor nodes and the propagation delays. The distance value is determined to calculate the difference between the two successive next hop forwarders or sensor nodes should be taken into the close proximity values to avoid its excessive latencies coordination in the sensor nodes. Therefore, the cluster heuristic algorithm is used to determine the forwarding set for each sensor nodes is used. The algorithm 2 is used to expansion of  $C_j$  which will proceeds to no node left until the  $EEL_i$  (Success) will exceeds the dead line. After this, the cluster is formed to next hop reliability is selected to the  $i$ th node.

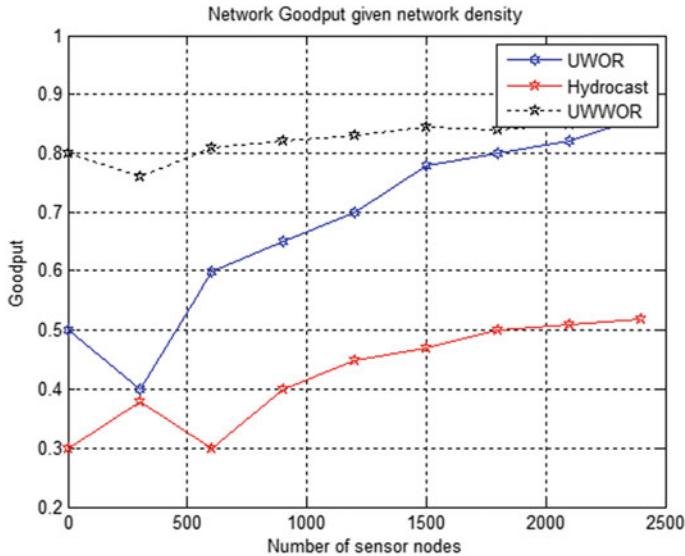
## 4 Performance Metrics Evaluation of the Proposed Routing Protocols

In this section, the performance metrics of this proposed novel routing schemes are compared with the existing routing protocol for UWSN. The sensor is having the hardware parameters both acoustic and RF modems. The transmission and reception powers are initially given as 50 and 20 W. We apply and generate different topologies and its average performance is analyzed by using NS2 simulator. For each run, select  $N$  events to generate packet from the source nodes to its destination through  $m$  intermediate nodes. There are 100 sensor nodes are employed randomly over the entire region. The deployment region is  $5000 \times 5000$  m. The packet size is 32 Bytes and its transmission range is 200 m. The most important performance metrics are evaluated in this work.

In order to understand the more energy consumption of the geo-opportunistic routing algorithm schemes of isolating the effects of the deadline constraints. In order to evaluate our proposed algorithm, the performance metrics of these methods are achieved by the other methods of some other existing routing protocols. The proposed Geo-opportunistic Routing algorithm which is used to jointly considered with deadline and delays produced during the propagation of the transmitted or the received signal for the various delay-sensitive of the UWSN applications. The parameters are analyzed with the end-to-end latency value estimated of the pre-processing as 0.95. In order to analyze the performance of the proposed algorithm affects the behavior of the wireless sensor networks.

### 4.1 Network Goodput

Network Goodput is defined as the amount of useful data packets received by the total packets transmitted across the networks. In this, proposed routing protocol outperforms than the existing routing protocol is equal to 1.2 times than the existing systems. This proposed routing protocol gives almost 80% than the existing routing protocols of underwater wireless sensor networks. It also reduces the latency across the networks reaching up to 3.2 times than the existing routing protocols. This is because of the less propagation delay between the nodes and having good coordination among the sensor nodes. This is due to the proper choosing of the coordination of the forwarding nodes and set of relay priorities over the existing routing protocols which is depicted in Fig. 1. The existing protocols are applied to avoid retransmission. latencies in the networks. In order to analyse the performance metrics of the existing routing protocols systems, the various node densities in this proposed routing protocols ability to scale the large-scale dimensions. The number of the sensor nodes from the 600 to 2400. This improved routing protocol yields 50–68% having high good put than the existing systems to improve the Geo-opportunistic routing to handle the wireless communications.

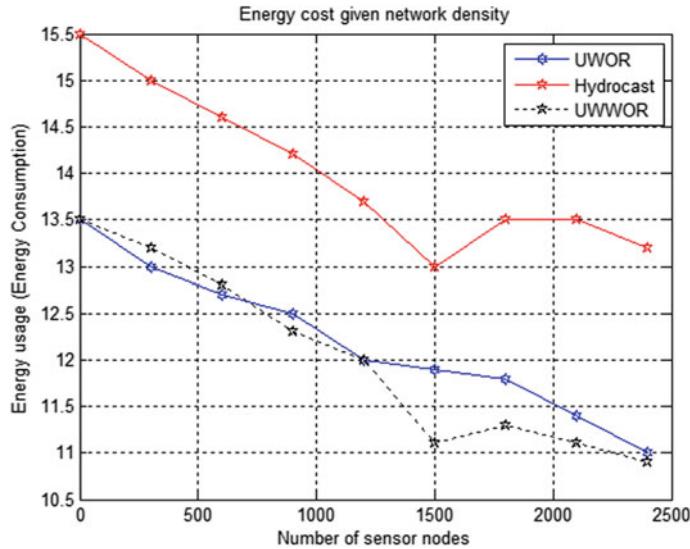


**Fig. 1** Network goodput given network density

## 4.2 Energy Consumption

The energy consumption is defined as the total energy consumed by the sensor nodes over the entire network. There are  $N$  number of the sensor nodes are deployed across the entire network. After varying the high-density networks into the low-density network (i.e., the density of the network), we evaluate the performance of the proposed routing protocols in both the small and large areas to maintain the density of the nodes. By rapid increasing of the sensor node density values, this will directly affect the number of nodes to forward the data packets from the source node to its destination node based on the position information. As the number of packets increase this will improve the consumption of the energy of each and every sensor node. The transmission or communication range of the sensors will imply the correlation between the number of hops and its distances of the sensor nodes. After the initial high performance for very small regions, the performance results will extend to wider the range and it will outperform than the existing schemes.

Figure 2 shows that the size of the packet increases it will be increasing the energy consumption of the sensor nodes and it will decrease the effect of the data packets in which it will reach the monitoring station or the sink before the deadline. The size of the data packet is 64 bytes is every good option for successful data packets at the sink.



**Fig. 2** Energy cost given network density

## 5 Conclusion

The paper discusses the data transmission problem in the Underwater Wireless sensor networks (UWSN): One in which the data packets is useful only as to arrives its destination or sink node before the expenditure allotted time. The energy expenditure is a key factor of any network of the Underwater scenario so it will be often deployed randomly in extremely harsh and highly dense underwater environments to access the battery charges. The geo-opportunistic routing is used to select the packet forwarders each sensor nodes with the neighbor's nodes in a reliable manner in terms of the energy cost and goodput. The performance metrics are analyzed with the existing protocols in many scenarios. The results will show the results is outperforming than the existing routing protocols. In this routing protocols, the time between the maintenance and its simply increase the network lifetime. This proposed routing protocols will improve the residual energy by letting them to progress for selecting the ore suitable forwarding path will be discarded that to reach the sink node. The simulated results show that the proposed routing protocol will have energy-s decongest the data packets would automatically increase the expectation of successful data packet reception in a very dense environment of the underwater wireless sensor networks.

## References

1. Jiang, M., Guo, Z., Hong, F., Ma, Y., Luo, H.: OceanSense: a practical wireless sensor network on the surface of the sea. In: Proceedings IEEE International Conference on Pervasive Computing and Communications (PerCom), pp. 1–5 (2019).
2. Noh, Y., Wang, P., Lee, U., Torres, D., Gerla, M.: DOTS: a propagation delay-aware opportunistic MAC protocol for underwater sensor networks. In: Proceedings 18th IEEE International Conference on Network Protocols (ICNP), pp. 183–192 (2010).
3. Pompili, D., Melodia, T., Akyildiz, I.F.: Routing algorithms for delay-insensitive and delay-sensitive applications in underwater sensor networks. In: Proceedings 12th ACM Annual International Conference on Mobile Computing and Networking (MobiCom), 2006, pp. 298–309 (2006).
4. Hsu, C.-C., Lai, K.-F., Chou, C.-F., Lin, K. C.-J.: ST-MAC: spatialtemporal MAC scheduling for underwater sensor networks. In: Proc. 28th IEEE International Conference on Computer Communications (INFOCOM), April 2009, pp. 1827–1835 (2009).
5. Xie, P., Cui, J.-H., Lao, L.: VBF: vector-based forwarding protocol for underwater sensor networks. In: Proceedings 5th International IFIP-TC6 Networking Conference NETWORKING, 2006, pp. 1216–1221 (2006).
6. Hsu, C.-C., et al.: Delay-sensitive opportunistic routing for underwater sensor networks. *IEEE Sens. J.* **15**(11) (2015).
7. Gopi, S., Govindan, K., Chander, D., Desai, U.B., Merchant, S.N.: E-PULRP: energy optimized path unaware layered routing protocol for underwater sensor networks. *IEEE Trans. Wirel. Commun.* **9**(11), 3391–3401 (2010)
8. Yan, H., Shi, Z.J., Cui, J.-H.: DBR: depth-based routing for underwater sensor networks. In: Proceedings of the 7th International IFIP-TC6 Networking Conference Ad Hoc Sensor Networks, Wireless Networks, Next Generation Internet (NETWORKING), 2008, pp. 72–86 (2008).
9. Yu, H., Yao, N., Gao, Z., Lu, Z., Chen, B.: Improved DBR routing protocol for underwater acoustic sensor networks. *Sens. Lett.* **12**(2), 230–235 (2014)
10. Coutinho, R.W.L., Boukerche, A., Loureiro, A.A.F.: PCR: A power control-based opportunistic routing for underwater sensor networks. In: Proceedings of the 21st ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, pp. 173–180 (2018).
11. Rahman, Z., Hashim, F., Rasid, M.F.A., Othman, M.: Totally opportunistic routing algorithm (TORA) for underwater wireless sensor network. *PLoS ONE* **13**(6) (2018).
12. Moradi, M., Rezazadeh, J., Ismail, A.S.: A reverse localization scheme for underwater acoustic sensor networks. *Sensors* **12**(4), 4352–4380 (2012)
13. Lee, S., Kim, K.: Localization with a mobile beacon in underwater acoustic sensor networks. *Sensors* **12**(5), 5486–5501 (2012)
14. Baranidharan, V., Sivaradje, G., Varadharajan, K.: Void node recovery based geographic-opportunistic routing for underwater sensor networks. *J. Adv. Res. Dyn. Control Syst.* **5**, 323–332 (2018)
15. Baranidharan V., Varadharajan, K.: Secure localization using coordinated gradient descent technique for underwater wireless sensor networks. *ICTACT J. Commun. Technol.* **9**(01), 1716–1720 (2018).
16. Baranidharan V., Sivaradje G., Varadharajan, K., Vignesh, S.: Clustered geographic-opportunistic routing protocol for underwater wireless sensor networks. *J. Appl. Res. Technol.* **18**, 62–68 (2020).
17. Shakeel, U., Jan, N., Khizar, Qasim, U., Khan, Z.A., Javaid, N.: DRADS: depth and reliability aware delay sensitive routing protocol for underwater WSNs. 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing, pp. 78–83 (2016).

# Early Detection of Pneumonia from Chest X-Ray Images Using Deep Learning Approach



**Prateek Sarangi, Pradosh Priyadarshan, Swagatika Mishra, Adyasha Rath, and Ganapati Panda**

**Abstract** With an objective to save the lives of children from pneumonia, many efforts are being made by doctors and medical practitioners. Early detection and subsequent appropriate care will reduce the number of deaths of the pneumonia affected children. In recent past, machine learning and artificial intelligence-based methods have been introduced for decision making in various applications including health care. With this motivation, the research work in mind this paper has been carried out, and an efficient deep learning model is developed for classification between healthy and pneumonia patients. Using standard database comprising of chest x-ray images, the proposed model is trained and tested. The validation results reveal that the proposed model provides consistent performance in terms of classification accuracy. The analysis of the simulation-based results reveals that for a batch size of 16 and for two number of hidden layers in the classification stage of the CNN model yields highest validation accuracy of 93.73%.

## 1 Introduction

The World Health Organization (WHO) has reported that pneumonia kills maximum number of children and elderly people across the globe. The fatality rate due to pneumonia is nearly four millions. The pneumonia causes inflammation in lungs which may lead to death, if the diagnosis is not done in time. Proper examination of chest x-ray is an important diagnosis method which is being followed worldwide. Detection of pneumonia by examination of chest x-ray is time consuming as well as

---

P. Sarangi (✉) · P. Priyadarshan · S. Mishra  
Veer Surendra Sai University of Technology, Burla, Odisha 768016, India

A. Rath  
Department of Computer Science and Engineering, Siksha O Anusandhan (Deemed to be)  
University, Bhubaneswar, Odisha 751030, India

G. Panda  
Department of Electronics and Tele Communication, C. V. Raman Global University,  
Bhubaneswar, Odisha 752054, India

less accurate. In addition, the analysis of x-ray images is a tedious and critical task. To alleviate this problem, computer-based diagnosis tools have been developed to analyze the chest x-ray images. But these tools are also not very effective as well as less informative. In the recent past, many machine learning techniques have been successfully applied for diagnosis of diseases and other healthcare problems. Appreciable numbers of research articles have been reported on detection and diagnosis of pneumonia using images of chest x-rays. Deep learning and convolution neural network have emerged as promising methods for many engineering application, finance, manufacturing and health care. Mostly, the deep learning approach is employed for classification, forecasting and detection purpose. It is observed that many articles have been reported in the literature on the diagnosis of pneumonia.

In the next section, a review of articles on machine learning-based pneumonia detection is presented.

## 2 Review of Literature

A texture-based identification scheme of interstitial pneumonia from multi-detector CT is presented in [1]. A KNN classifier is employed to identify normal, ground class and reticular categories. It is reported that the proposed system can detect, characterize and quantify interstitial diffuse parenchymal lung disease (DPLD). In [2], the authors have proposed evolutionary fuzzy cognitive maps for forecasting of pneumonia infection. The proposed approach has provided better accuracy of prediction of the infection using real-life patient data. For diagnosis of childhood pneumonia, cough analysis has been reported [3] using wavelet-based features. The authors have employed logistic regression classifier and have used Mel cepstral coefficients as input to the classifier for clinical decision. The simulation performance using real-life data is reported to be better than the WHO criteria. In another interesting paper [4], the authors have suggested a deep learning approach for identification of pneumonia. The chest x-ray images are used in the Mask-RCNN model for obtaining global and local features and subsequently identifying the pneumonia. Improved identification performance is obtained by the method developed in the paper. The ECG signal is used along with statistical evaluation of myocardial enzyme for treating pneumonia patients [5]. It is reported that abnormal changes in ECG show that myocardium is affected after pneumonia. The detection of creatine kinase, isozyme as well as troponin level in pneumonia patient shows high sensitivity and specificity. A deep learning model has been developed in [6] which can screen the CT images of COVID patients automatically. This model also classifies COVID, influenza-A viral pneumonia (IAVP) or healthy cases. An overall accuracy of 86.7% has been achieved by this model. For diagnosis of pediatric pneumonia, a deep residual network using transfer learning method is suggested [7]. The proposed method is reported to identify the appropriate area of image which indicates the presence of pneumonia. The test results using real-life datasets show higher accuracy and F1-score compared to other reported articles. In [8], a multi-dilation CNN-based model is developed for

the detection of pneumonia and COVID-19 using chest x-ray images. In this paper, the authors have employed transferable multi-receptive approach for optimization of the features. It is reported that the suggested approach provides satisfactory detection performance with high accuracy. The combined CNN and transfer learning are used [9] for detection of pneumonia from chest x-ray images. Six different models have been suggested by the authors. It is observed that model 2 and VGG 19 networks provide the best performance and hence can be used by medical practitioners. Finally, in [10], a systematic review and analysis of detection of pneumonia from chest x-ray images using deep learning-based approach is presented. The exploratory meta-analysis shows that the deep learning approach can reliably classify pneumonia as well as can distinguish between bacterial and viral pneumonia.

### **3 Research Gap, Motivation of Research, Research Objective and Organization of the Paper**

#### ***3.1 Research Gap***

The review of related articles on prediction and diagnosis of pneumonia reveals that many researchers across the globe are working on early detection and proper diagnosis of pneumonia particularly from images obtained from chest x-ray. It is also observed that many recent articles in this area employ deep learning and CNN-based approaches for prediction of this disease. Even though many acceptable results have been reported on the prediction of this disease, there is a further scope improvement in improving the accuracy of classification between healthy as well as patients suffering from pneumonia infections. Further, there is a need to suggest and validate a robust CNN prediction model to provide reliable and robust performance of pneumonia detection by varying the batch size at the input as well as varying the number of hidden layers associated with the CNN.

#### ***3.2 Motivation of Research***

Identifying the above-stated needs has motivated the authors to develop a well-performing CNN-based classification model which would provide consistent and reliable performance of detection of pneumonia. This model can be recommended to be used for Web-based application which can be conveniently used by doctors for prediction of pneumonia.

### ***3.3 Research Objective***

Based on the motivation of the research, the work of the proposed paper is carried out to fulfill the following objectives:

- To identify and develop an appropriate CNN model which would provide consistent detection performance.
- To train and validate the CNN model using standard datasets pertaining to chest x-ray images from standard datasets.
- To use an augmentation method to scale all images to a standard size of 150\*150 pixel values to be used in the CNN.
- To investigate on the effects of batch sizes as well as the number of hidden layers of the CNN. So that a best possible CNN model can be chosen which would provide the best validation accuracy.
- To outline the major contributions of the proposed deep learning-based detection model of pneumonia from standard chest x-ray images.

### ***3.4 Organization***

Keeping the above-cited research objectives in view, the organization of paper proceeds as follows: In Sect. 4, the methods and materials related to the investigation are presented. This section deals with the method employed for detection, details of the dataset and the pre-processing of the data. In Sect. 5, the details of the development of CNN model comprising of both training and validation are presented. This section also outlines the details of the CNN model and the purpose of using different layers. The details of the simulation-based experiments are explained in Sect. 6. This section also presents the various experimental results obtained from the simulation study using x-ray images of both healthy and pneumonia patients. In Sect. 7, the analysis of the results is presented and the contribution of the paper is outlined. Finally, the conclusion of the paper is presented which includes the weakness of the work as well as the scope of further extension of the current work.

## **4 Methods and Materials**

### ***4.1 Method Used***

To develop and implement the proposed pneumonia detection model, the convolution neural network has been chosen as the basic classification model. If the dataset is either huge or comprises of signals and images, then CNN or deep learning-based models are better candidate for classification and prediction purpose. It has been

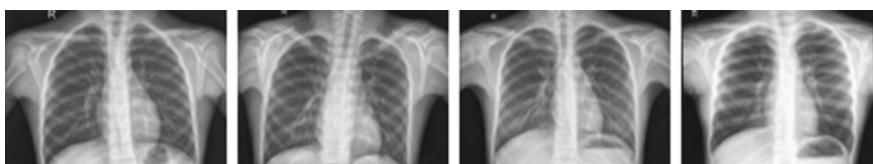
reported that CNN-based technique has been successfully applied for classification of skin cancer, hemorrhage identification, arrhythmia detection, diabetic retinopathy and pulmonary tuberculosis. In the present case, an appropriate CNN model is chosen for two primary purposes: features extraction through series of convolution operations and classification by employing few layers of fully connected artificial neural network.

#### 4.2 Materials Used

The materials required in the present paper are the chest x-ray images comprising of normal and pneumonia patients. These images are standards and have been taken from Kaggle. The details of the data are: A total of 5856 x-ray images of anterior-posterior chest have been taken from child patients between 1 and 5 years old. The entire image data have been grouped into training and validation groups. Out of total images, 3722 images have been used for training purpose and the remaining 2134 x-ray images have been employed for validation purpose. As an illustration, the chest x-ray images of four normal children are displayed in Fig. 1. Figure 2 shows the chest x-ray images of four pneumonia affected patients.

#### 4.3 Pre-processing of the Data

The images have been scaled appropriately using data augmentation method. The scaling has been carried out to bring all chest x-ray images to have equal sizes of



**Fig. 1** X-ray chest images of four patients without pneumonia

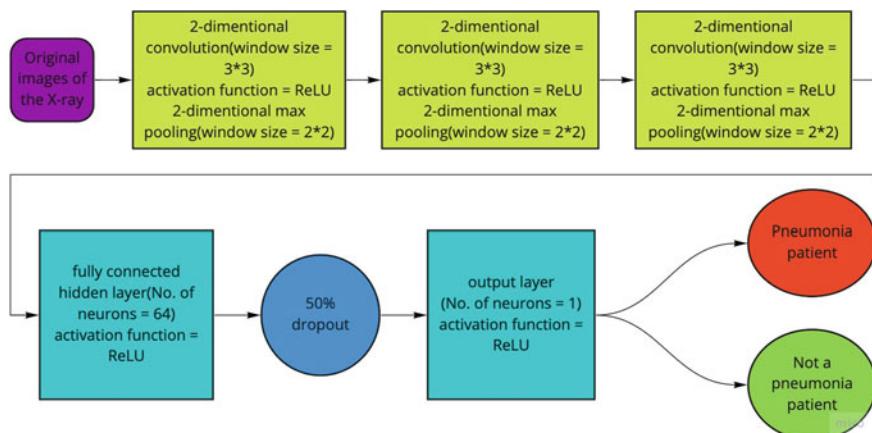


**Fig. 2** X-ray chest images of four patients with pneumonia

150\*150. The standardized images have been enabled the training in a single model, and hence the training process becomes faster. The image data of equal sizes are then fed to the proposed CNN model for training and validation purposes.

## 5 Development of CNN Detection Model

The overall proposed architecture of the CNN detection model is presented in Fig. 3. The CNN model essentially performs two key tasks: (i) extraction of features (ii) classification or detection of disease. The raw equal-sized chest x-ray images are fed to the CNN model. In the features extraction stage, each layer takes the output of the preceding layer as its input. The output of one layer is passed as an input to the succeeding layers. Such processing of image data in CNN is called sequential processing. The architecture for the CNN model consists of convolution, max pooling and then classification layer. The feature extractor contains 32 layers of Gaussian filters each of size 3\*3. It is then followed by max pooling block of size 2\*2, convolution block of 32 layer of Gaussian filter of size 3\*3, max pooling of size 2\*2, 64 layers of Gaussian filter and finally the max pooling block of size 2\*2. After the feature extraction stage, the output is flattened to form a single layer. It is then followed by a classification layer consisting of 64 densely connected hidden neurons. A random dropping of 50% of the neurons is used to deal with regularization in the classification layer. The ReLu activation function is used to generate the desired output class. In the next section, the simulation-based experiments have been carried out, and the classification results obtained are tabulated and analyzed.

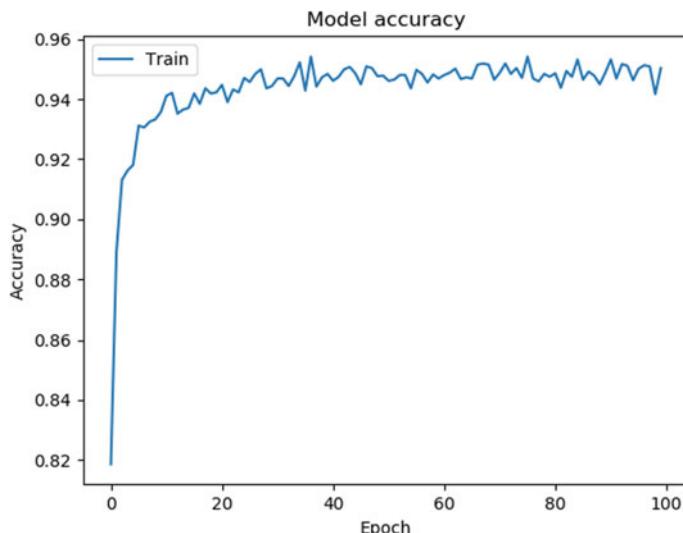


**Fig. 3** Architecture of CNN model used for detection of pneumonia of children

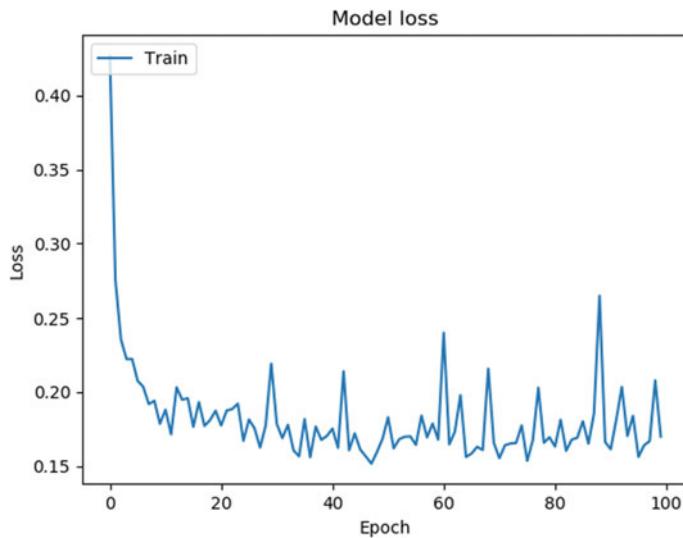
## 6 Simulation-Based Experiments

In this section, the block diagram of CNN model as shown in Fig. 3 is simulated. The detailed architecture of the CNN detection model is dealt in Sect. 5. The chest x-ray images obtained from the Kaggle pneumonia datasets are used as inputs to the model after the images are scaled to the same size of 150\*150. As discussed in the material, Sect. 4.2, a total of 5856 chest x-ray images are available in the dataset out of which 3722 number of images were used for training purpose. The remaining 2134 images were employed for the purpose of model validation. The training image sets comprising of both normal and pneumonia patients are fed to the model sequentially till the estimated class is achieved. The images are transformed into the desired features after passing through series of convolution layers followed by max pooling alternatively. The extracted and flatten features are passed through the classification layers comprising of 64 numbers of densely connected hidden neurons. Adopting the strategy of 50% dropouts and passing through ReLu activation function, the estimated class of each child is obtained. As mentioned earlier, the output is divided into two classes as healthy and having pneumonia disease. During the training phase, the accuracy of classification is evaluated with variation in number of epochs. RMSProp weight optimization is used in the model. The corresponding plot obtained from the simulation is plotted in Fig. 4. In the similar way, the loss function at the output during training phase is determined by varying the number of epochs. The corresponding plot is displayed in Fig. 5.

The performance of the model is also evaluated during testing or validation phase. In the present case, 2134 chest x-ray images have been used as input to the model



**Fig. 4** Plot of accuracy achieved by the CNN model during training phase



**Fig. 5** Plot of model loss obtained during training phase

**Table 1** Effect of batch size on the accuracy of classification during training and validation phase

Batch size	Training accuracy	Validation accuracy
4	93.72	89.47
8	94.22	91.64
16	95.31	93.73
32	91.75	87.21

**Table 2** Effect of the number of neurons in the hidden layer of the classification stage on the accuracy of classification

Number of hidden layers	Training accuracy	Validation accuracy
1	92.41	90.42
2	95.31	93.73
3	91.75	86.56
4	88.27	84.72

to find its output class. The validation accuracy as well as the training accuracy has been found for different batch sizes varying from 4 to 32. The accuracy performance in percentage both during training and validation periods are presented in Table 1. In addition, the numbers of hidden layers of the classification stage of the model have been varied from 1 to 4. In each case, the training and validation accuracy of classification is determined and listed in Table 2.

## 7 Analysis of the Results

It is observed from the plot of Fig. 4 that at around 35 epochs the accuracy obtained during training phase almost remains constant for input batch size of 16. The training accuracy at this stage is 95.31%. However, it is observed from Table 1 that the classification accuracy is lowest for batch size of 32. So in terms of percentage of accuracy, the batch sizes are 16, 8, 4 and 32, respectively. Similar observation is also made in case of validation accuracy. In other words, batch size of 16 yields highest accuracy of 93.73%. It is then followed by batch sizes of 8, 4 and 32. It is, in general, observed that the relation between training and validation accuracy with the number of batch sizes is consistent. Another interesting observation is that the validation accuracy for each batch size is less than training accuracy.

This is because the number of input test patterns used in the study is less. This difference can be reduced if the number of training input images is increased. Further, k-validation scheme can be adopted to assess the robustness of the proposed CNN model. It is observed from Table 2 that when the number of hidden layers is two then both the training and validation classification accuracy are the highest. The corresponding percentages of accuracy are 95.31 and 93.73%. In this case also the validation accuracy is lower than the training accuracy. To find the consistent performance of the proposed model, testing can be made using other similar standard datasets. In general, it is found that the highest validation accuracy from the proposed CNN model is 93.73% achieved for a batch size of 16. Similarly, when the number of hidden layers is two, the validation accuracy is found to be the highest. Since the validation accuracy is high, the model is not overfitted.

## 8 Conclusions

This paper has investigated on the development of a CNN-based model for detection of pneumonia using chest x-ray images of healthy and pneumonia affected children. The paper has employed standard chest x-ray images of children belonging to age group 1–5 years old comprising of healthy and diseased children. After successful training of the proposed model, the validation of the model has been carried out using test images. The effects of both batch size and number of hidden layers on the accuracy of classification have been investigated. It is demonstrated that with two hidden layers the validation accuracy is the highest (93.73%). Also, it is observed that the developed model offers the same highest validation accuracy when the batch size is 16. The robustness of this model can be determined by evaluating the performance of the model using other datasets as well as larger datasets. In this paper, the classification stage comprises of artificial neural networks. The performance of the model can also be determined by using LSTM or other classification techniques combined with convolution layers and the accuracy of classification can be obtained and compared.

## References

1. Korfiantis, P.D., Karahaliou, A.N., Kazantzi, A.D., Kalogeropoulou, C., Costaridou, L.I.: Texture-based identification and characterization of interstitial pneumonia patterns in lung multidetector CT. *IEEE Trans. Inf. Technol. Biomed.* **14**(3), 675–680 (2009)
2. Papageorgiou, E.I., Froelich, W.: Application of evolutionary fuzzy cognitive maps for prediction of pulmonary infections. *IEEE Trans. Inf. Technol. Biomed.* **16**(1), 143–149 (2011)
3. Kosasih, K., Abeyratne, U.R., Swarnkar, V., Triasih, R.: Wavelet augmented cough analysis for rapid childhood pneumonia diagnosis. *IEEE Trans. Biomed. Eng.* **62**(4), 1185–1194 (2014)
4. Jaiswal, A.K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., Rodrigues, J.J.: Identifying pneumonia in chest X-rays: a deep learning approach. *Measurement* **145**, 511–518 (2019)
5. Li, Z., Li, X., Zhu, Z., Zeng, S., Wang, Y., Wang, Y., Li, A.: Signal analysis of electrocardiogram and statistical evaluation of myocardial enzyme in the diagnosis and treatment of patients with pneumonia. *IEEE Access* **7**, 113751–113759 (2019)
6. Xu, X., Jiang, X., Ma, C., Du, P., Li, X., Lv, S., Yu, L., Ni, Q., Chen, Y., Su, J., Lang, G.: A deep learning system to screen novel corona virus disease 2019 pneumonia. *Engineering* (2020)
7. Liang, G., Zheng, L.: A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Comput. Methods Progr. Biomed.* **187**, 104964 (2020)
8. Mahmud, T., Rahman, M.A., Fattah, S.A.: CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multi-receptive feature optimization. *Comput. Biol. Med.* **122**, 103869 (2020)
9. Jain, R., Nagrath, P., Kataria, G., Kaushik, V.S., Hemanth, D.J.: Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement* **165**, 108046 (2020)
10. Li, Y., Zhang, Z., Dai, C., Dong, Q., Badrigilan, S.: Accuracy of deep learning for automated detection of pneumonia using chest X-Ray images: a systematic review and meta-analysis. *Comput. Biol. Med.* **123**, 103898 (2020)

# Detection of Network Anomaly Sequences Using Deep Recurrent Neural Networks



R. Ravinder Reddy, K. Ayyappa Reddy, C. Madan Kumar, and Y. Ramadevi

**Abstract** The enormous growth of the Internet and its usage creates many vulnerabilities, so network anomaly detection becomes a crucial problem these days. With this tremendous growth of the internet applications makes the security is a top priority for everyone. An attack might cause loss of valuable information/assets to any individual or the organization, attacker may be insider or outsider of the organization. The challenging part is to identify the attack real-time and act accordingly to prevent the losses, either in the form of data or money. In the early age of the network intrusion detection systems (NIDS) statistical and data mining techniques are used to detect the intrusions. These techniques have the limitations, these approach can't perform well with the huge data. Deep learning approaches performance increases with the data size. These techniques can predict the occurrence of the attack more accurately. With the availability of huge processing capability and enormous network data deep learning algorithms are become more flexible in detection of these anomalies accurately. In this work, the recurrent neural network algorithm and its variants are used to detect the network anomalies.

## 1 Introduction

The network anomaly detection frequently change their pattern and more dynamic in nature, the attacker always changes pattern every time prior to attack. The dynamic nature of this behavior is difficult in the existing systems, like statistical and data mining approaches. To capture these dynamics of the anomaly more number of

---

R. Ravinder Reddy (✉) · K. Ayyappa Reddy · Y. Ramadevi  
Chaitanya Bharathi Institute of Technology, Hyderabad, India  
e-mail: [ravinderreddy\\_cse@cbit.ac.in](mailto:ravinderreddy_cse@cbit.ac.in)

Y. Ramadevi  
e-mail: [yrd@cbit.ac.in](mailto:yrd@cbit.ac.in)

C. Madan Kumar  
KITSW, Warangal, India

combinations have to be identified in the system with the available data. The deep learning approaches with the increased number of hidden layers can identify these patterns in more promptly in the system environment. With the deep learning applications, network anomaly detection approaches become an effective among the other intrusion detection systems. It searches for a specific pattern of anomaly, and that pattern does not match the other profiles. As of now, because of the tremendous development in PC systems and applications, numerous difficulties emerge for digital security research. Assaults can be characterized as a lot of occasions which can bargain the standards of PC frameworks, trading off the accessibility, authority, secrecy, and trustworthiness of the framework in any way. The firewalls are utilized to recognize the principal level of assaults just, these frameworks can't identify present-day assault situations and can't examine arrange parcels inside and out. To safeguard these assaults we required a powerful framework which will shield the framework from the assaults. On account of these reasons, IDSs are created to accomplish high insurance for the security foundation. Interruption recognition frameworks are the guard dogs of data frameworks [1]. IDSs are frameworks planned and modified to mechanize the way toward observing occasions occurring in a PC framework or arrange and examining them for potential security issues.

Network Intrusion Detection System (NIDS) frameworks become hot examination issues. As system practices and examples change and interruptions advance, it has especially gotten important to move away from static and one-time datasets toward all the more powerfully created datasets which not just mirror the traffic arrangements and interruptions of that time, but at the same time are modifiable, extensible, and reproducible. Preparation of such suitable data many network monitoring labs has released different datasets which contain the required information about the attacks. Captured Raw data of the network contains few attributes only. Most of the data is available in the hexa-decimal format only. Retrieving the features from the raw network is difficult. Some of the labs have simulated the networks and generated the dataset for intrusion detection. Some of the datasets like NSL-KDD, UNSW\_NB15, and ISCX contain information about different patterns of different attacks. The KDDCUP99 dataset is so famous but it is the old dataset as compared with ISCX and UNSB. In this work, we mainly focused on training Deep Learning Models like LSTM, RNN, GRU, and FNN with UNSW\_NB15 dataset to classify the network data by analyzing different patterns in it as normal and anomaly [2].

One of the primary issues is the plenitude of information in contemporary cybersecurity datasets which requires insightful calculations, for example, Deep Learning calculations, for removing important data [2]. In particular, its application to IDS includes the requirement for a high measure of highlights with the target to choose the best approach and identify the chance of an assault. The issue is significant, in light of the fact that a high number of qualities in a dataset lead to a model over fitting, thus transforming into helpless outcomes on the approval datasets. In this paper, we are going to utilize a benchmark dataset UNSW\_NB15 which is created by Australian Center for Cyber Security (ACCS) utilizing IXIA Perfect Storm device in their Cyber Range Lab.

The KDD98, KDDCUP99, and NSL-KDD were benchmark datasets created and utilized 20 years back; be that as it may, late investigations have demonstrated that these datasets have matured, implying that these datasets no longer comprehensively reflect cutting edge organize traffic, henceforth this work utilizes UNSW-NB15 [3]. This dataset comprises 49 segments including the normal and the assault class. It comprises of the 9 assaults, to be specific, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shell code, and Worms.

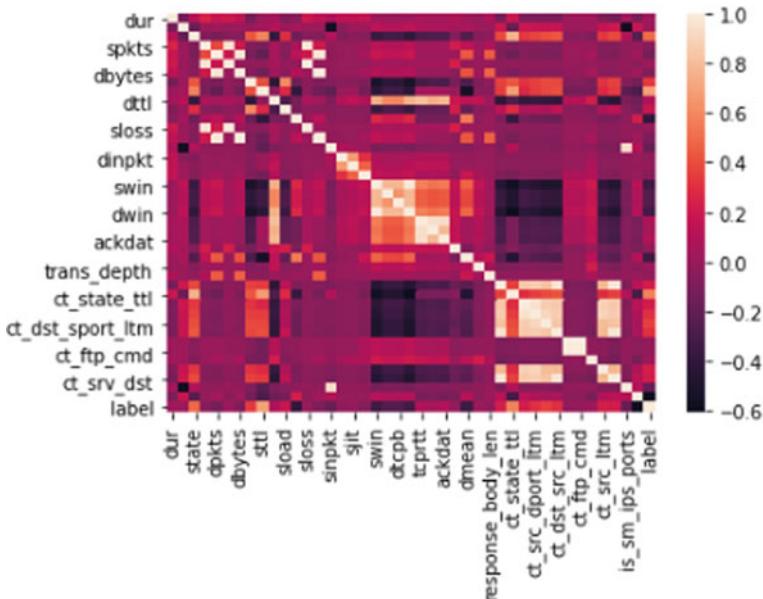
This work mostly centers on preparing the Deep Learning models like LSTM, RNN, GRU [4] with the UNSW\_NB15 Dataset. The Dataset comes in two sets preparing and testing. The preparation comprises of 175,341 records and the testing comprises of 82,332 records. A definite exploratory examination is done on the dataset utilizing T-SNE (t-disseminated Stochastic Neighbor Embedding) which shows the conveyance of assaults in two-dimensional planes.

## 2 Related Work

In the literature of the network anomaly detection, most of the works are done based on the signature-based approaches. Many of the researchers proposed statistical approaches in the initial days of the intrusion detection techniques. But later on with the machine learning and data mining approaches used to build the dynamic approaches like anomaly detection for adaptive models [5]. Later on, many of the techniques are used in this field, but significantly artificial neural networks show considerable improvement in the detection process. The success of deep learning techniques in image processing and other applications, motivated us to apply this for anomaly detection [6, 7].

With the huge success of artificial neural networks in different applications, it is applied to computer security. The main focus on the detection of intrusions in a network environment. The deep artificial neural networks are considered as an efficient approach these days for pattern classification and prediction. The main issue in these approaches are required high calculations and the long training cycles. But with the revolution in the hardware technologies like GPU, TPU's the artificial neural become popular in almost all applications. With the availability of high computing processing hardware given the scope for more research in the deep neural network techniques. The use of artificial neural networks systems to the security is for the most part centered around the discovery of interruptions in a system since neural network systems are viewed as an effective way to deal with design arrangement. The deep learning approaches are considerable good in network anomaly detection process [8–10].

The deep learning techniques require huge amounts of data, which is available from the public network. The nature of anomaly detection is dynamic, to extract the appropriate patterns from the system require huge learning data and suitable approach. The deep learning technique like recurrent neural networks works on the sequences in the data. These sequences identifies the appropriate anomaly patterns



**Fig. 1** Heat map of the UNSW\_NB15 dataset

and shows significance growth in detection rate. The recurrent neural network is used to represent the sequences in the communication entities and is used to identify the anomalies [11]. The variant of recurrent neural network of Gated Recurrent Unit Recurrent Neural Network (GRU-RNN) is used to detect the anomalies in software-defined networking (SDN) [12]. Hybrid intrusions and their corresponding signature are detected using the deep neural networks [13]. Deep learning approaches are used for automatic intrusion detection, it is very important to outline the security feature of the system [14]. Anomaly detection is common in cyberphysical systems, to detect these anomalies also RNN is used [15]. The hierarchical approach of IDS is used in the deep learning to detecting the anomalies in various levels of the detection process [16], in the hierarchy processes convolution neural network is also used to detect the anomalies [17].

The network intrusion detection problem existed from a long back, but there is no standard dataset till now, because of the dynamic nature of the problem. In the literature of the NIDS, there are three different datasets available for network anomaly detection are KDDCUP99, ISCX, and UNSW. To train the model appropriate datasets are required. Even though different cyber security proposals are depended on old open datasets, their outcomes are not similar because of various causes: various calculations think about various highlights, usage of pre-shifting activity, and the utilization of various parts among test and train dataset preparing informational collection. The UNSW\_NB15 Dataset is available in the split sets consisting of training and testing

records. The heat map of the Dataset is shown in Fig. 1 which shows the correlation of the attributes which lead to detection of anomaly class.

### 3 Pre Processing

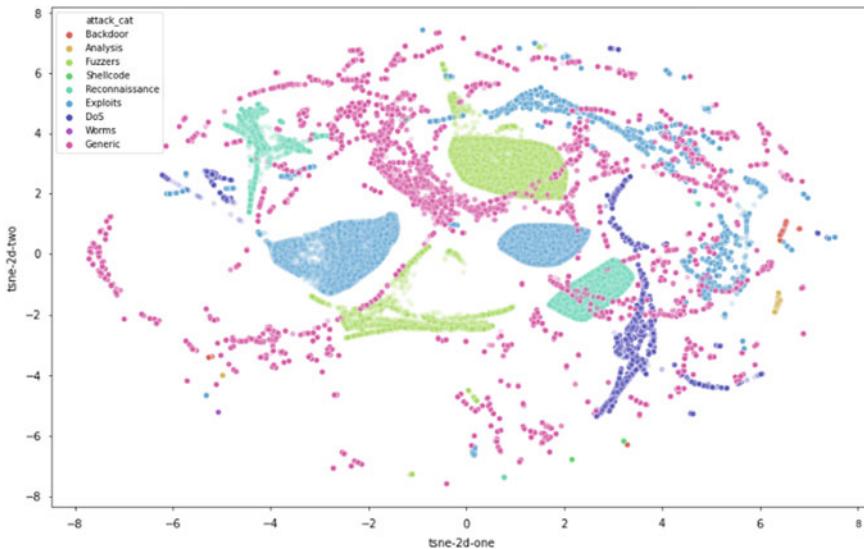
The existing anomaly detection datasets are using the different kinds of datasets, in the early age of the intrusion detection systems have used the feature selection approaches to identify the appropriate features for anomaly detection. In this work, the correlation of the features identified and mapped with the heat map. To reduce some dimensionality we adopted this approach in this work.

By observing the heat map in Fig. 1 we can see {sloss, sbytes}, {dloss, dbytes}, {is\_ftp\_login, ct\_ftp\_cmd} are correlated with each other. So removing any one of the elements from one set will not affect the results. The following are the columns that are affecting the label with positive correlation values (Table 1).

The exploratory analysis of the UNSW\_NB15 is also done using the t-SNE by projecting data onto the 2-D planes. Because of high dimensionality (41) of the UNSW-NB15 dataset, picturing the information is troublesome undertaking. In this manner, we utilize a method known as t-SNE (t-appropriated Stochastic Neighbor Embedding) to extend the highlights to a lower dimensional space by limiting the Kullback–Leibler uniqueness between the conveyance of higher dimensional highlights and circulation of lower dimensional anticipated highlights as appeared in Fig. 2. T-SNE consequently endeavors to discover designs in the information by recognizing groups dependent on closeness of information focuses with various highlights. The dataset contains the various kinds of values, all these are encoded and scaled.

**Table 1** Table showing the correlation values with the label column

Attribute name	Correlation value
stl	0.692741
ct_state_ttl	0.577704
state	0.497685
ct_dst_sport_ltm	0.357213
rate	0.337979
ct_src_dport_ltm	0.305579
ct_dst_src_ltm	0.303855
ct_src_ltm	0.238225
ct_dst_ltm	0.229887
ct_srv_src	0.229044
ct_srv_dst	0.228046



**Fig. 2** T-SNE plot of UNSW\_NB15 dataset

## 4 Methodology

As the detection process of anomaly is novel and finding the zero-day attacks is crucial with the increase of the network data. To detect these insights relations of the data to detect these attacks require more complex models. The Deep Learning (DL) models are designed using neural network architecture are more suitable to detect these attacks. Deep neural networks perform effective results in anomaly detection. The model will perform more combinations to find the anomaly patterns. Anomaly is happened on the sequence of events, to detect these sequences recurrent networks models are used. In the implementation process, we consider the recurrent neural network and its variants, each variant of the model has shown improved performance. In this work, we begin the implementation with the feed forward neural network and compared with the RNN, LSTM, and GRU [13, 14]. The models were built and run with the various parameters tuning. The implementation of the work is done in the following steps.

- Preprocessing the Dataset and converts the categorical data with Minmax scaler and Label encoder.
- Feed forward NN with 2 hidden layers.
- Feed forward NN with 14 hidden layers.
- Recurrent Neural Network (RNN).
- Long Short Term Memory (LSTM).
- Gated Recurrent Unit (GRU).
- Compare the performance measures of these models along with the ROC plots.

**A. Preprocessing the Dataset with Minmax scalar and Label encoder:**

While processing the data in the neural network, first the data has to be transformed properly. In the UNSB dataset contains some categorical data, it will be transformed to numeric. The columns ‘label’ and ‘attack\_cat’ are text labels which consist of Normal/Anomaly and Type of the attack. They are converted to integers using label encoder. The other columns whose ranges are varying from other are normalized using the minmax scalar. Some columns which contains missing values, which are replaced with ‘mean’ of the column value.

**B. Training and Testing the Feed forward NN with 2 hidden layers:**

Initial stage of the model for testing purpose, we build a normal Feed forward neural network with 2 hidden layers each of 100 nodes is built. The activation function ‘relu’ is used.

**C. Training and Testing the Feed forward NN with 14 hidden layers:**

Gradually we increase the model complexity further by adding the additional layers. A normal Feed forward neural network with 14 hidden layers each of 100 nodes is built with ‘relu’ as activation function. It is trained with 175,341 samples and tested with 82,327 samples each of 43 columns for 50 epochs.

**D. Training and Testing with RNN:**

A Recurrent Neural network with 4 Simple RNN layers of 64 nodes each and a dropout layer of 0.1 with the ‘Adam’ optimizer and ‘binary\_crossentropy’ is used as a loss function.

**E. Training and testing with Long Short Term Memory (LSTM):**

A LSTM Neural network with 4 Simple LSTM layers of 32 nodes each and a dropout layer of 0.1 and ‘Adam’ optimizer is used and ‘binary\_crossentropy’ is used as a loss function.

**F. Training and Testing with Gated Recurrent Unit(GRU):**

A GRU Neural network with 5 Simple GRU layers of 32 nodes each and a dropout layer of 0.1 with ‘Adam’ optimizer is used and ‘binary\_crossentropy’ is used as a loss function.

All these models are run with the UNSW\_NB dataset, and it contains the 175,341 training samples and tested with 82,327 samples each of 43 columns. Each of the model has run for 50 epochs.

## 5 Results and Discussions

After the pre-processing of the UNSW\_NB15 dataset, it is given to the model to verify the test and train accuracies along with the time consumed for each model. The results are shown in Table 2, which shows the considerable increase in accuracy of the models. These accuracies outperformed when it compared with NSL-KDD dataset as shown in Table 3.

As we can see from Tables 3 and 4, that the Feed forward neural network with 14 hidden layers has the highest accuracy. Apart from the Feed forward neural networks

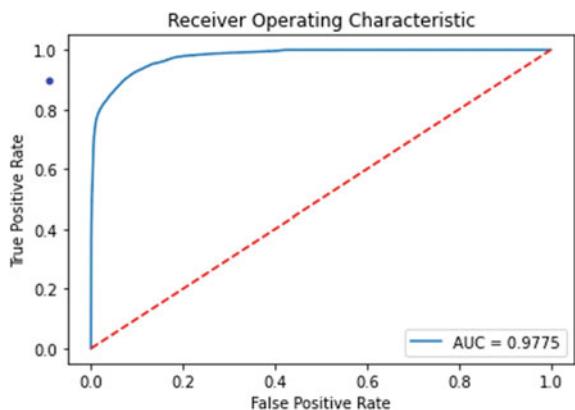
**Table 2** AUC score and Accuracies of different models

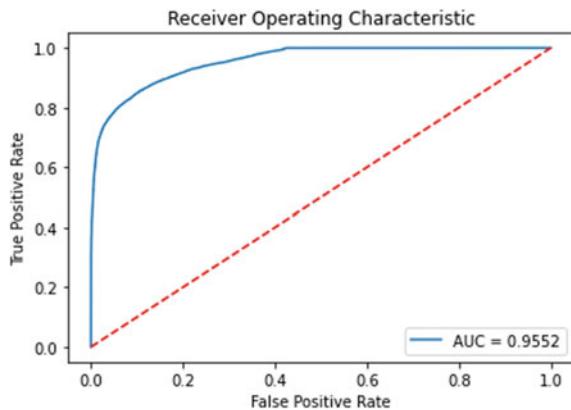
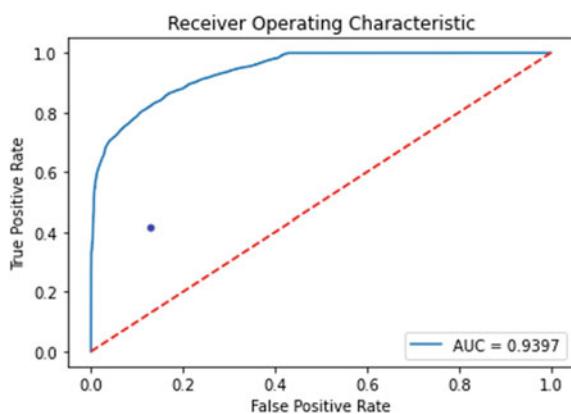
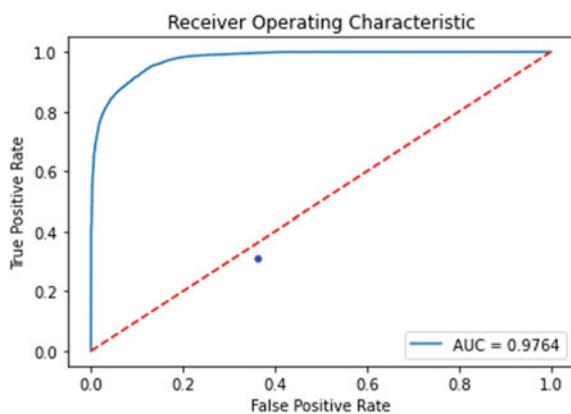
S. No.	Model name	AUC score	Accuracy	Time to train (min)
1	Feedforward with 2 hidden layers	93.98	82.07	<3
2	Feedforward with 14 hidden layers	97.75	87.62	<8
3	RNN with 4 RNN layers	93.97	82.43	<10
4	LSTM with 4 LSTM layers	95.52	83.92	<15
5	GRU with 5 GRU layers	97.63	85.59	<10

**Table 3** Comparing accuracies obtained with NSL-KDD & UNSW\_NB15

S. No	Model name	Accuracy with NSL-KDD	Accuracy With UNSW_NB15
1	Feedforward with 2 hidden layers	74.97	82.07
2	Feedforward with 14 hidden layers	78.29	87.62
3	RNN with 4 RNN layers	75.62	82.43
4	LSTM with 4 LSTM layers	76.22	83.92
5	GRU with 5 GRU layers	78.93	85.59

the GRU stands in the next place. The roc graphs are plotted for all the models shown in Figs. 3, 4, 5 and 6, the AUC score of GRU is the highest compared to the remaining models. All the classification algorithms take AUC score as the main criteria i.e. to reduce the false positive rates or increase the True positive rates so that obtained results are reliable.

**Fig. 3** ROC for NN

**Fig. 4** ROC curve LSTM**Fig. 5** ROC for RNN**Fig. 6** ROC for GRU

## 6 Conclusion and Future Work

In this work, The UNSW\_NB15 dataset is used to train different Deep Learning models to find the anomalous behavior in the network. Initially, the Exploratory data analysis of the dataset inferred that the attacks have some patterns which resulted in the formation of clusters in the t-SNE scatter plot. The observed patterns are trained to the Deep Learning models so that any similar pattern observed is marked as Anomaly. Concluding the work by saying that GRU got the Highest AUC score compared with all the other models which are trained. The accuracy can be further increased by using some feature selection and feature extraction methods with Regularization methods. Different optimizers like SGD and RMSPROP can also be used to increase the accuracy.

The future scope of the work is to extend the service to real-time network and divert the anomaly traffic to the honey pots to ensure cyber security by using more complex and well-trained networks with the real-time network traffic.

## References

1. Larriba-Novo, X.A., et al.: Evaluation of cyber security data set characteristics for their applicability to neural networks algorithms detecting cyber security anomalies. *IEEE Access* **8**, 9005–9014 (2020)
2. Moustafa, N., Slay, J.: UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In: 2015 Military Communications and Information Systems Conference (MilCIS). IEEE, 2015.
3. Bagui, S., et al.: Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset. *Securi. Priv.* **2**(6), e91 (2019)
4. Le, T.-T.-H., Kim, Y., Kim, H.: Network intrusion detection based on novel feature selection model and various recurrent neural networks. *Appl. Sci.* **9**(7), 1392 (2019)
5. Aydin, M.A.: Halim Zaim, A., Gökhan Ceylan, K.: A hybrid intrusion detection system design for computer network security. *Comput. Electr. Eng.* **35**(3), 517–526 (2009)
6. Fu, Y., et al.: An intelligent network attack detection method based on rnn. In: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE, 2018.
7. Ponkarthika, M., Saraswathy, V.R.: Network intrusion detection using deep neural networks. *Asian J. Sci. Technol.* **2**(2), 665–673 (2018)
8. Aldosari, M.S.: Unsupervised anomaly detection in sequences using long short term memory recurrent neural networks. Diss, 2016.
9. Yin, C., et al.: A deep learning approach for intrusion detection using recurrent neural networks. *Ieee Access* **5**, 21954–21961 (2017)
10. Naseer, S., et al.: Enhanced network anomaly detection based on deep neural networks. *IEEE Access* **6**, 48231–48246 (2018)
11. Radford, B.J., et al.: Network traffic anomaly detection using recurrent neural networks. arXiv preprint [arXiv:1803.10769](https://arxiv.org/abs/1803.10769) (2018).
12. Tang, T.A., et al.: Deep recurrent neural network for intrusion detection in sdn-based networks. In: 2018 4th IEEE Conference on Network Softwarization and Workshops (NetSoft). IEEE, 2018.
13. Kaur, S., Singh, M.: Hybrid intrusion detection and signature generation using deep recurrent neural networks. *Neural Comput. Appl.* 1–19 (2019).

14. Elsherif, A.: Automatic intrusion detection system using deep recurrent neural network paradigm (2018).
15. Goh, J., et al.: Anomaly detection in cyber physical systems using recurrent neural networks. In: 2017 IEEE 18th International Symposium on High Assurance Systems Engineering (HASE). IEEE, 2017.
16. Wang, W., et al.: HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. *IEEE Access* **6**, 1792–1806 (2017).
17. Kwon, D., et al.: An empirical study on network anomaly detection using convolutional neural networks. In: 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2018.

# Driver Drowsiness Detection Using Convolution Neural Networks



P. Ravi Teja, G. Anjana Gowri, G. Preethi Lalithya, R. Ajay, T. Anuradha, and C. S. Pavan Kumar

**Abstract** According to Central Road Research Institute, more than 30% of the road accidents in India are happening because of driver drowsiness especially in the early hours of the day. This is observed in most of the countries which is a serious threat to the society and has to be addressed at the earliest. The ratio of accidents due to driver drowsiness can be reduced if a drowsiness detection mechanism is in continuous active state in the vehicle, which helps drivers to get an alert and go back to a normal driving mode. The proposed approach focuses on building a drowsiness detection mechanism to alert the driver to avoid the catastrophe. In this work, the detection system can identify whether the driver's eyes were closed or open even in low light or dim light and how much time the eyes were in closed state. Based on the time the system will generate an alert. The proposed system which is built using convolution neural network achieved 89% accuracy in normal light and 78% accuracy in dim light conditions.

## 1 Introduction

Across the globe, each year, 1.35 million people are killed due to road accidents. Everyday almost 3700 people are killed in road traffic crashes. Out of these, around 20–30% of accidents happen due to the driver's fatigue and drowsiness. Single vehicle accidents such as hitting the divider, bumping into a tree, all these happens due to the driver hypo vigilance. According to the current studies, these road crashes can be reduced by using driver face monitoring system [1]. India is one of the busiest countries in the world in terms of road traffic. With rapid increase in the road traffic, road safety has become an important issue to be subjected. Road accidents are one of

---

P. Ravi Teja · G. Anjana Gowri · G. Preethi Lalithya · R. Ajay · T. Anuradha (✉) ·

C. S. Pavan Kumar

Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada, AP, India

C. S. Pavan Kumar

e-mail: [pavanchitturi@vrsiddhartha.ac.in](mailto:pavanchitturi@vrsiddhartha.ac.in)

the major causes of death in India [2]. On an average, more than 1214 crashes happen on a daily basis. There are various reasons which cause road accidents. Some of them are distracted driving, bad road conditions, tailgating, drunk driving, etc. Apart from the bad roads and non-functional street lights, continuous driving for several hours during the night is riskier. Driver fatigue makes the driver's responding capacity weak. He eventually feels drowsy and loses control over the vehicle. Generally, the driver gets exhausted after 2 h of continuous driving. Drowsiness generally arises at the early hours, after eating lunch, at midnight, etc. In addition to this, alcohol consumption, drugs intake, and using any kind of hypnotic medicines also lead to fatigue. According to WHO fact sheet, around 1.35 million die because of road accidents every year and it has set the target of reducing them by 50% by 2030 [3]. A solution to this problem is to alert and wake up the driver when he/she is feeling drowsy and falling asleep. Lot of research is happening on this issue using face monitoring system, eye blink ratio, etc. So, to prevent these accidents we will be making a driver drowsiness detection system. The existing algorithms provide most important benefits such as interpretability. The proposed research focuses on identifying driver drowsiness using Haar cascade features and convolution neural networks with increased correctness in drowsiness detection first by identifying the driver's face. The system is able to perfectly identify whether driver's eyes are closed or opened even if the driver wore spectacles.

## 2 Existing Approaches

Drowsiness detection based on driving patterns and deviation of the vehicle from actual line position is done in [4]. Drowsiness identification based on whether the eyes are open or closed and by detecting whether the driver is yawning or not is done in [5]. Here drowsiness is detected with high accuracy even in low light conditions by first detecting face and then eyes and mouth parts. Driver drowsiness detection in real-time by using deep learning methods in light weight model for android applications is done in [6]. Identifying the face region using Haar based features and classification of face images using adaboost algorithm is proposed in [7]. Drowsiness is detected by identifying the corners of eye lid and eye blinking and by using Harris corner detection algorithm in [8]. Drowsiness detection using convolution neural networks by classifying the face image as drowsy or non-drowsy is done in [9]. A review on drowsiness detection system of automobile drivers using soft computing techniques is done in [10]. Identification of facial features in real time and classification of facial images using SVM is done in [11]. A system called Dricare is proposed to identify drowsiness detection based on difference in facial features like yawning and eye blinking in [12]. Identifying Driver fatigue through eye blinking using signal processing techniques and CMOS technology is done in [13]. Usage of image processing techniques and Haar cascade samples for detecting eye blink and drowsiness is done in [14]. From the existing literature, it was observed that most of the researchers concentrated on identifying drowsiness but not concentrated

on alerting the driver when he/she feels drowsy. The proposed approach focuses on building a drowsiness detection mechanism to alert the driver by using a buzzer sound after identifying drowsiness condition.

### 3 Proposed Approach

The proposed method for drowsiness detection is done using the camera installed in the vehicle. It continuously gets video of driver. The video is divided into frames. In each frame, face is detected using HAAR cascade. After detecting the face, drowsiness is observed based on the eye open or close. The eye detection is done with the help of HAAR cascade features. If eye closing is continuous more than a threshold value, drowsiness will be detected. Figure 1 shows the block diagram of the proposed approach.

The entire work can be divided into 5 steps.

#### **Step 1—Capture the images as an input using a camera**

The classifier used to classify the images in the proposed research work is convolution neural network which require a lot of training data. The training data contains the face images of the driver in different angles with eyes open and closed. The images of the face and eyes are captured by the webcam installed in front of the driver seat.

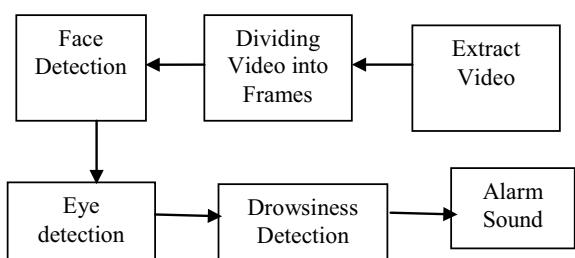
#### **Step 2—Detect the face and eyes from the image and create a region of Interest (ROI)**

After having the image, the second step is to identify the face from image. This is done using Haar cascade classifier proposed by Viola Jones in his research work on object detection [15]. In this classifier shown in Fig. 2, convolution like kernel is used to detect different features. Each feature is represented in the form of a value which is the difference between pixels under white and black regions. As the face is detected from image, in the same way, eyes are detected from the face. While implementing with OpenCV, the region of interest is detected with a bounding box around it.

#### **Step 3—Detecting the eye state**

Here the eye blinks are estimated using facial land mark algorithm [16] which calculates Eye Aspect Ratio (EAR) value to identify how far the eye is opened in each

**Fig. 1** Block diagram of proposed approach



**Fig. 2** Haar cascade classifier with white and black regions



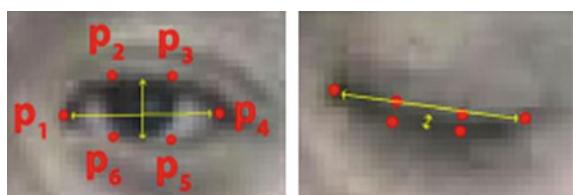
image. EAR value is calculated using Eq. 1 where  $p1, p2, p3, p4, p5$  and  $p6$  are different positions around the 2D image of the eye when it is closed and opened as shown in Fig. 3. The graph of EAR values for a single eye blink is shown in Fig. 4.

$$\text{EAR} = (\|p2 - p6\| + \|p3 - p5\|)/\|p1 - p4\| \quad (1)$$

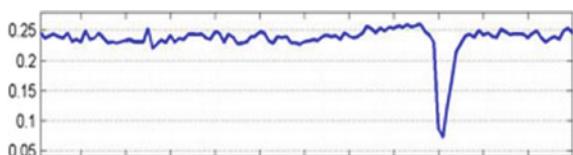
#### Step 4—Classification of eye blinks using CNN

Drowsiness can be detected by observing the duration of eye closure and number of eye blinks. A person's eye close duration in the wakeup state during eye blink generally will be 100–400 ms and in the sleepy state, it will be more. In the facial landmark algorithm [16], the eye blink patterns were classified using SVM algorithm. In the proposed approach, Convolution neural network (CNN) algorithm [17] was used for classification. CNN algorithm contains a sequence of convolution and pooling layers followed by fully connected layer and softmax layer and uses RELU activation function as shown in Fig. 5. The major operations in CNN are the convolution and pooling operations. In convolution operation, the values in input matrix are multiplied

**Fig. 3** 2D image of eye when open and closed with positions  $p1-p6$



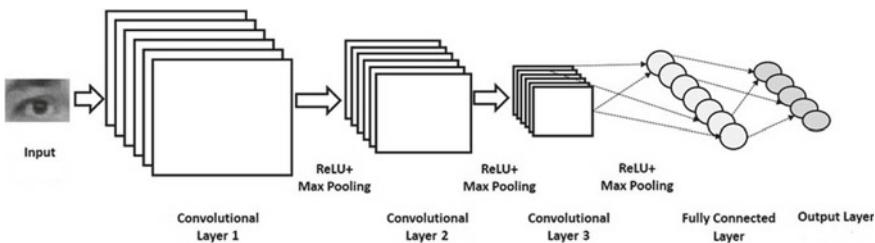
**Fig. 4** Plot of Eye Aspect Ratio values for a single eye blink



by the values in kernel matrix element by element. The kernel matrix will be moved over the entire input matrix to cover all columns and then all rows to generate output matrix as shown in Fig. 6. In max pooling, the input matrix will be considered as different square parts, and the maximum value from each of the square will be taken in output as shown in Fig. 7.

### Step 5—Score calculation for Drowsiness detection and alerting the driver

To determine the duration of eye closure, a counter is set which increases when both eyes closed and decreases when eyes are open. If the score is beyond a certain

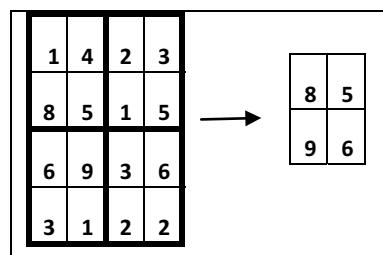


**Fig. 5** CNN architecture

$\begin{matrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 \end{matrix}$	$\times$	$\begin{matrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{matrix}$	$=$	$\begin{matrix} 3 & 3 & 4 \\ 2 & 4 & 3 \\ 2 & 3 & 4 \end{matrix}$
---	----------	---	-----	---

**Fig. 6** Convolution operation of  $5 \times 5$  input with  $3 \times 3$  kernel

**Fig. 7** Max pooling  $3 \times 3$  image with  $2 \times 2$  kernel



threshold value, it indicates driver is in drowsy mood. Then a buzzer sound is created in the program which alerts the driver when he/she feels sleepy.

The score is basically a value we will use to determine how long the person has closed his eyes. So if both eyes are closed, we will keep on increasing score and when eyes are open, we decrease the score. We are drawing the result on the screen which will display real-time status of the person. The average blink duration of a person is 100–400 ms (i.e. 0.1–0.4 of a second). Hence if a person is drowsy his eye closure must be beyond this interval. We set a time frame of 1 s. If the eyes remain closed for one or more seconds, drowsiness is detected and alert pop regarding this is triggered.

## 4 Experimental Work

In experimental work, images of a person were collected using the built-in web camera of the laptop. The images of the persons who sit in front of the camera in different angles were taken as input data. To get continuous images, infinite loop was taken in programming using OpenCV which captures each and every frame. As OpenCV works better with grey scale images, before using Haar cascade classifier, the image was converted to grey scale and resized to 24 \*24 pixels as our model was trained on 24 \*24 pixel images. After detecting the face using Haar classifier, EAR value was calculated in each frame to know how far the eye was closed. The CNN model architecture used in the experiment consists of the following layers: 1st Convolutional layer; 32 nodes, kernel size 3, 2nd Convolutional layer; 32 nodes, kernel size 3, 3rd Convolutional layer; 64 nodes, kernel size 3, Fully connected layer; 128 nodes, The final layer is also a fully connected layer with 2 nodes. In all the layers, a Relu activation function is used except the output layer where a softmax layer is used. Score is calculated based on how much time eyes are closed. In the experimentation, test score is a numerical value which is used to identify the duration of eye closure. If both eyes are closed, we will keep on increasing score and when eyes are open, we decrease the score and reset every time when eye closes. If the score reaches a threshold value, beep sound was generated. Here the threshold was taken as 20.

## 5 Experimental Results

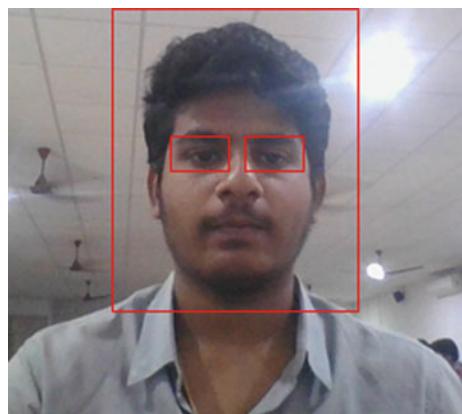
For experimentation, a person is sitting in front of the laptop on which the proposed experimental work is implemented in python and OpenCV. Figure 8 shows the rectangle box around the face after it was detected as ROI using Haar classifier. It returned the 2D coordinates and height and width of the object selected as ROI in the form of bounding box. Then the same procedure was used to detect eyes from face and

**Fig. 8** Displaying bounding box around face



this left and right eyes image data was given to CNN as input. Figure 9 shows the bounding boxes around left and right eyes separately.

**Fig. 9** Detecting eyes from ROI



**Fig. 10** Classifying the image as open when eyes were open



**Fig. 11** Classifying the image as closed when eyes were closed



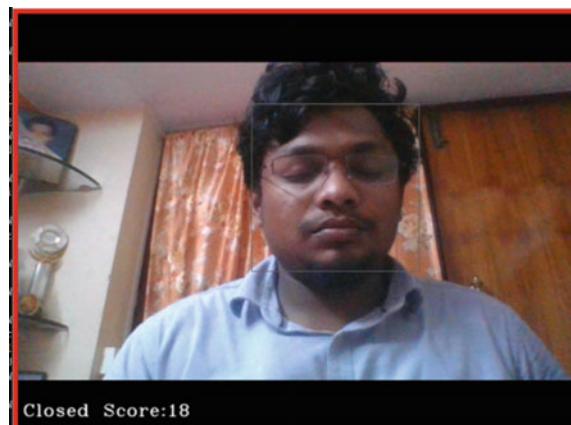
Figure 10 shows the output of CNN classifier. Here the image was classified as open when the person's eyes were open and here the score was 0. Figure 11 shows the image was classified as closed when eyes were closed. Here the score was increased starting from the classifier identified image as closed. When score reaches 20, an alarm sound was generated.

The experimentation was done in conditions like, when the driver wore the spectacles and in low light conditions. Figure 12 represents the image was classified correctly as open when the driver worn the spectacles in low light condition. Figure 13 shows image was classified correctly as closed when driver worn the spectacles in low light conditions. Table 1 shows the accuracies obtained in different experimental conditions. It was observed that the proposed work obtained considerable accuracies of more than 75% even when the driver having spectacles, when face turned sideward and in low light conditions also.

**Fig. 12.** Classifying image correctly as open with spectacles and in low light condition



**Fig. 13.** Classifying image correctly as closed with spectacles in low light condition



**Table 1** Accuracies obtained in different conditions

S. No.	Condition	Accuracy obtained (%)
1	Normal	89
2	Face turned sideward	86
3	Wearing glasses normal condition	84
4	Low light condition	78
5	Wearing glasses low light condition	76
6	Face turned sideward low light condition	75

## 6 Conclusions

An automatic driver drowsiness detection and alerting mechanism were developed using convolution neural networks and Haar cascade classifiers. The experimental setup created using OpenCV and Python programming correctly identified whether eyes were closed or opened even in low light conditions and if the person wore spectacles. It also correctly classified the images even if the person's face is turned side wards. The proposed system achieved 89% accuracy in normal light and 78% accuracy in dim light conditions.

## References

1. Sigari MH, Fathy M, Soryani M. A driver face monitoring system for fatigue and distraction detection. International journal of vehicular technology. 2013;2013.

2. Dandona, R., Kumar, G.A., Gururaj, G., James, S., Chakma, J.K., Thakur, J.S., Srivastava, A., Kumaresan, G., Glenn, S.D., Gupta, G., Krishnankutty, R.P.: Mortality due to road injuries in the states of India: the Global Burden of Disease Study 1990–2017. *the Lancet Public Health*. **5**(2), 86–98 (2020)
3. Road Traffic Injuries, 7 February 2020 [online] <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries> accessed on: 15th Feb 2020
4. Krajewski J, Sommer D, Trutschel U, Edwards D, Golz M. Steering wheel behavior based estimation of fatigue. The fifth international driving symposium on human factors in driver assessment, training and vehicle design 2009;118–124.
5. W. Tipprasert, T. Charoenpong, C. Chianrabutra and C. Sukjamsri, A Method of Driver's Eyes Closure and Yawning Detection for Drowsiness Analysis by Infrared Camera. In: 2019 First International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics (ICA-SYMP), Bangkok, Thailand, 2019, pp. 61–64, <https://doi.org/10.1109/ICA-SYMP.2019.8646001>
6. Jabbar, R., Al-Khalifa, K., Kharbeche, M., Alhajyaseen, W., Jafari, M., Jiang, S.: Real-time driver drowsiness detection for android application using deep neural networks techniques. *Procedia Computer Science*. **1**(130), 400–407 (2018)
7. Manu BN. Facial features monitoring for real time drowsiness detection. In 2016 12th International Conference on Innovations in Information Technology (IIT) 2016 Nov 28 (pp. 1–4). IEEE.
8. Rahman A, Sirshar M, Khan A. Real time drowsiness detection using eye blink monitoring. In 2015 National Software Engineering Conference (NSEC) 2015 Dec 17 (pp. 1–7). IEEE.
9. Dwivedi K, Biswaranjan K, Sethi A. Drowsy driver detection using representation learning. *Advance Computing Conference (IACC)*, IEEE 2014;995–999
10. Edison, T., Ulagapriya, K.: Saritha A Prediction of Drowsy Driver Detection by using Soft Computing. *Journal of Critical Reviews*. **7**(6), 2020 (2019)
11. Mundra, P & Todwal, V. (2018). Design Simulation and Performance Analysis of Real Time Facial Features Monitoring for Drowsiness Detection Using Support Vector Machine. *IJIREEICE*. **6**. 53–56. <https://doi.org/10.17148/IJIREEICE.2018.688>.
12. Deng, W., Wu, R.: Real-time driver-drowsiness detection system using facial features. *IEEE Access*. **21**(7), 118727–118738 (2019)
13. Yassine, N., Barker, S., Hayatleh, K., Choubey, B., Nagulapalli, R.: Simulation of driver fatigue monitoring via blink rate detection, using 65 nm CMOS technology. *Analog Integr. Circ. Sig. Process* **95**(3), 409–414 (2018)
14. Jayadevappa, Suryaprasad & D, Sandesh & V, Saraswathi & D, Swathi & S, Manjunath. (2013). Real Time Drowsy Driver Detection Using Haarcascade Samples. *Computer Science & Information Technology*. **3**. 45–54. <https://doi.org/10.5121/csit.2013.3805>.
15. Viola P, Jones M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001 2001 Dec 8 (Vol. 1, pp. I-I). IEEE
16. Tereza Soukupova and Jan Čech, Real-Time Eye Blink Detection using Facial Landmarks, 21st Computer Vision Winter Workshop, February 3–5, 2016
17. Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks-- The ELI5 way, Dec 15, 2018, [online] <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53> [accessed]: Nov, 2019

# Glaucoma Detection Using Morphological Filters and GLCM Features



Babita Pal, Vikrant Bhateja, Archita Johri, Deepika Pal,  
and Suresh Chandra Satapathy

**Abstract** Glaucoma, a neuropathic disorder occurs due to change in shape of optic disk, cup and increase in eye pressure, etc. If Glaucoma is not diagnosed in its initial stages, then it leads to permanent vision loss. Diagnosis of Glaucoma involves various steps including pre-processing or automated processing, features extraction, and classification. Results obtained after automated processing are further used for identification of various properties that has to be analyzed during features extraction. Features or characteristics extraction is used for reducing the image to the various features used for further analysis. Further, classification has been used for differentiating between Glaucomatous and non-Glaucomatous eye. Combination of statistical features and gray level co-occurrence matrix (GLCM) is used for features extraction. Classifier used for differentiating the two classes is support vector machine (SVM), K-nearest neighbor (KNN), and artificial neural network (ANN). Accuracy of classification for KNN, SVM, and ANN is 82.5%, 85%, and 93.7%, respectively.

## 1 Introduction

Computer-aided diagnosis (CAD) has been introduced in the field of biomedical imaging. CAD acts as aid to the medical professional to analyze and study the various diseases. CAD improves the method of diagnosis, as it reduces human intervene, thereby reducing the chances of error associated to it. CAD approach also helps to analyze large dataset in less interval of time. CAD has been proposed to easily diagnose Glaucoma. Glaucoma is one of the major ophthalmic diseases causing permanent vision loss if not treated in time. Its symptoms are quite common to that

---

B. Pal · V. Bhateja (✉) · A. Johri · D. Pal

Department of Electronics and Communication Engineering, Shri Ramswaroop Memorial Group of Professional Colleges (SRMGP), Faizabad Road, Lucknow, UP 226028, India

Dr. APJ Abdul Kalam Technical University, Lucknow, UP, India

S. C. Satapathy

KIIT University, Bhubaneswar, Odisha, India

of any other retinal disorder; hence, its diagnosis is difficult. Various parameters have to be considered during diagnosis of Glaucoma such as if not the disease is genetic. Other diseases such as hyperthyroidism, diabetes, etc., are possessed by the suspect [1]. Features extraction has been used for extracting various properties of retinal images. The purpose for features extraction is to reduce input to significant features for further analysis of the problem. It acts as an aid to reduce the complexity of the system. Various features for the measurement of several properties have been carried out. Classifiers are used for differentiation between various classes in the diagnosis of the outlined disease [2].

CAD of Glaucoma has been presented by various medical professionals. Cup to disk (CDR) ratio is calculated using diameter of cup to that of disk. The vertical diameter of OC and OD is calculated separately considering their shape similar to that of an ellipse; then their ratio is considered. The increase in CDR is an indicative of retinal diseases. Eye has four regions that are ISNT ratio is areas of blood vessels in inferior and superior regions to that of nasal and temporal regions. ISNT ratio is an important aspect for various retinal problems [3]. Various features considered by other researchers for analysis as CDR and ISNT were extracted. Texture and high order spectrum features that is correntropy were analyzed, and classification was performed using least square SVM. Correntropy features were estimated using Student *t* test algorithm [4]. Image projection features were presented. Both horizontal and vertical projection features were considered for the analysis [5]. CDR has been extracted and classified using SVM [6]. Contextual features along with mean and median were extracted, and fisher discriminant analysis was presented for classification by [7]. VCDR and FFT had been used with generalized matrix LVQ [8]. There were various flaws in these features used for further analysis. In order to combat the various problems, here in this paper, approach of GLCM features has been proposed. The rest of the paper is organized in the following sections: Sect. 2 describes the overview of CAD of Glaucoma, Sect. 3 discusses the obtained results, and Sect. 4 discusses the conclusion of the work.

## 2 Proposed Framework for Diagnosis of Glaucoma

CAD of Glaucoma proposes pre-processing of retinal images involving contrast enhancement of retinal images and segmentation and localization of ROI. Features extraction from the segmented ROI has been presented. Further testing and training of dataset have been proposed using SVM, KNN, and ANN classifiers to differentiate between Glaucomatous and non-Glaucomatous retinal images.

## 2.1 Pre-processing

Pre-processing is the process of improvement of quality of the images or ROI. Sometimes due to dark background, overlapping of blood vessels or presence of unwanted noise in retinal images creates hindrance in detection and processing of ROI. Hence, pre-processing is used to remove the noise from the images, and it improves the visibility of an image which is being used for further processing steps. Segmentation is the conversion of images into true and false pixels, i.e., binary images. There are two pre-processing steps which are given in following subsections. In order to obtain ROI, true pixels are considered, and rest unwanted components of the image are removed.

Image enhancement takes major role in pre-processing step which improves the visibility of an image. Morphological filters are easier and shape selective filter which makes the enhancement process easier and more accurate. Top-hat and Bottom-hat transform is the combination of basic operations which takes vital role for contrast enhancement. Image quality assessment (EME) is used for the quantitative measure of an enhanced image [9]. In order to easily detect ROI (optic cup and optic disk), localization and segmentation are necessary for both optic cup and optic disk extraction. Segmentation of ROI has been performed using combination of global thresholding and morphological operations. Ellipse fitting has been applied for localization of optic cup and optic disk [10].

## 2.2 Features Extraction and Classification

Features are defined as the features are a set of different quantifiable entity which is a function of one or more objects. All features could be categorized into two levels which are low-level features and high-level features. Pixel values and pixel co-ordinates are the characteristics of texture. Harlick's features come up with fourteen texture features which are extracted from co-occurrence matrix. Texture features could be categorized into first-order texture measures and second-order texture measures. First-order texture measures describe the statistical features which could be calculated by individual pixels of an image without using the pixel's neighbors' relationship, while second-order measures consider the relationship among the neighbors [11]. Classification is the process to differentiate between various classes that may or may not be differentiable in normal means. Differentiation of different classes in diagnosis of any particular disease is crucial step. Sometimes there are many similarities between ophthalmic diseases; hence, it may take more time or become difficult for the medical expert to diagnose the specific disease. Hence, classifier that has to be considered has capability to differentiate between classes even with minute variations. Classifier that had been considered for differentiating between Glaucomatous and non-Glaucomatous is SVM, KNN, and ANN [12–15].

### A. First Order Texture Features

**Table 1** First-order texture features [17, 18]

Features	Formulae	Key points
Mean ( $\mu$ )	$\mu = \sum_{i=1}^{L-1} k_i p(k_i)$	It is used to find gray values of pixels
Standard deviation (SD)	$SD = \sqrt{\sum_{i=1}^{L-1} k_i^2 p(k_i)^2 - (\sum_{i=1}^{L-1} k_i p(k_i))^2}$	Measure the scattering in GLCM and also explain the level of intensity in an image
Variance ( $V$ )	$V = \frac{1}{S^2} = \frac{1}{N} \left( \sum_{i=1}^{L-1} k_i - \mu \right)^2$	Measures the second central moment and also describes the sameness within the region of an image
Skewness ( $S_k$ )	$S_k = \frac{1}{S^3} \sum_{i=1}^{L-1} (k_i - \mu)^3$	Skewness is measured as third central moment. Used to show the asymmetry of data which occurs around the sample
Kurtosis ( $K$ )	$K = \frac{1}{S^4} \sum_{i=1}^{L-1} [k_i - \mu]^4 p(k_i) - 3$	Kurtosis could be explained as fourth central moment. Acuteness of gray level histogram

Where  $k$  denotes the gray value of  $i$ th pixel,  $P(k_i)$  denotes the normalized histogram, and  $L$  denotes the number of gray levels in an image

It is used to define the pixels value of an image. It gives the information about how pixel intensities is varied according to the position of texture. First-order calculation includes the intensity histogram and intensity features. Statistical features consist mean, standard deviation, variance, and skewness. Here, mean and standard deviation bothered with the properties of individual pixels [16] (Table 1).

## B. Second-Order Texture Features

It describes the texture, color, and contrast of an image. Second-order texture features include GLCM features [21]. Gray level co-occurrence matrix (GLCM) is a type of matrix in which number of gray level in an image is same as the number of rows and columns.

In co-occurrence matrix, each individual pixel's intensity of an image could be compared in four directions ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$  and  $135^\circ$ ) with its neighbor pixels. These features determine the texture, shape, color, and contrast of an image. It consists of various features which are contrast, correlation, energy, homogeneity, and entropy which is shown in Table 2.

**Table 2** Second-order texture features [17, 18]

Features	Formulae	Key points
Contrast ( $C$ )	$C = \sum_{i=1}^{L-1} p_{ij}(i - j)^2$	Local abnormalities of an image could be measured by contrast. 0 value indicates constant image
Correlation (Co)	$Co = \sum_{i=1}^{L-1} \frac{p_{ij}(i - \mu)(j - \mu)}{\sigma_x \sigma_y}$	It is used to define the relation between pixels with its neighborhood. For the gray tone linear dependencies in an image, correlation is being used
Energy ( $E$ )/angular Second Moment (ASM)	$E = \sum_{i,j=0}^{L-1} (p_i)^2$	It depicts the uniformity in a gray level region. If the energy value comes 1, it seems as constant image
Homogeneity ( $H$ )	$H = \sum_{i,j} \frac{p(i,j)}{(1 -  i - j )}$	Homogeneity feature defines the closeness of pixels. Homogeneity is 1 for diagonal GLCM
Entropy ( $E$ )	$En = - \sum_{i=0}^n P(x_i)(\log_2 P(x_i))$	Entropy feature predicts the complexities presents in the intensities of an image

where

$\mu_i$  and  $\mu_j$  denote mean

$\sigma_x, \sigma_y$  denotes standard deviation

$p(i, j)$  denotes  $(i, j)$ th entity in normalized GLCM

$P(x_i)$  denotes partial probability density function

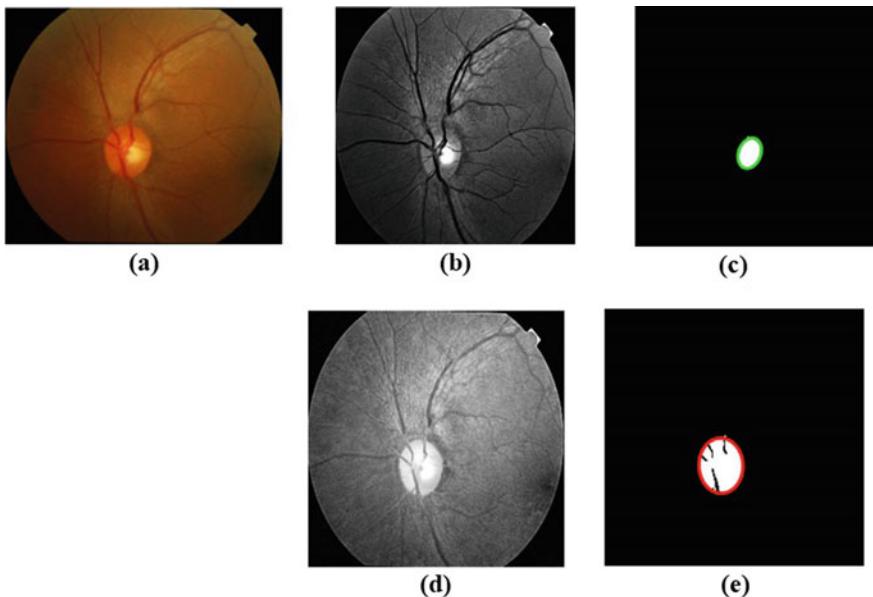
$n$  is the total number of pixels in an image

### 3 Results and Discussion

#### 3.1 Automated Processing of Retinal Images

CAD of Glaucoma has been performed on the retinal images obtained from publicly available database DRISHTI-GS [19]. Contrast enhancement and segmentation have been discussed earlier in [9, 10]. The simulated results for contrast enhancement, segmentation, and localization have been displayed in Fig. 1.

Contrast enhancement leads to improvement in visually of ROI observed in Fig. 1b and d. After pre-processing segmented ROI so obtained Fig. 1c and e, further features reextracted from them, and further classification is carried out.



**Fig. 1** Pre-processing of retinal images **a** Original retinal image, **b** green channel enhanced image **c** segmented localized OC, **d** red channel enhanced image, **e** segmented localized OD

### 3.2 Characteristic Extraction

Different features are extracted for the optic disk and optic cup which is being used for the further analysis or which will be used for the classification of Glaucomatous and non-Glaucomatous images. Total ten features and CDR [11] are extracted for OD and OC individually. Training and testing features of optic cup (OC) for some cases are shown in Table 3.

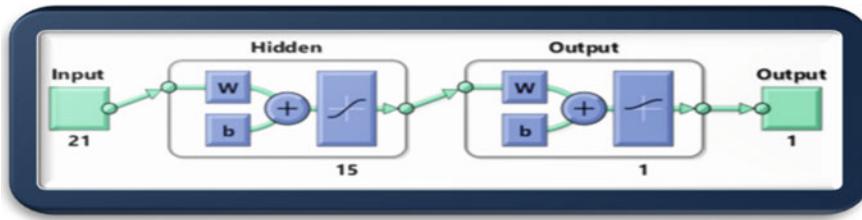
Similarly, training and testing features of OD are shown in Table 4. The features, which are extracted, are the sum-up of statistical feature and texture features. Total 101 images (Drishti GS-1 data set which is publicly available) features are extracted.

**Table 3** Features of OC

C	CO	E	HO	M	SD	V	$S_k$	K	EN
0.00156	0.94982	0.96706	0.99924	0.01592	0.12376	0.00163	1.8293	10.0535	0.1169
0.00158	0.95136	0.96506	0.9992	0.0169	0.1276	0.00172	2.10186	11.54212	0.12278
0.00164	0.9522	0.96348	0.99918	0.01768	0.13146	0.00182	1.908	9.75016	0.12442
0.0069	0.94774	0.97186	0.9993	0.01352	0.11346	0.00134	1.93496	11.00832	0.10208
0.00142	0.95232	0.96876	0.9993	0.01508	0.12162	0.00146	2.06756	10.36268	0.11264

**Table 4** Features of OD

<i>C</i>	CO	<i>E</i>	HO	<i>M</i>	SD	<i>V</i>	<i>S<sub>k</sub></i>	<i>K</i>	EN
0.003	0.93734	0.9459	0.99852	0.0262	0.15858	0.00258	3.4403	26.58494	0.1739
0.00278	0.9325	0.94326	0.9986	0.02786	0.1576	0.00286	3.01196	24.31004	0.17786
0.0163	0.95344	0.92616	0.99838	0.03654	0.18724	0.00382	4.05008	36.3416	0.22578
0.00464	0.92944	0.93038	0.99772	0.0337	0.178044	0.00348	3.346436	26.96898	0.20999
0.00306	0.94086	0.94396	0.99848	0.02714	0.16204	0.00272	3.64554	19.38016	0.17942

**Fig. 2** Simulated neural network

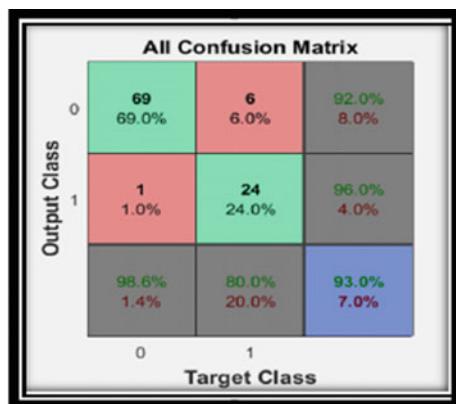
### 3.3 Performance Analysis of Classifiers

Performance evaluation of classifier accuracy, specificity, precision, and sensitivity parameters has been used [20]. Using TP, FP, TN, and FN, all the performance are evaluated for testing and training images. Total 100 images are taken for the simulation results of ANN classifier in which 80 images are used for training and rest of them are used for testing. Figure 2 shows the simulated neural network. A matrix with size of  $21 \times 100$  is taken as input for the classifier. 21 denotes number of features considered, and 100 denotes the total number of images.

ANN classifier used confusion matrix which is used to plot the training, testing, validation, and overall performance of the system. Confusion matrix is obtained by plot of actual and experimented value of training and testing images using extracted features. Figure 3 depicts the overall performance of ANN classifier, in which accuracy is obtained 70%, sensitivity is 28.7%, specificity is 84.62%, and precision is 50% for SVM.

Performance parameter obtained for KNN is 65% accuracy, 14.29% sensitivity, specificity is 92.31%, and the precision is 50%. Accuracy is obtained via ANN 93.7%, sensitivity 98.6%, specificity 80%, and precision 92%.

**Fig. 3** Confusion matrix for overall performance of ANN



## 4 Conclusion

Early detection and diagnosis of glaucoma are necessary to avoid permanent vision loss. For early detection, it is necessary to follow all the steps. Features extraction plays a vital role to define the individual features of an image. It categorizes features into two parts as follows: the texture features and statistical features. Here, gray level co-occurrence matrix is used. To classify the Glaucomatous and non-Glaucomatous images, classifier is used. From the result analysis, it is clear that ANN classifier gives more accuracy in comparison to the other classifiers.

## References

1. Kavitha, S., Zebardast, N., Palaniswamy, K., Wojciechowski, R., Chan, E.S., Friedman, D.S., Venkatesh, R., Ramulu, P.Y.: Family history is a strong risk factor for prevalent angle closure in a South Indian population. *Ophthalmology* **121**(11), 2091–2097 (2014)
2. Bhateja, V., Gautam, A., Tiwari, A., Satapathy, S.C., Nhu, N.G., Le, D.N.: Haralick features-based classification of mammograms using SVM. In: Proceedings of 4th International Conference on Information Systems Design and Intelligent Applications, India, pp. 787–795 (2018).
3. Nayak, J., Acharya U.R., Bhat, P.S., Shetty, N., Lim, T.C.: Automated diagnosis of glaucoma using digital fundus images. *J. Med. Syst.* **33**, 337–346 (2008)
4. Maheshwari, S., Pachori, R.B., Acharya, U.R.: Automated diagnosis of glaucoma using empirical wavelet transform and correntropy features extracted from fundus images. *IEEE J. Biomed. Health Inf.* **21**(3), 803–813 (2016)
5. Mahfouz, A.E., Fahmy, A.S.: Fast localization of optic disc using projection of image features. *IEEE Trans. Image Process.* **19**(12), 3285–3289 (2011)
6. Cheng, J., Liu, J., Xu, Y., Yin, F., Kee Wong, D.W., Tan, N.-M., Tao, D., Cheng, C.Y., Aung, T., Wong, T.Y.: Superpixel classification based optic disc and optic cup segmentation for glaucoma screening, *IEEE Trans. Med. Imaging* **10**(10), 1–15 (2013)
7. Zhou, W., Wu, C., Yi, Y., Duet, W.: Automatic detection of exudates in digital images using superpixel multi-feature classification. *IEEE Access* **5**(9), 17077–17088 (2017)

8. Guo, J., Azzopardi, G., Shi, C., Jansonius, N.M., Petkov, N.: Automatic determination of vertical cup to disc ratio in retinal fundus images for glaucoma screening. *IEEE Access* **7**(1), 8527–8541 (2019)
9. Johri, A., Pal, D., Bhateja, V., Pal, B.: Enhancement of retinal filters using morphological filters. In: Proceedings of 4th International Conference on Intelligent Computing and Communication, pp. 1–9 (2020)
10. Johri, A., Pal, B., Pal D., Bhateja, V.: Computer aided diagnosis of glaucoma using optic cup-disc ratio. In: Proceedings of 4th International Conference on Intelligent Computing and Communication, pp. 1–10 (2020).
11. Anuradha, K., Sankaranarayanan, K.: Statistical feature extraction to classify oral cancers. *J. Glob. Res. Comput. Sci.* **4**(2), 8–12 (2013)
12. Sheeba, O., George, J., Rajin P.K., Thomas, N., George, S.: Glaucoma detection using artificial neural network. *IACSIT Int. J. Eng. Technol.* **6**(2), 158–161 (2014).
13. Bhateja, V., Tiwari, A., Gautam,A.: Classification of mammograms using sigmoidal transformation and SVM. In: Smart Computing and Informatics, pp. 193–199. Springer, Singapore (2018)
14. Belgacem, R., Malek, I.T., Trabelsi H., Jabri, I.: A supervised machine learning algorithm SKVMs used for both classification and screening of glaucoma disease. *New Front. Ophthalmol.* **4**(4), 1–27 (2018)
15. Khassabi, Z., Shanbehzadeh J., Mahdavi, K.N.: A unified optic nerve head and optic cup segmentation using unsupervised neural networks for glaucoma screening. In: Proceedings of 40th International Conference of the IEEE Engineering in Medicine and Biology Society, Honolulu, HI, USA, pp. 5942–5945 (2018)
16. Karya, N., Padmaja, K.V.: Glaucoma detection using texture features extraction. In: Proceedings of IEEE 51st Asilomar Conference on Signals, Systems and Computers, Pacific Grove, USA, pp. 1471–1475 (2017).
17. Haralick, R.M., Shanmugam, K., Dinstein, Textural features for image classification. *IEEE Trans. Syst., Man Cybern.* **SMC-3**(6), 610–621 (1973)
18. Chaudhari, P., Agarwal, H., Bhateja, V.: Data augmentation for cancer classification in oncogenomics: an improved KNN based approach. *J. Evol. Intell.*, pp. 1–10 (2019)
19. Sivaswamy, J., Krishnadas, S.R., Joshi, G.D., Jain M., Tabish A.U.S.: DRISHTI-GS: retinal image dataset for optic nerve head (ONH) segmentation. *IEEE 11th International Symposium on Biomedical Imaging*, Beijing, China, pp. 1–4 (2014).
20. Roychowdhury, S., Koozekanani, D.D., Kuchinka S.N., Parhi, K.K.: Optic disc boundary and vessel origin segmentation of fundus images. *IEEE J. Biomed. Health Inf.* **20**(6), 1562–1574 (2015)
21. Tiwari, A., Bhateja, V., Gautam A., Satapathy, S.C.: ANN-based classification of mammograms using nonlinear pre-processing. In: Proceedings of 2nd International Conference on Micro-Electronics, Electromagnetics and Telecommunications, Singapore, pp. 375–382 (2018).

# Analysis of Encryption Algorithm for Data Security in Cloud Computing



Arijit Dutta, Akash Bhattacharyya, Chinmaya Misra,  
and Sudhangshu Sekhar Patra

**Abstract** Cloud computing offers various types of resources to its users. In recent years with the development of Cloud Computing, there has been an increase in the use of storage service from the cloud service providing organizations. However, there are still active issues regarding security, privacy, interoperability and reliability, which needs to be solved fast. Among these issues the issue regarding the security of cloud storage and how the same can be provided easily. This paper has helped in the study and proposal of a simple and secure as well as privacy-preserving design and architecture to cloud data sharing. The architecture has been based on the process of encryption and decryption algorithms for the protection of data stored in the cloud servers from unauthorized and illegal access.

## 1 Introduction

Cloud computing providers providing the customers with virtualized resources accompanied by various technologies like multi-tenancy, virtualization and web services. This makes the use of web applications for the cloud services one of the most important components of cloud computing [1]. The process gives rise to the concept of multi-tenancy in the process of cloud resources being shared. However, even with the provision of the right number of resources, there is the inclusion of exclusive risks among which the security risks being of the topmost priority [2]. The models are designed along with the characteristics and the proper technologies. The main characteristics, which needs to be present for a proper cloud environment, are multi-tenancy and virtualization [3]. Cloud services along with the technologies and deployment models are responsible for the introduction of cloud-related security risks and vulnerabilities. Similarly, it has been found that the virtual environment

---

A. Dutta · A. Bhattacharyya

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

C. Misra (✉) · S. S. Patra

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

gives rise to their own set of risks and various vulnerabilities including malicious interaction between virtual machines and virtual machine escape [4]. The implementation of multi-tenancy accessing the same data sector in the cloud also gives rise to the same issues with the security of data on the cloud. To provide dedicated resources for the users of the cloud is a complex work and would thus require a large amount of security. The following section is dedicated to the discussion of the security challenges that are currently being faced by the cloud computing environment.

## 2 Literature Survey

Rimal et al. [5] have been able to provide a classification of various cloud computing system. The main aim set forward by the author for this paper was to create a proper disciplined process for the scattered resources which will be incurring the least expensive with the highest throughput to achieve the best comfort in cloud computing. Jain [6] in his paper discusses the various cloud computing models, the research challenges cloud computing faces, as well as the security challenges of cloud computing. Padhy [7] has discussed the overview of the effect of the cloud computing environment and the various cloud services available.

Advanced encryption standard (AES) is a widely used symmetric block cipher. Asymmetric cryptographic key of 128 bit is used for this algorithm. AES has been seen to be able to be working in most of the cloud environment to provide security. The users first decide on the CSP and provide a requirement list to the CSP. The requirement is the service requirements of the user. During the migration of the data, all the data being uploaded would be encrypted with the AES algorithm. After encryption, the data is forwarded to the CSP for storage. A request for the reading of the data would be sent to the CSP; the data would be decrypted and then sent to the user.

The data encryption standard (DES) encryption methodology is a block cipher algorithm that uses a block size of 64 bits each. A block of 64-bit plain text is encrypted using a symmetric key algorithm to produce 64-bit ciphertext. The key used in this algorithm is 56 bits long even though a 64-bit key is provided.

One of the most common forms of public-key encryption key algorithms is RSA. The algorithm follows asymmetric key cryptography. The public key is used for the encryption method and is shared over the network with all. The private key is used for the decryption of the ciphertext and is not shared with anyone. Blowfish is another encryption algorithm, which uses symmetric key encryption methodology. The algorithm encrypts a 64-bit block of plain text with a key value of the variable length of 128–448 bits. Blowfish algorithm is best suited for applications and environment where the key value remains constant for a long period (communications linking encryption) and not for the applications where the key changes frequently (packet switching method of data interchange) (Table 1).

**Table 1** Comparative study of AES, DES, RSA and blowfish algorithm

Characteristics	AES	RSA	DES	Blowfish
Platform	Cloud environment	Cloud environment	Cloud environment	Cloud environment
Scalability	Positive	Negative	Positive	Positive
Security	Both user and providers are secured	Only users are secured	Both user and providers are secured	Both user and providers are secured
Data encryption capacity	The large quantity of data	Small data only	Lesser than AES	Lesser than AES
Authentication type	Best authentication process	Robust authentication	Less than AES	Almost similar to AES
Memory usage	RAM required is low	Highest requirement	Required more than AES	Can be executed within 5 KB
Execution time	Fastest	Slowest	Same as AES	Less time for execution

### 3 Proposed Method of Hybrid Encryption Model

This section discusses a proposed hybrid encryption algorithm, which would be able to provide a high amount of data security in the cloud environment. Based on the theory, two models have been proposed: two-tier model comprised of RSA and blowfish, three-tier model comprised of RSA, blowfish and DES algorithm. The three methods have been selected for the following reasons:

1. Fast working of DES algorithm with short key generation
2. Increased security with long key generation in blowfish
3. RSA has high compatibility with Internet processing and higher security.

The use of the proposed hybrid model would be able to reduce the complexity of the data encryption when the data is at rest as well as during transmission. The model makes use of both asymmetric key cryptography and symmetric key cryptography algorithms to provide a higher amount of data security to the cloud environment.

#### 3.1 Two-Tier Hybrid Model

The two-tier architecture model has been designed based on the theory of providing stronger strength to the users' data protection and security in the cloud environment. In the model, the process follows a simple collection of steps as discussed (Fig. 1):

1. The sender initiates the algorithm and invokes both the encryption algorithms: RSA and blowfish

A 2-tier RSA-Blowfish Encryption Algorithm	A 2-tier RSA-Blowfish Decryption Algorithm
<pre> Start SENDER initiate encrypt Sys Encrypt Sys INITIATE RSA, BWF RSA GENERATE RSAprivatekey &amp; RSApublickey BWF GENERATE BWFsecretkey If (Encrypt Sys != RSAkeys AND Encrypt Sys != BWFkey) {     Encrypt Sys REQUEST RSAprivatekey &amp; RSApublickey     Encrypt Sys REQUEST BWFkey     Encrypt Sys RECEIYE RSAprivatekey &amp; RSApublickey     Encrypt Sys RECEIVE BWFkey; } Else     SENDER UPLOAD docXYZ     Encrypt Sys E(docXYZ) &lt;- BWFkey     Encrypt Sys E(E(docXYZ) &lt;- BWFkey) &lt;- RSApublickey     STORE (E(E(docXYZ) &lt;- BWFkey) &lt;- RSApublickey) Endif;</pre>	<pre> Start RECEIVER INITIATE Decrypt Sys Decrypt Sys INITIATE RSA, BWF RECEIVER DOWNLOAD (E(E(docXYZ) &lt;- BWFkey) &lt;- RSApublickey) from STORE While (docXYZ ≠ 0) {     Decrypt Sys D(E(BWFkey)) &lt;- RSAprivatekey     Decrypt Sys D(E(docXYZ)) &lt;- BWFkey     RECEIVER gets ORIGINAL docXYZ } Endwhile;</pre>

**Fig. 1** Encryption and decryption algorithm for the two-tier hybrid model (Source [6])

2. RSA algorithm generates the public and private keys for the algorithm
3. Blowfish generates the secret key for the models' encryption system
4. The sender chooses the file to be encrypted
5. Blowfish algorithm first encrypts the file with the secret key generated
6. The RSA public key generated is then encrypted with the help of the blowfish secret key
7. Both the file and the encrypted key are stored in the cloud storage until requested.

During the decryption process, the model uses the following process:

1. The client requests the file to be read
2. The model has already stored the blowfish secret key as well as the RSA keys in the cloud system during encryption
3. The RSA private key is used to decrypt the blowfish secret key
4. The secret key is then used to decrypt the file which the client has requested.

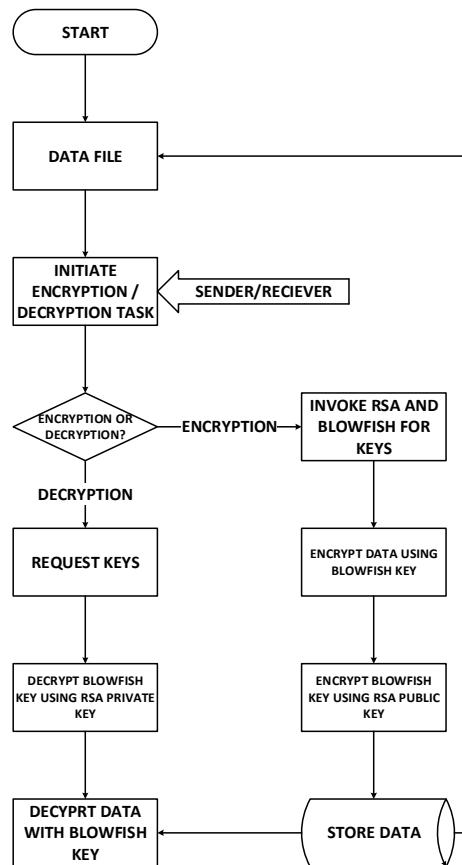
The security lies with the RSA private key, which is present only with the client. Without the proper private key, the blowfish secret key cannot be decrypted (Fig. 2).

### 3.2 Three-Tier Hybrid Model

The proposed method of a three-tier model is based on a study where DES and RSA were separately used to provide data security. The three-tier hybrid model is considered to be an improved version of the two-tier architecture. When a file is to be uploaded to the cloud, the model follows the following steps (Fig. 3):

1. Invoke all the three algorithms to generate the keys: blowfish secret encryption key, RSA public and private keys and DES secret encryption key
2. The file is first encrypted with the blowfish secret key
3. The output is again encrypted with the DES secret key
4. The blowfish secret key and DES secret key are encrypted with the help of RSA public key

**Fig. 2** Flowchart for two-tier hybrid model

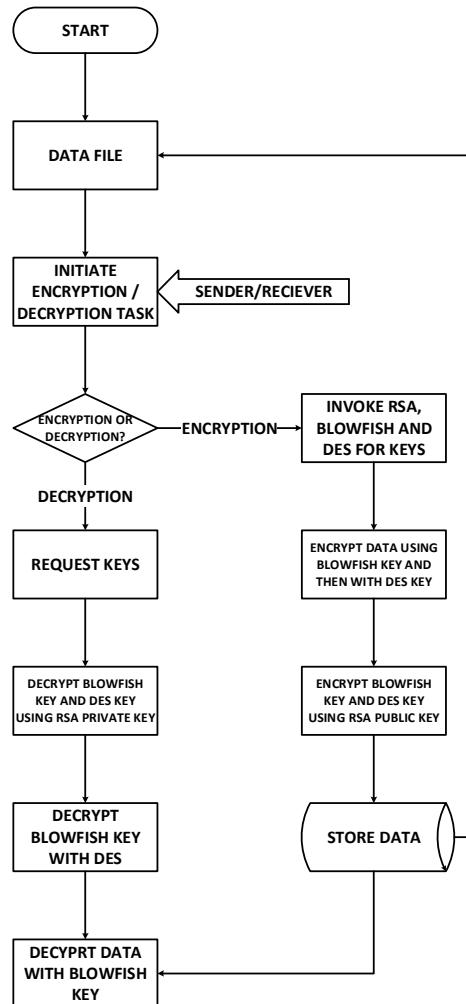


5. The two files are then stored in the cloud environment.

The process for the decryption method is similar to the two-tier architecture as follows:

1. As all the keys are already present in the cloud, the files are collected for decryption
2. The blowfish secret key and DES secret key are decrypted with the help of RSA private key of the client
3. The DES key is used to first decrypt the file requested
4. The output file is then decrypted with the help of blowfish secret key to obtain the original file uploaded by the client.

**Fig. 3** Flowchart for three-tier hybrid model. (*Source* Made by Author)



#### 4 Empirical Evaluation of the Method and Results

This section discusses the performance analysis of the two proposed hybrid model. The attributes selected for the analysis of the performance are security provision strength, time consumption calculated based on the encryption and decryption time on various file sizes of texts, image and video. Along with these, the size of the file was calculated before and after the encryption as well as after decryption. Several iterations were conducted to ensure the consistency of the results found.

The computation was completed based on three types of data files: text, image and video. The main objective was to determine which model would be the most time-efficient. The three files when encrypted or decrypted, the time required was

A 3-tier RSA-Blowfish-DES Encryption Algorithm	A 3-tier RSA-Blowfish-DES Decryption Algorithm
<pre> Start SENDER INITIATE Encrypt Sys Encrypt Sys INITIATE RSA, BWF, DES RSA GENERATE RSAprivatekey &amp; RSApublickey BWF GENERATE BWFseckeys DES GENERATE DESseckeys If ((Encrypt Sys != RSAkeys) AND (Encrypt Sys != BWFkey) AND (Encrypt Sys != DESkey)) {     Encrypt Sys REQUEST RSAprivatekey &amp; RSApublickey     Encrypt Sys REQUEST BWFseckeys     Encrypt Sys REQUEST DESseckeys     Encrypt Sys RECEIVE RSAprivatekey &amp; RSApublickey     Encrypt Sys RECEIVE BWFseckeys     Encrypt Sys RECEIVE DESseckeys } Else SENDER UPLOAD docXYZ to Encrypt Sys 1<sup>st</sup> Encrypt Sys E(docXYZ) &lt;- BWFseckeys 2<sup>nd</sup> Encrypt Sys E(E(docXYZ) &lt;- BWFseckeys) &lt;- DESseckeys 1<sup>st</sup> Encrypt Sys E(BWFseckeys) &lt;- RSApublickey 2<sup>nd</sup> Encrypt Sys E(DESseckeys) &lt;- RSApublickey STORE: ((E(E(docXYZ) &lt;- BWFseckeys) &lt;- DESseckeys),+ E(BWFseckeys) &lt;- RSApublickey + E(DESseckeys) &lt;- RSApublickey) in cloud data store </pre>	<pre> Start RECEIVER INITIATE Decrypt Sys Decrypt Sys INITIATE RSA, BWF, DES RECEIVER DOWNLOAD     ((E(E(docXYZ) &lt;- BWFseckeys) &lt;- DESseckeys),+     E(BWFseckeys) &lt;- RSApublickey +     E(DESseckeys) &lt;- RSApublickey) from STORE While (docXYZ ≠ 0) {     1<sup>st</sup> Decrypt Sys D(E(BWFseckeys)) &lt;- RSAprivatekey     2<sup>nd</sup> Decrypt Sys D(E(DESseckeys)) &lt;- RSAprivatekey     1<sup>st</sup> Decrypt Sys D(E(docXYZ)) &lt;- D(E(DESseckeys))     2<sup>nd</sup> Decrypt Sys D(D(E(docXYZ))) &lt;- D(E(BWFseckeys)) } RECEIVER gets ORIGINAL docXYZ Endwhile; </pre>

**Fig. 4** Encryption and decryption algorithm for the three-tier hybrid model (*Source [6]*)

found to be directly proportional to the size of the file being used. It was found that the time required for the completion of the process in the two-tier model was less compared to the three-tier model. This is due to the use of the three-encryption algorithm (Fig. 4).

The most security was found to be in the three-tier architecture due to the following three factors:

1. Three computational steps of encryption increase the security factor largely. The use of RSA has been proven the most secured in the recent encryption history. As the encryption process is conducted before the transmission of the files, the symmetric keys are well-secured during the transmission process.
2. RSA and blowfish have been found to provide larger key blocks. Thus when the encryption of the blowfish secret key with RSA would result in a larger key size, which is harder to break.
3. DES has a smaller key size, which compensates with the larger key size of the other two algorithms.

## 5 Conclusion

Based on the above three factors, the three-tier hybrid model can be considered to be the most secured algorithm than the two-tier model.

## References

1. Duan, Q., Yan, Y., Vasilakos, A.V.: A survey on service-oriented network virtualization toward convergence of networking and cloud computing. *IEEE Trans. Netw. Serv. Manage.* **9**(4), 373–392 (2012)
2. Guan, B., Wu, J., Wang, Y., Khan, S.U.: CIVSched: a communication-aware inter-VM scheduling technique for decreased network latency between co-located VMs. *IEEE Trans. Cloud Comput.* **2**(3), 320–332 (2014)
3. Xiao, Z., Xiao, Y.: Security and privacy in cloud computing. *IEEE Commun. Surv. Tutorials* **15**(2), 843–859 (2013)
4. Tari, Z.: Security and privacy in cloud computing. *IEEE Cloud Comput.* **1**(1), 54–57 (2014)
5. Rimal, B.P., Choi, E., Lumb, I.: A taxonomy and survey of cloud scheming systems. In: IEEE Fifth International Joint Conference on INC, IMS, and IDC, vol. 10, No. 2, pp. 44–51 (2009)
6. Jain, P.: Security Issues and their Solution in Cloud Computing. *Int. J. Comput. Bus. Res.* **3**(1), 1–7 (2010)
7. Padhy, R.P., Patra, M.R., Satapathy, S.C.: Cloud computing: security issues and research challenges. *Int. J. Comput. Sci. Inf. Technol. Secur.* **1**(2), 136–146 (2011)
8. Menzel, M., Ranjan, R., Wang, L., Khan, S.U., Chen, J.: CloudGenius: a hybrid decision support method for automating the migration of web application clusters to public clouds. *IEEE Trans. Comput.* (2014). <https://doi.org/10.1109/TC.2014.2317188>
9. Latif, R., Abbas, H., Assar, S., Ali, Q.: Cloud computing risk assessment: a systematic literature review. In: Future Information Technology, pp. 285–295. Springer, Berlin, Heidelberg (2014)
10. Fernandes, D.A.B., Soares, L.F.B., Gomes, J.V., Freire, M.M., Inácio, P.R.M.: Security issues in cloud environments: a survey. *Int. J. Inform. Sec.* **13**(2), 113–170 (2014)
11. Khan, A.N., Kiah, M.L.M., Ali, M., Madani, S.A., Shamshirband, S.: BSS: block-based sharing scheme for secure data storage services in mobile cloud environment. *J. Supercomput.* **70**(2), 946–976 (2014)
12. Rong, C., Nguyen, S.T., Jaatun, M.G.: Beyond lightning: a survey on security challenges in cloud computing. *Comput. Electr. Eng.* **39**(1), 47–54 (2013)
13. Subashini, S., Kavitha, V.: A survey on security issues in service delivery models of cloud computing. *J. Netw. Comput. Appl.* **34**(1), 1–11 (2011)
14. Modi, C., Patel, D., Borisaniya, B., Patel, A., Rajarajan, M.: A survey on security issues and solutions at different layers of cloud computing. *J. Supercomput.* **63**(2), 561–592 (2013)
15. Hashizume, K., Rosado, D.G., Fernández-Medina, E., Fernandez, E.B.: An analysis of security issues for cloud computing. *J. Internet Serv. Appl.* **4**(1), 1–13 (2013)
16. Che, J., Duan, Y., Zhang, T., Fan, J.: Study on the security models and strategies of cloud computing. *Proc. Eng.* **23**, 586–593 (2011)

# A Machine Learning Approach in Data Perturbation for Privacy-Preserving Data Mining



Jayanti Dansana and Adarsh Singh

**Abstract** Data mining is a process where we can extract relevant information or patterns from the collection of data. In this era of big data, every organization aims to handle huge amounts of data and perform data mining techniques in order to extract pieces of information or patterns for various work and decision making. To protect privileged data and leakage of private information, the clients use different privacy-preserving techniques such as perturbation that protects client's data from revealing private information. The job of perturbing data on the client side is a herculean task, and it gets more difficult with the increase in the size of data. In this paper, we proposed a machine learning regression model that has been trained in such a way that it predicts the perturb data from original data and it even contains a comparative study of different regression models and their accuracy in perturbing the data.

## 1 Introduction

Data come with huge complexity and size, and in order to bring out necessary pieces of information or patterns, different techniques are used. Organizations today not only face problems in application of different information extracting techniques but also face problems in collecting the relevant datasets from different clients. Therefore, it makes up a distributed system where the data from different parties are collected at the server side, and the collected data are then fed to different information extracting techniques for getting the needed information. During this process of sharing data over distributed system, the parties face the problem of preserving its privacy since the private data that the parties share with the other parties can get misused. In order to avoid such situations, privacy-preserving techniques were developed that can preserve the privacy of data getting shared on a distributed system, and these techniques are divided into three major categories such as cryptography, perturbation, and anonymization. Data perturbation techniques have widely been used by various

---

J. Dansana (✉) · A. Singh  
KIIT Deemed To Be University, Bhubaneswar, Odisha, India  
e-mail: [jayantifcs@kiit.ac.in](mailto:jayantifcs@kiit.ac.in)

parties for perturbing their data before sending it to the server side. To preserve their privacy and to avoid unnecessary threats to any individual, perturbation technique includes various methods for preserving the privacy. In this paper, we have studied the noise addition schema used for privacy preservation, which has widely been accepted, but the implementation bringing a huge workload over to the client-side which by design increases as we increase the size of data that needs to be perturbed. By using the machine learning approach, the workload on the client-side can drastically be reduced.

The rest of the paper is organized as follows: Sect. 2 recapitulates the various works in this area. Section 3 explains the proposed algorithm and its usage in distributed architecture. Section 4 explains the experimentation carried out using regression models. Finally, the paper is concluded with its contribution in the field of privacy-preserving data mining and its future work.

## 2 Literature Survey

### 2.1 Related Work

In previous years, distributed data collection came into the picture, and with the rise of the threat to privacy for client's data, different privacy-preserving algorithms have been developed which have been found to be really useful and effective in preserving the privacy of any data but surely with a trade-off between accuracy and privacy. These different approaches of preserving privacy data mining (PPDM) in a distributed scenario studied extensively by the author in [1]. This approach has been implemented and experimented, and its relevance has also been calculated in [2]. These different approaches of preserving privacy data mining (PPDM) in a distributed scenario has been studied extensively by the author in [1, 3] and been implemented and experimented with by [2]. These privacy-preserving techniques consist of different methods such as oblivious transfer—a protocol in which a sender sends one of the relevant pieces of information to the accepter but remains oblivious to the whole information sent to the accepter, this methodology has been well implemented by Ding and Klein [4]. Secure multi-party computation is a type of protocol which does not reveal any data other than the result, and this model of privacy preserving has been well implemented by Dansana et al. [5] has been analyzed. Secure multi-party computation — a protocol that does not reveal any data other than the result, and this model of privacy-preserving has been well implemented by Dansana et al. [5] and has been analyzed by Teo et al. [6].

Perturbation is a simple yet effective technique to preserve the privacy of data as this approach deals with manipulation, randomization, distortion, or deformation of data in such a way that it creates a great trade-off between the preserving privacy and getting accuracy results on the server size, and the technique of privacy-preserving involves different methods as discussed involves modification by means of noise

addition [9–12], data accretion [13, 14], data exchange [15, 16], repression [17, 18], original transformation [19, 20], and randomization [21].

Anonymization is a type of information suppressing technique involving steps like encryption of data or removal of identifiable information, comparatively studied by Rashid and Hegazy [22].

## 2.2 Overview of Noise Addition Schema in Decision Tree

The machine learning regression model proposed in this paper is trained by using the manually calculated perturbed data obtained from the noise addition perturbation technique [16] as the target class of the training dataset, with a set of engineered features.

This approach of perturbation [12] uses noise addition technique by adding specific noise to the numeric attributes after obtaining the decision tree of the original data. It involves different algorithms for noise addition in numerical attributes and categorical attributes, and it first checks if the attribute is numerical or categorical and then uses perturbation techniques for leaf reaching path attributes (PTLRPA) [16] and perturbation techniques for leaf wrong path attributes (PTLWPA) [16] for numerical data or categorical attribute perturbation technique (CAPT) [16] for categorical data (Table 1).

This was an Naive overview of noise addition schema which has been used manually to perturb the individual attribute in order to construct the target class of our training dataset upon which the regression model will be trained and eventually generate the predicted perturbed value of the testing dataset, and thereafter, this model will work in a distributed system where multiple parties will be using the models to perturb their data and then sending it to the server where the collected data will be used for data mining (Table 2).

**Table 1** Perturbed data

Kyphosis	Age	Number	Start
Absent	61.931577	2.931577	17.931577
Absent	37.876281	3.876281	16.876281
Absent	113.437496	2.437496	16.437496
Present	59.078497	6.078497	12.078497
Present	82.068515	5.068515	14.068515
Absent	148.380582	3.380582	16.380582
Absent	18.593798	5.593798	2.593798
Absent	1.635813	4.635813	12.635813
Absent	168.380032	3.380032	18.380032
Absent	1.87747	3.87747	16.87747

**Table 2** Unperturbed data

Kyphosis	Age	Number	Start
Absent	61	2	17
Absent	37	3	16
Absent	113	2	16
Present	59	6	12
Present	82	5	14
Absent	148	3	16
Absent	18	5	2
Absent	1	4	12
Absent	168	3	18
Absent	1	3	16

### 3 Privacy-Preserving Perturbation

In order to obtain the training data, we have used the noise addition schema in decision trees for privacy-preservation [12] for which we trained the CART decision tree algorithm on the diabetes dataset and went on calculating the perturbed form of each attribute through base traditional approach [12]. The perturbed data can be used as the target class whereas the original data will be used as the feature class for training the model.

#### 3.1 Construction of Training Dataset

The machine learning regression approach needs huge datasets to get trained and predict values, but for obtaining a good training dataset such that it does not over fit or under fit the model, we need to have relevant attributes and class label, and to construct the training data, we have kept four attributes

1. The numeric attribute
2. These three statistical attributes are probabilistic density function (PDF), mean and standard deviation.
3. Mean of the attribute.
4. The perturbed data manually calculated through noise addition schema [12]

In order to train the model, a large dataset is required. So, each attribute is aligned with their normal distribution, mean and its perturbed data horizontally in such a way that it increases the data size and even aligned it in such a way that it will be able to train the model effectively.

### **3.2 Model Planning**

In this phase, we planned the methods or the workflow we intend to use to construct the model for getting a feasible perturbation according to the training data and our problem statement. Based on the problem statement, four regression analyses are considered for experimentation.

1. Linear regression
2. Lasso regression
3. Random forest regression
4. Decision tree regression.

### **3.3 Model Building**

In the third phase, training will be done with the planned model with the constructed training data, making sure that it does not over fit or generalize our model and also considering whether its existing tools will suffice for running the models, or if it needs a more robust environment for executing the model and workflow. Thereafter, we used this trained model for finding the perturb value of our testing dataset which consists of only three attributes. The probabilistic density function (PDF) of each record for a normal probabilistic density function where mean is the attribute's mean and standard deviation is the attribute's standard deviation.

### **3.4 Algorithm for Model Perturbation**

Let  $X$  be the attribute that needs to be perturbed.

Step 1—Construct class label.

Calculate the perturbed data manually for distributing  $X$  through noise addition schema [12].

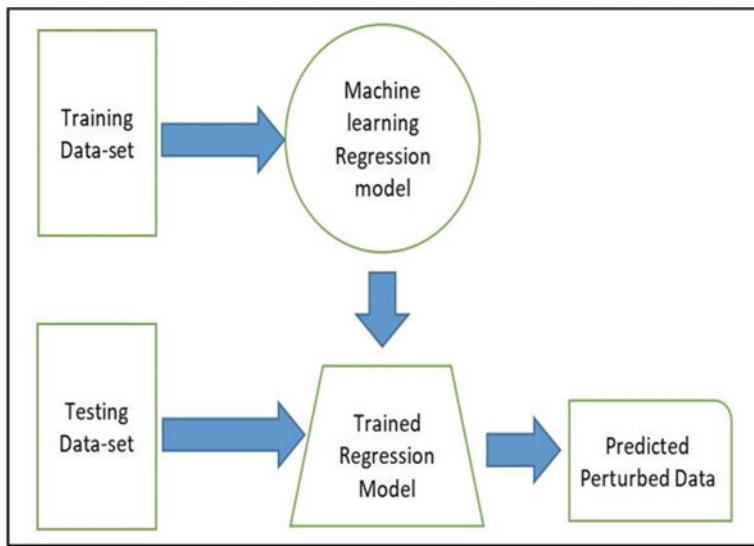
Step 2—Construction of training data: In this step, we have added new columns such as mean, standard deviation and probability density function of the records along with the attribute  $X$  in the feature class and the perturbed form of attribute  $X$  obtained from step-1 as the target class.

Step 3—Regression model planning: In this step, we plan to use the regression model considering the following parameters, i.e., type of problem, size of training data, accuracy, linearity, and training time.

Step 4—Model training: The model will be trained using the regression model considering the parameter mentioned in step-3.

Step-5—Predicting perturbation.

This step will generate predicted perturbed form of the original data by feeding our original dataset in testing data format, into our trained model from Step-4, which

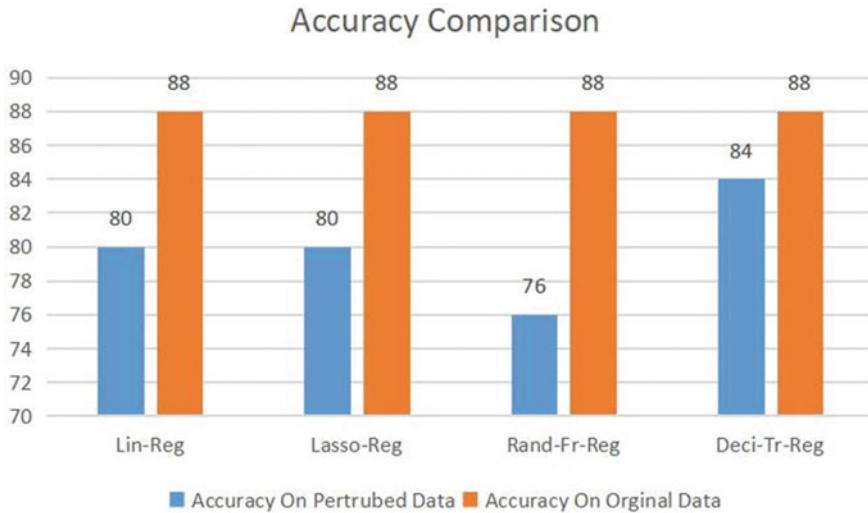


**Fig. 1** Training model

now can be sent to the server where it gets integrated with other client's data, and finally, data mining tools are applied on the collected data for information extraction (Fig. 1).

### 3.5 *Distributed Data Mining Architecture*

Data mining requires large amount of data in order to extract useful information and to help in decision making, but these large amount data are usually not available at one place, therefore, we seek distributed data mining in which the data are collected from individual parties at the server, where the data mining tools are applied, and this reduces the complexity, memory cost and preserves privacy through resource sharing [24]. In our proposed work, we have implemented our trained machine learning perturbation model to work for distributed data mining. Each party will have their own individual trained model which will perturb the data that they need to send to the server.



**Fig. 2** Comparison on different classifier accuracy

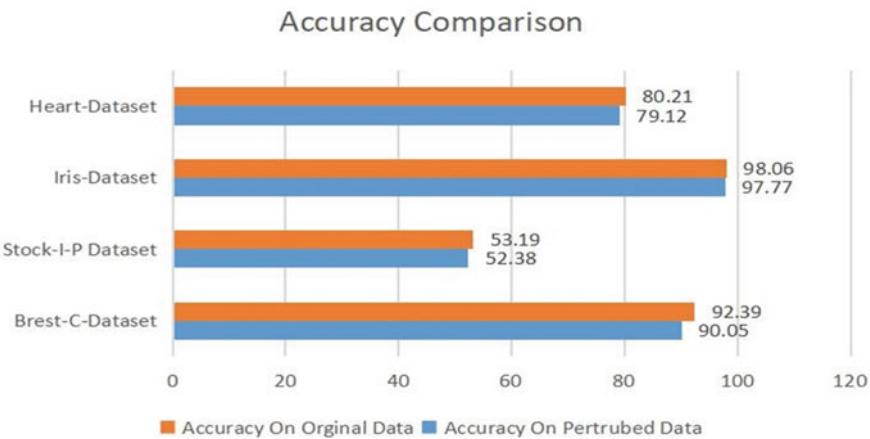
## 4 Experiments and Results

In the initial phase, we conducted our experiment on test dataset, to find the suitable model and thereafter with the other dataset. The model is trained on our training dataset.

Decision tree regression has performed the best out of four algorithms with minimum accuracy difference. There is a difference between the accuracies of decision tree classifier [22] on unperturbed dataset and perturbed dataset since it is not an ideal case, therefore, we cannot accept the perturbed dataset to act same as the unperturbed, and we need to have a trade-off between our privacy that has been preserved [23] and the accuracy of the dataset. We have done the accuracy comparison of decision tree classifier when trained [24] upon model's perturbed data and the original data on four different dataset mentioned so as to showcase the efficiency and effectiveness of this approach. Decision tree regressor algorithm has been used for model building, since it has shown desirable results on the dataset (Figs. 2 and 3; Tables 3 and 4).

## 5 Conclusion

The proposed machine learning regression model approach is an efficient way which reduces the time complexity for perturbation on the client side. This approach makes sure that the criticism faced by different methods of perturbation does not hold on when implemented through this technique. Thereafter, this regression model



**Fig. 3** Accuracy comparison on different data set

**Table 3** Decision tree classifier accuracy

Algorithms	Decision tree classifier accuracy (in %) (Perturbed data)	Decision tree classifier accuracy (in %) (Original data)
Linear regression	80	88
Lasso regression	80	
Random forest regressor	76	
Decision tree regressor	84	

**Table 4** Comparative accuracy on different data set using Lasso regression

Dataset	Decision tree classifier accuracy (in %) (perturbed data)	Decision tree classifier accuracy (in %) (original data)
Brest-cancer dataset	90.05	92.39
Stock-index prediction	52.38	53.19
Iris dataset	97.77	98.06
Heart disease	79.12	80.21

approach can be effectively used for data perturbation and preserving privacy in a distributed system for data mining. In the context of distributed data mining, this approach of data perturbation deals with the issues of privacy-preserving data mining. Future research, this approach of model perturbation can be used for implementing other better perturbation techniques with better training and hyperparameter tuning so as to propose a better framework in privacy.

## References

1. Shah, A., Gulati R.: Privacy-preserving data mining: techniques, classification and implication—A survey. *Int. J. Comput. Appl.* **137**(12) (2016)
2. Pinkas B.: Cryptographic techniques for privacy-preserving data mining ACM SIGKDD Explor. **4**(2), 12–19 (2002)
3. Shah, A., Gulati, R.: A survey on cryptographic techniques for privacy-preserving data mining. *Int. J. Data Wareh. Mining* **2**(1), 8–12 (2012)
4. Ding Y., Klein K.: Model-Driven application-level encryption for the privacy of e-health data. In: International Conference on Availability, Reliability and Security(2010)
5. Dansana, J., Kumar, R., Dey, D.: Modified Ck secure sum algorithm in horizontally partitioned databases. In: International Conference on Research and Development Prospects on Engineering Technology, ICRDPET-2013, vol. 5
6. Teo, S.G., Shuguo Han, V.L.: A study of efficiency and accuracy of secure multiparty protocol in privacy-preserving data mining. In: 26th International Conference on Advanced Information Networking and Applications Workshops, pp. 85–90 (2012)
7. Zhan, J., Matwin, S., Chang, L.W.: Privacy-preserving collaborative association rule mining, *J. Netw. Comput. Appl.* **30**(3), 1216–1227 (2007)
8. Kargupta, H., Datta, S., Wang, Q., Sivakumar, K.: On the privacy-preserving properties of random data perturbation techniques. In ICDM, pp. 99–106. IEEE Computer Society (2003)
9. Agrawal, R., Srikant, R.: Privacy-preserving data mining. In: The Proceedings of the ACM SIGMOD Conference, pp. 429–450 (2000)
10. Muralidhar, K., Sarathy, R.: A general additive data perturbation method for database security. *J. Manage. Sci.* **45**(10), 1399–1415 (2002)
11. Agrawal, D., Aggarwal, C.C.: On the design and quantification of privacy-preserving data mining algorithms. In: ACM PODS Conference, pp. 247–255 (2002)
12. Kadampur, M.A., Somayajulu, D.V.L.N.: A noise addition scheme in decision tree for privacy-preserving data mining. *J. Comput.* **2**(1), 137–144 (2010)
13. Li, Y., Zhu, S., Wang, L., Jajodia, S.: A privacy-enhanced micro aggregation method. In: Proceedings of 2nd International Symposium on Foundations of Information and Knowledge Systems, pp. 148–159 (2002)
14. Iyengar, V.S.: Transforming data to satisfy privacy constraints. In: Proceedings of SIGKDD'02, Edmonton, Alberta, Canada (2002)
15. Fienberg, S.E., McIntyre, J.: Data swapping: variations on a theme by Dalenius and Reiss. *J. Off. Stat.* **2**, 309–323 (2005)
16. Muralidhar, K., Sarathy, R.: Data shuffling—A new masking approach for numerical data. *Manage. Sci. Forthcoming* **52**(9), 658–670 (2006)
17. Hintoglu, A.A., Saygin, Y.: Suppressing microdata to prevent probabilistic classification based inference. In: Proceedings of Secure Data Management, 2nd VLDB Workshop, pp. 155–169 (2005)
18. Rizvi, S., Harista, J.R.: Maintaining data privacy in association rule mining. In: Proceedings of 28th VLDB Conference, Honk Kong China, pp. 682–693 (2002)
19. Mukherjee, S., Chen, S., Gangopadhyay, A.: A privacy-preserving technique for euclidean distance-based mining algorithms using fourier-related transforms. *VLDB J.* 293–315 (2006)
20. Xu, S., Lai, S.: Fast Fourier transform based data perturbation method for privacy protection. In: Proceedings of the IEEE Conference on Intelligence and Security Informatics, pp. 222–225. New Brunswick, New Jersey (2007)
21. Malik, M.B., Ghazi, M.A., Ali, R.: Privacy-preserving data mining techniques: current scenario and future prospect. In: Third International Conference on Computer and Communication Technology, pp. 26–32 (2012)
22. Rashid, A.H. Hegazy, A.: Protect privacy of medical informatics using K-anonymization model. In: The 7th International Conference on Informatics and Systems (INFOS), pp. 1–10. Cairo (2010)

23. Rizvi, S., Harista, J.: Maintaining data privacy in association rule mining. In: Proceedings of 28th VLDB Conference, pp. 82–693. Honk Kong, China (2002)
24. Dansana, J., Kumar, R., Rautaray, J.: Techniques for privacy preserving association rule mining in distributed database. IRACST Int. J. Comput. Sci. Inf. Technol. Secur. (IJCSITS) **2**(6) (2012)

# IoT Service-Based Crowdsourcing Ecosystem in Smart Cities



Arijit Dutta, Ruben Roy, Chinmaya Misra, and Kamakhya Singh

**Abstract** In this information era, online access gives netizens the opportunity to post or interact with each other. Online crowdsourcing is also known as citizen science. The rapid increase of service and demand for IoT devices drive the attention towards crowdsourcing, and the most eye-seeking part is that it involves the participation of the individual. It also can be stated as receiving services, ideas, or solutions to problems by getting connected to large groups of people via the Internet. Obviously, having a huge, well-disciplined network would be of prime importance in this case. Millions of devices are connected which may congest the network due to heavy traffic. To deal with the situation, researchers are explored different aspects to handle the situations. So, in this paper, we propose a reliable fog-based spatial-temporal crowdsourcing. Here, we try to tinker the data processing that takes place between fog nodes and cloud. We use the batch processing method for saving energy. However, how to allocate the tasks to proper fog nodes and eventually improve communication efficiency are critical.

## 1 Introduction

The Internet has become such a big thing in our lives; for any information, we look for Google and most of the time Google point towards Wikipedia where we get to know almost everything. To build such a platform having a huge amount of information is a big thing, the concept of crowdsourcing plays a vital role and makes it most comprehensive. Thanks to our growing connectivity due to which concepts like crowdsourcing come into the picture. It increases productivity and reduces labour expenses.

---

A. Dutta

School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, India

R. Roy

Department of IT, Government College of Engineering and Leather Technology, Kolkata, India

C. Misra (✉) · K. Singh

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

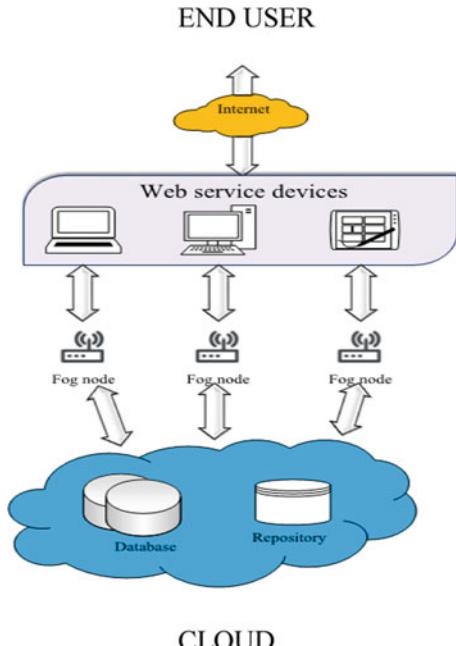
It is a sort of invitation of broad communities who are eventually customers and execute it on a large scale. Anybody can handle crowdsourcing, namely companies, governments, groups, and even individuals. Crowdsourcing works in four different ways [1, 2].

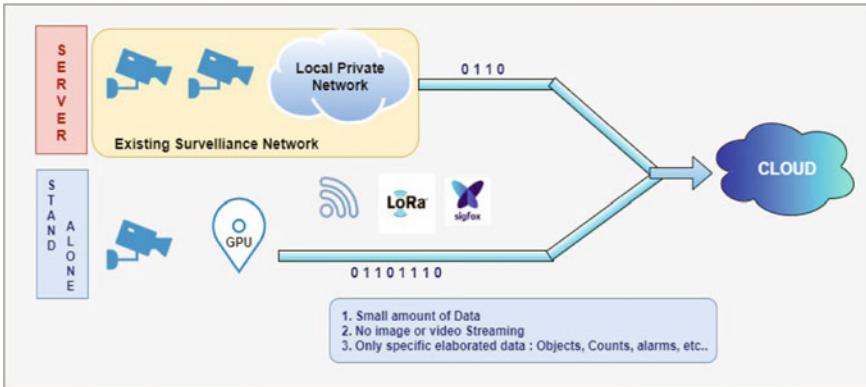
Firstly, it enables us to have a huge online labour force for our work. We identify our workers by some selection criteria or we can let the interested labour find our work. Secondly, we can post our work online for which we want a solution. Thirdly, there could be a scenario where we have the knowledge available with us, but in a much-scattered way. Hence, we would need the help of people to filter and organize it. Lastly, we also need the crowd to get feedback or opinions and ideas. People use crowdsourcing for different purposes. When we take the help of the crowd to collect and organize information, it is termed as accessing distributive knowledge (Fig. 1).

The process of collecting funds from the people believing in our cause is also facilitated by crowdsourcing and is termed as crowdfunding [3]. Each one can make small donations which adds up to produce a good sum. When we use crowdsourcing to get access to an on-demand scalable workforce, it is called crowd labour (Fig. 2).

In this world of information era, the use of different sensors and applications based on IoT gaining popularity rapidly[3]. By this year end, Cisco companies expecting that 50 billion devices will send data and this huge data volume may congest our network. To reduce the load of networking traffic, lots of network management scheme is introduced so that it can be managed in a better manner. Cloud and fog architecture provides better data management in this scenario [4, 5].

**Fig. 1** Fog-based smart city architecture





**Fig. 2** Graphical representation of the workflow of crowdsourcing in IoT solutions

We have observed that cloud computing scheme also facing some issues like QoS like latency, jitter, and delay or packet loss [6–8]. To improve this QoS, we have to satisfy the constraints like bandwidth constraint, energy consumption, and delay-sensitive real-life streaming handling. Fog and edge computing give us some light to handle these issues, and it has been all accepted by both industries and academics [9, 10].

Internet of things encounters volatile data streams, and because of obvious reasons, the situation is going to be more severe, data is to be processed change permanently, so the computational resources needed to be reconfigured and most importantly fog provides a better edge towards reducing latency than the cloud-based computation [2, 11].

From various survey results, it is found that every year we lose 3.1 trillion dollars due to delay of data and poor quality of data, The security, reliability, scalability, cost, and response time are some of the areas to keep an eye on. The bandwidth for the increased data increments by 50% every year. So data filtering is an important aspect to look after, and real-time-sensitive data are to be processed in fog nodes. We have different methods to process data like interactive processing, batch processing, and so on, and all these have their own advantages and disadvantages in different applications. Here in this paper, we are going to propose a batch processing model (Table 1).

## 2 Model Introduction

Let suppose that due to some reasons or deliberately we want a single processor working in the cloud, with this approach, we have to consider the selection of data for batch processing, data need to be analysed with a set of pattern, rules, and actions using recursive or compound analysis. A set of conditions come into the picture for

**Table 1** Themes based on research in smart cities

	Themes	Relevant studies
Smart mobility	Vehicle tracking	Lee et al. [7]
	Internet of vehicles	Zhu et al. [8]
Smart living	Public safety	Breetzke and Flowerday [9]
	Healthcare	Hussain et al. [10]
Smart environment	Air pollution, quality	Castelli et al. [11]
	Water quality	Corbett and Mellouli [12]
Smart citizens	Social interactions, communications	Sproull and Patterson [13]
Smart government	Social media	Díaz-Díaz and Pérez-González [14]
	Websites	Fietkiewicz et al. [15]
Smart economy	Smart business	Johnson et al. [16]
	M-commerce	Keegan et al. [17]
Smart architecture and technologies	Data exchange	Aguilar et al. [18]
	Data processing	Hashem et al. [19]
	Data storage	Huang et al. [20]

the aggregation of data, network definition, and meta information also plays a role here.

If we use interactive processing here, then there will be a huge amount of data in queue at waiting status so response time will be more and also energy consumption will be more because the processor is the active whole time. In the case of batch processing, the processor will be active only when an amount and type of data will be gathered from fog layer for final processing.

So according to our requirements, batch processing engines are designed to process a comparatively larger datasets efficiently. This approach is also feasible when a different type of data from different devices are processed together. Each fog node will take care of a specific dataset.

In this section, we have discussed bulk processing for our proposed IoT-based smart city cloud architecture. We have derived the response time, waiting time in buffer, and packet loss in the system. When IoT senses data and forwards it to the fog layer, the fog layer removes noise and reduced redundancy. The fog layer forwards to cloud layer for further storage.

- The arrivals of IoT data are independent and follow the Poisson distribution and they ate independently.

- The processing is independent and they follow the first come first serve (FCFS) basis, and they also follow the Poisson distribution and represented by  $\mu$ .
- The buffer size is limited and value is  $N$ .
- We have considered heterogeneous environment as in real life, the systems have different speed.

### 3 Algorithm for Saving Energy

Input: Request from IoT Devices

Output: Acknowledgement after Execution

Start;

```

If(buffer holding delay sensitive packets)
{
    Label:busy_sensitive
        /* The IoT Layer Request Fog layer for service*/
        Service_Activate_Fog();
    /* Fog Layer able to execute Request*/
        Acknowledgement= urgentProcessing();
        /*Fog sends the acknowledgement to IoT Layer*/
        RETURN(Acknowledgement);
    If(packet) goto busy_sensitive
    else partial_sleep()
}
    If(buffer holding delay tolerant packets)
{

    Label:busy_tolerant
    /* The IoT Layer create buffer of size 'A' and Request Cloud layer for
    service*/
    Service_Activate_Cloud();
    /* Cloud Execute Request receive Acknowledgement*/
        Acknowledgement= Batchprocessing();
    /*cloud sends the acknowledgement to IoT Layer*/
        RETURN(Acknowledgement);
    If(packet) goto busy_tolerant
    else full_sleep();
}
End

```

### 3.1 Performance and Cost Matrices

In this section, we have defined the probability of staying busy  $P(B)$  or the probability of staying full sleep mode  $P(V)$  or partial sleep mode  $P(W)$ .

$$P(B) = \sum_{i=1}^{\infty} \pi_{i,2}, P(V) = \sum_{i=0}^{\infty} \pi_{i,1} \quad \text{and} \quad P(W) = \sum_{i=0}^{\infty} \pi_{i,0}$$

We have derived an average number of data packets present in the buffer ( $L_s$ ), the average number of packets in the system ( $L_q$ ), the average waiting in the system ( $W_s$ ), the average waiting time in the queue ( $W_q$ ) are given by

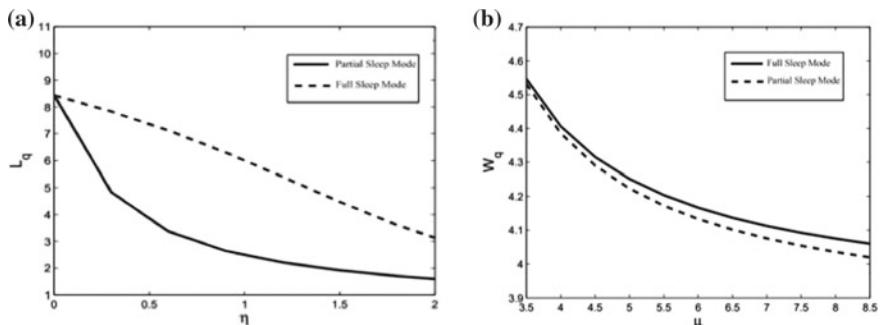
$$L_s = \sum_{i=1}^{\infty} i(\pi_{i,0} + \pi_{i,1} + \pi_{i,2}), L_q = \sum_{i=1}^{\infty} (i-1)\pi_{i,0} \sum_{i=1}^{\infty} i\pi_{i,1} + \sum_{i=1}^{\infty} (i-1)\pi_{i,2}$$

Using Little's algorithm,  $W_s$  can be defined as

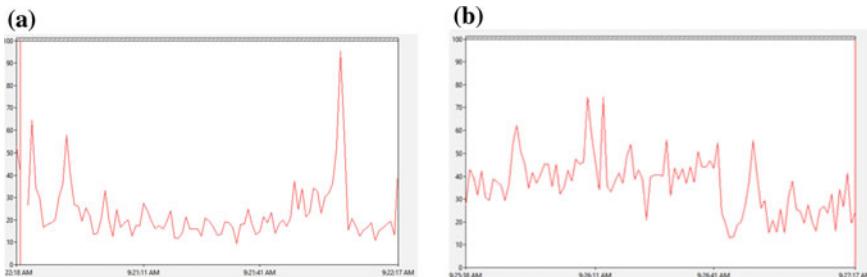
$$W_s = L_a/\lambda.$$

### 3.2 Numerical Results and Analysis

We have simulated our proposed system using Python 3.6-based simulator. In Fig. 3a, we have shown the impact of  $\eta$  on average waiting time ( $L_q$ ) of the packets in the buffer for full sleep mode and partially sleep mode. In the figure, it has been observed that the average waiting time in the buffer will be less in partial sleep mode compare to full sleep mode. Hence, it will be handled more effectively with real-time health



**Fig. 3** **a** Impact of  $\eta$  on  $L_q$  in the buffer residing in IoT plane. **b** Impact of  $\mu$  on  $W_q$  in the buffer residing in IoT plane



**Fig. 4** **a** Average waiting time in queue for processing 200 MB data using the proposed architecture. **b** Packet loss in the system for processing 200 MB data using the proposed architecture

data for processing in fog level and sends an acknowledgement to IoT layer. Delay-tolerant traffic will accumulate in the buffer and will be transmitted when batch size will be “A”. Power efficiency will be comparatively better as waiting time in the buffer is less, and in real life, fog-level service rate is usually less comparative to cloud level. Less power will be consumed for a lower service rate. The other parameters are  $\lambda = 3.0$ ,  $\mu = 4.0$ ,  $\phi = 0.2$ , and  $\psi = 0.1$ . The effect of service rate on waiting time in the queue has been depicted in Fig. 3b. From the figure, it can be observed that when  $\mu$  is less, the delay for both traffic delay will be less. But with the increase of service rate, the delay will be more in case of fully sleep mode compared to partial sleep mode. As it has been stated earlier that delay-tolerant traffic will use fully sleep mode and real-time urgent health data will incorporate partial sleep mode for reducing the delay in the buffer.

We have simulated our proposed system using Python 3.6-based simulator, and in Fig. 4, we have shown the average waiting time in the buffer for transfer of 200 MB data. We have run the simulator for arrival process 750 jobs per second and processor can process 700 jobs per sec. We have run the simulator for 2 min 30 s. Similarly in Fig. 4, we have depicted packet loss in the system.

## 4 Conclusions

In this era of IoT, there is a tremendous growth of wearable and home or city applications. All these applications need internet-connected devices which need real-time low latency services. For further processing and storing the huge data, cloud computing framework is very popular. To enhance the application’s efficiency, fog computing is also included. We have study the impact of arrival rate on packet loss and average waiting time in the queue. In future, we have the plan to handle big data and health data of enormous sizes. To reduce the waiting time in the queue, we have to make some changes in the framework.

## References

1. Rao, R.M., Fontaine, M., Veislari, R.: A reconfigurable architecture for packet based 5G transport networks. In: IEEE 5G World Forum, 5GWF 2018—Conference Proceedings, pp. 474–477 (2018)
2. Misra, C., Goswami, V.: Analysis of power saving class II traffic in IEEE 802.16 E with multiple sleep state and balking. *Found. Comput. Decis. Sci.* **40**(1), 53–66 (2015)
3. Barik, R., et al.: Fog2fog: augmenting scalability in fog computing for health GIS systems. In: 2017 IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE). IEEE (2017)
4. Tang, B., Chen, Z., Hefferman, G., Wei, T., He, H., Yang, Q.: A hierarchical distributed fog computing architecture for big data analysis in smart cities. In: Proceedings of the ASE BigData & SocialInformatics 2015, p. 28. ACM (2015)
5. Ema, R.R., Islam, T., Ahmed, M.H.: Suitability of using fog computing alongside cloud computing. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1–4. IEEE (2019)
6. Goswami, V., Misra, C.: Discrete-time modelling for performance analysis and optimisation of uplink traffic in IEEE 802.16 networks. *Int J Commun Netw Distrib Syst* **10**(3):243–257 (2013)
7. Lee, G., Mallipeddi, R., Lee, M.: Trajectory-based vehicle tracking at low frame rates. *Expert Syst. Appl.* **80**, 46–57 (2017)
8. Zhu, W., Gao, D., Zhao, W., Zhang, H., Chiang, H.: SDN-enabled hybrid emergency message transmission architecture in internet-of-vehicles. *Enterp. Inf. Syst.* **12**(4), 471–491 (2018)
9. Breetzke, T., Flowerday, S.V.: The usability of IVRs for smart city crowdsourcing in developing cities. *Electron. J. Inf. Syst. Dev. Countries* **73**(1), 1–14 (2016)
10. Hussain, A., Wenbi, R., Da Silva, A.L., Nadher, M., Mudhish, M.: Health and emergency-care platform for the elderly and disabled people in the smart city. *J. Syst. Softw.* **110**, 253–263 (2015)
11. Castelli, M., Gonçalves, I., Trujillo, L., Popović, A.: An evolutionary system for ozone concentration forecasting. *Inf. Sys. Front.* **19**(5), 1123–1132 (2017)
12. Corbett, J., Mellouli, S.: Winning the SDG battle in cities: How an integrated information ecosystem can contribute to the achievement of the 2030 sustainable development goals. *Inf. Syst. J.* **27**(4), 427–461 (2017)
13. Sproull, L., Patterson, J.F.: Making information cities livable. *Commun. ACM* **47**(2), 33–37 (2004)
14. Díaz-Díaz, R., Pérez-González, D.: Implementation of social media concepts for e-government: Case study of a social media tool for value co-creation and citizen participation. *J. Org. End User Comput.* **28**(3), 104–121 (2016)
15. Fietkiewicz, K.J., Mainka, A., Stock, W.G. (2017). eGovernment in cities of the knowledge society. An empirical investigation of smart cities' governmental websites. *Govern. Inf. Q.* **34**(1), 75–83.
16. Johnson, P., Iacob, M.E., Välia, M., van Sinderen, M., Magnusson, C., Ladhe, T.: A method for predicting the probability of business network profitability. *Inf. Syst. e-Business Manage.* **12**(4), 567–593 (2014)
17. Keegan, S., O'Hare, G., O'Grady, M.: Retail in the digital city. *Int. J. e-Business Res.* **8**(3), 18–32 (2012)
18. Aguilar, J., Sanchez, M.B., Jerez, M., Mendonca, M.: An extension of the MiSCi middleware for smart cities based on fog computing. *J. Inf. Technol. Res.* **10**(4), 23–41 (2017)
19. Hashem, I.A.T., Chang, V., Anuar, N.B., Adewole, K., Yaqoob, I., Gani, A., Ahmed, E., Chiroma, H.: The role of big data in smart city. *In. J. Inf. Manage.* **36**(5), 748–758 (2016)
20. Huang, K., Zhang, X., Wang, X.: Block-level message-locked encryption with polynomial commitment for IoT data. *J. Inf. Sci. Eng.* **33**(4), 891–905 (2017)

# On Interior, Exterior, and Boundary of Fuzzy Soft Multi-Set Topology



S. A. Naisal and K. Reji Kumar

**Abstract** In this paper, our focus is on the properties of the operations of fuzzy soft multi-sets. The notions of fuzzy soft multi-open sets, fuzzy soft multi-closed sets and soft multi-neighborhood on fuzzy sets are included, and some results on exterior, interior and boundary of fuzzy soft multi-set topology are presented.

## 1 Introduction

We experience different kinds of vagueness and uncertainties in our daily life. Changing nature of environment and the subjectivity of human consciousness is an important reason for uncertainties or vagueness. The classical set theory, which deals with exact and accurate concepts, is not enough to model vagueness and uncertainties. Fuzzy sets, soft sets and multi-sets help us to rectify the inability of the classical theory.

The concept of fuzzy sets was introduced by Zadeh [1], while soft set was first studied by Molodtsov [2] and the rough sets [3] by Pawlak. Molodtsov gives parametrization of soft sets [2]. He applied soft set theory, in many fields, such as the Riemann integral and game theory. Maji et al. [4] applied the soft sets in problems on decision making. Chen et al. [5] and Maji et al. [4] have defined and studied fuzzy soft sets by combining the notions of fuzzy sets and soft sets. Alkhazaleh et al. [6] combined the concepts of soft sets and multi-sets. Applications of soft multi-set weree given by Salleh and Alkhazaleh. Alkhazaleh and Salleh [7] also combined the concepts of fuzzy sets, soft sets and multi-sets.

In this paper, we focus on a detailed study of the concepts explained in [8]. We need the following definitions for which we owe to [2, 4, 7, 9].

In this discussion,  $U$  stands for the universal set. Soft set over  $U$  is defined in [2]. Consider the initial universe  $U$ , and the set of parameters  $E$ .  $P(U)$  represents

---

S. A. Naisal

Government Institute of Teacher Education, Alappuzha, Kerala, India

K. Reji Kumar (✉)

N.S.S. College, Cherthala, Alappuzha, Kerala, India

power set. Then the pair  $(G, B)$ , where  $B$  subset of  $U$  is a soft set, such that  $G$  is a mapping from  $B$  to  $P(U)$ . A multi-set  $M$  related to the set  $X$ , determined by  $C(M)$ , where  $C(M)$  is a function defined from  $X$  to  $N$ . Here,  $N$  represents the set of all non-negative integers. In the definition of soft sets if we replace  $P(U)$  by  $I^U$ , then the pair  $(G, B)$  is a fuzzy soft set over  $U$  [4].

Let  $\{U_i : i \in I\}$  be a collection of universes satisfying the condition that  $\cap_{i \in I} U_i = \emptyset$  and  $E_{U_i} : i \in I$  is a collection of sets of parameters. Also, let  $U = \prod_{i \in I} P(U_i)$ , where  $P(U_i)$  denotes the power set of  $U_i$  and  $E = \prod_{i \in I} E_{U_i}$  and  $B \subseteq E$ . Alkhazaleh et al. [7] defined the pair  $(G, B)$  as a fuzzy soft multi-set over  $U$ , where  $G$  is a mapping such as  $G : B \rightarrow U$ , for every  $b$  in  $B$ ,  $G(b) = (\{\frac{u}{\mu_{G(b)}(u)}\}; i \in I)$ . Also, in [7], the union and intersection of two fuzzy soft multi-sets are defined.

## 2 Some Topological Concepts Related to Fuzzy Soft Multi-Set Topology

In this section, we discuss the exterior, boundary and interior of fuzzy soft multi-sets and study some of its properties. We consider the two fuzzy soft multi-topological spaces  $(U, E, \tau_1)$  and  $(U, E, \tau_2)$ . If  $\tau_2 \supseteq \tau_1$ , then  $\tau_2$  is soft finer than  $\tau_1$ . If  $\tau_2 \supset \tau_1$ , then  $\tau_2$  is soft strictly finer than  $\tau_1$ . Let  $(U, E, \tau)$  be a fuzzy soft multi-topological space over  $(U, E)$ , and the elements of  $\tau$  are termed as fuzzy soft multi-open set in  $(U, E, \tau)$ .

Consider  $(U, E, \tau)$  as a fuzzy soft multi-topological space over  $(U, E)$ . Then the soft set  $(G, E)$  over  $B$  is termed as a fuzzy soft multi-closed set in  $B$ , if the relative complement of  $(G, B)$  belongs to  $(G, E)'$ .  $F_B^e$  is named as a fuzzy soft multi-point in  $(U, E)$  if we can develop an  $e \in (G, B)$  where  $G(e) \neq 0$  also for  $e' \in B - \{e\}$ ,  $G(e') = 0$ .

*Example 1* We consider the three universes  $U_1, U_2, U_3$ . Suppose  $P$  wishes to buy a home for accommodation, a four-wheeler car and a plot for business. Take a fuzzy soft multi-set  $(G, B)$  which narrates homes, four-wheeler cars and plots that  $P$  is going to take for family accommodation, transportation and to begin a new business.

Consider  $U_1 = \{h_1, h_2, h_3, h_4\}, U_2 = \{c_1, c_2, c_3\}, U_3 = \{p_1, p_2\}$  and let  $\{E_{U_1}, E_{U_2}, E_{U_3}\}$  be the collection of sets of decision parameters taking in consideration of above universe, where  $E_{U_1} = \{e_{U_{1,1}} = \text{expensive}, e_{U_{1,2}} = \text{cheap}, e_{U_{1,3}} = \text{wooden}, e_{U_{1,4}} = \text{in green surroundings}\}$   $E_{U_2} = \{e_{U_{2,1}} = \text{expensive}, e_{U_{2,2}} = \text{cheap}, e_{U_{2,3}} = \text{sporty}\}$   $E_{U_3} = \{e_{U_{3,1}} = \text{expensive}, e_{U_{3,2}} = \text{cheap}, e_{U_{3,3}} = \text{in town}, e_{U_{3,4}} = \text{in a depth population}\}$

Consider  $U = \prod_{i=1}^3 GS(U_i), E = \prod_{i=1}^3 E(U_i)$  and  $A \subseteq E$  such that  $A = \{a_1 = \{e_{U_{1,1}}, e_{U_{2,1}}, e_{U_{3,1}}\}, a_2 = \{e_{U_{1,1}}, e_{U_{2,2}}, e_{U_{3,1}}\}\}(G_1, B) = ((a_1, (\{\frac{h_1}{0.2}, \frac{h_2}{0.2}, \frac{h_3}{0.8}, \frac{h_4}{0}\}, \{\frac{c_1}{0.8}, \frac{c_2}{0.4}, \}, \{\frac{p_1}{0.8}, \frac{p_2}{0.7}\})), (a_2, (\{\frac{h_1}{0.9}, \frac{h_2}{0.5}, \frac{h_3}{0.5}, \frac{h_4}{0.2}\}, \{\frac{c_1}{0.7}, \frac{c_2}{0.8}, \frac{c_3}{0.5}, \}, \{\frac{p_1}{0.5}, \frac{p_2}{0.5}\}))(G_2, B) = ((a_1, ((U_1, U_2, U_3), (a_2, \phi))))\}$   $(G_3, B) = ((a_1, (\{\frac{h_1}{0.7}, \frac{h_2}{0.7}, \frac{h_3}{0.1}, \frac{h_4}{0.8}\}, \{\frac{c_1}{0.8}, \frac{c_2}{0.6}, \frac{c_3}{0.3}, \}, \{\frac{p_1}{0.5}, \frac{p_2}{0.4}\})), (a_2, ((\frac{h_1}{0.9}, \frac{h_2}{0.5}, \frac{h_3}{0.5}, \frac{h_4}{0.2}), \{\frac{c_1}{0}, \frac{c_2}{0.2}, \frac{c_3}{0.7}, \}, \{\frac{p_1}{0.8}, \frac{p_2}{0.7}\})))$  are fuzzy soft multi-sets over  $(U, E)$ .

**Example 2** In the above example if take  $(G, B) = \{G(e_1) = (a_1, (\{\frac{h_1}{0}, \frac{h_2}{0}, \frac{h_3}{0}, \frac{h_4}{0}\}, \{\frac{c_1}{0}, \frac{c_2}{0}, \frac{c_3}{0}, \}, \{\frac{p_1}{0}, \frac{p_2}{0}\})), (a_2, (\{\frac{h_1}{0.1}, \frac{h_2}{0.2}, \frac{h_3}{0.3}, \frac{h_4}{0.4}\}, \{\frac{c_1}{0.2}, \frac{c_2}{0.3}, \frac{c_3}{0.5}, \}, \{\frac{p_1}{0.5}, \frac{p_2}{0.4}\}))\}$ , were  $e_2 \in (G, B)$ ,  $G(e_2) \neq 0$  and for  $e' \in B - \{e_2\}$ ,  $G(e') = 0$ . Then  $G_B^e$  is a fuzzy soft multi-point in  $(U, E)$ .

A fuzzy soft multi-point  $G_B^e$  is said to be in fuzzy soft multi-set  $(G, B)$  if  $A \subseteq B$  and for  $e \in A$ ,  $F(e) \subseteq G(e)$ . In our further discussion, we denote fuzzy soft multi-set as FSMS.

Take a fuzzy soft multi-set  $(G, B)$  and a FSMS point  $G_B^e \in (U, E)$  in a FSMS topological spaces  $(U, E, \tau)$ . Now the open fuzzy soft multi-set  $(I, C)$  in  $(U, E, \tau)$  can be treated as a neighborhood of FSMS point  $G_B^e$ , if there exist FSM open set  $(G, B)$  such as  $G_B^e \in (G, B) \subseteq (I, C)$ . The FSMS  $(G, B)$  in a FSM topological space  $(U, E, \tau)$  is said to be a FSM neighborhood of the FSMS  $(G, B)$  if we can a FSM open set  $(H, c)$  such that  $(G, B) \subseteq (H, c) \subseteq (I, C)$ .

**Example 3** Consider a FSMS  $(G, B) = \{(G(e_1) = (a_1, (\{\frac{h_1}{0.1}, \frac{h_2}{0.7}, \frac{h_3}{0.8}, \frac{h_4}{0.7}\}, \{\frac{c_1}{0.6}, \frac{c_2}{0.8}, \frac{c_3}{0.7}, \}, \{\frac{p_1}{0.6}, \frac{p_2}{0.8}\}))), (G(e_2) = (a_1, (\{\frac{h_1}{0.2}, \frac{h_2}{0.5}, \frac{h_3}{0.8}, \frac{h_4}{0.7}\}, \{\frac{c_1}{0.8}, \frac{c_2}{0.7}, \frac{c_3}{0.7}, \}, \{\frac{p_1}{0.7}, \frac{p_2}{0.8}\})))\}$ , Also consider a fuzzy soft multi-set point. Take a fuzzy soft multi-topology  $\tau = \{\phi, E, (G, B)\}$ . Now take a fuzzy soft multi-set  $(H, B) = \{(H(e_1) = (a_1, (\{\frac{h_1}{0.3}, \frac{h_2}{0.8}, \frac{h_3}{0.8}, \frac{h_4}{0.8}\}, \{\frac{c_1}{0.7}, \frac{c_2}{0.8}, \frac{c_3}{0.8}, \}, \{\frac{p_1}{0.7}, \frac{p_2}{0.9}\})), (H(e_2) = (a_1, (\{\frac{h_1}{0.4}, \frac{h_2}{0.6}, \frac{h_3}{0.9}, \frac{h_4}{0.8}\}, \{\frac{c_1}{0.9}, \frac{c_2}{0.8}, \frac{c_3}{0.9}, \}, \{\frac{p_1}{0.8}, \frac{p_2}{0.9}\})))\}$ . Clearly that soft multi-set point defined in Example 2.6 lies in  $(G, B) \subseteq (H, B)$ . Thus  $(H, B)$  is the FSM neighborhood of the fuzzy soft multi-point. Also  $(G, B) \subseteq (H, c) \subseteq (I, C)$ .

A FSM  $(G, B)$  over  $(U, E)$  is said to be a FSM closed set in  $(U, E, \tau)$  if its compliment  $(G, B)'$  is a FSM open set in  $(U, E, \tau)$ .

**Example 4** Consider  $(F_1, A) = \{(a_1, (\{\frac{h_1}{0.2}, \frac{h_2}{0.2}, \frac{h_3}{0.8}, \frac{h_4}{0}\}, \{\frac{c_1}{0.8}, \frac{c_2}{0.5}, \frac{c_3}{0.4}, \}, \{\frac{p_1}{0.8}, \frac{p_2}{0.7}\}))\}$ , then  $(F_1, A)' = \{(a_1, (\{\frac{h_1}{0.8}, \frac{h_2}{0.8}, \frac{h_3}{0.2}, \frac{h_4}{1}\}, \{\frac{c_1}{0.2}, \frac{c_2}{0.5}, \frac{c_3}{0.6}, \}, \{\frac{p_1}{0.2}, \frac{p_2}{0.3}\}))\}$ , is a FSM open set in  $(U, E, \tau)$ . Thus,  $(F_1, A)$  is a FSM closed set in  $(U, E, \tau)$ .

Exterior of a FSMS is defined as the complement of the interior of FSM, and it is denoted by  $\text{exte}(G, B)$ . That is  $\text{exte}(G, B) = \text{inte}(G, B)^c$ .  $\text{exte}(G, B)$  is a largest open set contained in  $(G, B)^c$ . Also  $\text{bd}(G, B) = \text{cl}(G, B)^c = \text{cl}(G, B) \cap \text{cl}(G, B)^c$ .

**Theorem 1** Take  $(U, E, \tau)$  a FSM topological spaces. Also take two soft multi-sets  $(G, A)$  and  $(G, B)$ . Then

- (i)  $\text{exte}((G, A) \cup (G, B)) = \text{exte}((G, A) \cap (G, B))$
- (ii)  $\text{exte}((G, A) \cap (G, B)) \geq \text{exte}(G, A) \cup \text{exte}(G, B)$
- (iii)  $\text{inte}(\text{exte}((G, A) \cup (G, B))) = \text{exte}(G, A) \cap \text{exte}(G, B)$
- (iv)  $\text{inte}(\text{exte}((G, A) \cap (G, B))) \geq \text{exte}(G, A) \cup \text{exte}(G, B)$

**Proof** Proof of (i):  $\text{exte}((G, A) \cup (G, B)) = \text{inte}((G, A) \cup (G, B))^c = \text{inte}((G, A)^c \cap (G, B)^c) = \text{inte}((G, A)^c) \cap \text{inte}((G, B)^c) = \text{exte}(G, A) \cap \text{exte}(G, B)$ .

Proof of (ii):  $\text{exte}((G, A) \cap (G, B)) = \text{inte}((G, A) \cap (G, B))^c = \text{inte}((G, A)^c \cup (G, B)^c) \geq \text{inte}((G, A)^c) \cup \text{inte}((G, B)^c) = \text{exte}(G, A) \cup \text{exte}(G, B)$ .

Proof of (iii):  $\text{inte}(\text{exte}((G, A) \cup (G, B))) = \text{inte}(\text{exte}(G, A) \cap \text{exte}(G, B)) = \text{inte}(\text{int}((G, A)^c) \cap \text{inte}((G, B)^c)) = \text{inte}((G, A)^c) \cap \text{inte}((G, B)^c) = \text{exte}(G, A) \cap \text{exte}(G, B)$ .

Proof of (iv):  $\text{inte}(\text{exte}((G, A) \cap (G, B))) \geq \text{inte}(\text{exte}(G, A) \cup \text{exte}(G, B)) = \text{inte}(\text{exte}(G, A)) \cup \text{inte}(\text{exte}(G, B)) = \text{inte}(\text{int}((G, A)^c)) \cup \text{inte}(\text{int}((G, B)^c)) = \text{inte}((G, A)^c) \cup \text{inte}((G, B)^c) = \text{exte}(G, A) \cup \text{exte}(G, B)$

**Theorem 2** Take  $(U, E, \tau)$  as a FSMS topological spaces. Let  $(G, A)$  and  $(G, B)$  be two FSMS. Then

- (i)  $(\text{bd}(G, A))^c = \text{inte}(G, A) \cup \text{exte}(G, A)$
- (ii)  $\text{exte}((\text{bd}(G, A))^c) = \text{exte}(\text{exte}((G, A))) \cap \text{exte}(\text{int}((G, A))) \geq \phi$
- (iii)  $\text{inte}(G, A) \vee \text{bd}(G, A) = \text{cl}(G, A)$
- (iv)  $\text{bd}(G, A) = \text{cl}(G, A) \ominus \text{inte}(G, A)$
- (v)  $\text{int}((G, A)) = (G, A) \ominus \text{bd}(G, A)$

*Proof* Proof of (i):

$$\begin{aligned} (\text{bd}(G, B))^c &= (\text{cl}(G, A) \cap \text{cl}((G, A)^c))^c \\ &= (\text{cl}(G, A))^c \cup (\text{cl}((G, A)^c))^c \\ &= \text{int}((G, A)^c) \cup \text{int}(G, A) \\ &= \text{exte}(G, A) \cup \text{int}(G, A) \end{aligned}$$

Proof of (ii):

$$\begin{aligned} \text{exte}((\text{bd}(G, A))^c) &= \text{exte}(\text{exte}(G, A) \cup \text{int}(G, A)) \\ &= \text{exte}(\text{exte}(G, A) \cap \text{exte}(\text{int}(G, A))) \\ &\geq \text{int}(A) \cap (\text{bd}(A) \cup \text{exte}(A)) \\ &= \text{int} \cap (\text{int}(A))^c = \phi \end{aligned}$$

Proof of (iii):

$$\begin{aligned} \text{int}(G, A) \cup \text{bd}(G, A) &= \text{int}(G, A) \cup (\text{cl}(G, A) \cap \text{cl}(G, A)^c)) \\ &= \text{int}(G, A) \cup (\text{cl}(G, A) \cap (\text{int}(G, A) \cup \text{cl}(G, A)^c)) \\ &= \text{cl}(G, A) \cap (\text{int}(G, A) \cup (\text{int}(G, A)^c)) \\ &= \text{cl}(G, A). \end{aligned}$$

Proof of (iv): As we have  $\text{cl}(G, A) \ominus \text{int}(G, A) = \max\{\text{cl}(G, A) - \text{int}(G, A), 0\}$ . We have two cases to consider. Case 1: Let there maximum as 0  
So we have  $\text{cl}(G, A) = \text{int}(G, A)$ . Therefore

$$\begin{aligned}
bd(G, A) &= \text{inte}(G, A) \cap cl(G, A)^c \\
&= \text{inte}(G, A) \cap (\text{inte}(G, A))^c \\
&= \phi
\end{aligned}$$

Proof of (v): As we know that  $(G, A) \ominus bd(G, A) = \max\{(G, A) - bd(G, A), 0\}$   
So we have two cases

Case 1: Suppose  $\max\{(G, A) - bd(G, A), 0\} = 0$ , then the proof is trivial.

Case 2: Suppose  $\max\{(G, A) - bd(G, A), 0\} = (G, A) - bd(G, A)$ , then

$$\begin{aligned}
(G, A) \ominus bd(G, A) &= (G, A) \cap (bd(G, A))^c \\
&= (G, A) \cap (\text{inte}(G, A) \cup \text{exte}(G, A)) \\
&= (G, A) \cap (\text{inte}(G, A) \cup \text{inte}((G, A)^c)) \\
&= ((G, A) \cap \text{inte}(G, A)) \cup ((G, A) \cap \text{inte}((G, A)^c)) \\
&= \text{inte}(G, A) \cup \phi \\
&= \text{inte}(G, A)
\end{aligned}$$

**Theorem 3** Take  $(U, E, \tau)$  as a FSM topological spaces. Let there be a FSMS  $(G, A)$  in  $(U, E, \tau)$ , then we have  $(G, A)$  is open if and only if  $(G, A) \cap bd(G, A) = 0$ .

*Proof* Take  $(G, A)$  as an open FSM. So we have  $\text{inte}(G, A) = (G, A)$ .  $(G, A) \cap bd(G, A) = \text{inte}(G, A) \cap bd(G, A) = 0$ . For the converse part, take a fuzzy open soft multi-set  $(G, A)$  such as

$$\begin{aligned}
(G, B) \cap bd(G, B) = 0 &\Rightarrow (G, B) \cap (cl(G, B) \cap cl((G, B)^c)) = 0 \\
&\Rightarrow (G, B) \cap cl((G, B)^c) = 0 \\
&\Rightarrow cl((G, B)^c) \leq (G, B)^c \\
&\Rightarrow (G, B)^c \text{ is a closed soft multi set} \\
&\Rightarrow (G, B) \text{ is an open soft multi set.}
\end{aligned}$$

**Theorem 4** Take a fuzzy soft multi-set topological space  $(X, E, \tau)$ . Then  $(G, B)$ , a FSMS, is closed in  $(X, E, \tau)$  if and only if  $bd(G, B) \leq (G, B)$

*Proof* Take  $(G, B)$  as a FSM closed set in  $(X, E, \tau)$ . Then  $cl(G, B) = (G, B)$ .  $bd(G, B) = cl(G, B) \cap cl((G, B)^c) \leq cl(G, B) = (G, B)$ .

Now let

$$\begin{aligned} bd(G, B) \leq (G, B) &\Rightarrow bd(G, B) \cap (G, B)^c = 0 \\ &\Rightarrow bd((G, B)^c) \cap (G, B)^c = 0 \end{aligned}$$

This implies that  $(G, B)^c$  is FSM open set. Thus we have  $(G, B)$  as an FSM closed set.

**Theorem 5** *Take  $(G, B)$  as an FSMS in an FSM topological space  $(X, E, \tau)$ . Then  $bd(G, B) = 0$  if and only if  $(G, B)$  is clopen.*

*Proof* We take,

$$\begin{aligned} bd(G, B) = 0. &\Rightarrow cl(G, B) \cap cl((G, B)^c) = 0 \\ &\Rightarrow cl(G, B) \leq (cl((G, B)^c))^c \quad \Rightarrow cl(G, B) \leq int(G, B) \leq (G, B) \end{aligned}$$

So we get  $(G, B)$  is closed. Again take

$$\begin{aligned} bd(G, B) = 0. &\Rightarrow cl(G, B) \cap cl((G, B)^c) = 0 \\ &\Rightarrow cl(G, B) \cap (int(G, B))^c = 0 \quad \Rightarrow (G, B) \cap (int(G, B))^c = 0 \\ &\Rightarrow (G, B) \leq int(G, B) = 0 \end{aligned}$$

Thus  $(G, B)$  is open. We now consider the converse part. That is, the FSM set  $(G, B)$  be closed and open. Then we get  $bd(G, B) = cl(G, B) \cap cl(G, B)^c = cl(G, B) \cap (int(G, B))^c = (G, B) \cap (G, B)^c = 0$ .

**Theorem 6** *let  $(G, B)$  be a FSM set in  $(X, E, \tau)$ . Then we have the following.*

1.  $bnd(bnd(G, B)) \leq bnd(G, B)$
2.  $bnd(bnd(bnd(G, B))) = bnd(bnd(G, B))$

*Proof* Proof of (i): Consider a FSMS  $(G, B)$  in  $(X, E, \tau)$ .  $bnd(bnd(G, B)) = bnd(cl(G, B) \cap cl(G, B)^c) = (cl(cl(G, B) \cap cl((G, B)^c))) \cap (cl(cl(G, B) \cap cl(G, B)^c)^c) \leq (cl(G, B) \cap cl((G, B)^c)) \cap (cl(inte(G, B)^c \cup inte(G, B))) = bnd(G, B) \cap cl(set) = bnd(G, B) \cap (set) = bnd(G, B)$

Proof of (ii):  $bnd(bnd(bnd(G, B))) = cl(bnd(bnd(G, B))) \cap cl(bnd(bnd(G, B))^c) = bnd(bnd(G, B) \cap cl(bnd(bnd(G, B))))^c$ . Also  $(bnd(bnd(G, B)))^c = (cl(bnd(G, B)) \cap cl(bnd(G, B))^c)^c = (bnd(G, B) \cap cl(bnd(G, B))^c)^c = (bnd(G, B))^c \cup (cl(bnd(G, B))^c)^c$ . Now we take closures on both sides.

Then we have  $cl(bnd(bnd(G, B)))^c = (cl(bnd(G, B))^c) \cup cl(cl(bnd(G, B))^c) \geq (cl(bnd(G, B))^c) \cup (cl(bnd(G, B))^c)^c = E$ .

We know that,  $bnd(bnd(bnd(G, B))) = cl(bnd(bnd(G, B))) \cap cl(bnd(bnd(G, B))^c) = cl(bnd(bnd(G, B))) \cap E = bnd(bnd(G, B))$ .

**Theorem 7** *Let  $(X, E, \tau)$ , be an FSM topological space. For any  $(G, B)$ , FSM set,  $bnd(bnd(G, B))$  is a closed FSMS.*

*Proof* We will show that  $cl(bnd(bnd(G, B))) = bnd(bnd(G, B))$ .  $cl(bnd(bnd(G, B))) = cl(cl(bnd(G, B)) \cap cl(bnd((G, B)^c))) \leq cl(cl(bnd(G, B)) \cap cl(cl(bnd(G, B)^c))) \vee = cl(bnd(G, B)) \cap cl(bnd(G, B)^c) = bnd(bnd(G, B))$ . So the closure of  $bnd(bnd(G, B))$  is contained in itself and is so  $bnd(bnd(G, B))$  is a closed fuzzy soft multi-set.

Take two FSM topological spaces  $(F, A, \tau_A)$  and  $(G, B, \tau_B)$ . Then  $(F, A, \tau_A)$  is a product related to  $(G, B, \tau_B)$  if for any FSM set  $\eta$  of  $(F, A)$  and  $\zeta$  of  $(G, B)$ ,  $\lambda^c$  not  $\geq \eta$  and  $\mu^c$  not  $\geq \zeta$  imply  $\lambda^c \times I \cup I \times \mu^c \geq \eta \times \zeta$ , where  $\lambda \in \tau_A$  and  $\mu \in \tau_B$ , also there exist  $\lambda_1 \in \tau_A$  and  $\mu_1 \in \tau_B$  such that  $\lambda_1^c \geq \eta$  and  $\mu_1^c \geq \zeta$  and  $\lambda_1^c \geq \eta$  and  $\mu_1^c \geq \zeta$  also  $\lambda_1^c \times I \cup I \times \mu_1^c = \lambda_1^c \times I \cup I \times \mu^c$ .

**Theorem 8** Take  $(F, A, \tau_A)$  and  $(G, B, \tau_B)$  as product related to FSM topological spaces. Then for any FSM sets  $(G, B)$  and  $(G, B)$  in  $(F, A, \tau_A)$  and  $(G, B, \tau_B)$ , respectively, following holds.

1.  $cl((H, A) \times (G, B)) = cl(H, A) \times cl(G, B)$
2.  $inte((H, A) \times (G, B)) = inte(H, A) \times inte(G, B)$
3.  $ext((H, A) \times (G, B)) = exte(H, A) \times exte(G, B)$

*Proof* Take FSMS  $(H, A)$ 's in  $(F, A, \tau_A)$  and  $(G, B)$ 's in  $(G, B, \tau_B)$ . Then we have

1.  $\inf\{(H, A), (G, B)\} = \min\{\inf(H, A), \inf(G, B)\}$
2.  $\inf\{(H, A) \times I\} = \inf(G, B) \times I$
3.  $\inf\{I \times (H, A)\} = I \times \inf(G, B)$

First we will prove that  $cl((H, A) \times (G, B)) \geq cl(H, A) \times cl(G, B)$ .  $cl((H, A) \times (G, B)) = \inf\{(H, A) \times (G, B)^c / ((H, A) \times (G, B))^c \geq ((G, B) \times (G, B))\} = \inf\{(H, A)^c \times I \vee I \times (H, A)^c / (G, B)^c \times I \vee I \times (G, B)^c \geq ((G, B) \times (G, B))\} = \inf\{(G, B)^c \times I \vee I \times (G, B)^c / (G, B)^c \geq (G, B) \text{ or } (G, B)^c \geq (G, B)\} = \min(\inf\{(H, A)^c \times I \vee I \times (G, B)^c / (H, A)^c \geq (H, A)\}, \inf\{(H, A)^c \times I \vee I \times (G, B)^c / (H, A)^c \geq (G, B)\})$ . Also

$\inf\{(H, A)^c \times I \vee I \times (G, B)^c / (H, A)^c \geq (G, B)\} \geq \inf\{(G, B)^c \times I / (H, A)^c \geq (G, B)\} = \inf\{(H, A)^c / (H, A)^c \geq (H, A)\} \times I = cl(H, A) \times I$ .  $\inf\{(H, A)^c \times I \vee I \times (G, B)^c / (G, B)^c \geq (G, B)\} \geq \inf\{(G, B)^c \times I / (G, B)^c \geq (G, B)\} = I \times \inf\{(G, B)^c / (G, B)^c \geq (G, B)\} = I \times cl(G, B)$ . So we can have  $cl((H, A) \times (G, B)) \geq \min(cl(H, A) \times I, I \times cl(G, B)) = cl(H, A) \times cl(G, B)$ . It can be proved by using the facts that  $(int\alpha)^c = cl\alpha^c$ , and  $(cl\alpha)^c = int\alpha^c$

Take  $(X_i, E, \tau)$ ,  $i = 1, 2, \dots, n$  as a family of product related FSM topological spaces. Let  $(F_i, A)$  be a FSMS in  $(X_i, E, \tau)$ , then we have  $bd \prod_{i=1}^n (F_i, A) = (bd(F_1, A) \times cl(F_2, A) \times \dots, cl(F_n, A)) \vee (cl(F_1, A) \times bd(F_2, A) \times \dots, cl(F_n, A)) \vee \dots, \vee (cl(F_1, A) \times cl(F_2, A) \times \dots, bd(F_n, A))$ .

**Theorem 9** Let  $f : (X, E, \tau) \rightarrow (Y, E, \tau)$  be a FS multi continuous function, then we have  $bd(f^{-1}(F, A)) \leq f^{-1}(bd(G, B))$ , for any FSMS  $\alpha, A$  in  $(Y, E, \tau)$ .

*Proof* Consider  $f$  as a FSMS continuous function and take a FSMS  $(F, A)$  in  $(Y, E, \tau)$ . Then  $cl(G, B)$  is a FSM closed set in  $(Y, E, \tau)$ ,  $\Rightarrow f^{-1}(cl(G, B))$  is FSM closed in  $(X, E, \tau)$ . So we have  $bd(f^{-1}(G, B)) = cl(f^{-1}(G, B)) \wedge cl(f^{-1}(G, B))^c \leq cl(f^{-1}cl(G, B)) \wedge cl(f^{-1}cl(G, B)^c) = (f^{-1}cl(G, B)) \wedge (f^{-1}cl(G, B)^c) = f^{-1}(cl(G, B) \wedge cl(G, B)^c) = f^{-1}(bd(G, B))$ . Thus,  $bd(f^{-1}(G, B)) \leq f^{-1}(bd(G, B))$ .

**Theorem 10** Consider an FSMS in a soft multi-topological space  $(X, E, \tau)$ . Then  $ext(G, B)$  is empty if and only if every nonempty open FSMS in  $(X, E, \tau)$  contains a point of  $(G, B)$ .

*Proof* Suppose every nonempty open FSMS in  $(X, E, \tau)$  contains a point of  $(G, B)$ . So every  $(e_i, k/x) \in (G, B) \subseteq (X, E, \tau)$ . Then we have

$k \leq cl(G, B) \Rightarrow set \leq cl(G, B)$  if  $exte(G, B) = \phi \Leftrightarrow inte(G, B)^c = \phi \Leftrightarrow (cl(G, B))^c = \phi \Leftrightarrow cl(G, B)$  a full soft multi-set. But we have  $cl(G, B)$  less than or equal to a full soft multi-set. So  $exte(G, B) = \phi$ . Conversely, take  $exte(G, B) = \phi$ . Also take  $(G, P)$  as any open FSMS in  $(X, E, \tau)$ . We have to show that  $(G, P)$  contains a point of  $(G, B)$ . Take  $(e_i, k/x) \in (G, P)$ . As  $exte(G, B)$  is empty, no neighborhood of  $(e_i, k/x)$  is contained in  $(G, B)^c$ . That is, all neighborhoods of  $(e_i, k/x)$  are contained in  $(G, B)$ . Therefore,  $(G, P) \cap (G, B) \neq \phi$ . That is every nonempty open FSMS in  $(X, E, \tau)$  contains a point of  $(G, B)$ .

### 3 Conclusion

The research presented here is an extension of the work done on fuzzy soft multi-topological spaces. In this paper, some properties of fuzzy soft multi-topological spaces are established. Interior, exterior, boundary and closure for the fuzzy soft multi-topological spaces are studied and some important results on them are also presented. The work can be further extended to the important topological notions like connectedness, separation, nets, filters, etc. Applications of the concepts similar to that presented in the papers [10–12] are also interesting topic for future research.

### References

1. Zadeh, L.A.: Fuzzy sets. Inf. Control **8**, 352–378 (1965)
2. Molodtsov, D.: Soft set theory—First results. Comput. Math. Appl. **37**, 19–31 (1999)
3. Pawlak, Z.: Rough sets. Int. J. Inf. Comput. Sci. **11**, 341–356 (1982)
4. Maji, P.K., Biswas, R., Roy, A.R.: Fuzzy soft sets. J. Fuzzy Math. **9**(3), 589–602 (2001)
5. Chen, D., Tsang, E.C.C., Yeung, D.S., Wang, X.: The parameterized reduction of soft sets and its applications. Comput. Math. Appl. **49**, 757–763 (2005)
6. Alkhazaleh, S., Saleh, A.R., Hassan, N.: Soft multiset theory. Appl. Math. Sci. **5**(72), 3561–3573 (2011)
7. Alkhazaleh, S., Salleh, A.R.: Fuzzy soft multi-set theory. Abstract Appl. Anal. Article ID 350603 (2012). <https://doi.org/10.1155/2012/350603>.

8. Reji Kumar, K., Naisal, S.A.: Interior exterior and boundary of fuzzy soft multi topology in decision making. In: IEEE control instrumentation and computational techniques (2016). <https://doi.org/10.1109/ICCI CCT.2016.7987934>
9. Reji Kumar, K., Niasal S.A.: On fuzzy soft multi set topology. *Univ. J. Math. Math. Sci.* **10**(2), 69–93 (2017)
10. Reji Kumar, K., Naisal, S.A.: A mathematical modeling of Ayurveda Doshas: application of fuzzy soft topology. *Int. J. Pure Appl. Math.* **106**(8), 33–43 (2016)
11. Reji Kumar, K., Naisal, S.A.: A mathematical modeling of Ayurvedic Doshas in connection with viruses: application of fuzzy soft multi set theory. In: Presented in Kerala Science congress, Jan 2018
12. Reji Kumar, K., Niasal, S.A.: Intuitionistic fuzzy soft multiset and its application. *Int. J. Math. Combinator. Spec.* **1**, 155–163 (2018)

# Early Prediction of Pneumonia Using Convolutional Neural Network and X-Ray Images



C. Kishor Kumar Reddy, P. R. Anisha, and K. Apoorva

**Abstract** In general, a patient who might be suffering from pneumonia goes to the hospital to take an X-ray. They then wait for the doctor to tell them the results by checking the X-ray. The doctor then decides whether the patient has pneumonia or not. The results are not just concluded based on the X-ray but furthermore, tests were conducted on the patient to verify the results of the doctor. This process is very time-consuming and even if the patient has severe pneumonia, he/she has to wait several days to get the test results. But with the recent development of artificial intelligence, the computational powers of computers have increased. Thus, a computer can help in predicting pneumonia by just passing the X-ray image as an input to our model. The main objective of this paper is to help the doctors predict pneumonia more accurately using a deep learning model. The objective is not only to help the doctors but also the patients to help verify whether they have pneumonia or not. A convolutional neural network (CNN) model is built from scratch using python to extract features from a given chest X-ray image and classify it to determine if a person is infected with pneumonia or not. Finally, a web application is built where the user or patient can upload the X-ray image and view the result on the User Interface (UI).

## 1 Introduction

Pneumonia is an infection in one or both lungs which can be caused by bacteria, virus. Pneumonia can cause the lungs to fill up with fluid, which can be life-threatening and especially dangerous to people over 65 years of age, infants and also children. A large number of children die due to pneumonia every year worldwide. Some symptoms include cough, cold, fever, difficulty in breathing, etc. It is often diagnosed using X-rays. Radiologists often search for white spots in the X-ray of lungs which can be used as an indicator regarding pneumonia. The objective of this project is to build a deep learning project which can be used to directly predict whether a patient has pneumonia or not just by looking at the X-ray. This speeds up the entire process. It

---

C. Kishor Kumar Reddy (✉) · P. R. Anisha · K. Apoorva  
Stanley College of Engineering and Technology for Women, Hyderabad, India

can be used by both the doctor or the patient. Few merits of this prediction model are the diagnosis is usually pretty quick when compared to a traditional diagnosis, early diagnosis means that significant treatment can be started at earlier stages, the web application can be used by both doctors and patients alike, the User Interface (UI) is very user friendly. Few demerits of this prediction model are As the model is not 100 percent accurate, some wrong predictions can be made, a doctors second opinion may always be needed.

## 2 Literature Survey

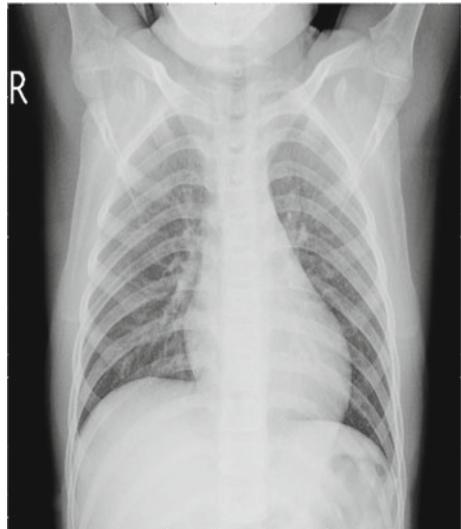
In this post on Understanding of Convolutional Neural Network (ConvNets or CNN), the writer explains how convolution layer is provided with input image, how parameters are chosen and how Iters with strides and padding is applied. It describes the procedure to reduce dimensionality size by performing pooling, to add sufficient convolutional layers, to feed a attenuated output into a fully connected layer, and use of activation function to output the class and classify images [1].

In this post on A Comprehensive Guide to Convolutional Neural Networks, the writer describes the need and use of Deep Learning and one of its widely used algorithm-Convolutional Neural Network. Writer also threw light on the uses of CNN, description of input image, the Kernel-Convolution Layer, pooling and its types, Fully Connected Layer (FC Layer), and various architectures of CNN [2].

Seetha et al. presented a paper Brain Tumor Classification using Convolutional Neural networks; This paper presents the process of efficient automatic brain tumor classification with low complexity and high accuracy and performance. It shows the benefits of CNN over Fuzzy C Means (FCM). Extraction of raw pixel value with depth, width, and height feature value is demonstrated. Usage of gradient decent based loss function for high accuracy is also explained [3].

Rachna Jain et al. presented a paper Pneumonia Detection in Chest X-ray Images using Convolutional Neural Networks and Transfer Learning; This paper presents the model building process of Deep learning-based pneumonia detection in x-ray images is done. Different models of deep learning and transfer learning are analyzed in this for the image classification application. An extensive analysis is carried out in this work with several experimental results [4].

In this post on Pneumonia Detection using Convolutional Neural Networks, the writer has put down the process of detecting pneumonia just by looking at the X-ray images. Writer explains ways to gathering the dataset (X-ray images from Kaggle), process of importing libraries needed, and then CNN usage for the dataset in a stepwise manner to obtain the desired result of predicting pneumonia. Accuracy increases with the changes in the number of layers used in the network or by changing the hyperparameters accordingly [5].

**Fig. 1** Normal X-ray

### 3 Dataset Description

#### 3.1 Data Collection

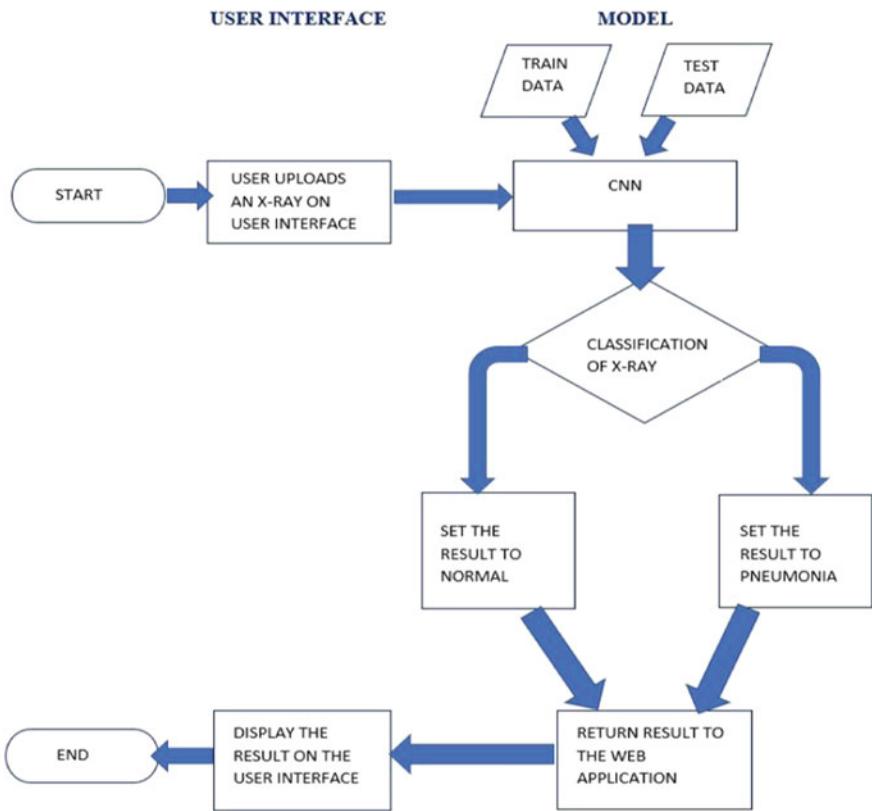
Since we are dealing with an image classification problem, we need to collect two categories of images: normal and pneumonia. In this project, we collected 5216 images for train-data (1341 normal images and 3875 pneumonia images) and 624 images for test-data (234 normal images and 390 pneumonia images).

An example of a normal X-ray and an X-ray suggesting pneumonia is given in Figs. 1 and 2.

#### 3.2 Data Preprocessing

To preprocess the previously collected X-ray images of lungs, we need to use the Keras ImageDataGenerator class to perform data augmentation. The Keras ImageDataGenerator works in the following manner.

1. Accepts a batch of images for training.
2. Takes each batch and applies a series of transformations on it. Transformations include rotation, shearing, resizing, etc.
3. Replaces the original batch with the new transformed batch.
4. Trains the convolutional neural network with this new batch.



**Fig. 2** X-ray suggesting Pneumonia

First, we import the `ImageDataGenerator`. Then we give the various parameters or arguments to the `ImageDataGenerator` class. Each image in a batch is transformed by many random translations but these translations are based on the parameters or arguments we give. Then we need to apply the `ImageDataGenerator` functionality to both the trainset and testset. Data preprocessing on the images is complete and we can move on to the next step [6].

#### 4 Proposed Algorithm

A person who wants to know whether they have pneumonia or not should first have to get an X-ray. However, even after getting an X-ray done, many tests are conducted on the patient to verify the findings of the X-ray. This can be time-consuming and some test results may take several days to arrive. This is not an ideal solution for someone who might be suffering from severe pneumonia and requires immediate

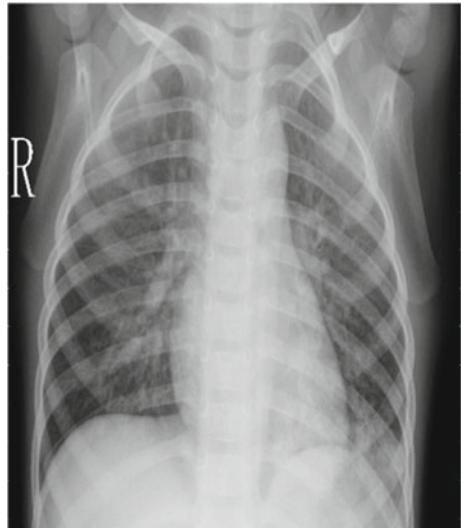
treatment. Instead of waiting to verify the X-ray readings with more tests, a deep learning (DL) convolutional neural network (CNN) can be built which can take a X-ray as input and immediately predict whether the patient has pneumonia or not. The neural network is trained with thousands of normal -rays and X-rays indicating pneumonia so it can accurately predict the result with a new X-ray. The doctor or patient can simply upload the image onto the User Interface (UI) and see the result.

A Convolutional Neural Network (CNN) is a Deep Learning algorithm which takes an input image, assigns importance (learnable weights and biases) to various aspects/objects in the image and is able to differentiate one from the other. CNN can successfully capture the Spatial and Temporal dependencies in an image with the help of necessary filters. The architecture of CNN performs a better fitting to the image dataset as the number of parameters involved is less and weights can be reused. Thus, images can be understood in a much better manner.

CNN trains and tests a set of images. Each image has to go through a series of convolution layers, pooling, and flattening. Convolution extracts the features from an input image using a filter and gives a feature map as an output. Convolution of an image with different filters can yield different feature maps. ReLU stands for Rectified Linear Unit for a non-linear operation. ReLU is used to introduce non-linearity into the convolutional neural network [7].

The next step is performed by the pooling layers. Pooling is used to reduce the dimensions or parameters if the image is too large. In our project, MaxPooling is used. MaxPooling takes the largest element from the rectified feature map of the image [8, 9]. Finally, Flattening is done i.e. the image is converted into a column vector. The flattened output is then given to the neural network. The neural network then starts classifying the images using an activation function. Some activation functions are sigmoid, softmax, ReLU, etc. The proposed architecture is shown in Fig. 3.

**Fig. 3** Proposed architecture



## 4.1 Model Building

1. The first step in model building is importing the libraries needed to form the layers.
2. The second step involves initializing the model. In this project, we use the Sequential class for initialization. The Sequential class is generally used to define a linear initialization of network layers which together constitute a proper model. We will add layers to the model using the add() method.
3. For our convolutional neural network (CNN), we will be adding three layers. The three layers are the Convolution layer, the Pooling layer, and finally the Flattening layer. The Convolution layer is the first layer which is used to extract features from an input image. It is a mathematical operation that can take two inputs such as an image matrix and a filter (kernel). The Pooling layer reduces the spatial size of representation to reduce the number of parameters and computation in the entire network. In this project, MaxPooling is used. The Flatten layer allows us to change the shape of the data from a 2D matrix into another format. This format can be understood by the dense layers.
4. Now, the dense layers are to be added. We initialize the units to 128( $64 \times 2$ ), use activation function ReLU, and go for uniform initialization. This layer acts as the hidden layer. The next dense layer to be added acts as the output layer. For this layer, we initialize units to be one since we only have two classes or categories. Initialization is uniform and activation function is sigmoid as it is a classification situation.

### Configuring the Learning Process

5. As model architecture is built, the model can now be compiled. Compilation normally requires three arguments: optimizer, loss and metrics. For this binary classification problem, the optimizer chosen is adam, the loss is binary\_crossentropy and the metrics is accuracy.

### Train the Model

6. After model compilation, we can train the model. The fit\_generator() method is used for this purpose. There are many parameters to be used here.

*Steps\_per\_epoch* is calculated by diving the train set size by batch size. The train size in this project is 5216 and batch size is taken to be 32. So, our calculated steps\_per\_epoch would be 163.

*Validation\_steps* is calculated by dividing the test set size by batch size. The test size in this project is 624 and batch size is taken to be 32. So, our calculated validation\_steps is 20.

*Epochs* taken is 10.

We finally obtain a model with accuracy 94.48.

### Save the Model

7. To use the created model for future prediction purposes on a web application or on another notebook file, we can save our model using save() with an extension.h5.

## Prediction

8. Now that the model is saved, we can use it in a notebook to predict for new X-rays. First, we need to import two new libraries for prediction to load an image
9. Now, we can use load\_model to load the previously saved model. Two predictions are: One normal X-ray and the other is an X-ray suggesting pneumonia.

## 4.2 Application Building

Now that our model is saved and we know that it is giving correct predictions, we can embed this model into a web application (Flask App). We begin by creating a main Flask App folder called Pneumonia Prediction Web in which we have many subfolders. The model pneumonia.h5 is also put in this folder. The main code is a python program called app.py which contains our main prediction code which deals with loading the model and the prediction part. The CSS and JavaScript code in the static folder deals with the design and the display of the image to be uploaded and the final result to be displayed. The static folder also contains the images we use as backgrounds in our web application. We also have an empty folder initialized called uploads.

As soon as an image is uploaded on the web application, that image is also sent to the uploads folder. In the templates folder, the base.html file handles the entire design and layout of the web application. CSS is also used to improve the web page further.

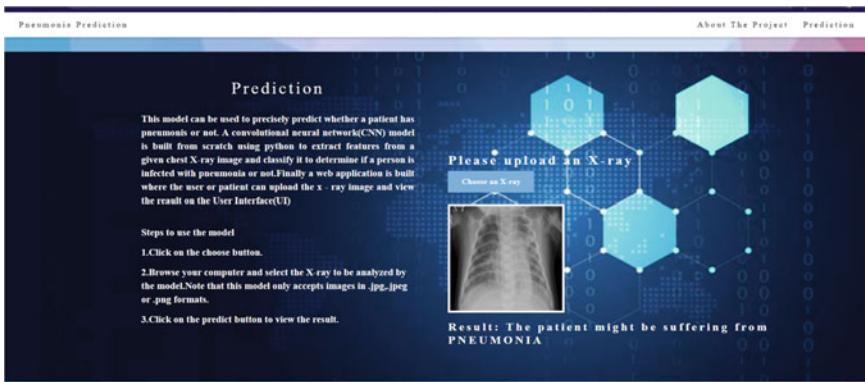
## 5 Results and Conclusion

The model is trained for 1341 normal X-rays and 3875 pneumonia-affected X-rays. It has been trained for 10 epochs, where the final accuracy for training is 94.48% and for validation is 89.26%. The final loss value for training is 0.1448 and for validation is 0.2985.

Early detection and treatment of pneumonia can reduce mortality rates among children significantly in countries having a high prevalence. Hence, this trained deep learning model can be used by doctors and patients since the tests after taking an x-ray usually takes time, it is not good for patients with severe Pneumonia. Hence this model is very useful to predict earlier than usual and give a concrete result before



**Fig. 4** Prediction for a normal X-ray



**Fig. 5** Prediction for an X-ray suggesting pneumonia

the detailed test results arrive. Early treatments due to early diagnosis can save many lives [10].

When an X-ray is uploaded it either displays “The patient’s X-rays are normal” or “The patient might be suffering from pneumonia”. The two scenarios are illustrated in the images in Figs. 4 and 5

## References

1. A Medium Corporation Homepage: <https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>

2. A Towards Data Science Homepage: [https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53#:text=Convolutional%20Neural%20Network%20\(ConvNet,differentiate%20one%20from%20the%20other](https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53#:text=Convolutional%20Neural%20Network%20(ConvNet,differentiate%20one%20from%20the%20other)
3. Seetha, J., Raja, S.S.: Brain tumor classification using convolutional neural networks. *Biomed. Pharmacol. J.* **11**(3), 14571461 (2018)
4. Jain, R., Nagrath, P., Kataria, G., Sirish Kaushik, V., Jude Hemanth, D.: Pneumonia detection in chest X-ray images using convolutional neural networks and transfer learning. *Measurement* **165** (2020) (Elsevier Measurement )
5. A Towards Data Science Homepage: <https://towardsdatascience.com/pneumonia-detection-using-convolutional-neural-network-12b94aeb1206>
6. Hashmi, M.F., Katiyar, S., Keskar, A.G., Bokde, N.D., Geem, Z.W.: Efcient pneumonia detection in chest xray images using deep transfer learning. *Diagnostics (Basel)* **10**(6), 417 (2020)
7. Jaiswal, A.K., Tiwari, P., Kumar, S., Gupta, D., Khanna, A., Rodrigues, J.: Identifying pneumonia in chest X-rays: a deep learning approach. *Res. Gate* (2020)
8. Kadam, K., Dr. Ahirrao, S., Kaur, H., Dr. Phansalkar, S., Dr. Pawar, A.: Deep learning approach for prediction of pneumonia. *Int. J. Sci. Technol. Res.* **8**(10):29862989 (2019)
9. El-Sohly, A., Hsiao, C.-B., Goodnough, S., Serghani, J., Grant, B.J.B.: Predicting active pulmonary tuberculosis using an articial neural network. *Chest* **116**(4), 968973 (1999)
10. Abiyev, R.H., Maaitah, M.K.S.: Deep convolutional neural networks for chest diseases detection. *J. Healthcare Eng.* (2018)

# Predicting the Energy Output of Wind Turbine Based on Weather Condition



P. R. Anisha, C. Kishor Kumar Reddy, and Nuzhat Yasmeen

**Abstract** Wind energy plays an increasing role in the supply of energy worldwide. The energy output of a wind farm is highly dependent on the weather conditions present at its site. If the output can be predicted more accurately, energy suppliers can coordinate the collaborative production of different energy sources more efficiently to avoid costly overproduction. In this paper, we take a computer science perspective on energy prediction based on weather data and analyze the important parameters as well as their correlation on the energy output. To deal with the interaction of the different parameters, we use random forest regression of machine learning algorithms. Our studies are carried out on publicly available weather and energy data for a wind farm. We report on the correlation of the different variables for the energy output. The model obtained for energy prediction gives a very reliable prediction of the energy output for supplied weather data.

## 1 Introduction

We predict wind power output using regression algorithm programming. We use public weather data for predicting energy. We report on a combination of different variations of wind turbine power output depending on the weather. Our model provides the most reliable predictions of power output. We predict power outages up to 95% accuracy [1, 2].

Wind power plays a vital role in the global energy supply. The power output of a wind turbine depends largely on the weather conditions present in its area. If the output can be measured more accurately, then energy suppliers can integrate and co-operate with different energy sources efficiently to avoid cost overproduction. In this paper, we have taken a computer science perspective on wind turbine power forecasts based on weather data and analysis of key parameters and their interactions in energy output. To deal with the interaction of different parameters, we use a variety

---

P. R. Anisha · C. Kishor Kumar Reddy (✉) · N. Yasmeen

Department of CSE, Stanley College of Engineering and Technology for Women, Hyderabad,  
India

of regression methods based on genetic planning tools. Our studies are conducted on publicly available weather and wind turbine power data. We report on the integration of different power output variables [3]. The obtained power forecasting model provides the most reliable prediction of the output of newly acquired weather data.

Renewable energy such as wind and solar energy plays a vital role in the global energy supply. This trend will continue as global energy demand increases, and the use of nuclear energy and traditional energy sources such as coal and oil is considered unsafe or lead to significant CO<sub>2</sub> emissions. Wind energy plays a key in the field of renewable energy. The turbine power generation capacity has increased significantly over the years. In Europe, for example, the volume of power for wind turbine production has doubled since 2005. However, wind power production is difficult to predict as it depends on the volatile weather conditions present in the wind turbine [4].

In particular, wind speed is important for energy production based on wind, and wind speeds can vary greatly at different times. Energy suppliers are interested in accurate forecasting, as they can avoid over-negotiating agreements with the interactive production of traditional power plants and climate-dependent energy sources. Our goal is to put climate data into energy generation. We want to show that even the publicly available data of weather stations near the wind turbine can be used to provide good forecasts of power output. In addition, we are examining the effect of various weather conditions on wind power. We are particularly interested in combining various elements that reflect weather conditions such as wind speed, Theoretical\_Power\_Curve (KWh), wind direction (°) and weather. On the other hand, short-term forecasts are often based on meteorological data, and study methods are used. Often, a person also has an interest in the work itself and the effect of different variations determining the outcome. We want to study the impact of different variations on the power output of a wind turbine. Indeed, the wind speed obtained by a wind turbine is a crucial parameter [5, 6]. Other parameters affecting energy emissions, for example, wind speed, theoretical power curve (KWh), wind direction (°) and weather condition. The genetic program is a type of evolutionary algorithm that can be used to search for tasks that include data mapping data. The beauty of this approach is that it comes with descriptive climate data that clarifies energy emissions. This discourse can be further analyzed to study the effect of the various variables that determine the outcome. To make such a computer, we use different regression algorithms based on the data we have used. We will also use the dataset to make national analysis studying the interplay between different variables and their impact on the accuracy of the forecast [7, 8].

## 2 Literature Survey

The purpose of this paper is to predict wind turbine emissions depending on the weather. The main purpose of the current study is to predict wind turbine emissions according to weather conditions. This can be very helpful in predicting weather

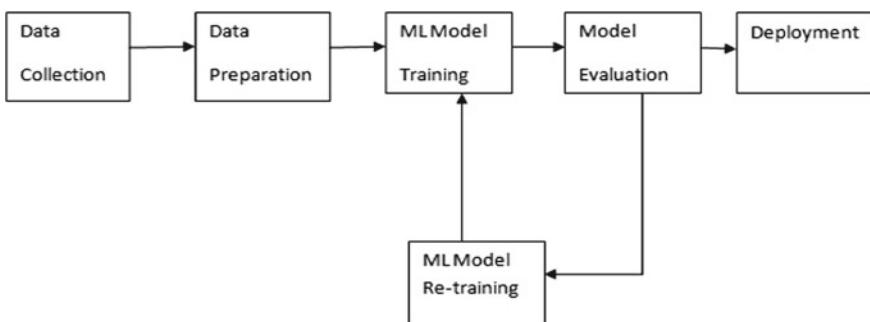
patterns. Parameters such as wind speed, wind direction, thinking power curve and weather conditions play the role of important to predict the power of a wind turbine [9–12].

## 2.1 Proposed Solution

It is important to analyze the factor using number of well-known approaches of machine learning algorithms like linear regression, decision tree and random forest to improve the efficiency of energy of wind turbine. The energy output of wind turbine is predicted based on the weather conditions like sunny, windy, rainy and cloudy. By using the machine learning algorithms. By giving the weather condition, wind speed, wind direction and theoretical power curve we can predict the energy of the wind turbine. By random forest regression. In this model, we considered various inputs like wind speed, wind energy, theoretical power curve and weather conditions. By considering all these inputs, we will get the energy output of the wind turbine. So one can easily predict the energy of the wind turbine based on the weather conditions. Our model in the dataset will be trained on different weather conditions, and it should give us a good estimate for energy of wind turbine (Fig. 1) [13].

## 3 Theoretical Analysis

- step1 Importing libraries: We will be using the following libraries: pandas and numpy. It offers data structures and operations for manipulating numerical tables and time series.
- step2 Reading the dataset and preparing it for processing. First, we read the csv using pandas read\_csv function. Our energy output depends on input value weather condition.



**Fig. 1** Block diagram

- step3 Data preprocessing—taking care of missing values, label encoding, one-hot encoding, feature scaling, splitting the data into train and test sets.
- step4 Model building—training and testing model in various regression algorithms, model evaluation.

### **3.1 Spyder**

- step1 building an Index.HTML file
- step2 building Python code
- step3 flask app—importing libraries—importing flask module in the project, an python file called app.py, pickle library to load the model. Templates folder which contains index.HTML file, static folder contains css and which contains style.css.
- step4 routing to the html page, showcasing prediction on UI Anaconda prompt: To run the app, we navigate to the app.py and type python.py command in the localhost we can view web page.

## **4 Experimental Investigation**

We have created the dataset which consists of 2999 rows and 5 columns. It consists of columns such as wind speed (m/s), theoretical power curve (KWh), wind direction (°), weather condition and energy (j) [14–16].

Wind Speed (m/s): The wind speed at the hub height of the turbine (the wind speed that turbine use for electricity generation).

Theoretical Power Curve (KWh): The theoretical power values that the turbine generates with that wind speed which is given by the turbine manufacturer.

Wind Direction (°): The wind direction at the hub height of the turbine (wind turbines turn to this direction automatically).

### **4.1 Data Preprocessing**

Firstly, we have to import the libraries for numpy, pandas as

```
import numpy as np
import pandas as pd
```

Then, we have to read the dataset using csv file. We have to check for any missing values which are there in the dataset. If there are no missing values, it will give as

**Table 1** Comparison of all algorithm [18–20]

S. No.	Model	Accuracy (%)
1	Multilinear regression	88
2	Decision tree regression	94
3	Random forest regression	95
4	Polynomial regression	95
5	Support vector regression	78

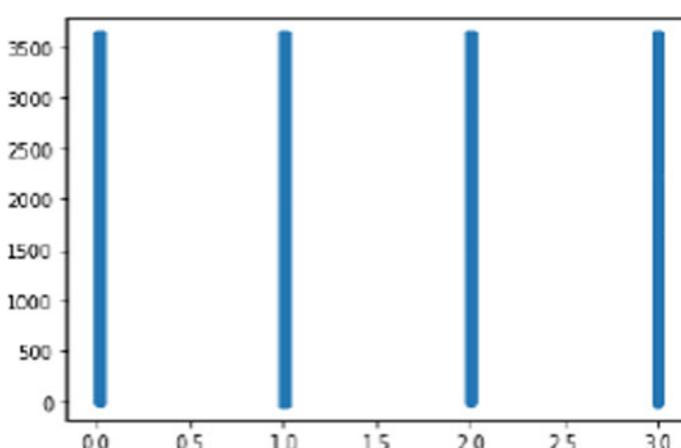
False for all columns. After checking the missing values, we have split the values into  $x$  and  $y$ . ‘ $X$ ’ consists of the inputs, and ‘ $Y$ ’ consists of output.

## 4.2 Model Building

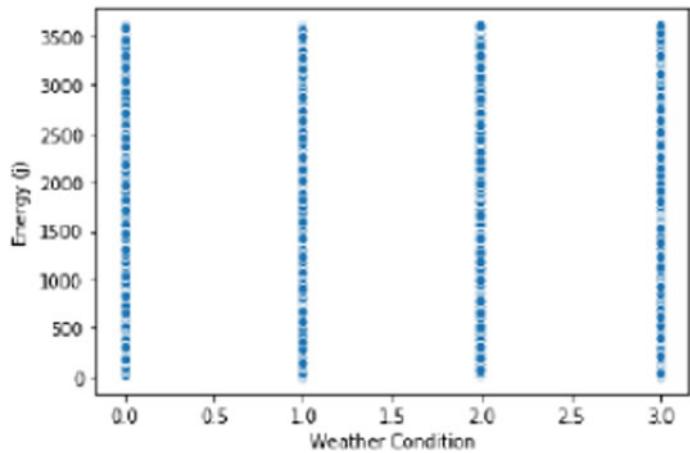
To build the model, we have to import `train_test_split`. We should declare `x_train`, `x_test`, `y_train`, `y_test` with test size as 0.2, and random state should be zero. We have to import the multilinear regression, decision tree regression, random forest regression, polynomial regression and support vector regression. For the dataset, we used more accuracy is given to the random forest regression. After this, we have to save the mode with pickle file (Table 1) [17].

## 4.3 Data Visualization

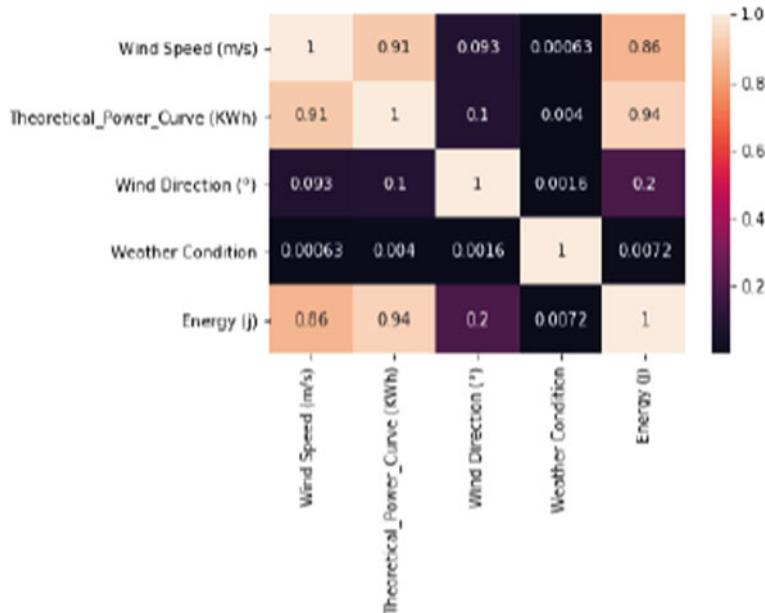
See Figs. 2, 3 and 4.



**Fig. 2** Matplotlib graph between weather condition and energy ( $j$ )



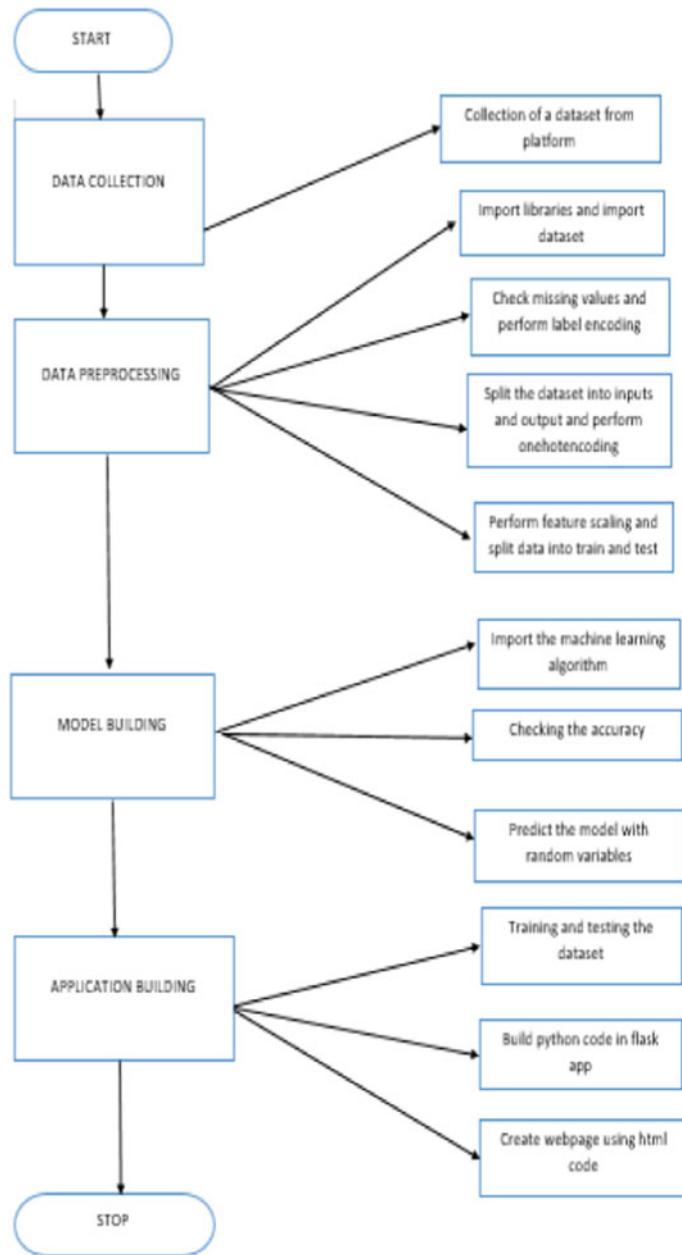
**Fig. 3** Scatter plot



**Fig. 4** Heatmap

## 5 Proposed Architecture

See Fig. 5.



**Fig. 5** Proposed architecture

## 6 Results and Conclusions

We have used random forest regression algorithm to predict the energy of wind turbines based on weather conditions and flask app for application building in flask app. We have used developed Python code and created HTML web pages.

By this, we can predict the energy of the wind turbine based on the weather conditions. This model predicts the output of energy of wind turbines based on weather conditions like sunny, cloudy, windy and rainy. The wind energy output can be predicted from available weather data with accuracy 95% by random forest regression algorithm. By giving the inputs like wind speed, wind direction, theoretical power curve and weather conditions, we can predict the output energy of wind turbine. This is so simple that it can be used by everybody for predicting the energy in wind turbines by weather conditions.

In the future of wind power, wind turbine energy is a clean, renewable way of generating electricity (See close-up ‘Harnessing the Wind’). In the future, provided costs are checked in, and the primary focus will be offshore development.

## References

1. Evolved Analytics LLC. DataModeler 8.0. Evolved Analytics LLC (2010)
2. Foley, A.M., Leahya, P.G., Marvugliac, A., McKeogha, E.J.: Current methods and advances in forecasting of wind power generation. *Renew. Energy* **37**, 1–8 (2012)
3. Jursa, R., Rohrig, K.: Short-term wind power forecasting using evolutionary algorithms for the automated specification of artificial intelligence models. *Int. J. Forecast.* **24**, 694–709 (2008)
4. Kotanchek, M., Smits, G., Vladislavleva, E.: Pursuing the Pareto paradigm tournaments, algorithm variations & ordinal optimization. In: *Genetic Programming Theory and Practice IV*, vol. 5 of *Genetic and Evolutionary Computation*, Chapter 12, pp. 167–186. Springer, 11–13 May 2006
5. Li, S., Wunsch, D.C., Ohair, E.A., Giesselmann, M.G.: Using neural networks to estimate wind turbine. *J. Guid. Control Dyn.* **16**(3), 276–282 (2001)
6. Milligan, M., Porter, K., DeMeo, E., Denholm, P., Holttinen, H., Kirby, B., Mille, N., Mills, A., OMalley, M., Schuerger, M., Soder, L.: Wind power myths debunked. *IEEE Power Energy Soc.* (2009)
7. Koza, J.R.: *Genetic Programming II: Automatic Discovery of Reusable Programs*. MIT Press, Cambridge, MA (1994)
8. Kramer, O., Gieseke, F.: Analysis of wind energy time series with kernel methods and neural networks. In: *Seventh International Conference on Natural Computation* (2011) (to appear)
9. Kramer, O., Gieseke, F.: Short-term wind energy forecasting using support vector regression. In: *International Conference on Soft Computing Models in Industrial and Environmental Applications*, pp. 271–280. Springer (2011)
10. Kusiak, A., Zheng, H., Song, Z.: Short-term prediction of wind farm power: a data mining approach. *IEEE Trans. Energy Convers.* **24**(1), 125–136 (2009)
11. Poli, R., Langdon, W.B., McPhee, N.F.: *A Field Guide to Genetic Programming*. lulu.com (2008)
12. Schmidt, M., Lipson, H.: Age-fitness Pareto optimization. In: *Genetic Programming Theory and Practice VIII*, *Genetic and Evolutionary Computation*, Chapter 8, pp 129–146. Springer (2010)

13. Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. Software available at <https://www.csie.ntu.edu.tw/cjlin/libsvm> (2001)
14. Corchado, E., Arroyo, A., Tricio, V.: Soft computing models to identify typical meteorological days. In: Logic Journal of the IGPL. Oxford University Press (2010)
15. Costa, A., Crespo, A., Navarro, J., Lizcano, G., Madsen, H., Feitosa, E.: A review on the young history of the wind power short-term prediction. *Renew. Sustain. Energy Rev.* **12**(6), 1725–1744 (2008)
16. Evangelista, P.F., Embrechts, M.J., Szymanski, B.K.: Taming the curse of dimensionality in kernels and novelty detection. In: Applied Soft Computing Technologies: The Challenge of Complexity, pp. 431–444. Springer (2006)
17. Herrero, A., Corchado, E., Gastaldo, P., Zunino, R.: Neural projection techniques for the visual inspection of network traffic. *Neurocomputing* **72**(16–18), 3649–3658 (2009)
18. Kusiak, A., Li, W.: The prediction and diagnosis of wind turbine faults. *Renew. Energy* **36**(1), 16–23 (2011)
19. Lew, D., Milligan, M., Jordan, G., Freeman, L., Miller, N., Clark, K., Piwko, R.: How do wind and solar power affect grid operations: The western wind and solar integration study. In: 8th International Workshop on Large Scale Integration of Wind Power and on Transmission Networks for Offshore Wind Farms (2009)
20. Li, G., Shi, J., Zhou, J.: Bayesian adaptive combination of short-term wind speed forecasts from neural network models. *Renew. Energy* **36**(1), 352–359 (2011)

# A Study and Early Identification of Leaf Diseases in Plants Using Convolutional Neural Network



R. Madana Mohana, C. Kishor Kumar Reddy, and P. R. Anisha

**Abstract** This work proposes a method of detection of plant diseases. While flora and crops are stricken by pests it influences the rural production of the USA. Normally, farmers or professionals look at the plant life with the naked eye for detection and identification of ailment. But this technique can be time processing, luxurious and faulty. Automatic detection using photograph processing strategies provides fast and correct consequences. This paper makes a speciality of growing the newest plant disease recognition model, based totally on plant leaf image category, by means of using deep convolutional networks.

## 1 Introduction

The fundamental reason for survival of living beings is plants, and they are useful to every living organism in some or the other way. Moreover, almost every thing that has existence on earth is directly or indirectly related to plants. Plants play a major role in food cycle and environmental cycle. Though it has so much importance, not everyone in concrete houses know the problems that those life savers are facing. Farmers who actually strive day and night to bring food on our plate face many problems while cultivating crops. One of the biggest problems in agriculture is diseased plants [1].

Most of the farmers in many countries are not so literate enough to know the exact illness the plant is suffering with [2, 3]. They have to first collect the sample of effected plant and then submit to some in laboratory and wait for results. All this is a lengthy process, and if results are known early, then there is chance of saving the plant or else one plant may effect whole crop and there might be much bigger lose. An automated system which can detect the disease in least time and with availability

---

R. Madana Mohana (✉)

Department of Computer Science and Engineering, Bharat Institute of Engineering and Technology, Hyderabad, India

C. Kishor Kumar Reddy · P. R. Anisha

Department of Computer Science and Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, India

to every individual would makes this process of disease detection easier and time saving so that there will be modernization as well a better solution for existing problem [4–6].

### **1.1 Why no to Manual Detection of Disease? [7]**

- Its time taking
- No accuracy
- No guarantee of the exact result.

### **1.2 How Detection Through Machine Learning Justifies the Process! [7]**

- Time saving
- Full accurate results
- Prediction into healthy and diseased leaf is exact.

## **2 Literature Review**

The type of diseased plant remains using the neural community algorithm  $k$ . Muthukannan and others participated. Leaf spot lesions can be classified based on the disease and the diseased leaf using different neural community algorithms. The method of classifying the leaves of diseased plants using the forward neural community (FFNN) studies the vector quantization (LVIC) and radial foundation feature network (RBF) by processing form and shape functions from the affected leaf image used. The simulation effect shows the effectiveness of the proposed scheme. With the help of these illustrations, a fully based machine can be built for the first rate improvement of the crop in the Indian economy [1].

Begin by capturing the Malvikaranjan picks in the paper detection and leaf sickness section using synthetic neural networks. Colors such as Hvv capabilities are extracted from the end result of function segmentation, and the artificial nervous system (N) works effectively with the help of selecting function values that can accurately distinguish between nutrient and disease samples. Experimental results have shown that the taxonomic performance by the set-taking feature is high with 80% accuracy. In the present work, a method for early detection of cotton leaf diseases has been proposed and as it should be, several photo processing strategies and the use of synthetic neural community (N) [2].

The purpose of paper leaf sickness is to classify the use of synthetic neural networks written by Ciaficahishakis et al. The use of the image processing method

is to capture and evaluate facts from leaf picks for the dangerous or dangerous classification of drugs that have been converted into plant leaf. Through the photo processing method, an algorithm of adjusted evaluation, segmentation and feature extraction is used to take the picture, and artificial neural networks are used to get the data results. Multilayer forceps neural networks multilayer perceptron and radial foundation feature RBFs are networking systems used for healthy or unhealthy use of leaves. In the final experiment, the end result indicates that the RBF network is more active than the MLP community [3].

Srdjansladojevic et al., In a deeply based deep neural network of plant diseases by classification of leaf images. Through the use of deep hardwood networks, mainly based on leaf photographic classification, plant disease is associated with a whole new technology for reputation pattern improvement. The new methods of training and the anthropology they used contributed to the quick and clean gadget implementation in practice. The developed model can identify the leaves of 13 different types of plants with clean leaves, which have the ability to distinguish plant leaves from their surroundings. All the important steps required to implement this disease popularity model were fully defined at some point in the paper, starting with the submission of photographs with the intention of creating a database and assessed with the help of agricultural experts. Kaif will use the in-depth study framework developed by the Berkeley Vision and Learning Center to conduct in-depth CNN education. Experimental results on the advanced model were performed between 91% and ninety-eight%, with an accuracy of 96.3%, for the personal glory test [8].

The use of organizational networks in the plant disease class and the unfavorable networks of generators and others. Emanuel used a public dataset of 86,147 images of diseased and nutritious flowers, a deeply mixed network and semi-monitored methods to raise awareness on crop species and clutter. A well-done experiment with unwanted data turned into rsnet. It was changed to a rating of more than eighty% over the school period in 5 ages with 1 e-five tuition fee [9].

In-depth knowledge of specimens for plant disorder identification and prognosis on this paper, and others. Constantinos p. Fantino Confucius Neural Community Fashions was developed to diagnose and diagnose plant disease, with in-depth knowledge of the method using simple leaf snapshots of healthy and diseased flora. Fashion education was completed using an open database of 87,848 photographs, including 25 different flowers in fifty-eight wonderful trainings of [plant, disease] blends, including healthy flowers. Many model architects are educated, 99. Fifty-three% of successful price [plant, disease] compounds (or healthy plants) are found in the overall performance. The highly achieved price model makes it a very useful advice or early warning tool and a well-suited technology to assist the plant disorder detection machine involved in working in real farming conditions [10].

Seravark Wallizin et al. The use of soybean plant detectable neural networks within the paper explains CNN's potential for plant disorder class for leaf picks taken under natural environments. The model is mainly based on the linate structure to eliminate soybean plant disease. Class. The 12,673 specimens contained four references to leaf snapshots, including healthy leaf images obtained from the Plantvilade base. Photographs were taken in an unrestricted environment. The demonstrated model

achieves ninety-nine% classification accuracy, which clearly shows that CNN can capture the required capabilities and classify plant diseases from snapshots taken in herbal environments [11].

Real-time Tomato Popular Disease and Pest Popularity and Deep Knowledge-Based Comprehensive Detector for others. We must not forget the three largest houses of Alvaro Fuels detectors: a strong neurological community (rapid R-NN)-based primarily in the initial field, region-based and absolutely good network (R-FCN) and female shot multibox detector (SSD) “meta” for the purpose of this work. It has a deep knowledge of architecture. “We combine each of these meta-architectures with the “Deep Trait Extractor,” which includes WagNet and residual network (Resnet)” [12].

### 3 Dataset Description

Dataset contains many images of various kinds of healthy and unhealthy leaves specifically we collected the images of potato, tomato, and pepper leaves, collected from repository. The entire dataset is divided into training and testing set in 80:20 ratio. This means 80% of the dataset is used for training and 20% of the dataset is used for testing purpose [13, 14]. The dataset has 15 classes of above mentioned three plants as follows:

- Potato\_healthy
- Potato\_early blight
- Potato\_late blight
- Pepper\_bell healthy
- Pepper\_bell bacterial spot
- Tomato\_healthy
- Tomato\_bacterial spot
- Tomato\_leaf mold
- Tomato\_septoria leaf spot
- Tomato\_early blight
- Tomato\_late blight
- Tomato\_yellow leaf curl virus
- Tomato\_mosaic virus
- Tomato\_spider mites\_two spotted spider mite
- Tomato\_target spot.

#### Healthy Potatoes

Potato leaves (*Ipomoea batatas*) are loaded with various nutrients, vitamins, dietary fiber, and essential fatty acids. It contains a large amount of protein, minerals Vitamin B, Beta carotene, Lutein, and antioxidants [15].

**Fig. 1** Healthy potato leaf

### Late Potato Disease

The leaf spots start out as small, pale green, with unusual odd spots. The spots usually have green to yellow rings around them. Spots are not limited to arteries but can grow on them. In cool, humid climates, the spots quickly grow brown to beautify dark areas.

### Potatoes The First Leaf Disease

Premature injury (EB) is a potato disease caused by the fungus *Alternariásolani*. It is found wherever potatoes are grown. The disease mainly affects leaves and stems, but under favorable weather conditions, and if left unmanaged, can lead to significant reductions and improve the risk of tuberculosis [16] (Figs. 1, 2 and 3).

### Pepper Bell Healthy Leaf

Pepper plants are the basis of many vegetable gardens. It is easy to grow and add good taste to countless containers. Medium varieties, such as iron peppers, are essential for many salads and snacks. Pepper plants are easy to grow, but at the same time, the problem will arise. It is good to get used to other issues with pepper if this happens. If you can see the problem, it is easy to find a solution (Figs. 4, 5, 6 and 7).

### Pepper bell Bacterial Spot Disease

Leaf spots from the underside of old leaves as small bumps and on the upper part of the leaf as small water-filled spots are a sign of bacterial habitat. This is an important case of cholera in Maryland. It also occasionally attacks tomatoes. Eventually, the spots grow gray to tan centers with black borders. Sores develop in warm, humid weather. The leaves may turn yellow, turn brown, and fall off. Sores can also grow

**Fig. 2** Potato late blight leaf



**Fig. 3** Potato early blight leaf



on the stems. Fruits form small, raised areas that do not affect the quality of the food. Highly infected leaves will shrink and lead to sunburn of peppers. The bacterial foliage is still pollinated by the shiny rain and works with wet, infected plants.

**Fig. 4** Pepper bell healthy Leaf



**Fig. 5** Pepper bell bacterial spot leaf



### **Tomato Healthy Leaf**

Tomato leaves have a variety of uses that include more than just harvesting ripe fruit to eat. Grow these useful plants in your garden to use all of their useful properties. Tomatoes are used raw and cooked, whole, sliced, and even fresh form to add flavor and acidity to a variety of kitchens [17].

**Fig. 6** Tomato healthy leaf



**Fig. 7** Tomato target spot leaf



### **Tomato Target Spot Disease**

The identified area, or early injury, is one of the most common diseases affecting the leaves and stems of potatoes and tomatoes. The disease is caused by the fungus *Alternaria solani*. In caterpillars, the disease develops from small circles to brownish-brown to black spots. These areas grow, oval to angular, and are often enclosed within

**Fig. 8** Tomato mosaic virus leaf



the main arteries of the tracts. They reach 6 mm in diameter. Under ideal conditions, individual spots can grow up to 10–12 mm. They look like skin, and the growth of the closest rings in each area gives the disease its name, its target location. When the disease is severe, the spots may come together and cause congestion at the top of the leaf tips and leaf death.

#### **Tomato Leaf Disease**

Symptoms of tomato infection can be found at any stage of growth and all parts of the plant can be infected. ToMV causes yellow markings on the leaves and fruit of tomatoes. Symptoms of tomato infection appear as normal height or mosaic appearance on the leaves (Figs. 8, 9, 10, and 11).

#### **Tomato Curl Virus Disease**

Infected tomato plants initially show vertical and vertical or vertical plant growth; infected plants at the beginning of growth will show greater resilience. However, the most diagnostic symptoms are those that are leafy. The leaves of infected plants are small and curved and showed a strong and balanced curve between the oval. The internodes of infected plants are shortened and, with dry growth, the plants often take on a hot appearance, sometimes called ‘bonsai’ or broccoli growth. Flowers planted on infected plants usually do not grow and fall (invisible).

#### **Tomato Bacterial Leaf Disease:**

The spots are small, numerous, dipped in the upper leaf of the leaf, and then slightly raised below. Single spots up to 3 mm; conjunctivitis, especially on leaf tips and genitals, and the leaves look hot and yellow; sometimes, the lower leaves die and fall off. The spots on the stalks and stalks are not circular. Dots appear on small fruits like enlarged scabs.

**Fig. 9** Tomato yellow curl virus leaf



**Fig. 10** Tomato bacterial spot leaf



### **Tomato Blight Early Blight**

Early injuries are a common tomato disease caused by the fungus *Alternaria solani*. It can affect almost every part of the tomato plant, including the leaves, stems, and fruits. The plants may not die, but they will be weak and will produce less tomatoes than usual.

**Fig. 11** Tomato early blight leaf



## 4 Proposed Methodology

Our proposed system is a convolution neural network (CNN). It is a special type of artificial neural network that uses perceptron. It is a supervised deep learning technique. It is used for image-based applications. Components of CNN are the convolution layer, nonlinear layer, pooling layer, and fully connected layer [18].

## 5 Convolution Layer

It is the first layer on CNN that extracts the features from the input image. In this layer, a mathematical operation takes place which takes two inputs they are image matrix and filter (or kernel).

## 6 Non-Linear Layer

In this layer, activation function will be added to the convolution operation. This layer will model our class label.

## 7 Pooling Layer

This layer reduces the number of parameters when the image resolution is too large. Pooling can be of three types:

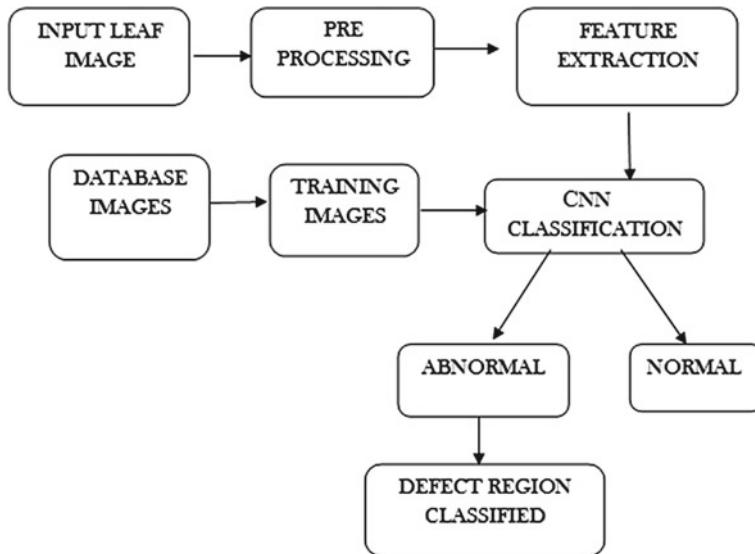
- Max pooling
- Average pooling
- Sum pooling.

It also avoids redundant scanning of already identified parts of the image.

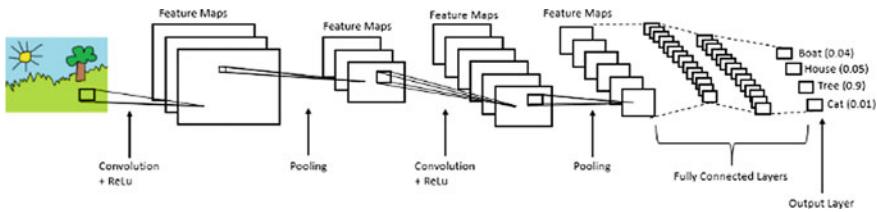
## 8 Fully Connected Layer

In this layer, we will combine all the features to create an output model. Finally, an activation function like sigmoid or softmax is used to distinguish the output, as shown in Figs. 12 and 13.

The proposed methodology is shown in the below flowchart.



**Fig. 12** CNN architecture



**Fig. 13** Proposed architecture

## 9 Results and Discussion

During our development process, we have carried out a series of testing to check the differences between the given input and the expected output of our system. Finally, to see whether or not the CNN works, different set of images and labels (other than the training image) were passed through the CNN, and the output results are seen and compared to calculate accuracy and loss. Positive and negative test cases are described below in Tables 1 and table 2, respectively.

The dataset consists of 40,000 images, out of which 32,000 samples are used for training and remaining 8,000 samples are used for testing, and these selection of samples is done randomly in order to obtain accurate result. After training the model, we acquired 96.77% of accuracy in 25 epochs.

Figure 14 shows the confusion matrix for the tomato using F-CNN(Full CNN).

## 10 Conclusion

Although there are many automated algorithms or models for disease detection of leaves, they are partially contributing to the purpose they have been designed. Our proposed CNN model does not include the lengthy procedure of image processing techniques, and it includes the characteristics of feature extraction which replaces tedious work of image processing, where the features get extracted which helps in decision making for output. We hope our model makes a suggestive contribution to the agricultural sector.

**Table 1** Positive test cases

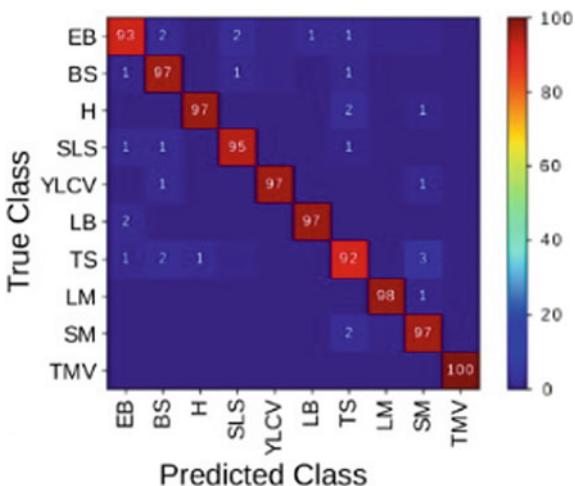
Test case ID	Test case objective	Steps	Input data	Expected output	Actual output	Status
TC_01	Test for potato leaf	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Potato_healthy	Potato_healthy	Pass
TC_02	Test for tomato leaf	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Tomato_healthy	Tomato_healthy	Pass
TC_03	Test for pepper bell healthy	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Pepper_bell_healthy	Pepper_bell_healthy	Pass

**Table 2** Negative test cases

Test case ID	Test case objective	Steps	Input data	Expected output	Actual output	Status
TC_01	Test for potato early blight leaf	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Potato_early_blight	Unable to predict	Fail
TC_02	Test for tomato bacterial spot	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Tomato_Bacterial_spot	File not chosen	Fail
TC_03	Test for pepper bell bacterial spot	1. Click on choose file 2. Select leaf from the folder 3. Click on detect button		Pepper_bell_bacterial_spot	Unable to predict	Fail

**Fig. 14** Confusion matrix for tomato plant.

Abbreviations: Early blight (EB), bacterial spot (BS), healthy (H), septoria leaf spot (SLS), yellow leaf curl virus (YLCV), late blight (LB), target spot (TS), leaf mold (LM), spider mite (SM), and tomato mosaic virus (TMV)



## References

- Muthukannan, K., Latha, P., Pon Selvi, R., Nisha, P.: Diseased plant leaves using neural network algorithms. ARPN J. Eng. Appl. Sci. (2015).
- Ranjan, M., Weginwar, M.R., Joshi, N., Prof. Ingole, A.B.: Detection and classification of leaf disease using artificial neural network international. J. Tech. Res. Appl. (2015)
- SyafiqahIshaka, M.H.F.R.: Leaf disease classification using artificial neural network. Jurnal Teknologi (Sci. Eng.) (2015)
- Shetye, H., Rane, T., Pawar, T., Dandwate, A.: An analysis of methodologies for leaf disease detection techniques (2016)
- Ravindra Naik, M., Sivappagari, C.M.R.: Plant leaf and disease detection by using HSV features and SVM classifier (2016)
- Ananthi, S., Vishnu, S.: Detection and classification of plant leaf disease (2012)
- Ramesh Kumar, S.A., Ramesh Kumar, K.: A Study on paddy crops disease prediction using data mining techniques (2013)
- Culibrk, D., Stefanovic, D.: Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification. Hindawi Publishing Corporation Computational Intelligence and Neuroscience (2016)
- Plant Disease Classification Using Convolutional Networks and Generative Adversarial Networks. Emanuel Cortes Stanford University.
- Soybean Plant Disease Identification Using Convolutional Neural Network SeraworkWallelign-Jimma Institute of Technology, Ethiopia LAB-STICC, Association for the Advancement of Artificial Intelligence.
- Fuentes, A., Yoon, S., Kim, S.C., Sun, D.: Park Sensors, a robust deep-learning-based detector for real-time tomato plant diseases and pests recognition (2017)
- Liu, B., Zhang, Y., He, D.J., Li, Y.: Identification of apple leaf diseases based on deep convolutional neural networks symmetry (2018)
- Mohanty, S.P., Hughes, D., Salathé, M.: Using deep learning for image-based plant disease detection (2016)
- Dhakal, A., Prof. Dr. Shakya, S.: Image-based plant disease detection with deep learning (2016)
- Ghaiwat, S.N., Arora, P.: Detection and classification of plant leaf diseases using image processing techniques: a review (2014)

16. Kajale, R.R.: Detection and reorganization of plant leaf disease using image processing and Android O.S (2015)
17. Deokar, A.S., Pophale, A., Patil, S., Nazarkar, P., Mungase, S.: Plant disease identification using content based image retrieval techniques based on Android system (2016)
18. Jagadeesh, D., Rajesh, P., Yakkundimath, A.S.B.: Identification and classification of fungal disease affected on agriculture/horticulture crops using image processing techniques. In: IEEE International Conference on Computational Intelligence and Computing Research (2014)

# Distributed and Energy Balanced Routing for Heterogeneous Wireless Sensor Network



Shivani S. Bhasgi and Sujatha Terdal

**Abstract** Wireless sensor networks are energy constraint. To minimize the energy usage, data is forwarded through short paths leading to uneven energy usage and causing network partitioning. Hence, energy balancing should also be considered along with energy consumption. Recently, energy harvesting nodes have gained popularity in solving the limited battery life of sensors. Energy harvesting nodes harvest energy from solar, wind, etc. and replenish their battery. But harvesting is sporadic and costly. Hence, in this paper, we propose a hybrid network with both energy harvesting and normal nodes. A distributed routing protocol considering energy usage and balancing is formed along with fault tolerance. Energy density, distance between the nodes and residual energy, is considered to choose relay nodes. Network is rerouted if any node failure is detected in the routing path. The proposed algorithm is compared with existing method. Results show that proposed DEBH performs efficiently when compared with previous method.

## 1 Introduction

Wireless sensor networks have gained a lot of popularity recently. They are used in diverse applications such as temperature monitoring, battlefield and windmills [3] but have limited battery. The batteries cannot be changed as the sensors are deployed in harsh environments [4]. Several protocols have been developed to reduce energy consumption [6]. But most of these techniques do not consider energy balancing which leads to uneven energy usage [9]. In this paper, we consider both and propose an efficient routing protocol using swarm intelligence technique artificial fish swarm optimization (AFSO) [7]. AFSO is used to select next hop node. A weight function is formed using density, remaining energy and distance between the nodes.

Energy harvesting nodes have a separate unit which harvests energy from environment and recharges its battery. Harvesting helps to improve lifetime of sensors but have their own disadvantages such as harvesting is sporadic and expensive. Hence,

---

S. S. Bhasgi (✉) · S. Terdal

Poojya Dodappa Appa College of Engineering, Gulbarga, Karnataka, India

we use hybrid nodes consisting of energy harvesting nodes as well as normal sensor nodes.

These next sections of this paper are organized as follows: Sect. 2 explains previous works, Sect. 3 explains proposed method, and Sect. 4 the simulation results are discussed followed by Sect. 5 conclusions.

## 2 Literature Survey

In papers [5, 9], distributed balanced routing algorithm is implemented using cost function which considers residual energy of nodes. But the algorithms may select a node which does not have any other node in its communication range which will waste energy, and data will not be delivered to sink. Hence, in our paper, while selecting next hop we consider density of nodes which considers residual energy of all the neighbors in its communication range.

In [1, 2, 10, 12], artificial fish swarm optimization technique is used, which has proved to improve the network lifetime [2, 12], improve clustering [1], improve coverage and enhance lifetime [10].

Using harvesting nodes can cause uneven energy and change in power so there is need for different network protocol. Extension of LEACH to make it suitable for harvesting is proposed in [8, 15]. Solar energy-aware wireless sensor nodes are used in [13, 14]. But energy harvesting is depended on the sun.

Several existing techniques have only considered energy reduction and have forwarded data through short paths leading to unbalanced energy causing network partitioning. Faulty nodes are ignored. Energy harvesting nodes are depended on energies like solar and wind, the absence of these due to climatic changes may cause failure of nodes. Most of the existing works do not consider hybrid networks which are cost friendly and effective. In this work, all these issues are considered, and an effective algorithm is implemented.

## 3 Proposed Method

We assume a hybrid network with harvesting nodes and normal nodes. All nodes are stationary. Nodes have information about their neighbors such as residual energy, distance and hop counts. At the beginning, the sink sends a packet to all the nodes with counter set to 0. After receiving it, the sensor nodes store the value increment it by 1 and forward to its neighbors. The data is forwarded by using next hop node (HN). To select the HN, fish swarm optimization algorithm (AFSO) is used. We use a weight function weight ( $n_k, n_j$ ) which considers density and residual energy of the node to balance the energy consumption and short distance nodes to save energy.

The weight function is calculated using the following parameters,

**Energy Density (ED)** The data should be forwarded through higher density region as energy of all the nodes should be balanced. The density is calculated as,

$$\text{Ed}(\mathbf{n}_k) = \sum_{j=1}^{CR_i} \text{RE}_{n_j} (\text{HN}(\mathbf{n}_k) < \text{HN}(\mathbf{n}_j)) + \text{RE}(\mathbf{n}_k) \quad (1)$$

To calculate the density of  $n_k$ , the residual energy (RE) of all the nodes in communication range of  $n_k$  with hop distance to sink less than  $n_k$  is considered. Density should be high so,

$$\text{Weight}(\mathbf{n}_k, \mathbf{n}_j) \propto \text{Ed}(\mathbf{n}_k) \quad (2)$$

**Residual Energy (RE)** If we only consider high density and do not consider RE of  $n_k$  and if RE of  $n_k$  is less, then it will cause uneven energy usage. Hence, residual energy of  $n_k$  is equally important as its density. RE should be high so,

$$\text{Weight}(\mathbf{n}_k, \mathbf{n}_j) \propto \text{RE}(\mathbf{n}_k) \quad (3)$$

**Distance (dist( $\mathbf{n}_k, \mathbf{n}_j$ ))** For reducing the energy consumption, the distance between  $n_k$  and  $n_j$  should be less.

$$\text{Weight}(\mathbf{n}_k, \mathbf{n}_j) \propto \frac{1}{\text{dist}(\mathbf{n}_k, \mathbf{n}_j)} \quad (4)$$

Combining Eqs. (2)–(4)

$$\text{Weight}(\mathbf{n}_k, \mathbf{n}_j) \propto \frac{\text{Ed}(\mathbf{n}_k) * \text{RE}(\mathbf{n}_k)}{\text{dist}(\mathbf{n}_k, \mathbf{n}_j)} \quad (5)$$

$$\text{Weight}(\mathbf{n}_k, \mathbf{n}_j) = \frac{\text{Ed}(\mathbf{n}_k) * \text{RE}(\mathbf{n}_k)}{\text{dist}(\mathbf{n}_k, \mathbf{n}_j)} \quad (6)$$

The weight function can be calculated as in Eq. (6).  $n_k$  selects the node with highest weight value as next hope node for data transmission.

#### Algorithm 1: Next Hop Node Selection

```

1. Start
2.if(dis(nk, sink) ≤ dmax) then // here dmax is the maximum
range of the node nk
3.next hop=sink
else
4.Nj= Neighbour(nk)
5.max=-1.0 // max is the maximum cost
6.while(Nj > 1) do
7. Select a neighbour nk.
8. if( weight(nk nj) > max) // calculating weight using equa-
tion(6)
9. form next hop node through AFSO
10.next hop= nj
11.max=weight(nk nj)
12.end if
13.end while

```

The pseudocode for next hop node selection is written in algorithm 1. If sink is in range of node, it sends data directly else next hop node is selected by AFSO using weight function. If any node does not reply to the acknowledgment message of its neighbor, then it is considered as failed and rerouting is executed.

### **Algorithm 2: Energy Harvesting**

```

1.Start
2.if (replenish==1) then// i.e if node is energy harvesting node
   3. if(NODE_ENERGY[k]< ETH)then// if node energy is
      less than threshold value
        4. NODE_ENERGY[k] = IE// Then node energy is re-
      plenished
      else
   5. if (replenish==0) then // if node is normal node
      6. if(NODE_ENERGY[k]< Ethreshold)then // and its energy
      is less than threshold
        7. ROUTE_DISCOVERY // rerouting is done
      8.end if
   9.Stop

```

Energy harvesting is shown in algorithm 2. If any nodes energy falls below the given ETH value and if it is an energy harvesting node, then its energy is replenished

or else rerouting is executed. Here, energy harvesting nodes are preferred over normal nodes for forwarding data.

## 4 Simulation Results

Performance of the proposed algorithm is compared with existing algorithm [11]. The algorithm is simulated in network simulator 2. The simulation parameters are listed in Table 1. The graphs are plotted against number of rounds versus lowest energy node, nodes available for balancing and remaining energy. Here, a round is considered as time taken from forming route, sensing and forwarding data to sink.

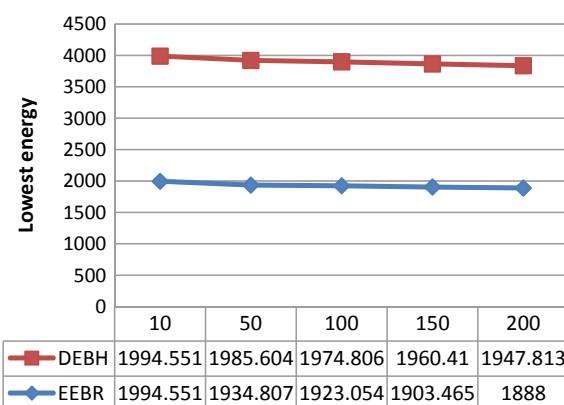
Figure 1 displays graph of lowest energy nodes versus rounds. It can be seen that the energy of EEBR drops, whereas the lowest energy node of DEBH has higher energy than EEBR this is because data is forwarded through energy-efficient path and the proposed algorithm used harvesting nodes.

From Fig. 2, it can be seen that there are more nodes for balancing even after 60 rounds. Tough DEBH and EEBR both considered density, and remaining energy EEBR has a smaller number of nodes for balancing because it did not use hybrid network with harvesting nodes.

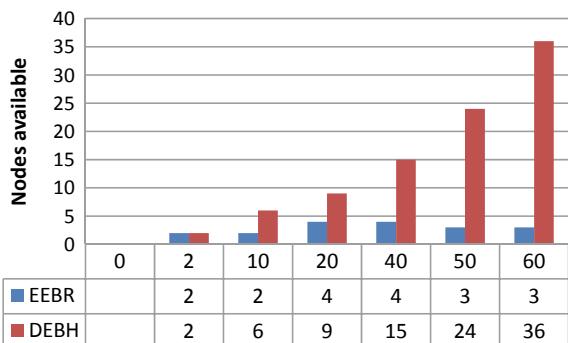
**Table 1** Simulation parameters

Parameters	Values
No. of nodes	50
$E_{\text{initi}}$ (initial energy of each node)	2000 J
Transmission power	2 J
Receiving power	2 J
$D_{\text{max}}$ (communication range)	100 m
Packet size	512 bytes

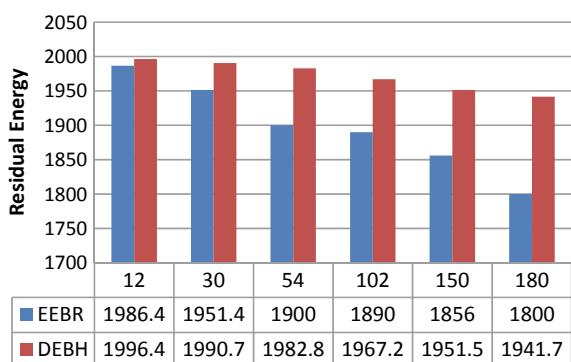
**Fig. 1** Lowest energy node versus rounds



**Fig. 2** No. of nodes available for balancing versus rounds



**Fig. 3** Residual energy versus rounds



Residual energy is high in DEBH even after 180 rounds (as in Fig. 3) but the residual energy of EEBR goes on decreasing. DEBH energy is high because hybrid nodes consisting of harvesting nodes and normal nodes are considered. The energy harvesting nodes are preferred as relay nodes over normal nodes once the energy falls below threshold value.

## 5 Conclusion

In this work, distributed and energy balanced routing for heterogeneous wireless sensor network is proposed. A hybrid network with harvesting nodes and normal nodes is considered. The proposed algorithm focuses on reducing energy usage and balancing the network when selecting route for data transmission. The route is selected using weight which considers density, remaining energy and distance between nodes. The density and remaining energy are important in balancing the network, while distance between nodes will help in reducing energy consumption. The harvesting nodes energy gets replenished if it falls below threshold value this helps in extending the network lifetime. The simulation results show that the proposed

method performs better than existing method in terms of residual energy, balancing, etc.

## References

1. Azizi, R., Sedghi, H., Shoja, H., Sepas-moghaddam, A.: A novel energy aware node clustering algorithm for wireless sensor networks using. **7**, 103–115 (2015)
2. Helmy, A.O., Ahmed, S., Hassenian, A.E.: Artificial fish swarm algorithm for energy-efficient routing technique. 509–519 (2015). <https://doi.org/10.1007/978-3-319-11313-5>
3. Kandris, D., Nakas, C., Vomvas, D., Koulouras, G.: Applications of wireless sensor networks : An up-to-date survey (2020)
4. Li, J., Liao, G., Wang, F., Li, J.: Maximum lifetime routing based on fuzzy set theory in wireless sensor networks. *J. Softw.* **8**, 2321–2328 (2013). <https://doi.org/10.4304/jsw.8.9.2321-2328>
5. Liu, A., Ren, J., Li, X., Chen, Z., Shen, X.S.: Design principles and improvement of cost function based energy aware routing algorithms for wireless sensor networks. *Comput. Netw.* **56**, 1951–1967 (2012). <https://doi.org/10.1016/j.comnet.2012.01.023>
6. Lonare, M.S.: A survey on energy efficient routing protocols in wireless sensor network. 4–8 (2013). 01.0401/ijact.2014.08.14
7. Maria, A., Rocha, A.C., Martins, T.F.M.C., Fernandes, E.M.G.P.: An augmented Lagrangian fish swarm based method for global optimization. *J. Comput. Appl. Math.* **235**, 4611–4620 (2011). <https://doi.org/10.1016/j.cam.2010.04.020>
8. Meng, J., Zhang, X., Dong, Y., Lin, X.: Adaptive energy-harvesting aware clustering routing protocol for Wireless Sensor Networks. In: 2012 7th International ICST Conference on Communications and Networking in China, CHINACOM 2012—Proceedings, pp. 742–747 (2012). <https://doi.org/10.1109/ChinaCom.2012.6417582>
9. Ok, C., Lee, S., Mitra, P., Kumara, S.: Computers & Industrial Engineering Distributed energy balanced routing for wireless sensor networks. *Comput. Ind. Eng.* **57**, 125–135 (2009). <https://doi.org/10.1016/j.cie.2009.01.013>
10. Rushdy, E., Attia, M., Abdalla, M.I. Energy aware optimized hierarchical routing. 614–623 (2018). <https://doi.org/10.1007/978-3-319-74690-6>
11. Singh, D., Kuila, P., Jana, P. K. (2014). A distributed energy efficient and energy balanced routing algorithm for wireless sensor networks. In: 2014 International Conference on Advances in Computing, Communications and Informatics, pp. 1657–1663. [https://doi.org/10.1109/ICA\\_CCI.2014.6968288](https://doi.org/10.1109/ICA_CCI.2014.6968288)
12. Song, X., Wang, C., Wang, J., Zhang, B.: 2010 International Conference on Computer Design And Applications (ICCDA 2010) A Hierarchical Routing Protocol Based on AFSO algorithm for WSN Y Y, vol. 2, pp. 635–639 (2010)
13. Voigt, T., Dunkels, A., Alonso, J., Ritter, H., Schiller, J.: Solar-aware clustering in wireless sensor networks. *Iscc* (2004). <https://doi.org/10.1109/ISCC.2004.1358411>
14. Voigt, T., Ritter, H.: Utilizing solar power in wireless sensor networks. In: *Networks, 2003 LCN'03*, pp. 416–422 (2003). <https://doi.org/10.1109/LCN.2003.1243167>
15. Xiao, M., Zhang, X., Dong, Y. (2013). An effective routing protocol for energy harvesting wireless sensor networks. In: 2013 IEEE Wireless Communication Network Conference, pp. 2080–2084. <https://doi.org/10.1109/WCNC.2013.6554883>

# ESRRAK-Efficient Self-Route Recovery in Wireless Sensor Networks Using ACO Aggregation and K-Means Algorithm



Abhijit Halkai and Sujatha Terdal

**Abstract** Designing of the application-specific protocol in WSN leads to several issues such as energy capacity, network failure, environmental hazards, and energy unbalancing. WSN nodes are deployed in a very harsh environment which makes replacing batteries a complicated task; hence, they are prone to energy drainage and network failure due to continuous data transmission. In this paper, we propose a cross-layer framework with duty cycling for route recovery due to failures in routing by integrating energy harvesting techniques, ACO, and K-Means. By rigorous simulations, in NS-2, it proves that our scheme outperforms the existing EECH algorithm.

## 1 Introduction

Wireless sensor networks (WSNs) are considered as very important in the networking field because of their salient features like cheaper, efficient sensing, adaptable, and smaller size. The design of wireless sensor networks is application-specific [1] due to different applications and topologies. WSN nodes are randomly distributed in a remote location to sense physical data such as vibrations, pressure, noise, temperature, and humidity. Wireless sensor networks have several applications such as process management, healthcare monitoring, environmental sensing, structural health monitoring (SHM), and target tracking. In structural health monitoring, continuous monitoring of the health is done by the automated system using WSN for reliable operation. Designing of application-specific protocol in WSN leads to several issues such as energy capacity, network failure, environmental hazards, and energy unbalancing. WSN nodes are deployed in a very harsh environment which makes replacing batteries a complicated task; hence, they are prone to energy drainage and network failure due to continuous data transmission. Wireless sensor network is an emerging area of research, so it has forced the developers to design algorithms for solving the above issues. Instead of each sensor sending the data individually to the base station (BS), which causes depletion of energy, data can be sent in an aggregated

---

A. Halkai · S. Terdal

Poojya Dodappa Appa College of Engineering, Kalaburagi, Karnataka, India

way by clustering. In this paper, a new cross-layer scheme is designed where each node is equipped with energy harvesting knowledge, a new MAC is designed with duty cycling, and route recovery is done by clustering through ACO and K-Means.

## 2 Literature Survey

Sensor networks illustrate substantial advancement when compared to traditional network. The nodes are randomly deployed in far monitoring regions to record the changes in environment. Secondly, only the sensors which sense the data can be deployed in the topology of sensors, and communication is carefully managed [2]. The sensor is a device composed of sensing unit, processing unit, transceiver, and a power unit; any other device added is based on different applications that are specific [3]. The power unit is fixed with a rechargeable battery which provides the necessary power for sensing and data transfer; the provided battery has limited power, wherein there is a need to recharge it continuously. When deployed in harsh environments, it becomes impractical for humans to reach and recharge. Techniques that are energy efficient should be designed on several challenges [4]. Limited energy, network lifetime, limited abilities, secure communications, cluster formation, cluster head election, synchronization, data aggregation, repair mechanisms, and QoS are the main limitations faced during the cluster formation. Hierarchical routing and cluster-based routing are efficient techniques to maintain the usage of energy in sensor nodes by involving them in multi-hop communication [5]. A survey of clustering algorithms for WSNs presents taxonomy and classification of typical clustering schemes and then summarizes different clustering algorithms for WSNs. Different clustering schemes with special emphasis on their cluster head selection strategies are based on the classification of the deterministic scheme, adaptive scheme, and combined metric scheme. The clustering algorithms available for WSNs [6] are classified based on the cluster formation parameters and cluster head election criteria. Low-energy adaptive clustering hierarchy (LEACH) [7] is broken up into lots of rounds, where each round is separated into two phases, the setup phase, and the steady-state phase. In the setup phase, the clusters are organized, while in the steady-state phase, data is delivered to the base station, and it considers only energy which is not feasible. In Efficient hierarchical clustering (EEHC) [8], cluster head with probability “p” and announces its election to the neighboring nodes within its communication range. All the nodes that are within “k”-hops distance from a “volunteer” cluster head are supposed to receive the election message either directly or through intermediate forwarding. The nature of the ant colony optimization technique of ant nature can also be used for clustering purpose which is also presented in [9] ANTCLUST algorithm, where the partial solution for better efficiency is produced based on the pheromone value of the artificial ant, wherein ants are considered to be WSN nodes. TL-LEACH uses the following two techniques to achieve energy and latency efficiency: randomized, adaptive, self-configuring cluster formation, and localized control for data transfers, and the only locally calculated parameter weight is defined for cluster head election in

DWEHC. Traditional routing protocols are no longer suitable for energy harvesting wireless sensor networks (EH-WSN). Hence, new duty cycling MAC is designed with energy harvesting clustering by ACO and K-Means.

### 3 ESRRAK Design and Implementation

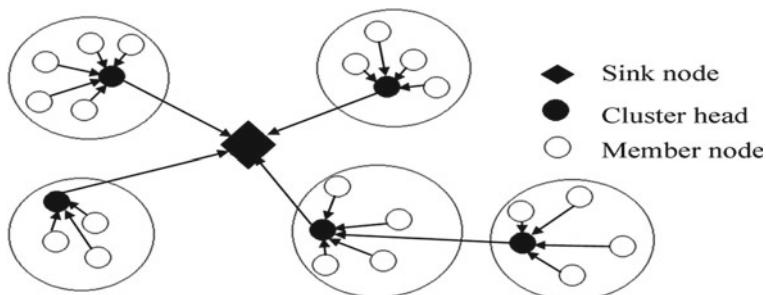
The sensor nodes  $N_1, N_2, \dots, N_N$  are deployed in the seashore structural monitoring with heterogeneous energies  $\alpha$  and  $\beta$  and energy harvesting capability at base station with infinite energy. The detailed flowchart of proposed algorithm is given in Fig. 2

#### 3.1 Network Formation

$N$  numbers of sensor nodes are deployed in a two-dimensional  $M \times M$  area of interest manually. These nodes communicate through the wireless channel. The energy harvested node as sink is set to be fixed with instantaneous recharge by natural sources. The radio range of each node is fixed, wherein the connectivity with other nodes is established within that range. The formation of network and data transmission through cluster heads to sink node occurs as shown in Fig. 1

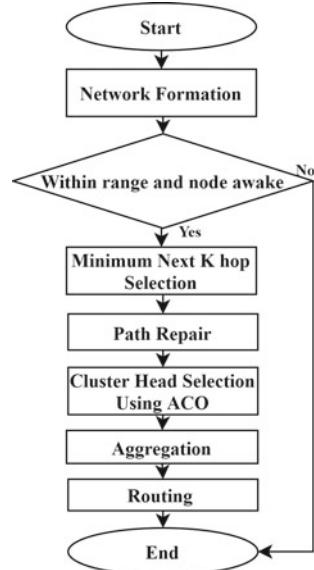
##### Algorithm 1: Network Formation

- *Initialize heterogeneous energy settings for all nodes.*
- *Calculatedistance and RSSI between each neighboring node.*
- *If(calculated distance is less than the range).*
- *Connectivity is possible.*
- *Elseconnectivity is never established.*



**Fig. 1** Network formation

**Fig. 2** Flowchart of proposed algorithm



**Table 1** Hop count beacon message format

No. of next hops	Previous hop node
------------------	-------------------

### 3.2 Minimum K-Hop Calculation

When connectivity is established between the nodes, it is measured by true 1 or false 0. The sink node disseminates the beacon message in the network hierarchically from one level to another calculating the no of k-hops from a sink to cluster head.

All the transactions of the hops calculation are recorded in HCM Table. Initially, the dissemination starts from the sink where the beacon HCM message contains 0 hops and the previous node as the sink as given in Table 1. This message when received by the next level node contains 1 hop and the previous node as sink. At each level, number of hops goes on incrementing, and the least hop to the sink is finally stored in the HCM Table of the respective node. This process is carried until the last node is reached. When each node receives an HCM message, transmission energy is also consumed (Fig. 2).

### 3.3 Path Repair

Path repair is carried in every transmission until the maximum transmissions T\_MAX is reached. If path repair is set to 1, the node is inserted in the queue where the index

of the queue is decremented excluding the node. In node failure condition, it helps in reforming the network without network crash.

### 3.4 Cluster Head Selection

In this clustering approach, sensors with more pheromone values are selected which is directly proportional to the lifetime of the ants. More the pheromone trails and visibility, more the chance of choosing as the optimized solution. Visibility is the parameter that refers to the number of nodes that will be covered if the node is added into the cluster. Artificial ants are used to find a solution for the best optimization problems of cluster head selection.

$$P_r = P_r \times \frac{E_{Residual}}{E_{initial}}, \text{ pheromone} \propto P_r \quad (1)$$

$E_{Residual}$  is energy left in node after each transmission, and  $E_{initial}$  is initial energy of the node in the cluster. The detailed flowchart of ACO clustering is presented in Fig. 3

#### Algorithm 2: Ant Colony Optimization (ACO) Cluster Head Selection

- If (event is detected at nodes and connectivity exists).
- Iterate from  $i$  to  $N$  times /\*  $i = 0$  and  $N = \text{No of Nodes}$  \*/.
- Select  $i$ th node as cluster head based on ant lifetime.
- Select neighbor in the range of  $i$ th cluster head with maximum pheromone value until all nodes are covered reducing more energy utilization on single node.
- Find final set of cluster heads with maximum pheromone trails.

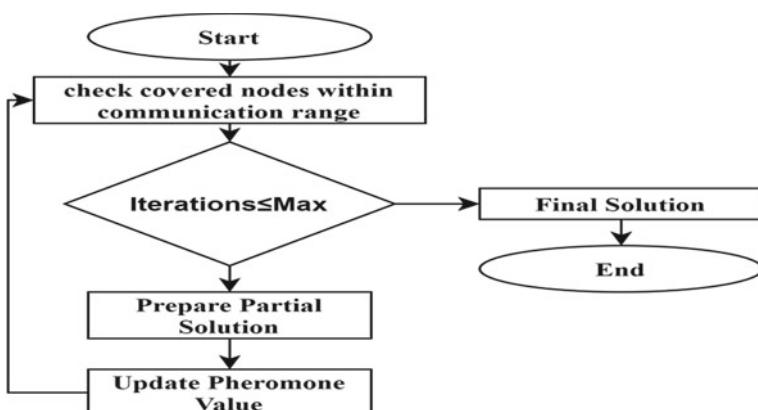


Fig. 3 Flowchart of ACO clustering

### 3.5 Data Aggregation and Routing

#### Algorithm 3: Data Aggregation

- If (event is detected at node and is not Head.)
- Transmit data from source Sensing Node to Cluster Head.

#### Algorithm 4: Data Routing

- Iterate from  $i$  to  $N$  /\*  $i = 0$  and  $N = \text{No Of Nodes}$  \*/.
- Get Distance of the relaying node to SINK.
- If (Node Failed).
- Reroute and find new next relaying hop node.
- Else if (connectivity exists).
- Find most optimal least  $K$ -hops relaying node considering the least distance with respect to SINK.

## 4 Energy Performance Modeling

The nodes in the architecture are defined as the heterogeneous energy nodes from 0 to  $N$ .  $N$  is the maximum no of nodes in the network. The initial energy of nodes is defined as  $E_{\text{initial}}$ , and the total energy of network is given by  $E_{\text{Total}}$  where  $i$  is the node.

$$E_{\text{Total}} = \sum_{i=0}^N E_{\text{initial}}(i) \quad (2)$$

$$E_{\text{sensing}} = E_{\text{initial}} - E_{tx} \quad (3)$$

$$E_{\text{Ch}} = E_{\text{initial}} - E_{tx} - E_{rx} - E_{\text{aggre}} \quad (4)$$

$$E_{\text{Intermediate}} = E_{\text{initial}} - E_{tx} - E_{rx} \quad (5)$$

$E_{\text{Residual}} = E_{\text{sensing}}$  for sensing nodes, and  $E_{\text{Residual}} = E_{\text{Ch}}$  for cluster head.

$E_{\text{Residual}} = E_{\text{Intermediate}}$  for intermediated nodes between cluster head and base station.

$$\text{Total Energy (TE) in network} = \sum_{i=0}^N E_{\text{Residual}}(i) \quad (6)$$

$E_{\text{Residual}}(i)$  is the remaining energy of node  $i$ .  $E_{tx}$  is expenditure in one transmission, and  $E_{rx}$  is expenditure in one reception.  $E_{\text{aggre}}$  is energy expenditure in

aggregation of data from neighbor sensing nodes.  $E_{\text{Intermediate}}$  is energy expenditure for intermediate routing nodes from cluster head to base station.  $E_{\text{Ch}}$  is energy expenditure for cluster head.

The efficiency is given as

$$E_{\text{ef}} = \frac{\text{TE}}{E_{\text{initial}}} \times 100 \quad (7)$$

TE is the total energy in the network.

## 5 Results and Discussion

The simulation is carried out in NS-2.35, and results are extracted through trace file (Table 2).

The primary aim of a sensor network is to send the devastating information to the base station in balanced energy utilization on all cluster heads, and also in an energy-efficient way, in proposed algorithm, the cluster head is instantaneous change during all transmissions. In Fig. 4, it shows that in EEHC [8] same cluster head is elected more than once increasing the energy load. But in proposed algorithm, every node in cluster gets equal chance of selecting as cluster head reducing more energy utilization on single cluster head.

As observed in Fig. 5, the network lifetime is compared between two algorithms, with that the proposed algorithm network lifetime being little higher than EEHC [8]. Network lifetime is measured as the total residual energy of the network. Transmissions are recorded from 1 to 7, and at each transmission, total energy of network is calculated. In EEHC [8], the energy in the network goes on depleting more when

**Table 2** Simulation parameters

No. of nodes ( $N$ )	50
Area	$600 \times 600 \text{ m}^2$
Range	140 m
Initial energy ( $E_{\text{initial}}$ )	200 J
Max hops	10
Energy for transmission ( $E_{\text{tx}}$ )	3 J
Energy for reception ( $E_{\text{rx}}$ )	3 J
Energy for aggregation ( $E_{\text{aggre}}$ )	1 J
Max packets	3
Time for transmission ( $t_{\text{TX}}$ )	0.01
No. of clusters and size	3, 5
No. of events	7

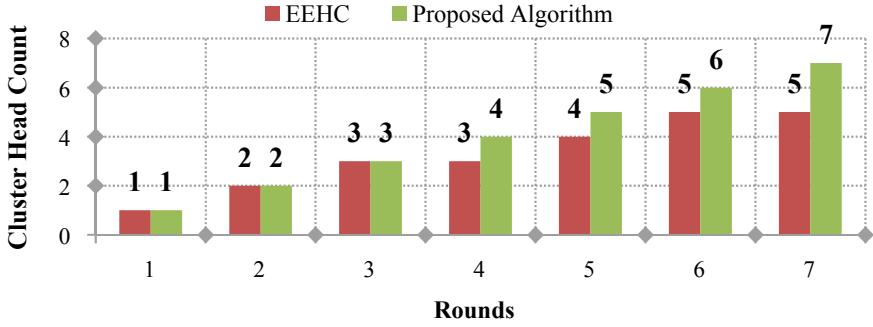


Fig. 4 Rounds versus cluster head count

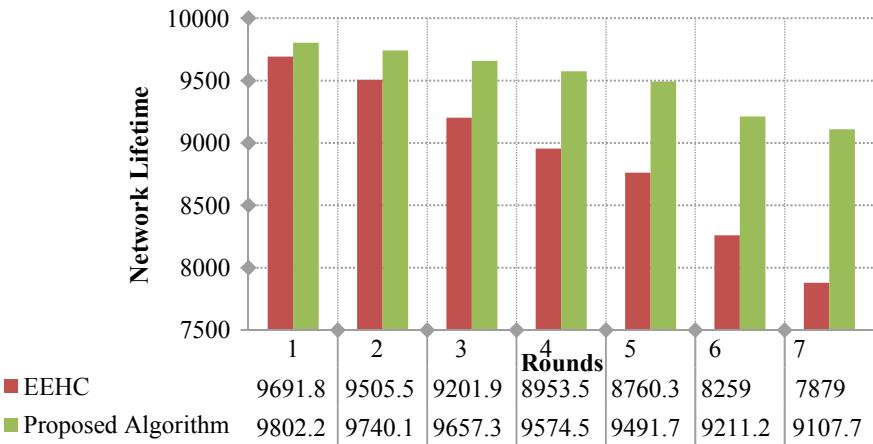


Fig. 5 Rounds versus network lifetime

compared with proposed algorithm. A comparison of both algorithms shows that total network lifetime of proposed algorithm is more than EEHC [8] by 6%.

## 6 Conclusion

In this cross-layer scheme, we propose a framework where the data is transferred to the base station efficiently without any increased energy load on the cluster head resulting in more energy consumption. To find the optimal repaired path to the BS, more pheromone value path is opted by node. The proposed algorithm is capable of sending information to the base station in energy-efficient and energy-balanced way. The link failure reorganizing capabilities form more effectiveness. The rigorous simulations provide the effectiveness and enhancement when compared to EEHC [8].

Shared load on all nodes leads in low energy consumption, thus increasing network lifetime, establishing chance for every node to be cluster head. In result, ESRRAK is proved to be more efficient than EECH in network lifetime and cluster load.

## Reference

1. Bhende, M., Wagh, S., Utpat, A.: IEEE 2014 International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India (2014)
2. Intanagonwiwat, C., Govindan, R., Estrin, D.: Directed diffusion: a scalable and robust communication paradigm for sensor networks. In: Proceedings of the ACM Mobi-Com'00, Boston, MA, pp. 56–67 (2000)
3. Akyildiz, I.F., Su\*, W., Sankarasubramaniam, Y., Cayirci, E.: Broadband and Wireless Networking Laboratory, School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA
4. Wei, C., Yang, J., Gao, Y., Zhang, Z.: Cluster-based routing protocols in wireless sensor networks: a survey. In: 2011 International Conference on Computer Science and Network Technology
5. Kumar, V.: Energy efficient clustering algorithms in wireless sensor networks: a survey. IJCSI Int. J. Comput. Sci. **8**(5), No. 2 1694–0814 (2011). ISSN (Online)
6. Arboleda, L.M.C., Nasser, N.: Comparison of clustering algorithms and protocols for wireless sensor networks. In: Proceedings of IEEE CCECE/CGEI, Ottawa, ON, Canada, 7–10 May 2006, pp. 1787–1792
7. Heinzelman, W.R., Chandrakasan, A., Balakrishnan, H.: Energy-efficient communication protocol for wireless microsensor networks. In: Proceedings of the 33rd Annual Hawaii International Conference on System Sciences, Maui, HI, USA, pp. 10–19, 4–7 Jan 2000
8. Sabet, M., Naji, H.R.: A decentralized energy efficient hierarchical cluster-based routing algorithm for wireless sensor networks. AEU Int. J. Electron. Commun. **69**(5), 790–799 (2015). ISSN 1434–8411, <https://doi.org/10.1016/j.aeue.2015.01.002>
9. Mane, S.U.: Nature inspired techniques for data clustering. In: 2014 International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)

# An Explanation of Personal Variations on the Basis of Model Theory or RKT



K. Reji Kumar

**Abstract** Human beings show significant variations from person to person. Variation in human consciousness is one of the reasons for the difference. Human actions are controlled by mental processes, which involve a comparison of models formed in the mind. A comparison of models can cause differences in consciousness. In this paper, using the model-based theory or RKT of consciousness, we analyze the reason for personal variations in performing different actions or activities and explain the reasons for the variations.

## 1 Introduction

Scientists are trying to explain the intriguing phenomenon of consciousness. The current status of the research is reviewed by Perlovsky, in his article [1]. Mathematical modeling of consciousness is an important step toward finding a solution to many problems related to human consciousness. The multifaceted problem of consciousness can be solved only through the collaborative effort of researchers from many disciplines. All the fundamental ideas in the field need to be defined and explained in a scientific manner. Some important theories that have been proposed by the pioneers in the field are mentioned here. Baars et al. proposed the *global workspace theory* (GWT) [2, 3], which is based on an interaction between bottom-up and top-down attentional modulation mechanisms. It allows a specific percept to become conscious throughout a broadcasting process. The *representation theory* is another theory that is mainly of two types—the *first order* and *higher order*. The essence of representation theories is that consciousness is directly associated with mental representations. Moreover, it has no connection with physical states. Higher-order representation theory is based on the thoughts of Locke and Kant [4]. Dretske [5] and Tye [6] are the advocates of the first-order theory, which argues that the perceptual representations formed in the sensory regions determine a conscious mental state.

---

K. Reji Kumar (✉)  
Department of Mathematics, N S S College, Cherthala, Kerala, India

The *integrated information theory* (IIT), which was introduced by Tononi and collaborators [7–9] can measure the level of consciousness of a system using some mathematical methods. *Quantum theories of consciousness* try to explain the consciousness using the hypotheses that implied quantum mechanisms. The four-dimensional approach to consciousness tries to explain the conscious experiences on the basis of space–time intervals [10]. Koehler’s mathematical approach is also notable in this regard [11].

During the period 2008–2010, the author [12] studied the problem of consciousness in a different direction and attempted to set the study of consciousness on the fundamental units named models. The whole approach proposed uses the methods of mathematical modeling. A set of axioms of consciousness were suggested (model axioms) to give the study a mathematical framework [13]. It is also argued that the whole edifice of consciousness is built with models. To explain it further, the models are categorized based on their increasing level of complexity. Initially, three types of models were suggested, which are named  $\alpha$  models,  $\beta$  models, and  $\gamma$  models [13, 14]. The  $\alpha$  models are very fundamental and are generated in the mind through the input information received through the sensory organs. But,  $\beta$  models and  $\gamma$  models are the models generated in our mind due to the complex model processing activities taking place in our mind. The knowledge of day and night in the mind of a child is an example of a  $\beta$  model. But, the knowledge of the same concept in the mind of a scientist is more developed and advanced. So, it can be called a  $\gamma$  model. While explaining this, it is worth mentioning that there are no precise boundaries between any model and the next model in the model hierarchy.

The concepts are further subjected to the treatments from mathematics to show that complex models can be developed using simple models [12]. Using the model axioms and models as the fundamental units, Reji Kumar explained the phenomenon of combinatorial complexity of consciousness [15] and the situations leading to variations in the level of consciousness of two different individuals. A clear and successful theory of consciousness must explain all the phenomena associated with it and must be ready to face all types of questions raised in this connection. Moreover, a theory that puts forward the claim that it is final and complete will contain a contradiction to its existence, in it. This philosophical aspect was a matter of highest concern while proposing the theory of models to explain the consciousness. Modeling theory is used to study various forms of knowledge, and it is used to explain the difference between Mathematics, Science, and other forms of knowledge [16].

Consciousness is a highly subjective phenomenon. It is related to our experiences. Consciousness and experience are more or less similar. We experience something because we have consciousness. Conversely, if we have consciousness, we can experience objects and events around us and the state of our mind as well. The subjectivity of consciousness is mathematically explained in [14, 17].

Our explanations of a phenomenon is a kind of representation, which is called a model in the language of science. A particular phenomenon has a different representation in a different consciousness. Consciousness is the totality of all representations in the mind of a person. At the same time, a phenomenon can have different representations on different occasions in a consciousness. For example, the representation of

a knife until it causes a wound on the body may be different from its representation after the wound. Some models representing pain will also be added to the representation. Models of gravity of the earth are different in the consciousness of a physicist at different stages of the development of his consciousness. More clearly, when he was a child, he may not have any scientific models in his mind representing gravity, which gradually develops during his years of education until it gets a meaningful shape as he becomes a physicist. Still, it can undergo changes according to the changes in the field of knowledge and his own experiences.

We all know that, as a living organism, we do activities and the activities that we do can be classified broadly as conscious and unconscious. The unconscious activity is one that is done without the interference of our consciousness. For example breathing and sleeping. Even though we can experience breathing, it is very hard to experience sleep. We can fully experience our conscious activities, and it is completely controlled by us. In this paper, we develop a mathematical model to explain the interactions between our consciousness and the activities. It is expected to answer all the related questions based on the model proposed.

In the next section, the conscious action of an individual is modeled using mathematical modeling. We need the following basic concepts in Reji Kumar's model theory (RKT) [18] to make the subsequent discussion more clear and understandable.  $C_x$  represents the collection of all models that constitute the consciousness of an individual, say  $x$  and a model in the consciousness by  $m$ . The collection of all models, which generates a model  $m$ , is denoted by  $m_c$ .  $m_c = [m_1, m_2, \dots, m_n]$  represents the collection of all models  $m_1, m_2, \dots, m_n$  that make a model  $m$ . The model  $m_i$  is a sub-model of the model  $m$  if the model  $m$  contains the whole model  $m_i$ , which is extended to  $m$  by including other models, and the model  $m_i$  remains without any change. If  $m_i$  is a sub-model of the model  $m$ , then we write  $m_i \leq m$ . At the same time, the model  $m$  is called a super-model of  $m_i$ , which is denoted by  $m \geq m_i$ . An extension is an operation performed on a model. The model  $m$  is a generalization of the models  $m_1$  and  $m_2$ , if  $m \leq m_1$  and  $m \leq m_2$ . The null model ( $\emptyset$ ) is introduced in the theory to represent any reality. The universal model ( $U$ ) stands for a model that includes all models. The universal model represents the complete reality. These two models serve the special purpose that they make the theory meaningful. We can use capital letters to represent a collection of models that will help us to differentiate them from single models, which are represented by small letters.

## 2 Consciousness and Actions

Throughout the foregoing discussion, we assume that our conscious actions are controlled by the three relevant areas in our brain. One is for planning and evaluating the activities, the second is for developing and transmitting the models of actions, and the third is for recording the completed actions. It is an extension of a preliminary study in which only two components—one for planning and the other completed actions—were considered. The models belonging to the above-mentioned three categories are

represented by the three collections of models. The first collection is ( $M_P$ ) which contains all the models of planning for the activity, and the second collection is ( $M_A$ ) which contains all the models of the activity performed. Signals are transferred from the portion  $M_P$  to the muscles to perform the action. Each action performed is stored in the portion  $M_A$  for evaluation. The models in  $M_P$  are compared to the models in  $M_A$ , to get an assessment of the level of the activity performed at the planning and evaluating area. A new set of models ( $M'_P$ ) is generated based on the evaluation, and it replaces  $M_P$  in the next level of implementation.

Suppose  $M_P = [p_1, p_2, p_3, \dots, p_n]$  and  $M_A = [a_1, a_2, a_3, \dots, a_n]$ . Here,  $a_i$  is the model action performed corresponding to the model  $p_i$  of action planned. We introduce a function  $f$  for comparison. The function  $f : M_P \bullet M_A \rightarrow [0, 1]$ , such that  $f(p_i, a_i) = 1$  if the models perfectly agree with each other. On the other hand,  $f(p_i, a_i) = 0$  if they completely disagree. The collection  $M_P \bullet M_A = [(p_1, a_1), (p_2, a_2), (p_3, a_3), \dots, (p_n, a_n)]$ .

A possible way of comparison of two models and its consequence is described in [15]. As an example, for comparing  $M_P$  with  $M_A$ , each  $p_i$  is compared to every model in  $M_A$ . The relationships that exist among the collection of models are also a model. Such relationships are also compared to the relationships that are associated with the models in the collection  $M_A$ . Next, we give a mathematical explanation for variations in the level of activities performed by two individuals. Moreover, differences can be observed if the same activity is performed by the same persons on different occasions. We can explain all these differences in the context of the mathematical modeling technique used in this paper.

### 3 Explanation for the Variations in the Actions Performed

An activity performed by a human being is the totality of the mental preparation, the physical conditions, and the conditions driven by external factors. In this study, we limit our attention to the internal activities of the mind. It is further explained using two sets of assumed data given in Tables 1 and 2. We define the value of comparison as  $V = \sum_i f(p_i, a_i)$ . We can call this value the *value of agreement*. Depending on the values obtained, by comparison, corrective measures are taken and modified models are generated to improve the performance of the action. This is again done by sending signals from the planning area to the area for developing activities after making necessary changes in the models. A new collection of models  $M'_P$  is developed and used to replace the existing collection, and this procedure is repeated continuously.

The value of the agreement for the first comparison is 4.6 and that of the second comparison is 0.36. So, the first performance is highly efficient than the second. Based on the method of comparison, we can explain the difference in the performance of different individuals in completing the same task under similar circumstances. When two models are compared, the consciousness will not make a comparison of all constituent models. It will continue the comparison until a certain level, which

**Table 1** Data showing high level of similarity

	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$
$p_1$	1	0	0	0	0
$p_2$	0	0.8	0	0	0
$p_3$	0	0	0.8	0	0
$p_4$	0	0	0	1	0
$p_5$	0	0	0	0	1

**Table 2** Data showing very low level of similarity

	$a_1$	$a_2$	$a_3$	$a_4$	$a_5$
$p_1$	0.1	0	0	0	0
$p_2$	0	0	0	0	0
$p_3$	0	0	0.15	0	0
$p_4$	0	0	0	0.1	0
$p_5$	0	0	0	0	0.01

is decided by the consciousness itself. As the number of comparisons increases the efficiency of the performance also increases. Due to this, the consciousness may take false judgment that the performance is good even though it is very poor.

In Table 1, slight disagreement is shown by only two comparisons. So, the corresponding models may get modification in the next level of implementation. But in Table 2, all sub-models show significant difference. So, all of them must be modified. Thus, we have  $M'_P = [p_1, p'_2, p'_3, p_4, p_5]$  in the first case and  $M'_P = [p'_1, p'_2, p'_3, p'_4, p'_5]$  in the second case.

## 4 Conclusion

In this paper, we have discussed the variations in consciousness due to the difference in comparison of models related to the performance of activities. It is a commonly accepted fact that consciousness is highly subjective. But, mathematical modeling methods can be used to explain the intriguing phenomenon of consciousness because mathematics is an objective area of knowledge, which allows no room for any level of ambiguity. The main focus of this paper is to suggest a mathematical technique to model the mental activities related to the performance of human actions and the variations exhibited in its accomplishment. The method discussed here is completely based on the theory of modeling (Reji Kumar's theory or RKT). As a model based on some mathematical methods, it can further be modified and improved after having a comparison with reality.

## References

1. Perlovsky, L.I.: Towards physics of the mind, concepts, emotions, consciousness and symbols. *Phys. Life Rev.* **5**, 2355 (2006)
2. Baars, B.J.: In the theatre of consciousness: global workspace theory, a rigorous scientific theory of consciousness. *J. Conscious. Stud.* (1997)
3. Baars, B.J., Franklin, S.: Consciousness is computational: the LIDA model of global workspace theory. *Int. J. Mach. Conscious.* **1**(1), 2332 (2009). <https://doi.org/10.1142/S1793843009000050>
4. Mehta, N., Mashour, G.A.: General and specific consciousness: a first-order representation a list approach. *Front. Psychol.* (2013). <https://doi.org/10.3389/fpsyg.2013.00407>
5. Dretske, F.: Naturalizing the Mind. Bradford/MIT Press, Cambridge, MA (1995)
6. Tye, M.: Consciousness, Color, and Content. MIT Press, Cambridge (2000)
7. Balduzzi, D., Tononi, G.: Integrated information in discrete dynamical systems: Motivation and theoretical framework. *PLoS Comput. Biol.* **4**(6) (2008)
8. Edelman, G., Tononi, G.: A Universe of Consciousness. Basic Books, New York (2000)
9. Oizumi, M., Albantakis, L., Tononi, G.: From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS Comput. Biol.* **10**(5), e1003588 (2014). <https://doi.org/10.1371/journal.pcbi.1003588>
10. Sieb, R.A.: Human conscious experience is four-dimensional and has a neural correlate modeled by Einsteins special theory of relativity. *NeuroQuantology* **14**, 630644 (2016). <https://doi.org/10.14704/nq.2016.14.3.983>
11. Koehler, G.: Q-consciousness: Where is the flow? *Nonlinear Dyn. Psychol. Life Sci.* **15**(3), 335357 (2011)
12. Reji Kumar, K.: Mathematical modeling of consciousness. Project Report Submitted to DRDO, India (2010)
13. Reji Kumar, K.: Modeling of consciousness—Classification of models. *Adv. Stud. Biol.* **3**, 141–146 (2010)
14. Reji Kumar, K.: Modeling of consciousness—A model based approach. *Far East J. Appl. Math.* **1**, 1–14 (2011)
15. Reji Kumar, K.: How does consciousness overcome combinatorial complexity? *Adv. Intell. Syst. Comput.* Springer India **325**, 167–172 (2015). [https://doi.org/10.1007/978-81-322-2135-7\\_19](https://doi.org/10.1007/978-81-322-2135-7_19)
16. Reji Kumar, K.: Modeling of consciousness: mathematics. Science and other forms of knowledge. Manuscript (2009)
17. Reji Kumar, K.: Mathematical modeling of consciousness: subjectivity of mind. In: Proceedings of the International Conference on Circuit, Power and Computing Technologies (ICCPCT) (2016). <https://doi.org/10.1109/ICCPCT.2016.7530157>
18. Reji Kumar, K.: Mathematical foundation of information processing: a consciousness based study. *I J C T A* (International Science Press) **8**(5), 1989–1995 (2015)

# Fingerprint Enhancement Using Fuzzy Logic and Deep Neural Network



Sridevi Sarraju and Franklin Bein

**Abstract** Fingerprint recognition analysis is one of the most leading preferred prodigious biometric advancements which have drawn generous consideration in biometrics. In this work, fingerprint intensification is performed which is defined by fuzzy logic technique and recognize the matching image with its unique characteristics extracted and classify the features extracted from a fuzzy enhanced image along with three major types of neural networks which are feedforward artificial neural network, neural network, and recurrent neural network in order to classify the unique features extracted from a fingerprint image. This work efficiently expresses the results with fuzzy logic enhancement and neural network classifiers. Its principle goal is to improve the image using fuzzy and extricate the spurious minutiae detected and classify the different features generated using GLCM and DWT. This work displays a framework of unique finger impression classification based on characteristics for extricating different features and three types of neural network for classification. Fuzzy technique is used for the fuzzy-based image enhancement to urge the clear see of the unique finger impression.

## 1 Introduction

Fingerprint image intensification is the procedure to enhance the distorted images to encourage the recognizable proof. The motivation behind the work is to enrich the quality of the distorted condition image generated from any fingerprint sensor, as Images can be corrupted due to various conditions and one of the principle issues is the resolution of the fingerprint sensor generating noisy images. High-quality pictures are vital for exact coordinating of unique finger impression pictures. But unique mark pictures are seldom of idealized refinement. As it may be corrupted or debased due to varieties of the skin, impression state, and condition. In this way, unique finger

---

S. Sarraju · F. Bein (✉)

School of Computer Science and Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea  
e-mail: [bien@unist.ac.kr](mailto:bien@unist.ac.kr)

impression images must be improved before utilized. The idea behind this work fingerprint image intensification process is to improve the quality of distorted and noisy fingerprint images generated from a low-cost fingerprint sensor. Execution of current fingerprint acknowledgment frameworks is vigorously influenced by the precision of their characteristic extraction evaluation. These days, there are more ways to deal with fingerprint analysis with worthy outcomes. Issues begin to emerge in low-quality conditions where the dominant part of the conventional strategies dependent on examining the surface of fingerprint cannot handle this issue so effectively as neural networks. Fuzzy logic technique is implemented first to remediate the distorted picture and enhance it with the implementation of GLCM and DWT2 algorithm features of an image which is extracted, post to which three types of neural network classification is performed to analyze the accuracy of the image generated from the extracted feature parameters and match the test and trained result with the implementation of neural networks and classify the outcome results. The three neural networks used are artificial neural network (ANN), neural network (NN), and recurrent neural network (RNN). This algorithm works efficiently to identify the fingerprint matching from the predefined trained images from the fuzzy enhanced image generated.

Experiments are performed on 100 images in MATLAB 2019 to make sure the extraction process should not get the false minutiae and preserve the true extracted features fuzzy-based image enhancement method makes sure the feature traits of the image is intensified. Better improvement proves the quality improvement further incrementing the highest accuracy determined in the classification further. This work can be used in wide area of applications in biometrics as it is a combined work of distorted fingerprints enhancement, false feature removal, true feature extraction, matching of the images for identification purpose, and classification using neural networks. Experiments show results which are quite promising and give a direction of the subsequent further analysis in future work.

## **1.1 Proposed Method**

There have been several researched carried on the topic and one of the traditional methods is Gabor filter method the most famous and traditional technique in finger-print enhancement. Distortions of the fingerprint image in low resolution, cost-effective sensors often lead to the challenge in training the computer interface to reconstruct the noisy pieces of the information which leads to the following problems as spurious minutiae will be generated as true minutiae can be lost in the process, errors in positioning of the extracted features, the ridge and bifurcations are not always well defined, classification of the extracted features is quite challenging. Also there are certain drawbacks related to the traditional Gabor filter method as it is unknown fact that what could be the accuracy of the enhancement and what is the percentage of the evaluation, classification needs to be performed using different classifiers to understand the evaluation of the analysis performed on the enhancement

and noise reduction which should also extract the true value and should be able to eliminate the false values.

In order to overcome the drawbacks from the traditional Gabor filter method and evaluate the percentage of accuracy associated with the enhancement evaluation, there must be some technique implemented to find the answers of improvement. Classification is an important stage to classify the test results to compare between the different classifiers and find out the best possible solution among all the classifiers. The overall outcome of the results should eliminate the noisy part of information and preserve the accurate features and in order to do that we need improved feature extraction techniques to have large part of information which can be evaluated and analyzed. Matching should be performed between the test and trained data in order to have the fingerprint identification. Out of all the initial step should be the enhancement part and the implementation are based on fuzzy logic with GLCM and DWT 2 dimensional as feature extracting algorithms is implemented. Deep neural networks have been turned out to be used for the classification purposes and the outcome is estimated measured across different neural networks post to which matching is performed between the test and results obtained.

## ***1.2 Pre-processing***

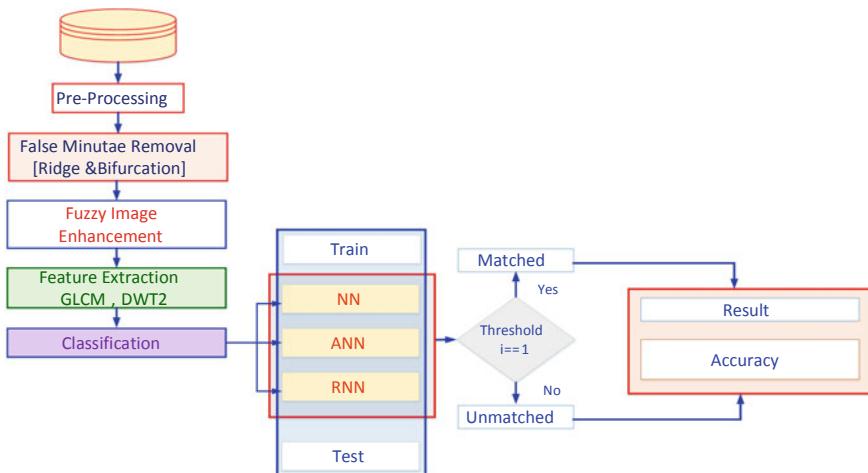
The pre-processing begins with ROI binary image conversion from the input image upon which morphological operation takes place which converts the image to thinned image. Then minutiae and feature extraction are performed, and ridges and bifurcations are determined, and then false detection of these features is detected and removed. Once again true ridge and bifurcation findings are calculated and preserve the true features and false minutiae filtering is applied to the results of feature extraction. After the successful completion of all the pre-processing steps, the image is applied to the fuzzy logic for the enhancement.

## ***1.3 Fuzzy Logic***

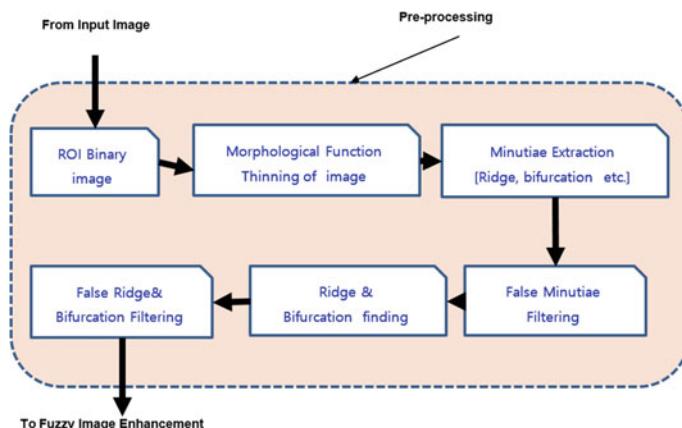
Fuzzy is a rationale or control arrangement of a n-esteemed rationale framework which uses the degrees of state “degrees of truth of the information sources and delivers yields which rely upon the conditions of the data sources and pace of progress of these states as opposed to the standard thing genuine or fake. It acknowledges the deliberate factors as info and changes over the numerical qualities to semantic factors. With the application of fuzzy logic technique, the image is pre-processed and enhanced first, and features of the minutiae are extracted with morphological operations in order to identify the matching true minutiae and make comparison between trained and test fingerprint image from the database and classification is performed using three types of the neural network. In pre-processing, original image

is taken of a particular size suitable for further pre-processing. Algorithm is explained in detail (Figs. 1, 2, and 3).

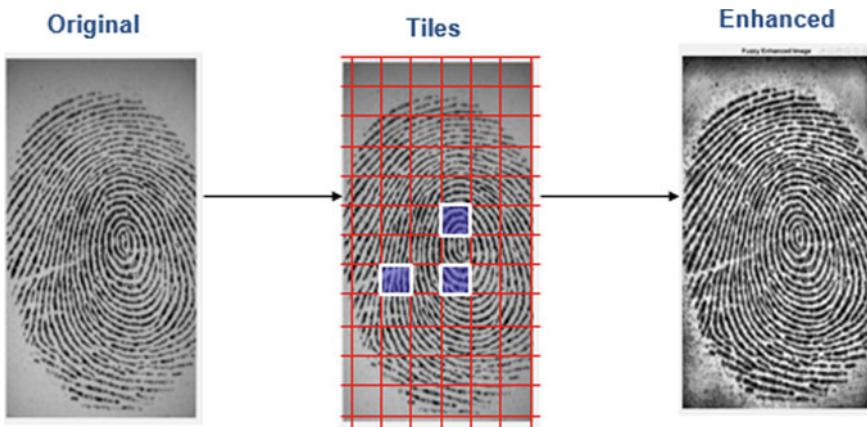
- Pre-processed picture is exposed to the proposed strategy for fuzzy-based upgrade.
- Fuzzy upgrades the nature of the given image by wiping out clamor and improving the magnitude fragments.
- It works on minute information locales known as tiles, as opposed to working on the whole picture.
- Each tile differentiation is improved with the goal that the histogram of the yield area coordinates around to the predetermined histogram.



**Fig. 1** Architecture: proposed method



**Fig. 2** Algorithm



**Fig. 3** Fuzzy enhancement tile operation

- The adjacent tiles are then joined utilizing bilinear insertion so as to dispose of misleadingly initiated limits. First the input image is specified to a binary range [0 255] Discover the histogram of the image
- Divide the outcome by number of pixels
- Calculate cumulative sum
- Convert the image into the int value

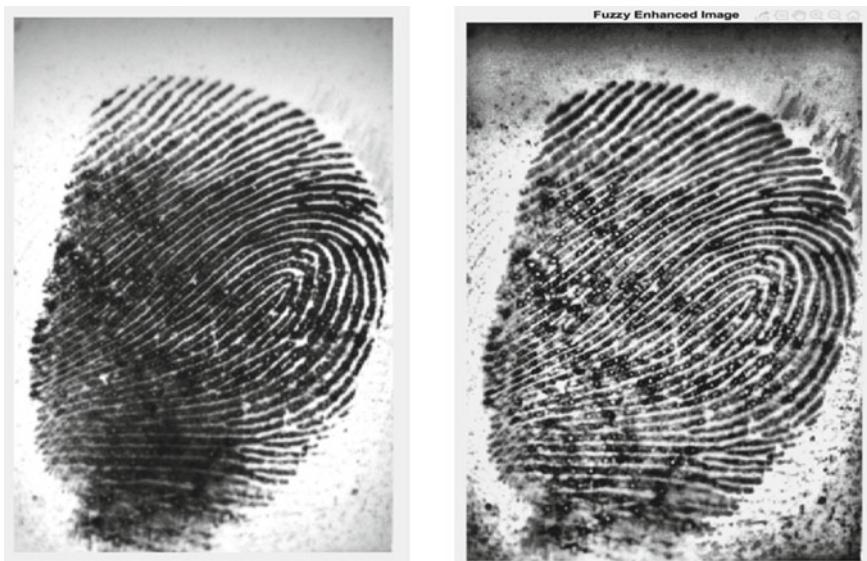
Example in Figs. 4, 5 and 6 it is noticed that from the original input image the fuzzy enhanced image has intensified the image and the features which are hidden, and unseen are clearly visible in the final output and the intensity of the image is also



**Fig. 4** Fuzzy enhancement result 1, shows conversion of noisy to enhanced image



**Fig. 5** Fuzzy enhancement result 2, enhancement constructs the missed and noisy lines



**Fig. 6** Fuzzy enhancement result 2, dark shades are enhanced

improved. Missing ridges and bifurcations are also visible and there is a noticeable difference seen. In all the images the intensity of the image is improved and brighter.

## 2 Deep Neural Network Classification

Deep learning is an incredible breakthrough in the field of artificial intelligence. Its applications have been widely utilized in every one of the fields. In this work, we have utilized different deep learning neural network Classification technique, which mainly helps to analyze the computational value and compare the results to find the best possible outcome using feedforward artificial neural network, neural network, and recurrent neural network. NNs works well with the enormous data as it can be Headings. Trained with extra layers unlike machine learning where training can be performed till a specific degree of extent. The three types of NNs have been utilized in this work so as to measure the accuracy of the results using different classifiers and find the best possible solution. The three types of NNs utilized in this thesis work are artificial neural network (ANN), two-layer neural network (NN), and recurrent neural network (RNN). Results show considerable amount of improvement in all the classifiers and also matching of the image is performed to test the accuracy scores of the trained versus test results.

The neural networks which function as a human brain nervous system have input layer and output layer where we get the final predicted output. The outcome of the activation function checks in the event that the specific neuron gets actuated or not and, at that point enacted neuron transmits information to the neuron of the following layer over the channels. The output value obtained is usually termed as a probability value. ANNs operate on the feedforward neural network principle. The reason of using ANN is especially in pattern recognition since patterns are difficult to characterize. ANNs work well with the distorted data and easy to characterize. Mean Square error is implemented as a Performance function. Training, testing, and validation are done and result of test worth is contrasted and the trained evaluation and shows whether the true parameter estimation of test parameter after the spurious minutiae removal is matched or not with the trained values. RNN is utilized capriciously to insinuate two wide classes of frameworks with a similar general structure.

### 2.1 Feature Extraction GLCM and DWT2

Gray Level Co-occurrence Matrix (GLCM) has been used for strong and exact classification. This efficient method helps to evaluate the extracted features and removal of spurious minutiae. One of the most important reasons for using this algorithm is its distinctive way of summarizing fundamental characteristics. Co-occurrence matrices are often used for extraction because they are composed of regular texture and its strong, effective, and distinctive way of characterizing the features. Appropriation

of pixel dark levels can be depicted by second-request insights like the likelihood of two pixels having specific dim levels at specific spatial connections. This data can be abridged in two-dimensional dim level co-event grids, which can be processed for different separations and directions. It calculates the Euclidean distance across the two centroids of the image and evaluates the features. Discrete wavelet transformation extracts features using high pass and low pass filter; in this algorithm, single level-2-dimensional wavelet transformation is been used along with the Haar wavelet which is a succession of rescaled square molded capacities all together or premise. DWT2 as a feature extraction algorithm is used due to its n levels of data filtering technique it receives approximation coefficient and detailed coefficients. Features extracted from DWT2 coefficients are considered useful to input to classifier.

DWT2 returns the approximation detailed coefficient matrices such as cA, cH, cV, and cD termed as horizontal, vertical, diagonal, and approximation coefficients.

Various features extracted from the image are subjected to pre-processing part first upon which it is subdivided such that there are division and subdivision of the blocks according to the pixelwise operation. The Euclidean distance is what relates to regular experience and recognitions. That is, the sort of single or multi-level dimensional direct metric presence where the partition between any two points in space identifies with the length of a straight line drawn between them. In the wake of figuring the separation between two sets of focuses, for which there is least separation those will be coordinated.

### 3 Results

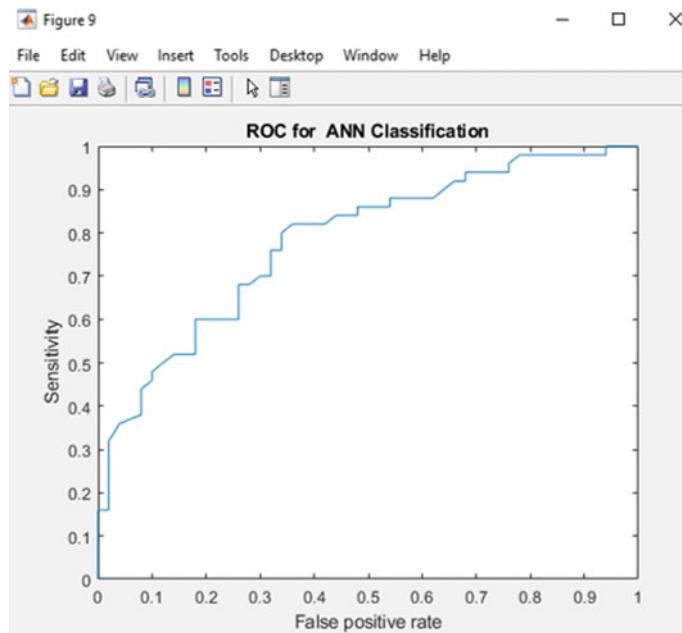
Results in Table 1 show the percentage of the accuracy obtained by all the classifiers, artificial neural network scored the highest possible accuracy scoring 96% when compared to the remaining classifiers, a simple neural network scored the second highest scoring 93%, and recurrent neural network got the least accuracy of 85% comparing to the remaining classifiers.

#### 3.1 Receiver Operating Characteristics (ROC)

ROC curve illustrates the effectiveness of the initiated strategy. In classification to evaluate the execution performance visually this curve comes into picture. The curve

**Table 1** Accuracy obtained in classification

Type of classifier	Percentage of accuracy (%)
Artificial neural network (ANN)	96
Neural network (NN)	93
Recurrent neural network (RNN)	85



**Fig. 7** ROC curve ANN classifier result

is plotted for all the three classifiers showing the best results in all specially ANN. For each image we test in MATLAB ROC curve is plotted for all the classifiers. A good result shows the more value closer to 1, so if the value is close to 1 the higher is the accuracy to be considered (Figs. 7, 8, and 9).

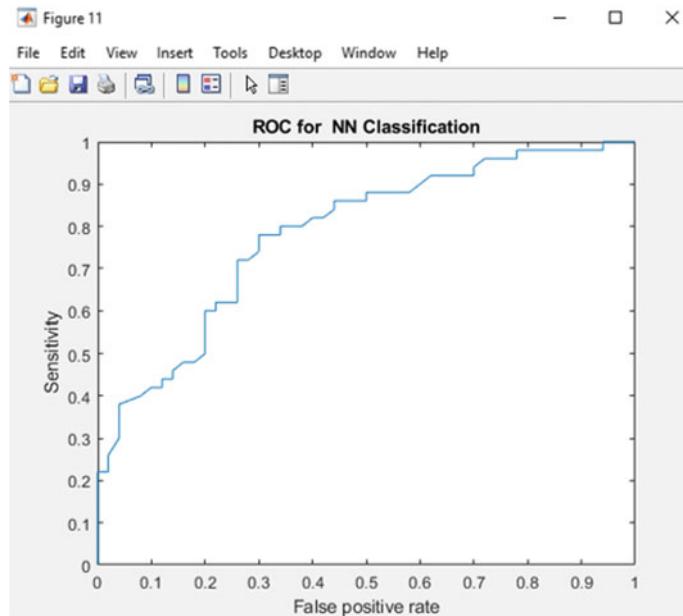
The formula to measure the false positive rate and sensitivity is described below.

$$\text{FPR} = \text{FP}/(\text{FP} + \text{TN})$$

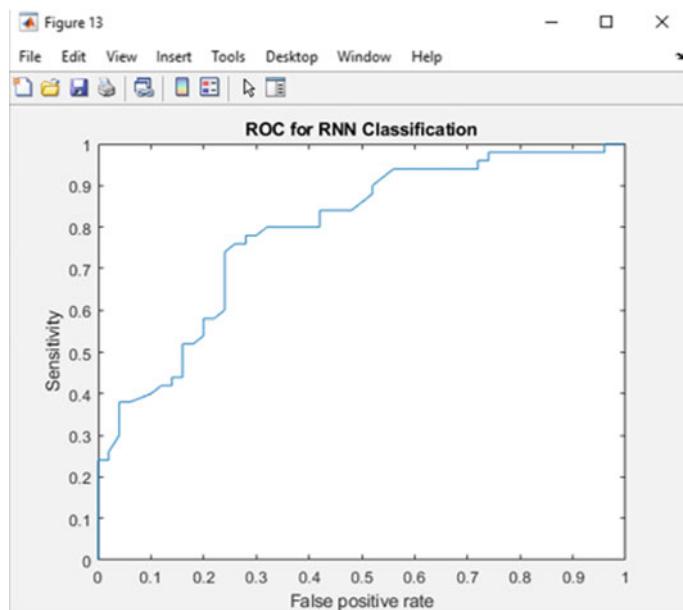
$$\text{Sensitivity} = \text{TP}/(\text{TP} + \text{FN})$$

### 3.2 Equal Error Rate (ERR)

Table 2 describes the EER metrices of all the classifiers and once again ANN proves to be having the best results when compared to the remaining classifiers. The more the value close to 0, the more efficient is the performance.



**Fig. 8** ROC curve NN classifier result



**Fig. 9** ROC curve NN classifier result

**Table 2** ERR results

Type of neural networks	Equal error rate (EER)
Artificial neural network (ANN)	0.0400
Neural network (NN)	0.0700
Recurrent neural network (RNN)	1.5000

## 4 Conclusion

The algorithm helps in precise mapping and classification of the fingerprint images. Utilizing morphological operations false minutiae is removed and true features are preserved using GLCM and 2-dimensional DWT operational algorithms. Fuzzy logic-based image enhancement technique ensures good quality fingerprint images in succession increase the classification and detection of the highest accuracy. The overall classification is comparatively taken less computational time. A feature particulars-based matching is utilized to acquire the matching across the training and the testing images. The process flow begins with pre-processing first and then filtering which acts as a double filtering technique and operated twice in order to make sure the spurious part is detected properly and eliminated preserving the true values characteristics. After performing the morphological function image is set to fuzzy-based enhancement. Total of 100 images were taken in the dataset for the training purposes. The feature extraction algorithm like GLCM and DWT2 extracts different features as the more the features are the better the classification will be performed. The classification results give the highest result after using ANN classification. While performing the classification fingerprint matching is been performed to compare the test versus trained features of the image. The advantages of the work include Better image enhancement technique ensures high-quality fingerprint images which will in turn increase the classification and detection accuracy. Precise mapping and classification of fingerprint images.

**Acknowledgements** This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-20192017-0-01635) supervised by the IITP (Institute for Information and Communication Technology Promotion).

# Gaussian Filter-Based Speech Segmentation Algorithm for Gujarati Language



Priyanka Vishwas Gujarathi and Sandip Raosaheb Patil

**Abstract** Automatic speech segmentation is a main step in speech signal production and analysis process. Great advancement in speech synthesis has already been made using concatenative algorithms. Syllable is most suitable speech unit for concatenative speech synthesis because it does not require extensive prosodic models and provide better co-articulations than other sound units. To get natural sounding, output speech segmentation plays very important role. Speech segmentation is the process of dividing speech signal in to smaller units of sound. So accurate selection of speech unit and detection of boundaries are very important. In this research work, Gujarati language is used for segmentation and database is created. This paper suggests a method of syllable segmentation to detect boundaries of syllable by means of start point of syllable and end point of syllable. Performance parameters such as accuracy and peak signal to noise ratio (PSNR) are evaluated. Producing natural sounding speech signal in different Indian languages is a very demanding and ongoing problem.

## 1 Introduction

Text to speech (TTS) system is a process of converting written raw text data into spoken speech waveform. The TTS method typically involves two steps: text processing and speech generation [1]. Segmentation is the process of dividing the continuous speech signal into a separate sub-word unit, i.e., finding correct boundaries of particular signal. Most of the research work is carried out to segment the speech into units like word, sub-word, syllable and phonemes using various segmentation approaches. Pronunciation duration is very important problem for speech segmentation, mainly in languages like Tamil [2]. Deterministic annealing expectation–maximization (DAEM) algorithm is used to over EM algorithm [3]. Group

---

P. V. Gujarathi (✉)

JSPM's Rajarshi Shahu College of Engineering, Tathawade, Pune 411033, India

S. R. Patil

Bharati Vidyapeeth's College of Engineering for Women, Dhankawadi, Pune 411043, India

delay-based segmentation and spectral transition measure (STM)-based segmentation method are used [4]. Gaussian filter-based method is used [5]. Gaussian component sequence produces better synthetic speech than EM algorithm [6]. When the target raw text is given for converting into speech signal, corresponding smaller units from database are concatenated (String together) to build a continuous speech waveform. Speech corpus development is the process of converting the human voice signal into smaller speech units [7]. For database, a raw text data is selected from text database, which has rich collection sentences. It may be from News bulletins, story books, novels, magazines, interviews, etc. Then, with the help of a good native speaker of particular language database is created [7]. Syllable units can capture better co-articulation than phones. In Indian language about 35 consonants and 18 vowels are involved [8]. Speech signal is generated by using two components vocal cords and vocal tract. When we speak, we utter different words. Aksharas are the essential units of the writing system which is syllabic in nature [9]. For Indian languages, syllable-like units are a much better than other sound units [10]. Co-articulation effect and periodic mismatch between two adjacent units given for Telugu language [11]. Basically, word is formed by using vowels, and then, it is combined with consonants to form syllables which ultimately give a word. For example, the word Gujarati is composed of four syllables: gu, ja, ra and ti. Usually, a list of VC, CV, VCV or CVC words is used and for longer words CVVC, VCCV or CCCVCCC combinations are also required. So producing all syllable combination is difficult. Various criteria have to be taken into consideration while building the speech database, speaker selection, optimal text selection and so on [12]. A fraction-based waveform concatenation technique produces intelligible speech segments [13]. Speech synthesis research is mostly focused on traditional unit selection, statistical parametric speech synthesis (SPSS) [14]. Segmentation plays vital role in speech recognition and synthesis systems. An automatic segmentation of tamil speech into syllable using vowel onset point (VOP) and spectral transition measure (STM) [15]. VOPs detected using auto-associative neural network (AANN) [16]. Automatic syllable segmentation of Pumi speech is done with GABPNN, DWPTMFCC, fractal dimension [17]. To get natural sounding, output syllable boundaries must be accurately detected. Paper is organized as database creation for Gujarati language, speech segmentation at syllable level, experimental results.

## 2 Methodology

### 2.1 *Gujarati Language*

Gujarati is an Indo-Aryan language and spoken mostly by the Gujarati people, and Gujarati is part of the Indo-European language family. There are about 1652 Indian languages; 49 million Gujarati speakers are available. Gujarati language is made up

**Table 1** Commonly used Gujarati words and corresponding English word

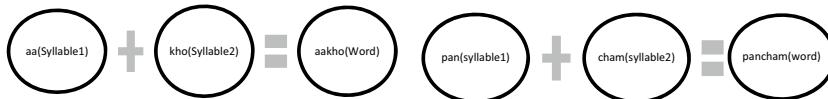
Guj Word	Eng Word	Guj Word	Eng Word	Guj Word	Eng Word	Guj Word	Eng Word	Guj Word	Eng Word
એક	Ek	આ	Aa	તેને	Tene	તમે	Tame	પૈસા	paisa
બે	Be	ચોપડી	chopadi	તમને	tamne	કઈ	kai	જગા	Jaga
જ	chha	કલામ	kalam	કેટલા	ketla	તમારું	Tamaru	દુર	Dur
નવ	nav	મને	mane	તે	Te	શુ	shu	નામ	Naam
દસ	Das	મારે	mare	કેટલી	ketali	કણો	kayo	સુઈ	Sui
સત્તર	sattar	જમણ	jaman	છે	chhe	હતા	hata	સાથે	shathe
સાત	sāt	છાપુ	chapu	કેવો	kevo	ગયા	gaya	ખોજું	khokhu
અગ્રિયાર	agiyār	કાયુ	kayu	હતો	hato	આપણો	aapsho	પરીક્ષા	pariksha
બાર	bār	ક્ષાં	kya	આપ્યા	aapya	આવણો	aavasho	લખી	lakhi
તેર	tēr	સારુ	saru	પધારો	padharo	મારી	mari	પાણી	Pani

of 13 vowels and 34 consonants. Table 1 gives some commonly used Gujarati word along with its English word.

## 2.2 Syllable Unit

Speech signal in Indian language is based on basic sound units which are from consonant, consonant vowel, consonant consonant vowel, vowel consonant, consonant vowel consonant combinations. Speech units like syllable perform better for Indian languages than rest all speech units like diphone, phone and half phone [18]. Syllable is represented as  $C^*VC^*$ , where  $C$  is a consonant,  $V$  is a vowel, and  $C^*$  indicates there may be no or more consonant present.

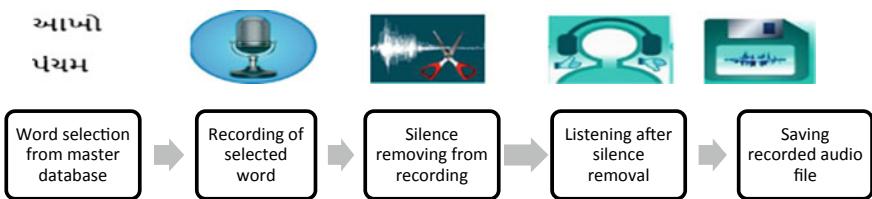
Some grapheme to syllable conversion rules are given in [11]. Example for formation of words from syllables is given: For Gujarati word (1) aakho (આક્ષો) syllable will be aa + kho = aakho. (2) pancham (પંચમ) = pan + cham



Example for formation of word from syllables

## 2.3 Text Database Collection and Speaker Selection (Speech Database)

Text database (5000 + Words) was collected from newspaper, magazines, story books, dictionaries, etc. For speech database creation, speech samples were collected from different speakers, and different parameters were considered like variations in



**Fig. 1** Database creation basic blocks

pitch, amplitude, duration and clarity. Audibility test [Mean Opinion Score (MOS)] was carried out to select speaker, and the speaker with the maximum MOS was preferred for recording of Gujarati speech database (Fig. 1).

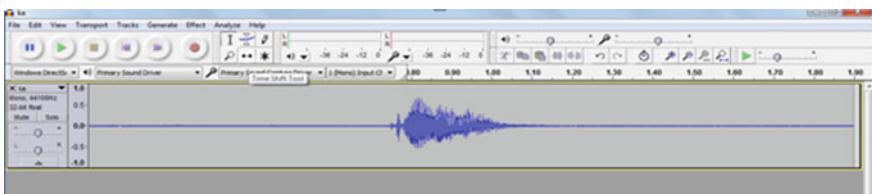
### 2.3.1 Recording Phoneme ka.wav ( $\text{કા}.\text{wav}$ ) and Silence Part Removal From Phoneme by Audacity Software

Figure 2 waveform of phoneme (ka.wav) using audacity software phoneme before silence removal

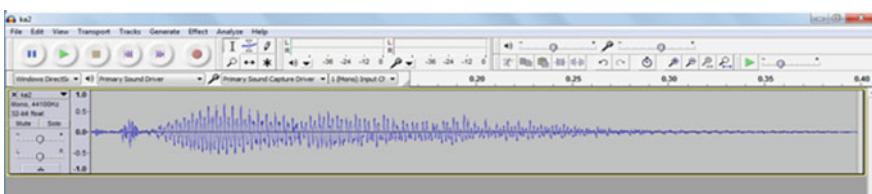
Figure 3: waveform of phoneme ka.wav using audacity software phoneme after silence removal

Figure 4 waveform of phoneme (ka.wav) using MATLAB software phoneme before silence removal

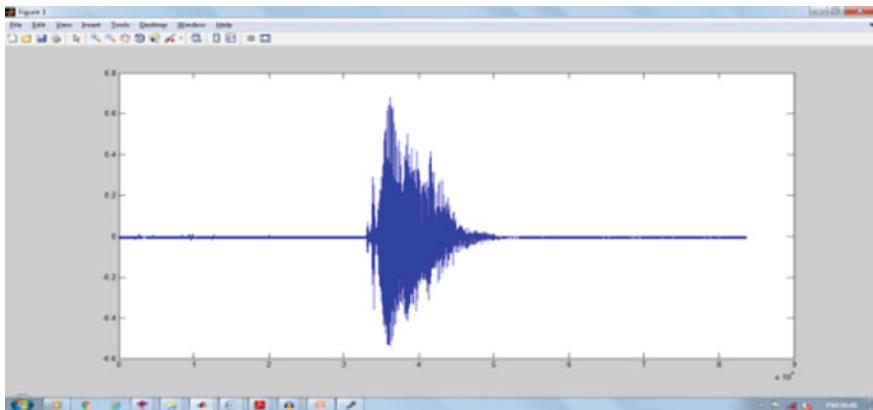
Figure 5: waveform of phoneme ka.wav using MATLAB software phoneme after silence removal



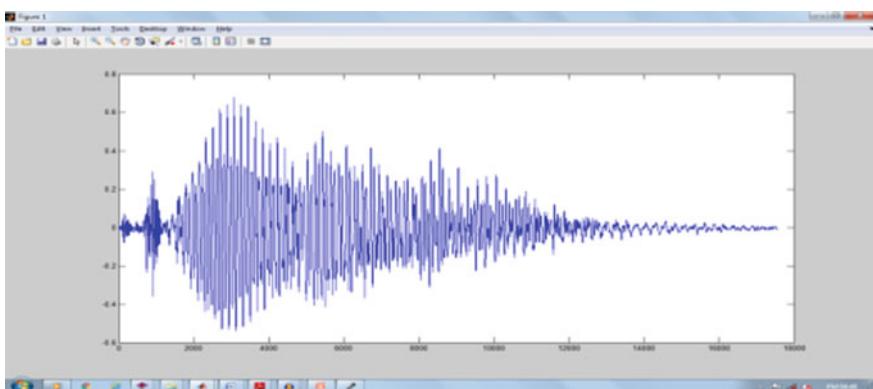
**Fig. 2** Phoneme ka.wav



**Fig. 3** Phoneme ka.wav



**Fig. 4** Phoneme ka.wav



**Fig. 5** Phoneme ka.wav

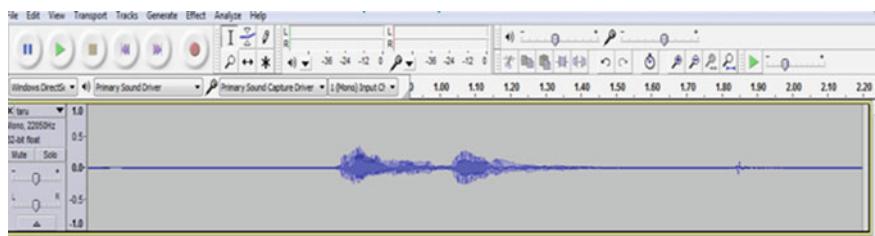
### **2.3.2 Recording Word taru.wav (တရာ့) and Silence Part Removal From Phoneme by Audacity Software**

See Figs. 6 and 7.

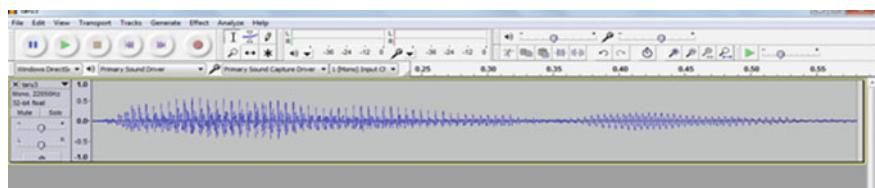
### **2.3.3 Manual Segmentation Using Praat Software: For Word taru.wav Manual Segmentation Is Done**

See Fig. 8.

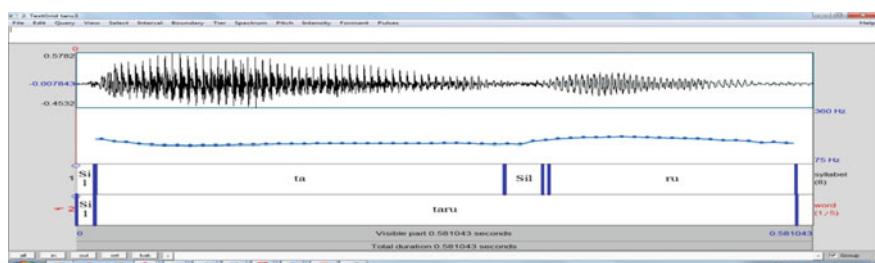
Manual segmentation is very monotonous and lengthy task. It is also difficult to arrive at a common labeling strategy for speech signal across different researchers [19]. So to avoid these problems, automatic segmentation is used. Before saving audio



**Fig. 6** Word taru.wav before silence removal



**Fig. 7** Word taru.wav after silence removal



**Fig. 8** Word taru.wav

file to particular location, we must listen to it. Reasons to listen recorded sound file before saving in proper location: (1) To check clarity of sound. (2) To check silence removal is made properly or not. (3) To check recorded sound is for selected word or not. (4) To check correctness of pronunciation. (5) Saving audio file.

## 2.4 Labeling and Speech Corpus

In this process, given text (syllables) and corresponding speech signals have to be aligned with each other. All the experiments in this paper are conducted on authors database of 1000 Gujarati words. Basically, to cover all the possible syllables is very difficult because number of syllables is not fixed. If sound units are not there in

database, then syllables for those sound units are split into phones. Phone units are used for concatenation to produce synthesized speech signal [5].

### 3 Speech Segmentation at Syllable Level

#### 3.1 Speech Segmentation Using Gaussian Filter Method: [5]

Gaussian function is used in numerous research areas. It is a smoothing operator. The Gaussian filter is a non-uniform low-pass filter (LPF). The Gaussian filter impulse response is given,

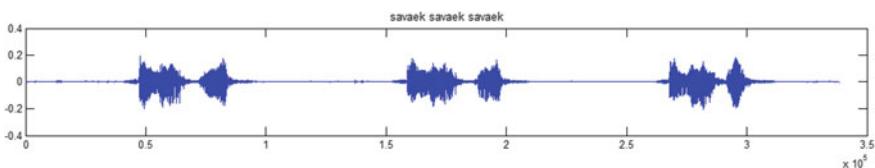
$$g(t) = e^{-\alpha^2 t^2} \quad (1)$$

For a given Gujarati speech input signal, short term energy contour is calculated using overlapped window. Then, filtered STE contour is calculated using Gaussian filter. After that locations of the minima's are indicated by using start point and end point of syllables in the Gaussian filtered STE contour. Calculation of start point (SP) and end point (EP) will give boundaries of syllables. According to SP and EP, calculated syllables are extracted from given input. In figure boundaries are indicated by red color star.

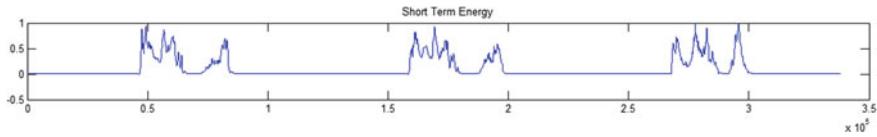
### 4 Experimental Results

For Gujarati word “savaek savaek savaek” syllables are extracted using SP and EP calculation:

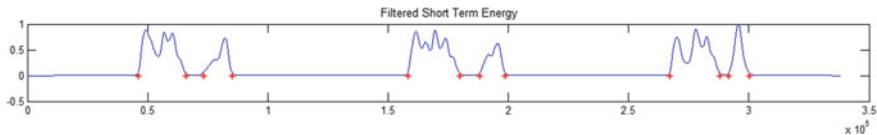
1. Speech waveform is plotted for word “savaek savaek savaek” સવાએક સવાએક સવાએક is given and on x-axis: Time (Sample No), y axis: Amplitude (Fig. 9).
2. Short-term energy contour (STE) is calculated for word “savaek savaek savaek” (Fig. 10).
3. Using gaussian filtering we get final output, i.e., filtered STE output. Syllables for word savaek = sava + ek (Fig. 11).



**Fig. 9** Gujarati word uttered savaek savaek savaek



**Fig. 10** STE For uttered savaek savaek savaek



**Fig. 11** Filtered short-term energy

So in our input savaek is uttered three times so we will get three start points and three end points.

"sava + ek" "sava + ek" "sava + ek" == "savaek" "savaek" "savaek."

Red marks in figure shows start point and end point of syllables. Here, we have taken example savaek so in this first syllable is started at sample 46,122 and ends at 65,862. So boundaries of syllable is calculated to get proper syllable (Table 2).

4. For validation subjective listening test is carried out with list of questionnaires. Voice pleasantness: In this test, voice signal is described in terms of voice pleasantness. Subjective listening test carried out with 20 listeners, and nearly, 4.6 score is obtained out of 5 which means voice was pleasant while listening. Correctly detected boundaries are a measure of segmentation accuracy, and this is referred to as the hit rate (HR)

**Table 2** Start point and end point of syllables from word savaek, Ekavan

Word uttered	Word	Syllable	Syllable	Start point	End point
Same word uttered three times: savaek savaek savaek સવાએક સવાએક સવાએક	Savaek	Syllable 1	Sava	46,122	65,862
		Syllable 2	Ek	73,173	85,251
	Savaek	Syllable 1	Sava	158,143	179,927
		Syllable 2	Ek	188,095	198,765
	Savaek	Syllable 1	Sava	267,044	287,924
		Syllable 2	Ek	291,679	300,295
Same word uttered three times: Ekavan Ekavan Ekavan એકવાન એકવાન એકવાન	Ekavan	Syllable 1	Eka	26,587	31,683
		Syllable 2	Van	37,017	59,577
	Ekavan	Syllable 1	Eka	124,253	128,491
		Syllable 2	Van	134,281	158,595
	Ekavan	Syllable 1	Eka	212,401	217,829
		Syllable 2	Van	222,251	235,689

Nhit = Number of boundaries correctly detected Nref = Total number of boundaries in the reference

Then,

$$\text{HR} = \text{Nhit}/\text{Nref} * 100 \quad (2)$$

The efficiency of the segmented syllable has been evaluated using the PSNR measure which is defined as follows,

$$\text{PSNR} = 20 * \log 10 [\text{Max signal Value}/] \quad (3)$$

The PSNR performance is checked on this algorithm, and HR value is obtained in percentage 79.32. Experimental results shows this method works better for author's database as it is recorded taking into consideration proper Gujarati pronunciations. The speech files those gave incorrect syllables were rerecorded.

## 5 Conclusion

Segmentation of speech signals into linguistic units has enormous applications in the fields of speech synthesis. The main purpose of this work is to provide automatic segmentation algorithm for Gujarati speech database at syllable level. Result is evaluated by subjective listening test (MOS) and PSNR performance. From subjective listening test (MOS), we can conclude that the produced synthesized speech is preserving naturalness. The further research work direction still there is the scope for accurate segmentation at boundaries to improve quality of speech signal.

## References

1. Geeta, S., Muralidhara, B.L.: Syllable as the Basic Unit for Kannada Speech Synthesis. IEEE WiSPNET 2017 Conference
2. Geethaa, K., Vadivel, R.: Syllable segmentation of tamil speech signals using vowel onset point and spectral transition measure. automatic control and computer sciences (2018)
3. Shah, N.J., Patil, H.A., Madhavi, M.C., Sailor, H.B., Patel, T.B.: Deterministic annealing EM algorithm for developing TTS system in Gujarati. 2014 IEEE
4. Patil, H.A., Patel, T., Talesara, S., Shah, N., Sailor, H., Vachhani, B., Akhani, J., Kanakiya, B., Gaur, Y., Prajapati, V.: Algorithms for speech segmentation at syllable-level for text-to-speech synthesis system in Gujarati. In: 2013 International Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)
5. Talesara, S., Patil, H.A., Patel, T., Sailor, H., Shah, N.: A novel gaussian filter-based automatic labeling of speech data for TTS system in Gujarati Language. In: 2013 International Conference on Asian Language Processing
6. Takamichi, S., Toda, T., Shiga, Y., Sakti, S., Neubig, G., Nakamura, S.: Parameter generation methods with rich context models for high-quality and flexible text-to-speech synthesis. IEEE J. Sel. Top. Sign. Process. **8**(2) (2014)

7. Kiruthiga, S., Krishnamoorthy, K.: Design issues in developing speech corpus for Indian languages—a survey. In: 2012 International Conference on Computer Communication and Informatics (ICCCI-2012), Jan. 10–12, 2012, Coimbatore, India
8. Bellur, A., Badri Narayan, K., Raghava Krishnan, K., Murthy, H.A.: Prosody modeling for syllable-based concatenative speech synthesis of Hindi and Tamil IEEE (2011)
9. Shreekanth, T., Udayashankara, V., Arun Kumar, C.: An unit selection based hindi text to speech synthesis system using syllable as a basic unit. *IOSR J. VLSI Sign. Process.* (IOSR-JVSP) (2014)
10. Kurian, A.S., Narayan, B., Nagarajan Madasamy, Bellur, A., Krishan, R.: Indian language screen reader and syllable based festival text to speech synthesis system. *Assoc. Comput. Linguist.* 63–72 (2011)
11. Saraswathi, S., Geetha, T.V.: Design of language models at various phases of Tamil speech recognition system. *Int. J. Eng. Sci. Technol.* (2010)
12. Patil, H.A., Patel, T.B., Shah, N.J., Sailor, H.B., Krishnan, R., Kasthuri, G.R., Nagarajan, T., Christina, L., Kumar, N., Raghavendra, V., Kishore, S.P., Prasanna, S.R.M., Adiga, N., Ranbir Singh, S., Anand, K., Kumar, P., Chandra Singh, B., Binil Kumar, S.L., Bhadran, T.G., Sajini, T., Saha, A., Basu, T., Sreenivasa Rao, K., Narendra, N.P., Sao, A.K., Kumar, R., Talukdar, P., Acharyaa, P., Chandra, S., Lata, S., Murthy, H.A.: A syllable-based framework for unit selection synthesis in 13 Indian languages. In: 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)
13. Panda, S.P., Nayak, A.K.: A waveform concatenation technique for text-to-speech synthesis. *Speech Technol.* (2017)
14. Black, A.W., Zen, H., Tokuda, K.: Statistical parametric speech synthesis. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
15. Geethaa, K., Vadivel, R.: Syllable segmentation of tamil speech signals using vowel onset point and spectral transition measure. *Autom. Control Comput. Sci.* (2018)
16. Gangashetty, S.V., Sekhar, C.C., Yegnanarayana, B.: Spotting multilingual consonant-vowel units of speech using neural network models. In: International Conference on Nonlinear Analyses and Algorithms for Speech Processing; Lect. Notes Comput. Sci. (2005)
17. Fu, M., Pan, W., Hu, W., Chen, S.: Syllable segmentation of pumi speech with GABPNN based on fractal dimension and DWPTMFCC. In: 2016 International Conference on Information System and Artificial Intelligence
18. Kishore, S., Black, A., Kumar, R., Sangal, R.: Experiments with unit selection speech databases for Indian languages. National seminar on Language Technology Tools: Implementation of Telugu, October 2003, Hyderabad, India
19. Kamakshi Prasad, V., Nagarajan, T., Murthy, H.A.: Automatic segmentation of continuous speech using minimum phase group delay functions. *Speech Communication* 2004

# Smart Farming Technology with AI & Block Chain: A Review



Deepali Jawale and Sandeep Malik

**Abstract** Agriculture being backbone for the survival of mankind needs proper facilities with great production to increase the economy of the farmers. In our proposed System details will be provided by farmer for analyzing details like soil type (fertility and nutrients), weather conditions, etc. which will help them analyze and Classify Crop and Provide which pesticide and fertilizer to be used and improve crop yield by providing information related to crop need. The system will also avoid data replication using Blockchain technique easing farmers to take better decisions regarding their farms or growth of crops. Having detailed information about soil, farmers can cultivate multiple crops in their farms proposed by system. The System will also give weather prediction to farmers helping them from indulging in any unwanted situation and have enough yields.

## 1 Introduction

In the World full of Technology where each field is digitized, agriculture is the sector which is promoted where minimum efforts yield maximum production. Even today some farmers lack basic facilities such awareness about soil quality, their maintenance such as right amount of fertilizers and pesticides using in the fields, growth of multiple crops using one soil structure, government introduced schemes, etc.

Rapid climatic changes have been witnessed during the last few years. The indication is in the form of an irregular monsoon in India. There are unpredictable levels of rainfall in different parts of the country and farmers are usually unaware. Detecting nutrient deficiencies will help to deliver correct fertilizers and reduce pesticide use

---

D. Jawale (✉) · S. Malik  
Oriental University, Indore, MP, India

S. Malik  
e-mail: [sandeepmalik@orientaluniversity.in](mailto:sandeepmalik@orientaluniversity.in)

as weaker crops are vulnerable to pests. Having Right amount of fertilizers at right place will lead to increased efficiency in plant growth.

The Farmers need proper guidelines related to their cultivation of crops in their fields. It can be done based on proper prediction based on classification accurate parameters and results derived from them.

In our System Based on data the system will suggest which type of crops can be cultivated in future to increase productivity. Based on classification of parameters taken results can be derived for farmers helping them to implement into their farms. The data related to crop will be stored on Blockchain to attain security, avoid redundancy, and predicting accurate results. System will predict the future Multi cropping pattern and also suggest which pesticides and fertilizer can be used for particular crop. The System will forecast the weather conditions to avoid any unwanted conditions for farmers.

## 2 Literature Review

Many Researchers presented a comprehensive framework of agricultural surveillance, implementation. This segment explains analysis of the literature.

Shirsath Rakesh et al. [1] the proposed decision support system is useful to help farmers pick a crop for cultivation mapping using different soil parameters such as soil type, soil PH-value, average weather needed, water consumption required, temperature range, etc. This method has been used to improve crop production by providing basic details and the crop list. Their user-friendly android application suggests most likely seasonal and soil-type matching crops for farmers so that they can grow more suitable crops and increase the production ratio.

Lambros Lambrinos et al. [2] they also developed a decision support system for precision/smart farming that uses data from various sources to provide pertinent data. More precisely, they combine data from a variety of sensors acquired through the LoRaWAN network with weather and crop information in order to make informed decisions that are currently focused primarily on the use of water and crop safety against adverse weather conditions, which are increasingly troubling farmers worldwide as a result of climate change.

Md Ashifuddin Mondal et al. [3] the goal of this analysis is to proffer an IoT-based smart farming framework for tackling adverse situations. Smart agriculture can be embraced and offers high precision crop manage, useful information mobilization, and computerized farming techniques. This work introduces an intelligent observing system which controls temperature and soil humidity for agricultural fields. On the basis of these, principles it takes necessary action after processing the sensed data without human intervention.

Konlakorn Wongpatikaseree et al. [4] The propound of this investigation is to elaborate a traceability process, to compile and present recognized data from a smart farm. Via the Internet of Things (IoT) was implemented using various sensors to examine environmental information in the smart farm. All outcome data is measured

and presented using a method of traceability. More information, especially regarding the quality of the planting process, is given to customers by scanning the Quick Response Code (QR) before purchasing an agricultural product.

Eko Ariawan et al. [5] this paper introduced a smart micro-farm model, based on the IoT concept, designed for information visualization, information acquisition, and information monitoring. To successfully build device it integrates microcontroller, sensors, and cloud services. Implementing this observation model in the field will greatly contribute to increasing the harvested algae yield. The technologies perform the primary role of connection among digital system and the physical world. Their analysis has focused primarily on IoT framework design for the smart micro-farm. The sensors used are capable of observing the environment and of supporting the automation process.

Aditi Shah et al. [6] This paper focuses on confirmation of macro-nutrients classification using picture processing and machine learning technique. This will give farmers and concept of crop safety and precautionary action taken.

Yemeserach Mekonnen et al. [7] it explains the complete design, dispatch, and functionality of the systems. The test bed contains distributed WSN which observes various parameters for agriculture and the environment. Wireless data transmission and acquisition of sensors are handled via ZigBee protocol via IoT gateway router.

Gumaste et al. [8] they proposed a smart farming approach that would use key technologies, such as FFT and genetic algorithms. Farmers must be recorded with an Internet connection, from the web or via the android mobile app. The cloud warehouse is used to store farmers' and knowledge about the environment. The current weather conditions are obtained from the farmer's position using Internet and GPS coordinates for coming weather forecasts.

Gandhi et al. [9] this paper provides an application of data mining technology in decision-making. System consist of, artificial neural networks, bayesian networks, and vector supporting machines. Techniques are used for providing the relationships between different climates and other crop production factors. This analysis indicates that further studies are needed to understand how these techniques can be used with complex agricultural data sets for crop yield prediction, using GIS technology to combine seasonal and spatial factors.

Martinez-Ojeda et al. [10] this work has therefore been undertaken to identify the problems of primer land cultivation and precision agriculture techniques. This study elaborates the innovation of a self-recommended crop program that tracks soil quality and provides a database list of chosen general crops. The device utilized various sensors that measured pH, soil moisture, soil temperature, and soil fertility.

Akshay Sankpal et al. [11] proper nutrient ratios for the crop to grow through the fertilizers are required. It should be understood which fertilizers are needed to be applied to the soil. Several researchers are working to classify the key nutrients of nitrogen (N), phosphorous (P), and potassium (K) in soil. This paper discusses sensing systems and other portable methods for assessing NPK soil nutrients.

Patil et al. [12] this paper analyzed and evaluated the integration of IOT technology with sensor technology and wireless networks on the basis of the current agricultural system scenario. The Remote Monitoring System (RMS) is proposed as

an Internet and wireless communication hybrid solution. The main objective is to collect data on the real-time agricultural production climate that allows quick access to agricultural facilities, such as short message service (SMS) alerts and advice on weather conditions, crops, etc.

Lihua Xu et al. [13] the predictive model of usable nitrogen (AN) and phosphorus (AP) content was developed using the partial least square (PLS) method based on the analysis of the laboratory reflectance spectrum of 118 soil samples. The findings suggested that after continuum removal and re-sampling, the association of soil nutrients (AN and AP) and soil spectral characteristics were enhanced.

Yoon et al. [14] communication modules for smart farming using low-power Bluetooth and Low Power Wide Area Networks (LPWAN) were built in this paper, including the wired communication network used in the actual farm. In addition, the system uses the MQ Telemetry Transport (MQTT) communication type, which is a dedicated IoT protocol, to incorporate monitoring and control functions, thus improving the possibility of developing agricultural IoT.

Pudumalar et al. [15] precision agriculture is a modern farming technique that uses research data on soil characteristics, soil types, crop yield data collection and suggests farmers the right crop based on specific parameters for their sites. This reduces the wrong choice on a crop and increase in productivity. In this paper, this problem is solved by proposing a recommendation system through an ensemble model with majority voting technique using Random tree, CHAID, K-Nearest Neighbor, and Naive Bayes as learners to recommend a crop for the site-specific parameters with high accuracy and efficiency.

Dabre et al. [16] the developed decision support model combines different environmental factors such as humidity, temperature, pest sound frequency, soil nutrition, and soil pH to formulate and evaluate individual requirements for crop watering. System will also provide warning and recommendation on individual crop requirements for water-soluble fertilizers and pesticides. Before prediction, soil condition can be dissected for pre-processing to decide the water retention ratio, soil nutrition ratio, so that system precision can be elevated.

### 3 Existing System

Currently as such no system is invented which helps in accurate prediction of crops based on soil in the farmers. The farmers are unaware about the soil fertility, how to have more productivity of the crops etc. The Soil in the farms needs proper nutrients and less chemicals and pesticides mixed soil. The goal of every farmer is to attain maximum production with minimum efforts and proper knowledge, as they are currently lacking. That crop is suitable for different soils and how multiple crops can be grown with the help of particular soil needs to be updated with the farmers? If the weather is inaccurate or is changing continuously then the farmers being unable able to take proper steps as he is not able to judge the weather forecast. Currently in market many new fertilizers and pesticides are introduced claiming to be beneficial

for the growth of crops, but lack of knowledge dupes the farmer and they incur heavy losses.

## 4 Proposed System

Soil, water, air, and plants are essential natural resources that help human beings produce food and fiber. We also preserve the habitats that ultimately depend on every life on Earth. Soil serves as a substrate for growing plants; a drain for food, water, and chemicals; a filter for nutrients; and a biological medium for waste breakdown.

In the Proposed System farmer will upload the soil details in the system for getting the prediction regarding the crops needed to be grown on that soil. Maybe multiple crops can be grown on a particular soil helping farmer in increased productivity. After uploading the details on system pre-processing is done. Next step is Feature Extraction where input data are categorized into set of features. These Features help in discriminating between various categories of input. After Feature Extraction on data system suggests soil suitable for cultivating multiple crops and the results are predicted to user.

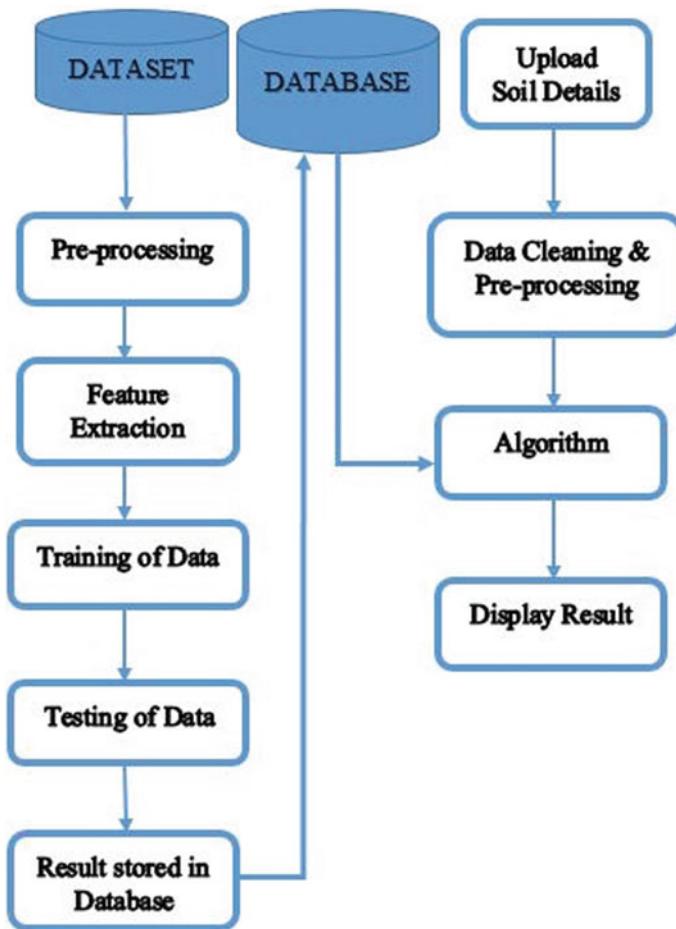
In the system real-time data will be taken as per the historical record, weather situation and the artificial intelligence algorithm the prediction related to crop cultivation will be made. Weather suggestions will also forecast the future of the crops. Based on the suggested crop further fertilizers and pesticides will be proposed and the system will generate a result for the user (Fig. 1).

## 5 Expected Result

We will evaluate the factors such as temperature, humidity, pesticides, minerals, and water level present in soil to have better growth of crops in soil. Analysing these factors help in taking appropriate steps for maintaining moisture of the soil. Fertilizers to be given in proper method and at proper time. Pesticides should be minimum as it will hamper the crops Based on these factors we will evaluate the soil and suggest which crops are best to be grown on soil.

## 6 Conclusion

We have proposed system details will be provided by farmer regarding the soil in his farm, weather details helping to develop and predict the type of crop suitable for farming. System will analyze and Classify Crop for recommending pesticide and fertilizer to be used for improving crop yield by providing information related to



**Fig. 1** Framework of proposed system

crop needs. The system will also provide an alternative for crop growth if case of failure of enough current yields.

We have proposed system details will be provided by farmer regarding the soil in his farm, weather details helping to develop and predict the type of crop suitable for farming. System will analyze and Classify Crop for recommending pesticide and fertilizer to be used for improving crop yield by providing information related to crop needs. The system will also provide an alternative for crop growth if case of failure of enough current yields.

## References

1. Shirasath, R., Khadke, N., More, D., Patil, P., Patil, H.: Agriculture Decision Support System using Data Mining, pp. 1–5. IEEE (2017)
2. Lambrinos, L.: Internet of Things in Agriculture: A Decision Support System for Precision Farming, pp. 889–892. IEEE (2019)
3. Ashifuddin Mondal, M., Rehena, Z.: IoT Based Intelligent Agriculture Field Monitoring System, pp. 625–629. IEEE (2018)
4. Wongpatikaseree, K., Kanka, P., Ratikan, A.: Developing Smart Farm and Traceability System for Agricultural Products using IoT Technology, pp. 180–184. IEEE (2018)
5. Ariawan, E., Makalew, A.S.: Smart Micro Farm: Sustainable Algae Spirulina Growth Monitoring System, pp. 587–591. IEEE (2018)
6. Shah, A., Gupta, P., Ajgar Y.M.: Macro-Nutrient Deficiency Identification in Plants Using Image Processing and Machine-Learning, pp. 1–4. IEEE (2018)
7. Mekonnen, Y., Burton, L., Sarwat, A., Bhansali, S.: IoT Sensor Network Approach for Smart Farming: An Application in Food, Energy and Water System, pp. 1–5. IEEE (2018)
8. Gumaste, S.S., Kadam, A.J.: Future Weather Prediction Using Genetic Algorithm and FFT for Smart Farming, pp. 1–6. IEEE (2016)
9. Gandhi, N., Armstrong, L.J.: A Review of the Application of Data Mining Techniques for Decision Making in Agriculture, pp. 1–6. IEEE (2016)
10. Martinez-Ojeda, C.O., Amado, T.M., Dela Cruz, J.C.: In Field Proximal Soil Sensing For Real Time Crop Recommendation Using Fuzzy Logic Model, pp. 1–5. IEEE (2019)
11. Sankpal, A., Warhadé, K.K.: Review of Optoelectronic Detection Methods for the Analysis of Soil Nutrients, Vol. 02, pp. 26–31 (2015)
12. Patil, K.A., Kale, N.R.: A Model for Smart Agriculture Using IoT, pp. 543–545. IEEE (2016)
13. Xu, L., Xie, D.: Prediction for Available Nitrogen and Available Phosphorus by Using Hyperspectral Data. IEEE (2016)
14. Yoon, C., Huh, M., Kang, S., Park, J., Lee, C.: Implement Smart farm with IoT Technology, pp. 749–752. IEEE (2018)
15. Pudumalar, S., Ramanujam, E., Rajashree, R.H., Kavya, C., Kiruthika, T., Nisha, J.: Crop Recommendation System for Precision Agriculture, pp. 32–36. IEEE (2017)
16. Dabre, K.R., Lopes, H.R., D'monte, S.S.: Intelligent Decision Support System for Smart Agriculture, pp. 1–6. IEEE (2018)

# Design and Development of Electronic System for Predicting Nutrient Deficiency in Plants



Amruta Chore and Dolly Thankachan

**Abstract** Plants require adequate nutrient content for a total as well as natural life cycle. Six macronutrients, such as nitrogen, calcium, phosphorus, potassium, sulfur, and magnesium are essential for the natural and healthy rise of plants. Regular activities with a lack of nutrients in plants lead to transportation difficulties and ultimately affect crop. Plants show a definite lack of nutrient on their leaves with notable differences in pattern. Our research suggested is to provide an automated and economically viable method for detecting defects nutritional conditions. Our system uses helpful information to forecast performance of crops. The dataset for deficient leaves and healthy leaves develop with the help of the RGB Color Extraction Analysis Technique, Disclosure of texture in real time, Identification of bottom edge, etc. This dataset will allow supervised machine learning to predict and identify accurate shortages of vitamins and healthy plants to prohibit growth rates.

## 1 Introduction

Plants required a proper mix of nutrients to reside, growth, and reproduction. It represents symptoms of being unhealthy when plants are malnourished. Two sources and micronutrients of plant resources fall into macronutrients. In relatively higher quantities, macronutrients are the elements needed. These include nitrogen, sodium, arsenic, calcium, magnesium, and phosphorus; Plants require the micronutrients in small amounts such as carbon, boron, manganese, zinc, copper, chlorine, and molybdenum. Macronutrients as well as microelements typically separate roots from the earth to get additional requirements plant roots need to obtain nutrients from the soils. Third, the soil must be proper enough to permit the roots to absorb nutrients and sustain it. Correcting ineffective methods of irrigation also reduce the symptoms of deficient nutrients. As Climat differences creates deficiency in various plants. [1] Fixing ineffective irrigation methods also eliminates symptoms of nutrient deficiencies. The temperature of the soil must decrease with a given range to ensure

---

A. Chore (✉) · D. Thankachan  
Oriental University, Indore, Madhya Pradesh 453555, India

the absorption of nutrients. The best combination of temperature, pH, and humidity varies for various species of plants. These nutrients are naturally exists in the soil but may not be accessible to plants. Information of soil pH, composition, and past can be very. Effective in determining which nutrients may be decreased. Phosphorus and copper are the only elements usually absent in soils in Arizona. Most of the others may be ignored in some situations but the drawbacks are quite unusual.

## 2 Literature Review

This section explains the literature review. For learning the present system, different papers have studied.

Susanto et al. [2] this paper found out nutrient content in wheat leaves by defining color types of leaves pictures taken on field with several lighting circumstances. They proposed the advancement of DSELM fusion and genetic algorithm (GA) to regularize plant images and to decrease color disparity produced due to sunlight intensity. In the picture segmentation, they applied the DSELM to distinguish wheat leaves from a dynamic background. Mean, variance, skewness, and kurtosis the 4 moments are takeout and used as forecasters in the nutrient approximation. The results have shown superior quality and processing speed.

Jin et al. [3] precise and high-performance extraction of phenotypic crop characteristics, as a key phase in molecular breeding, is of great significance in that production. Automatic stem-leaf segmentation, though, remains a major challenge as a requirement for certain correct extractions of phenotypic traits. Current research focuses on the analysis of 2-D image-based separation that is adaptive to illumination. With lively laser scanning and strong penetrating capabilities that pass through 2-D to 3-D phenotyping, precise 3-D information can be obtained through Light Detection and Ranging (LiDAR).

Noinongyao et al. [4] this paper suggested an image analysis approach to identify unusual regions that are induced by nutritional shortages on plant leaves. The suggested solution analyzes a histogram of normal leaf colors for the detection of irregularities on trees. This is divisible into three main acts. Firstly, the color characteristics of the leaf area are computed in an input image.

Hosseini et al. [5] presented design of picture deblurring in the appearance of one-time convolution filtering. Used Gaussian LPF to distinguish the image noise removal difficulty for image edge deblurring. Proposed an unsighted method to find the PSF statistics for 2 Gaussian and Laplacian model, planned for testing and authenticate the competency in given technique using 2054 originally blurred pictures across 6 imaging applications & 7 state-of-the-art de-convolution technique

Mustafa Merchant et al. [6] discussed as Indian national fruit, its leaves are enor-mance affected by a number of nutrient deficiencies such as nitrogen, phosphorus, potassium and copper. Mango leaves nutrients alter color. These leaves are consid-ered defective. This research has found the numerous nutrient deficiencies in mango

leaves. At the beginning, a data set is created by, obtain the various mango leave features.

Mitsugi et al. [7] suggested the consumption of plasma to eliminate soil-borne pathogens & worms as a method in least chemicals in farming. Ozone dispersion handling method used & real farming place for soil disinfection. By calculating the soil acidity and nitrogen nutrients, the ozone presence in soil measured. After that, a part of the field infected with the Streptomyces, taken along the ozone dispersion method. Then, radish seeds planted in the ozone area & control area. The result was radishes showed improved growth compared to the control & were not contaminated from outside.

Krithika et al. [8] the aim was to find diseases of the salad cucumber leaf at the first stage. The natural diseases existing in salad cucumber are Alternaria leaf blight, Bacterial wilt, Cucumber green mottle montage, Leaf Miner, Leaf spot, Cucumber Mosaic Virus (CMV) disease, etc. In this work, the use of K-means clustering, an unsupervised algorithm with Support Vector Machine (SVM) used to provide this problem.

Salazar et al. [9], this article provided an automatic system for understanding the root condition of avocado. This method uses k-means to divide leaves from identical backgrounds from pictures taken in ground under semicontrolled circumstances in s-v space at the super pixel level & a light neural network for classifying collected histograms from segmented plants into the following parts: Healthy, Fe insufficiency, Mg insufficiency, and red spider plague. The presented strategy divides the leaf from literature with an typical F-score of 0.98 and categorized the leaf state with a total correctness of 96.8%.

Chouhan et al. [10] the plant is essential for any living organism. Plants suffer from different kinds of diseases alike a human or other living thing. Such diseases are detrimental to crops, as they can influence the development of trees, seeds, fruits and leaves, etc. which can even cause the plant to die. BRBFNN method was designed to find and grouping of plant leaf diseases. The findings shown higher performance in diagnosing leaf.

Shaha et al. [11] intended that plants require sufficient nutritional content for a full and balanced lifecycle. Adequate amounts of six macronutrients such as Nitrogen, Calcium, Phosphorous, Potassium, Sulphur, and Magnesium are more essential for natural and balanced plant development. The lack of nutrients causes problems in plants' everyday operations and reduces the yield. Surfce based image processing, CNN is generally is used for finding a nutritional deficiency [12–14]. RNN fusion & genetic algorithm used to remove dificiency from Wheat plants and increase growth [15]. Also pattern recognition and color property analysis, picture analysis algorithm found important for different defective images [16, 17]. Venation Pattern also one of the technique to increase the plants vein growth [18]. Reducing nutrient in the medium shows nutritional absorption by the plant [19]. Biochemical bands also reduces the dificiency in plants [20].

### 3 Scope of Research

Nitrogen, phosphorus, and potassium are important and essential plant nutrients. Using the nutrient present in the herb, the machine can predict future crop output and give suggestions on how to enhance crop quality. Crop production is growing and has a direct impact on the farmer's economic life. Our system makes appropriate decisions on the quantity of pesticide depending on the deficiency of the element. This system works as a real-time framework to help farmers improve their crop production without any Collins. This approach is not restricted to any single weed, it also measures nutrients on any herb or fruit and it will be recommended accordingly. The various nutrient like magnesium, nitrogen, phosphorus, potassium, copper, etc. present in the crops, if any one falls down it will badly effect crops, so system helps to maintain all nutrient according to crops. It predicts the future market rate based on the previous record and estimates the future product risk.

#### 3.1 Objectives

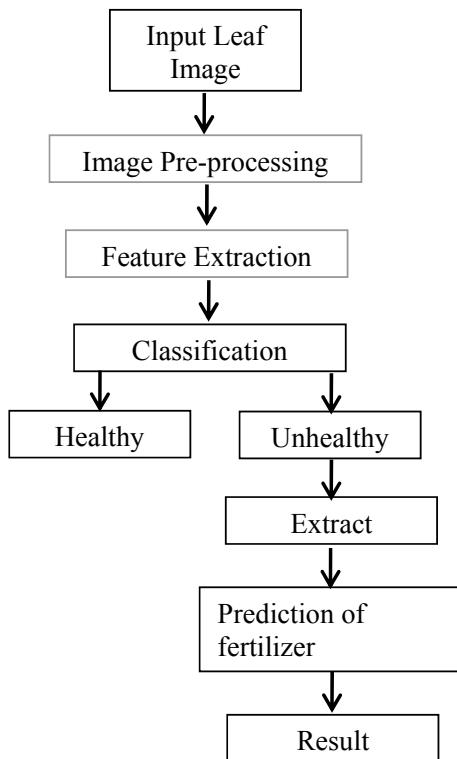
Following are the objective of the system

- Identify the nutrient in leaf with intensity and suggest how much amount of fertilizer required for crops.
- With the help of nutrient present in the leaf, system predicts the quality of future crops and gives suggestion on how to improve quality of crops.
- System provides the yield of crops using previous records and quality of crops. If crops production increases that directly, affect farmer's economical life.
- Make appropriate suggestion for quantity of pesticide based on nutrient intensity.
- System work as real-time application to help the farmer to increase their production of crops without any Collins.
- This system is not limited to any specific leaf only, it detects the nutrient on any plant or fruits leaf and give suggestion accordingly.
- System forecast the market rate of crops based on previous record with the help of machine learning approach.
- System predicts the future demanding product that has been demanding in future.

### 4 Proposed System

Figure 1 shows the Deficiency identification using machine learning approach and it is explained below:

**Fig. 1** Deficiency identification using machine learning



### Image Acquisition

- First, we have to shape the Supervised Machine Learning dataset. Violently 700 photos are needed for the healthy plant and six nutrient absences, around 100 for each.
- For every defective and stable leaf to take an image of the white background in usual brilliance, digital camera is used.

### Image Pre-processing

The image taken can contain some unnecessary noise or detail. Subtracting the context brings on the role of meaning. Noise is taken if present & the value portion, i.e., leaf, is improved for additional isolation & examination of deficiencies. By using Mean filtration to reduce noise and to provide a smooth picture. Mean filter eliminates abrupt pixel value shifts by substituting each pixel value with the nearest usual pixel value. This is center on the kernel that specifies the size and shape of the region to be verified. Amplitude is measured for Image Improvement using histogram equalization.

### Feature Extraction

Then the already processed image is taken to retrieve the first feature = extraction. The characteristics are red (12), green (G), blue (B), G/R, and GB band ratios. As contrast is firstly dominant on good leaf green is color. It also tests the average color spectrum of R, G, and B from 0 to 255.

### Edge Detection

If the value of the green color in the given input image is not dominant, a nutrient deficiency is likely in such cases, the area for error detection shall be the edge detection. Different edge detectors as Laplacian of Gaussian, Roberts, Prewitt, Sobel, Zero crossing, canny, etc. Roberts, Prewitt & Sobel used to discover derivative and Zero crossing, canny and LoG used to discover second-order derivatives. The gradient is the derivative of the 1<sup>st</sup> scale used to calculate changes in the amplitude of the signal gradient.

### Classification

In ML, classification is supervised learning procedure where the input is already known and the output depends on the output data. Classification is supervised learning procedure in ML in that data is already known and success is based on feedback from study, i.e., output is analyzed. We are using decision tree here for deficiency grouping. Picture will be pipe and the extraction method will be used. Such parameters will now be compared to the input dataset where the real parameters will fit the data set.

## 5 Expected Outcome

We will find all types of nutrient deficiencies in any type of plant or fruit leaf and provides the pesticides and fertilizers suggestions accordingly to have better crop production and to get good quality of crops. We will also provide the forecasting of market rate of crop, yield of crop, and future demanding product with the help of availability of previous records. Along with this system also provides the soil enrichment so that farmer can produce any crop in any region, which directly affects economical life of the farmer.

## 6 Conclusion

It is high time to focus on the highest yield to satisfy the increasing needs of the population. This can only happen if plants have enough space to grow. Plant nutrient quality is often overlooked although the value should be added. This paper emphasizes macronutrient recognition through image processing as well as machine learning methods. This will reduce farmer's work and give time to think on crop production. This will also be useful in vertical farms, where plants are given nutrient supplements.

## References

1. Makkar, T.Y.: A computer vision based comparative analysis of dual nutrients (Boron, Calcium) deficiency detection system for apple fruit. In: 4th International Conference on Computing Communication and Automation (ICCCA) (2018)
2. Susanto, B.S., Wu, D., Lok Woo, S.S. Dlay: Computational deep intelligence vision sensing for nutrient content estimation in agricultural automation. *IEEE Trans. Autom. Sci. Eng.* 1–15 (2018)
3. Jin, S., Su, Y., Wu, F., Pang, S.: Stem-leaf segmentation and phenotype trait extraction of individual maize using terrestrial LiDAR data. *IEEE Trans. Geosci. Remote Sensing* **57**(3) (2019)
4. Noinongyao, P., Watchareeruetai, U., Khatiwiriy, P., Waatanapiboonsuk, C.: Separation of abnormal regions on black gram leaves using image analysis. In: 14th International Joint Conference on Computer Science and Software Engineering (JCSSE) (2017)
5. Hossein, M.S., Plataniotis, K.N.: Convolutional deblurring for natural imaging. *IEEE Trans. Image Process.* (2019)
6. Merchant, M., Paradkar, V.D., Satish Khanna, M., Gokhale, S.: Mango leaf deficiency detection using digital image processing and machine learning. In: International Conference for Convergence in Technology, pp. 1–3 (2018)
7. Mitsugi, F.: Practical ozone disinfection of soil via surface barrier discharge to control scab diseases on radishes. *IEEE Trans. Plasma Sci.* (2019)
8. Krithika, P., Veni, S.: Leaf disease detection on cucumber leaves using multiclass support vector machine. In: IEEE International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), pp. 1276–1281 (2017)
9. Salazar-Reque, I.F., Pacheco, A., Rodriguez, R.Y., Lezama, J.G., Huamán, S.G.: An image processing method to automatically identify Avocado leaf state. In: Symposium on Signal Processing and Artificial Vision (STSIVA) (2019)
10. Chouhan, S.S., Koul, A., Pratap Singh, U., Jain, S.: Bacterial foraging optimization based Radial Basis Function Network (RBFNN) for identification and classification of plant leaf diseases: an automatic Approach towards Plant Pathology. *IEEE Access* (2017)
11. Shah, A., Gupta, P., Ajgar, Y.M.: Micro nutrient deficiency identification in plants using image processing and machine learning. In: 3rd International Conference for convergence in technology (2018)
12. Jae-Won, C., Trung, T.T., Thien, T.L.H., Geon-Soo, P., Van Dang, C., Jongwook, K.: A nutrient deficiency prediction method using deep learning on development of tomato fruits. In: International Conference on Fuzzy Theory and Its Applications iFUZZY (2018)
13. Watchareeruetai, U., Noinongyao, P.: Identification of plant nutrient deficiencies using convolutional neural networks. *IEECON 2018*, Krabi, Thailand (2018)
14. Han, K.A.M., Watchareeruetai, U.: Classification of nutrient deficiency in black gram using deep convolutional neural networks. In: 16th International Joint Conference on Computer Science and Software Engineering IEEE (2019)
15. Sulistyo, S.B., Woo, W.L., Dlay, S.S.: Regularized neural networks fusion and genetic algorithm based on-field nitrogen status estimation of wheat plants. *IEEE Trans. Ind. Inf.* (2017)
16. Latteet, M.V., Shidhal, S., Anami, B.S.: Multiple nutrient deficiency detection in paddy leaf images using color and pattern analysis. In: International Conference on Communication and Signal Processing, India, 6–8 Apr 2016
17. Wang, H., Li, G., Ma, Z., Li, X.: Image recognition of plant diseases based on backpropagation networks. In: 5th International Congress on Image and Signal Processing CISP (2012)
18. Arrasco, C., Khlebnikov-Núñez, S., Oncevay-Marcos, A., Beltran-Castanon, C.: Leaf vena-tion enhancing for texture feature extraction in a plant classification task. In: Latin American Conference on Computational Intelligence (LACCI), Guadalajara, Jalisco, Mexico, IEEE, 7–9 Nov 2018

19. Jiang, H., Ali, M.D.A., Jiao, Y., Dong, L.: In-situ, real-time monitoring of nutrient uptake on plant chip integrated with nutrient sensor. In: 19th International Conference on Solid-State Sensors, Actuators and Microsystems (2017)
20. Van Deventer, H., Cho, M.A., Mutanga, O., Naidoo, L., Dudeni-Tlhone, N.: Reducing leaf-level hyperspectral data to 22 components of biochemical and biophysical bands optimizes tree species discrimination. *IEEE J. Selected Topics Appl. Earth Observ. Remote Sensing* **8**(6) (2015)

# Classification of Hyperspectral Images with Various Spatial Features



Sandhya Shinde and Hemant Patidar

**Abstract** The remote sensing images cannot be captured via camera as the images are large in size. So we are introducing a system where the hyperspectral images can be captured through satellite can be identified and classified. The system will also identify the objects forbidden to enter in the restricted area. Our system can detect image from every angle and resolution. Our system can detect any object from earth observation (EO).

## 1 Introduction

In the remote sensing world, hyperspectral image (HSI) is often used to take full advantage of hundreds of distinct stream structures over an inclined picture. Hyperspectral picture requires accurate and vigorous recognition technique to acquire the picture properties. Hyperspectral picture was considered a particularly difficult issue because of the complicated image scene design (i.e., blended pixels, vast quantities of information and minimal practice testing), and so, in recent decades, many attempts have been made to tackle this problem.

The hyperspectral images can identify the images larger in size captured through satellite. HSIs are defined by hundreds of bands collected in adjacent spectral ranges and short spectrum intervals. They provide an extensive area of data for proper identification and classification of objects on the ground. In recent years, analyst has devoted significant attention to classifying HSIs for different applications.

A spatial background and spectral information should be incorporated to enhance classification efficiency. It was suggested that a variety of techniques could incorporate spatial characteristics into the HSI classification.

Detecting object in remote sensing images, in the area of aerial or satellite picture search, is a fundamental yet challenging problem, enhances a wide range of applications, and is becoming increasingly important in recent times. Deep learning has

---

S. Shinde (✉) · H. Patidar  
Oriental University, Indore, Madhya Pradesh 453555, India

made substantial advances in various techniques for image processing, and detection of objects using deep learning technique has currently attracted much research interest.

In order to overcome these problems, we are introducing an application-oriented system which classified all the objects detected through remote sensing imagery system. The system will first preprocess these images where images will be filtered and noise will be removed. The feature extraction process includes reducing the dimensions of the image. The next step will include the classification of images being categorized accordingly. The results will be displayed in the parameters of recall rate of the system, precision, accuracy, and confusion matrix of the picture.

## 2 Literature Survey

This segment gives an overview of various approaches that can be used for the classification of hyperspectral pictures as well as brief explanations of the algorithms used by researchers.

Tayeb Alipourfard et al. [1] using subspace dimension reduction techniques, CNN architectures have been suggested for the classification of HSI with high-quality characteristics. His benefit of the suggested framework is similar for both datasets, and the DR approach to preprocessing the traditional number of training pixels is appropriate for learning parameters. The proposed method has shown considerable classification efficiency by means of fixed architecture and a potent DR under the situation of small training sampling.

Shrish Bajpai et al. [2] components were extracted from decomposed signal or time series to boost the repairing of the initial signal with noise suppression that makes the pixel value smooth, which results in the process of classification which increases the parameters of precision.

Zhao Boya et al. [3], a multi-scale method of representation of features known as BUFPN was implemented to allow full use of the function by linking feature maps and combining lower-scale feature map and top-scale feature map into one. Additionally, distinct anchor outline is created on characteristic maps to clarify the obstacle of objects with various sizes and facet ratios.

Qishuo Gao et al. [4] for the HSI classification, a conservative smoothing algorithm is suggested, using the spatial accuracy of the adjacent pixels in the initial picture cube. The suggested technique uses the spectral similarities among pixels to allow weights to various neighboring pixels within a given local area in order to avert automatic over-smoothing. Under this scheme, it can be disclosed qualitative spatial details.

Lin He et al. [5], they developed a rapid DLRGF approach for the classification of HSI by remotely sensed spectral-space. The proposed approach provides an effective method for choosing spatial-spectral artifacts and provides important new

technology in how Gabor filtering is adapted to the specific characteristics of hyperspectral images, which involve spatial texture and spectral differential to support enhance the accuracy of classification.

Hao Huijun et al. [6], they suggested that the existing networks apply robust resolution features to small object prediction. They create new function of pyramid framework to predict the low-level culture that is efficient of complex tasks with vision. They also focus on the adoption and combination of current developments in deep learning technology, allowing for the re-formulation of the feature extraction portion based on the SSD paradigm that maximizes computational efficiency.

Licheng Jiao et al. [7], they propose a novel SAR image recovery method, one of the EO products, depending on semantic classification and calculation of regional similarities. They establish the SAR-oriented similarity measure IIRM depending on the IRM measure of classical similitude.

Ke Li et al. [8], they developed a new deep learning, coherent approach for remote image sensing detection of objects. The architecture for single object detection integrates the RPN's contextual functionality with the fusion network to identify the proposals. The suggested architecture provides compelling output of ten-class object detection data, which is accessible to the public.

Miaomiao Liang et al. [9], they suggest a new feature extraction technique for HSIC called deep spectral–spatial fusion (DMS3F2), which learns representative and discriminatory features by taking full benefit of the spectral–spatial information of HSI.

Lichao Mou et al. [10], they introduced a new end-to-end ConvDeconv residual network framework for unregulated spectral–spatial extraction of HSI.

Atif Mughees et al. [11], they have introduced a new SDBN approach to leverage spatial contextual knowledge through hyper-segmentation for efficient HSI classification. Every hyper-segment (HS)'s size and shape can be flexibly adapted to the actual spatial HSI structures, resulting in efficient use of spatial contexts. Then SDBN uses the multilayer DBN to efficiently exploit the spatial characteristics and spectral characteristics based on HS within and between segmented hyper segments.

Haodong Pan et al. [12] within the single-shot multibox detector (SSD) system present the adaptive dense pyramid network (ADFPNet) novel for object detection. The proposed network would identify historical artifacts by obtaining feature maps at different scales with complex multi-scale and responsive fields.

Swarnajyoti Patra et al. [13] a novel technique for classifying HSIs with small labeling samples is presented. The methodology suggested is broken down into two steps. Phase I generates the corresponding pattern by extracting spectral–spatial characteristics for each HSI pixel. The spectral–spatial patterns developed by EMPs are used as inputs with a specific number of labeled samples for the Phase II classification task.

### 3 Existing System

Currently, the images taken through camera are not adequate to capture larger images and identify the objects located in sea. The need to accurately identify the objects and classify them accordingly is important as sometimes suspicious objects cannot be traced down by radar.

### 4 Purposed System

In our purposed system, we define the process of our introduced system. The system will firstly upload the datasets and train them. The datasets are trained and uploaded into system. These datasets are then preprocessed where they are filtered and the noise from the image is removed using image denoising approach. Image denoising refers to the processing of a digital image polluted with noise. Image denoising is classified further into ICA filter and Gabor filter. Gabor filter is applied for texture analysis, which means that it originally analyzes whether there is any particular frequency material in the picture in particular directions in an area around the point or region of analysis and ICA filter has ICA components that are combined signals from multiple ECG channels, which means that they serve as spatial filters on the content (Fig. 1).

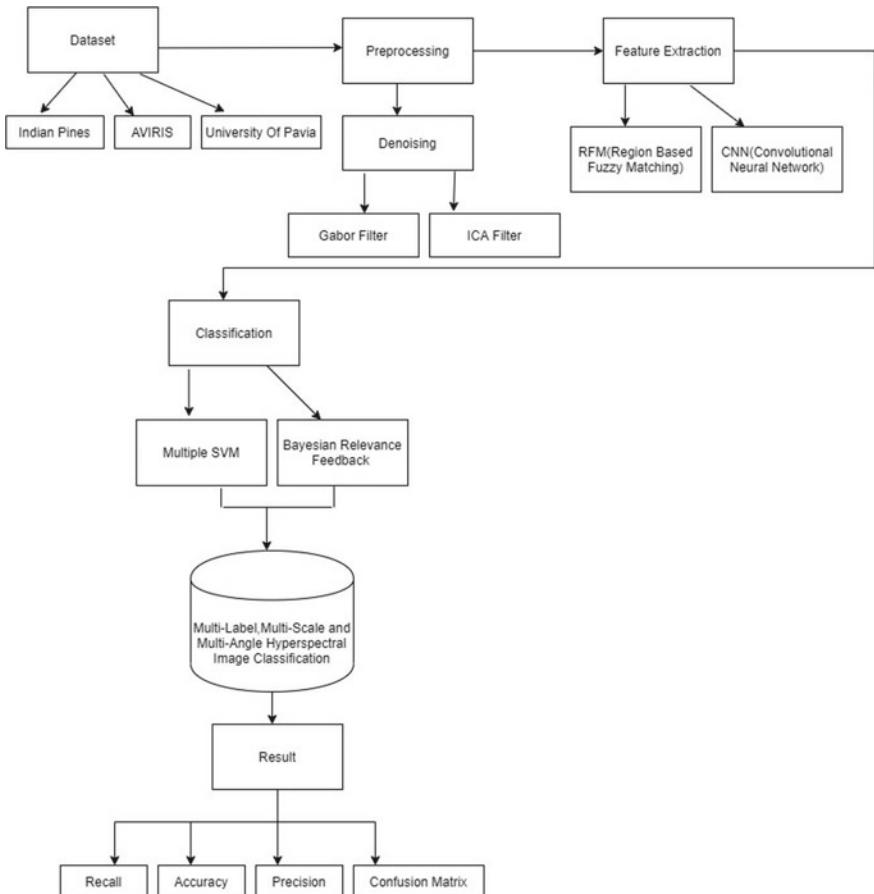
After preprocessing, the features of the image are extracted using feature extraction method. Feature extraction uses method like RFM and CNN for calculating similarity of two images. The next step is classification of SAR images. The classification is done by Bayesian relevance feedback and multiple SVM. The selected pictures are first scaled to different scales and then rotated in different angles so that, multi-scale-multi-angle dataset is formed of HIS/SAR images. The results will be displayed in the parameters of recall, accuracy, precision, and confusion matrix.

### 5 Algorithm

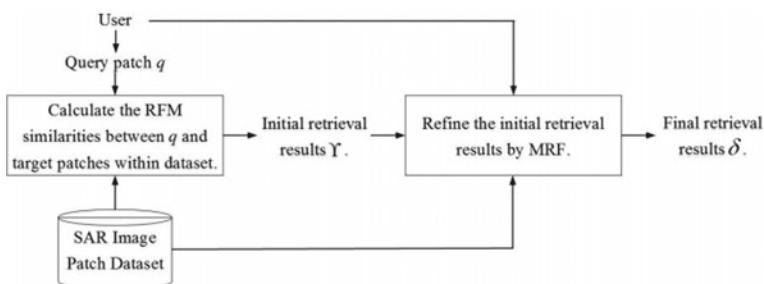
#### 1. Region-based fuzzy matching (RFM) measure

The IIRM divides them into various function spaces, i.e., brightness texture and edge feature spaces, to test the similarity among two I1 and I2 SAR image patches. First, IIRM divides a picture patch into a no of non-overlapping blocks to speed up segmentation for the feature space of the brightness texture. A block here indicates a fixed-size rectangle field, and the size is  $4 \times 4$  in the IIRM algorithm (Fig. 2).

Each block is defined by a four-dimensional vector  $[AA, AB, BA, XY]^T$ , where AA, AB and BA specify the energy of the one-level Daubechies 4 or hair wavelet transforms on the block, and XY specifies the mean gray value inside the block. The



**Fig. 1** Architecture of proposed work



**Fig. 2** Architecture of suggested content-based SAR image retrieval technique [14, 15]

adaptive k-means algorithm divides all the feature vectors into several groups, where every group correlates to a brightness texture field.

## 2. Multiple SVM Classification

The easiest classification problem SVM can address is the linearly separable binary classification. Instead of a set of training sampling,  $\{(x_i, y_i)\} N_i = 1, x_i \in R^d, y_i = \{-1, 1\}$ , SVM divides these sampling into two classes by the ideal hyper plane, running as wide as possible between the two classes with the gap to the nearest training sampling.

### Pseudocode

1. a do for all labels Class
2. Y → Pictures in class a
3. Sets Picture level labels in Y to 1
4. For all labels in Class c,  $a \neq c$  do
5. R → Select 2 k images randomly from Class c
6. Set image class marks in R to -1
7. Y + = R
8. end for
9. Build Si using the images in Y for classification of images
10. end for
11. RETURN All Si s

## 6 Bayesian Relevance Feedback

We set up a structure depending on Bayesian posterior probability which is used for relevant feedback. For the function of each class, the multivariate normal distribution is a very common and practical model. Let me be the collection of any and all images in the database and we'll say  $x$  is a random vector with dimensions  $d$ . Relating normal distribution density, multivariate by:

$$p(x) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left[ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right] \quad (1)$$

where  $\mu$  is the mean vector and  $\Sigma$  is the matrix of covariance.

The consumer has approved n images with function vectors in a given loop ( $\times 1, \dots, \times d$ ). We will measure  $\mu$  and  $\Sigma$  using the samples chosen by the consumer using the corresponding maximum likelihood estimators.

## 7 Expected Results

In our system, we will conduct research on the SAR images which will help in identifying and classifying the images. If some unknown objects enter within the region, it will also be captured in our radar system. For achieving these results, we are implementing multiple SVM and Bayesian relevance feedback technique.

## 8 Conclusion

We proposed an application-oriented method in this paper to identify the SAR images. The object detection system integrates input from area-based fuzzy matching test, SVM, and Bayesian relevance to excerpt the characteristics of the HSIs. The proposed methodology would yield more reliable results than prior techniques. The suggested network, where multi-scale extraction is used to acquire both the multi-scale spatial and spectral characteristics, is allocated both spatial information and relevant neighboring bands simultaneously.

## References

1. Alipourfard, T., Arefi, H., Mahmoudi, S.: A novel deep learning framework by combination of subspace-based feature extraction and convolutional neural networks for hyperspectral images classification. In: IGARSS 2018, IEEE International Geoscience and Remote Sensing Symposium, Valencia, pp. 4780–4783 (2018)
2. Bajpai, S., Singh, H.V., Kidwai, N.R.: Feature extraction & classification of hyperspectral images using singular spectrum analysis & multinomial logistic regression classifiers. In: 2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT), Aligarh, pp. 97–100 (2017)
3. Boya, Z., Baojun, Z., Linbo, T., Chen, W.: Multi-scale object detection by bottom-up feature pyramid network. In: IEEE, vol. 2019, no. 21, pp. 7480–7483 (2019)
4. Gao, Q., Lim, S., Jia, X.: Spectral–spatial hyperspectral image classification using a multiscale conservative smoothing scheme and adaptive sparse representation. Trans Geosci Remote Sens **57**(10), 7718–7730 (2019)
5. He, L., Li, J., Plaza, A., Li, Y.: Discriminative Low-Rank Gabor Filtering for Spectral-Spatial Hyperspectral Image Classification. Trans Geosci Remote Sens **55**(3), 1381–1395 (2017)
6. Huijun, H., Ronghua, Y., Zhongyu, C., Zhonglong, Z.: Multi-scale pyramid feature maps for object detection. In: 2017 16th International Symposium on Distributed Computing and Applications to Business, Engineering and Science (DCABES), Anyang, pp. 237–240 (2017)
7. Jiao, L., Tang, X., Hou, B., Wang, S.: SAR Images Retrieval Based on Semantic Classification and Region-Based Similarity Measure for Earth Observation. J Sel Top Appl Earth Obs Remote Sens **8**(8), 3876–3891 (2015)
8. Li, K., Cheng, G., Bu, S., You, X.: Rotation-insensitive and context-augmented object detection in remote sensing images. Trans Geosci Remote Sens **56**(4), 2337–2348 (2018). <https://doi.org/10.1109/TGRS.2017.2778300>

9. Liang, M., Jiao, L., Yang, S., Liu, F., Hou, B., Chen, H.: Deep multiscale spectral-spatial feature fusion for hyperspectral images classification. *J Sel Top Appl Earth Obs Remote Sens* **11**(8), 2911–2924 (2018)
10. Mou, L., Ghamisi, P., Zhu, X.: Fully conv-deconv network for unsupervised spectral-spatial feature extraction of hyperspectral imagery via residual learning. In: IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, pp. 5181–5184 (2017)
11. Mughees, A., Ali, A., Tao, L.: Hyperspectral image classification via shape-adaptive deep learning. In: IEEE International Conference on Image Processing (ICIP), Beijing, pp. 375–379 (2017)
12. Pan, H., Chen, G., Jiang, J.: Adaptively dense feature pyramid network for object detection. *IEEE Access* **7**, 81132–81144 (2019)
13. Patra, S., Bhardwaj, K., Bruzzone, L.: A Spectral-spatial multicriteria active learning technique for hyperspectral image classification. *J Sel Top Appl Earth Obs Remote Sens* **10**(12), 5213–5227 (2017)
14. Sun, X., Zhang, L., Yang, H., Wu, T., Cen, Y., Guo, Y.: Enhancement of Spectral Resolution for Remotely Sensed Multispectral Image. *J Sel Top Appl Earth Obs Remote Sens* **8**(5), 2198–2211 (2015)
15. Tang, X., Jiao, L., Emery, WJ.: SAR Image Content Retrieval Based on Fuzzy Similarity and Relevance Feedback. *J Sel Top Appl Earth Obs Remote Sens* **10**(5), 1824–1842 (2017)

# Detecting and Classifying Various Diseases in Plants



Rashmi Deshpande and Hemant Patidar

**Abstract** In today's era where production of crops is increasing day-by-day, it is also seen that harmful disease is also growing increasingly. Some of the disease being unidentifiable is causing more harm to the production of crops. Such harm not only causes harms crops but also farmers economical income is affected. To overcome all these problems, we are providing an overall solution in which all the problems related to farming which are faced by farmer. In the proposed system, it will identify the leaf disease harming crop and providing solution along with precautions to be taken. The system will not be limited to a particular crop leaf rather than it will detect leaf disease of all crops. The system will also recommend farmer quantity of pesticides to be used for crops. It will help to increase the production of crops decreasing their economic loss.

## 1 Introduction

Nowadays, a new idea of smart farming has been applied where field conditions are controlled and tracked using the self-operating systems. The disease's self-recognition is focused on recognizing the signs of illness. So that knowledge about the incidence of the disease could be given to the farmers quickly and accurately. It helps in reducing monitoring of huge fields by farmers. The disease may vary from crop to crop which includes change in shape, size, color, or texture. The emergence of the disease on the plant will develop in substantial loss agricultural product. Awareness for disease of leaf is essential for decision-making in crop situations and decision-making regarding disease managing.

A plant disease is a situation caused by an infectious agent or factor to the ecosystem. Fungi, bacteria, and viruses are among the species that can cause this disease. Parasites such as insects and mites can also be found eating the plant tissues resulting in diseased sections.

---

R. Deshpande (✉) · H. Patidar  
Oriental University, Indore, India

Nevertheless, diseases are important cause of agronomic reduction in India. Farmers face multiple problems controlling crop diseases. The identification of the illness is crucial part in field of agriculture, and this requires cautious analysis and careful monitoring to avoid the heavy losses. The disease cannot be passed from a damaged plant to the uninfected plant. A plant infection is a shift occurring for the worse of the plant's normal functioning, or the portion that may be affected by the disease that causes factors such as bacteria, fungi, nematodes viruses. Some of the plant illness can be an act of spreading from infected plant to the unaffected plant over broader parts of the plant.

In this system, we are suggesting a system which will not only identify the leaf disease but will also recommend precautions regarding it. It will suggest the quantity of pesticides to the farmers to be utilized for the crops. The proposed system will identify the intensity of the leaf disease for increasing the production of crops. It will also help in increasing crop production to reduce the losses for farmers. The will act as real-time application for helping farmers to increase their production of crops without any Collins. The system does not restrict itself for a particular crop leaf; it can detect all types' leaf disease for all the plants and recommend it precautions.

## 2 Literature Review

Several researchers had carried out research in this area. The corresponding literature review of the proposed study follows:

R. Anand et al. [1] this work sets out a system for detecting plant leaf disease and a cautious disease detection strategy. The proposed research aims at diagnosing brinjal leaf disease by means of artificial neural techniques and image processing. Brinjal diseases are a critical issue which results in a sharp decrease in the production of brinjal.

L. Shanmugam et al. [2] this paper describes automatic disease detection by means of remote sensing images. Because of various crop diseases, farmers face losses. Their system has two phases: first part handles training of safe and diseased data sets. The second phase handles crop observing and illness identification; and immediately intimates farmers with an initial warning message.

S. D. Khirade et al. [3] this paper examined the methods used to identify plant diseases using photographs of their leaves. This paper also addressed a certain segmentation algorithm and extraction function used in the detection of plant disease.

C. G. Dhaware et al. [4] this paper discusses picture preprocessing methods, image segmentation algorithms used for computerized recognition and works on different plant leaf illness classification algorithms that can be used for the classification of leaf disease.

R. P. Narmadha et al. [5] this paper aims at explaining paddy diseases. Some of the paddy disease involves brown spot disease (BPD), narrow brown spot disease (NBSD) that impedes expansion and health of paddy. Such technique was designed to

automatically eradicate noise, human error, and also reduce the time taken to weight the effect of paddy leaf illness as well as increase precision.

V. Pooja et al. [6] this paper introduces a technique for analyzing and arranging diseases with assistance of picture processing and machine learning equipment. First identification and capture of contaminated area is carried out, and the latter image processing is performed.

A. Devaraj et al. [7] this system is intended to create a reaction to a software system that regularly recognizes and classifies disease. Disease detection includes the step such as preprocessing, segmentation, extraction, and classification. The leaf photographs are taken to classify the diseases of plants.

R. M. Prakash et al. [8] in this paper, the image processing methods are mentioned for classify plant leaves illness. The aim of this paper is to integrate picture analysis and classification methods for identifying and classifying leaf illness.

P. Revathi et al. [9] this suggested work is planted on the segmentation method for picture edge revelation in which the captured pictures are first prepared for enhancement. In order to get specific regions (disease spots), the picture segmentation of the (R, G, B) color function is then performed. Later, pictures features such as border, form, light and texture are obtained to identify illness spots and monitor recommendations for plagues.

Bharat Mishra et al. [10] this paper presents a survey using image processing method on various leaf disease detection technologies and classifies them related on the type of investigation tool and appliance. In addition, the prevalent techniques used in the leaf illness detection process are studied seriously and addressed briefly; compression and presentation of accessible approaches are inspected.

N. Agrawal et al. [11] this paper develops multiclass SVM algorithms for classification of grape leaf diseases present in grapes plants. Uses resizing, enhancing, and smoothing for image preprocessing of the given system.

C. S. Sumathi et al. [12] based on their photographs, they introduced an automated system for identifying plant leaf. Plant classification depends on recognizing leaves which are commonly used in medicine and agriculture.

S. G. Wu et al. [13] in this paper, they use probabilistic neural network (PNN) with picture and data processing techniques to apply a general direction automatic leaf recognition for the classification of plants.

V. P. Gaikwad et al. [14] the goal of this paper is to establish software clarification that will automatically inspect and determine plant disease. This comprises four stages: first stage image acquisition, second step is image preprocessing, third step is image segmentation, and fourth step is color, form, and size extraction. Here, they used neural network-based classifier.

Shivani K. Tichkule et al. [15] this paper gives an overview of how various plant diseases are detected using methods of image processing. The use of image processing methods for detecting and recognizing illness is useful in agricultural applications.

Pranjali B. et al. [16] the goal of this paper is to purpose a new progress to exposure of grape leaf illness using image processing to reduce the damage and

optimize its computerized benefits. In this context, classification is rendered using various classifications of SVM and the ANN.

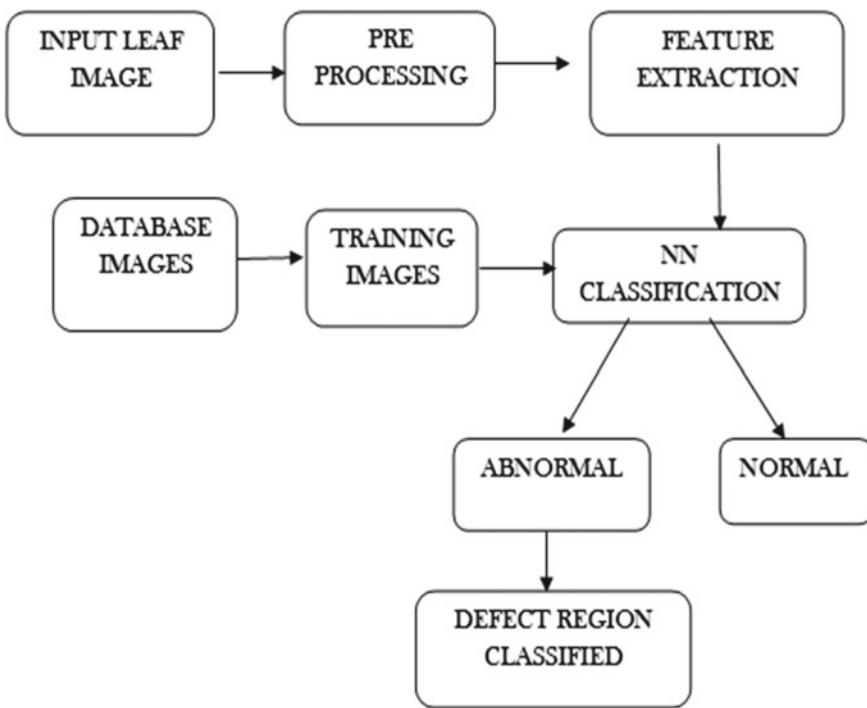
### 3 Existing System

Currently most of the farmers are using old methods for cultivating crops in field. Many crops get destroyed due to improper pesticides mixed in it which farmers cannot identify them. Sometimes the leaf diseases are not detected by the farmers reducing the production of crops. Due to decrease in crop production, the farmers face huge losses economically. Even they are unaware about the quantity of pesticides to be used for the healthy production of crops. The farmers do not have any real-time application to monitor and suggest farmers to increase their production of crops. The farmers cannot detect leaf disease with naked eyes and take precautions accordingly.

### 4 Proposed System

In the proposed system, the user will perform the process for identification of infected leaf disease. The user will give input leaf image to the structure. The input picture will be preprocessed further (Fig. 1).

With the help of feature extraction, the image functions are extracted after preprocessing. Application plays a very important part in the processing of images. Techniques for extracting features are applied to obtain features that will be useful in image detection and recognition. The extraction of features is the method of extracting the most important data from the raw data. Extraction of the function is to find the set of parameters that specifically and uniquely describe a character's form. In the extraction process of features, each character is represented by a vector, which becomes its identity. The main goal of extracting function is to remove a set of features, optimizing the identification rate with the least amount of elements and producing similar feature set for the same symbol's variety of instances. After feature extraction classification, images are done using NN technique. While using NN technique the leaf is classified into normal leaf and infected leaf. After detecting the infected leaf, it also classifies the region which is infected in leaf. Identification not only helps in disease detection but system will also recommend the precautions to be taken while curing the infected part of the leaf. The proposed system will be limited to only particular crop leaf rather than it will detect all types' leaves and their disease. The system will be working as real-time application for helping farmers to increase the production of their crops without collagens. The system will also give appropriate suggestion for using right amount of pesticides in field. The system will help famers in increasing crop production to improve their economic condition.



**Fig. 1** Proposed system architecture

#### **4.1 Steps Required for Proposed System**

##### **4.1.1 Image Acquisition**

At each phase of disease evolution, the color picture of the infected leaf was captured with camera. For image acquisition, a color camera was used. Pictures are stored in.jpeg format. In the present method, leaves damaged by five types of illness bacterial leaf spot, Cercospora leaf spot, leaf curl, plant mosaic, powdery mildew were used for identifying and classification.

##### **4.1.2 Image Preprocessing**

All the leaf images were preprocessed which mainly includes resizing and filtering of images. All leaf images were resized into  $256 \times 256$  pixels. The color transformation structure was obtained by converting the RGB into HIS color space. Hue–saturation–intensity (HSI) space is also a common color space, as it is focused on the perception of human color.

#### 4.1.3 Image Segmentation

“A process that splits the image into meaningful regions” is known as segmentation. It gives us regions and objects. Segmentation has two classes: foreground and background. There are many segmentation techniques which are useful for finding region of interest (ROI). Out of them K-means clustering is choosing for segmentation of disease-affected image.

#### 4.1.4 Feature Extraction

After segmentation, specific features are extracted from the infected area to classify the region infected with the disease. Texture, color, and shape-based features are typically based on definition of infected regions.

##### 4.1.5 Shape-Based

Form is an important visual attribute and one of the basic characteristics for defining image content. Description of the form content cannot be described precisely because it is difficult to measure the similarity between shapes.

##### 4.1.6 Color-Based

One of the most commonly used visual features in the classification of images is the color function. There are many benefits to the images distinguished by color effects.

##### 4.1.7 Texture-Based

Texture can be defined by texel and structure intensity, color. Texture characteristics are determined in statistical texture analysis from the statistical distribution of detected sequence of intensities at defined positions relative to each other in the picture.

First order (one pixel)—mean, variance, skewness, kurtosis)

Second order(pair of pixels)—angular second moment, contrast, correlation, homogeneity, entropy).

Higher order (two or more pixel)—values occur at specific locations relative to each other.

## 5 Algorithms

### 1. GLCM

A GLCM is a matrix where the number of rows and columns in the picture is equal to the number of gray levels, G. The matrix variable  $P(i, j, d)$  is the relative frequency with which two pixels, separated by distance  $d$  and in the direction defined by the particular angle (nearly), one with intensity  $I$  and the other with intensity  $j$ . In this matrix, we preserve symptoms of disease by measuring from the literature review the function values of correlation, heat, homogeneity, contrast, and entropy.

## 2. K-Means Cluster Algorithm

K-means is one of the clearly unattended learning algorithms that solve the well-known clustering problem. The method follows a simple and easy way of classifying a given set of data through several clusters (assuming  $k$  clusters).

## 6 Expected Results

We will evaluate the system and identify the leaf disease with its intensity of the disease. The system will help to evaluate and increase the production of crops reducing the losses of farmers. The system will suggest the quantity of pesticides for crops based on the disease intensity of crops. The system will also work as real-time application helping farmer to increase their product without any Collins. The proposed system will not be limited to a particular crop leaf but will detect any crop leaf disease.

## 7 Conclusion

The proposed system has overall solution proposed for farmer giving most benefit required for increasing crop production and decreasing losses incurred during farming. The farmers will be getting a real-time system implemented for the proper production of crops and benefiting the farmer. It will also recommend farmers to use appropriate pesticides for infected leaves. The system is not limited to a particular leaf detection rather it can detect all types of leaf disease and suggests the precaution accordingly.

## References

1. Anand, R., Veni, S., Aravindh, J.: An application of image processing techniques for detection of diseases on brinjal leaves using K-means clustering method, pp. 1–6 (2016)
2. Shanmugam, L., Adline, A.L.A., Aishwarya, N., Krithika, G.: Disease detection in crops using remote sensing images, pp. 112–115 (2017)
3. Khirade, S.D., Patil, A.B.: Plant disease detection using image processing, pp. 768–771 (2015)

4. Dhaware, C.G., Wanjale, K.H.: A modern approach for plant leaf disease classification which depends on leaf image processing, pp. 1–4 (2017)
5. Narmadha, R.P., Arulvadiu, G.: Detection and measurement of paddy leaf disease symptoms using image processing, pp. 1–4 (2017)
6. Pooja, V., Das, R., Kanchana, V.: Identification of plant diseases using image processing techniques, pp. 130–133 (2017)
7. Devaraj, A., Rathan, K., Jaahnavi, S., Indira, K.: Identification of Plant Disease using Image Processing Technique, pp. 749–753 (2019)
8. Prakash, R.M., Saraswathy, G.P., Ramalakshmi, G., Mangaleswari, K.H., Kaviya, T.: Detection of leaf diseases and classification using digital image processing, pp. 1–4 (2017)
9. Revathi, P., Hemalatha, M.: Advance computing enrichment evaluation of cotton leaf spot disease detection using Image Edge detection, pp. 1–5 (2012)
10. Mishra, B., Lambertm M., Nema, S., Nema, S.: Recent technologies of leaf disease detection using image processing approach—a review (2017)
11. Agrawal, N., Singhai, J., Agarwal, D.K.: Grape leaf disease detection and classification using multi-class support vector machine, pp. 238–244 (2017)
12. Sumathi, C.S., Senthil Kumar, A.V.: Neural network based plant identification using leaf characteristics Fusion. *Int. J. Comput. Appl.* **89**(5), 31–35 (2013)
13. Wu, S.G., Bao, F.S., Xu, E.Y., Wang, Y., Chang, Y., Xiang, Q.: A leaf recognition algorithm for plant classification using probabilistic neural network, pp. 11–16 (2007)
14. Gaikwad, V.P., Musande, V.: Wheat disease detection using image processing, pp. 110–112 (2017)
15. Shivan, K., Gawali, H.: Plant diseases detection using image processing techniques (2016)
16. Padol, P.B., Sawant, S.D.: Fusion classification technique used to detect downy and powdery Mildew grape leaf diseases, pp. 298–301 (2016)

# Offline Handwritten Dogra Script Recognition Using Convolutional Neural Network



Reya Sharma, Baijnath Kaushik, and Naveen Kumar Gondhi

**Abstract** The handwritten optical character recognition is a challenging and active branch of pattern recognition. The recognition of Indian scripts is potentially a complex problem due to several challenging issues like complex shapes of characters, similar shaped characters, positioning of diacritics, dots, etc. This paper focuses on the recognition of handwritten Dogra script. This paper also contributes to the construction of a handwritten Dogra script dataset. The handwritten document images are manually collected from several sources like archival departments, libraries, and museums and then pre-processed and segmented in order to obtain the final handwritten character dataset. The resultant handwritten Dogra dataset is evaluated using a convolutional neural network classifier, reporting a promising recognition accuracy of 88.95%.

## 1 Introduction

Optical character recognition (OCR) is one of the most challenging and fascinating area of pattern recognition, artificial intelligence and image processing. It is the study of detecting segments and identifying characters from the input image and transforming them into corresponding machine understandable format. This field immensely contributes toward the advancement of man–machine interface. It has huge variety of applications like zip code reading, finance, taxation, number plate recognition, education, bank check processing and many more related fields [1].

---

R. Sharma (✉) · B. Kaushik · N. K. Gondhi  
Shri Mata Vaishno Devi University, Katra, J&K, India  
e-mail: [sharmareya32@gmail.com](mailto:sharmareya32@gmail.com)

B. Kaushik  
e-mail: [baijnath.kaushik@smvdu.ac.in](mailto:baijnath.kaushik@smvdu.ac.in)

N. K. Gondhi  
e-mail: [naveen.gondhi@smvdu.ac.in](mailto:naveen.gondhi@smvdu.ac.in)

The handwritten text recognition problem is considered potentially more challenging as compared to the recognition of printed text. The primary challenge that contributes toward the difficulty handwriting recognition is the variation in writer specific preferences while drawing and joining the characters. Another factor contributing to this difficulty is the morphological complexity of cursive scripts [2]. While a huge amount of research has been done on Latin, Roman and Chinese script, handwritten cursive Indian script recognition is still at its infancy.

From the recognition point of view, a significant research work has been reported on Indian scripts like Bangla and Devanagari. But no research work has ever been reported on Dogra script till date. So, this paper is a first attempt in this direction, and it also provides a standard handwritten Dogra script dataset which can be used to obtain benchmark results for further research in this field.

The handwritten character recognition process is further classified into following phases [3]. The first phase of character recognition deals with the acquisition of input images; after this, we have the pre-processing phase which is further followed by the segmentation phase. Then the next phase is the feature extraction phase, and finally, the last phase of character recognition is the classification or recognition phase which may be further followed by the post-processing phase.

In this work, the convolutional neural network (CNN) is used for the recognition and classification of handwritten Dogra script. CNN has huge number of applications [4, 5] in the field of image and computer vision like image classification, natural language processing, object detection, face, character, speech recognition, etc.

The remainder of this paper is structured as follows: Section 2 describes the significant work reported on handwritten Indian scripts. Section 3 gives the detailed description of the dataset and its construction including the pre-processing and segmentation stage. Section 4 explains the methodology involved in the recognition of handwritten characters. Section 5 provides the experimental analysis and evaluations of results. Finally, Sect. 6 concludes the research work and gives the future directions for further research in this field.

## 2 Related Work

This section describes some of significant work reported on handwritten Indian scripts. Desai [6] has presented a feed-forward neural network technique for the recognition of handwritten Gujarati numerals. In this work, four profile features are obtained and fed as input to the classifier. Structural features are used along with a back-propagation network for the classification of handwritten Gurumukhi characters [7].

Zoning technique was used by Sharma and Jhajj [8] for the classification of handwritten Gurumukhi characters. In this work, support vector machine (SVM) classifier and KNN classifier are used for the recognition purpose. It is found that better results are obtained by using SVM classifier along with the polynomial kernel. Roy et al. [9]

have proposed a new deep CNN model for the recognition and identification of handwritten Bangla compound characters. A significantly high classification accuracy of 90.33% is obtained by deep CNN architecture.

Shelke et al. [10] have presented a fuzzy-based recognition system for the identification of handwritten Devanagari characters. In this paper, a two-stage recognition scheme is designed, in which the first stage has fuzzy inference model and the second stage uses the structural parameters. A feed-forward neural network is also used for the final recognition stages, and it provides a recognition accuracy of 96.95%.

KNN classifier is used by Sahare et al. in [11] for the classification of Indian document images. In this work, character segmentation technique is proposed and validated using SVM classifier. Geometric features are extracted for the identification of characters. These obtained geometric features are then finally fed to the KNN classifier for the final task of recognition. This combination provides a good recognition accuracy in comparison to other state-of-the-art results.

Raj et al. [12] have proposed structural and statistical feature extraction methods for the recognition of handwritten Tamil characters. The structural and statistical features are represented by PM quad tree in a hierarchical way. The obtained features are then classified or recognized with the help of SVM classifier.

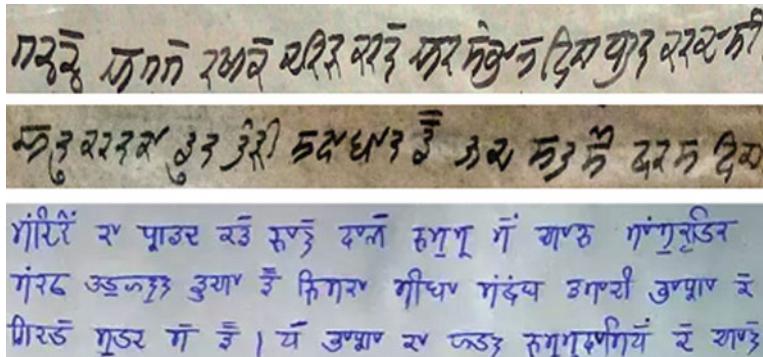
### 3 Dataset

The Dogra script is a regional script of India, historically related with Dogri language. It is a Brahmi-based alpha-syllabary, derived from Takri script [13]. The Dogra script has 10 independent vowels, 10 dependent vowels and 34 consonants, which are used in the present work. A significant contribution of this work is the development handwritten Dogra script database. The handwritten Dogra script documents are scanned at 200 dpi. The digitized documents are then pre-processed and segmented in order to obtain the individual character set database for the final recognition.

#### 3.1 Acquisition of Input Image

This step performs the transformation of machine printed or handwritten documents into digitalized form with the help of electronic devices such as scanners, cameras or tablets.

The Dogra script was the official script Jammu and Kashmir state during the regime of Maharaja Ranbir Singh. The Dogra script was used for several administrative purposes as well as upon postcards, stamp papers, postage stamps, currency notes and also for literary activities. Therefore, huge amount of data is available for Dogra script in the form of handwritten ancient manuscripts, temple inscriptions, etc. Figure 1 depicts some samples extracted from handwritten Dogra documents. So, we have surveyed several archival departments, libraries and museums of J&K



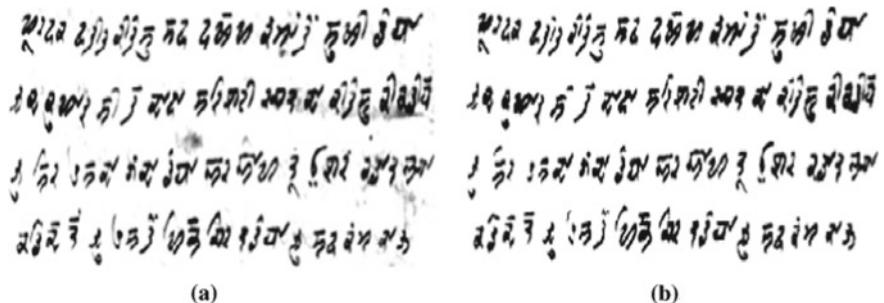
**Fig. 1** Samples extracted from Dogra documents

in order to collect handwritten documents. In this work, the handwritten documents are scanned and digitized at 200 dpi.

### 3.2 Pre-Processing

The pre-processing phase plays a significant role in the recognition of handwritten characters. This crucial step includes several operations like binarization, noise removal, smoothing, base line detection, skew detection, filtering, skeletonization or thinning, etc. [14].

The handwritten document images are firstly binarized by converting input rgb text images into gray-scale images and finally into corresponding binary images. An appropriate thresholding is used for obtaining the binary images (with level 1 representing background and level 0 representing foreground). The proper threshold value is selected by using the traditional Otsu algorithm [15]. Figure 2a, b represents the pre-processing phase.

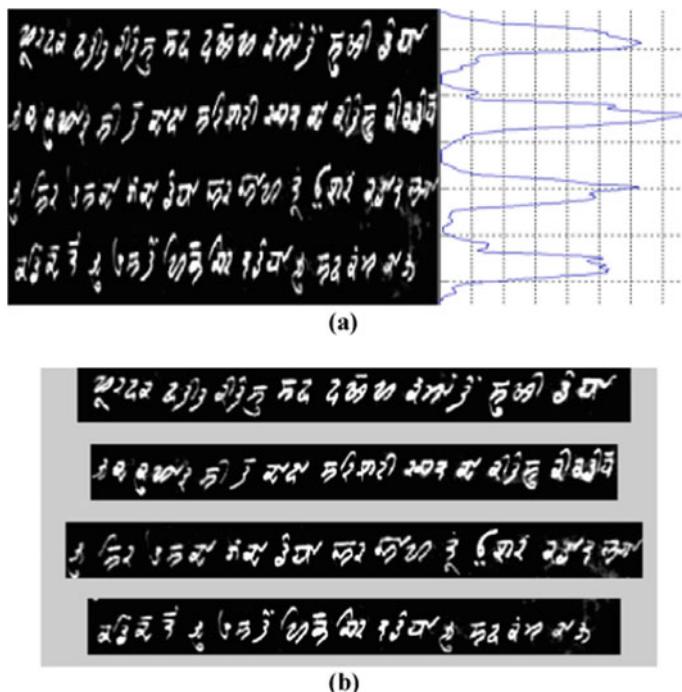


**Fig. 2** **a** Input image with noise. **b** Input image without noise

### 3.3 Segmentation

The segmentation phase is responsible for dividing the pre-processed input data into sub-components in order to recognize what exactly is contained in the input image [16]. In the present work, the lines of text are extracted or segmented from the document image using horizontal histogram projection profile [17] as displayed in Fig. 3a, b.

After obtaining the segmented lines, the individual characters are obtained by applying bounding box [18] technique as shown in Fig. 4. Finally, all the individual character images are normalized to uniform size of  $32 \times 32$  pixels.



**Fig. 3** a Horizontal histogram projection profile. b Segmented text lines



**Fig. 4** Segmented individual characters

## 4 Methodology

The convolutional neural network (CNN) is one of the most well-known deep learning architectures [19, 20]. In this work, we have used CNN for the identification of handwritten Dogra script. This proposed CNN architecture has two convolutional layers and with each convolutional layer followed by a max-pooling layer. The final layer is the soft-max layer as shown in Eq. 1.

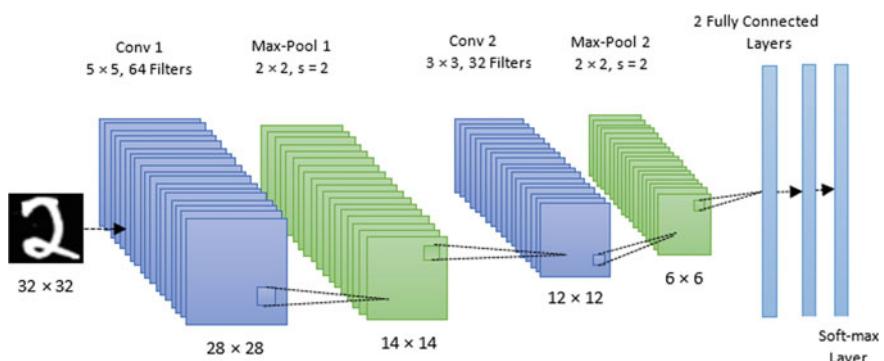
$$\text{softmax}(z_n) = \frac{e^{z_n}}{\sum_i e^{z_i}} \quad (1)$$

The first convolutional layer has 64 filters with each filter having size  $5 \times 5$  and stride of 1. This layer extracts features directly from the handwritten input character image and ReLu activation function is used as shown in Eq. 2. The output obtained from the ReLu function is then fed to max-pooling layer.

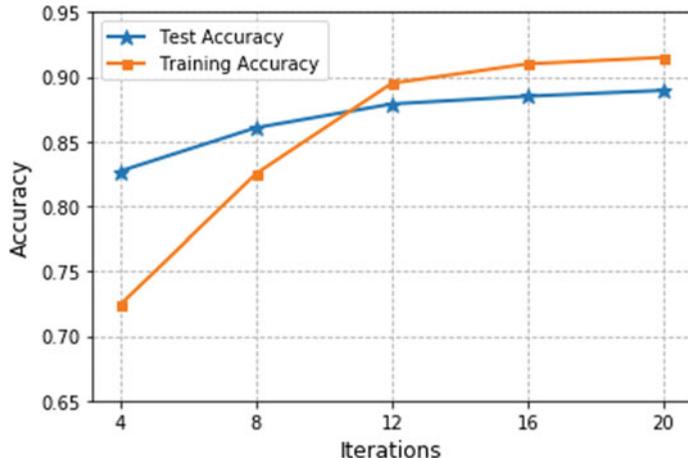
$$\text{Relu}(z) = \max(0, z) \quad (2)$$

The first max-pooling layer has size  $2 \times 2$ , with stride = 2. The output of first pooling layer has dimension of  $(14 \times 14 \times 64)$ , and this output is then passed to second convolutional layer having size  $3 \times 3$  with stride = 1, and 32 filters. Then, the second max-pooling has size  $2 \times 2$ , with stride = 2. The output features obtained from second max-pooling layer has dimension  $(6 \times 6 \times 32)$ , which is finally fed to the fully connected layer.

In this architecture, we have two fully connected layers. These fully connected layers are actually responsible for the final task of classification. In order to prevent the over-fitting of data dropout is applied at each layer. Dropout basically represents a regularization technique in which we omit randomly some layer's units for a given iteration. The final layer of the proposed model is the softmax layer. The CNN architecture is presented in Fig. 5.



**Fig. 5** CNN model for Dogra character recognition



**Fig. 6** Recognition accuracy on Dogra dataset

## 5 Experimental Results

We have trained our CNN model for 20 iterations on the handwritten Dogra character dataset. The dataset consists of 10 independent vowels, 10 dependent vowels and 34 consonants; therefore, we have 54 output classes. We have normalized all character images to uniform size of  $32 \times 32$  pixels, so each character image is represented as 1024-dimensional input.

The dataset has been randomly shuffled and then partitioned into three parts. The 75% of the handwritten character images are used for training the model. The remaining 25% character dataset is partitioned into validation set and testing set, with validation set having 5% of the handwritten character images and rest of the 20% data is used for testing the model. The trained CNN model gives a recognition accuracy of 88.95% when tested on handwritten Dogra character dataset as shown in Fig. 6.

## 6 Conclusion

In this paper, an effective deep learning model, i.e., CNN has been presented in order to recognize handwritten Dogra characters. This paper also contributes to the construction of handwritten Dogra script dataset. For constructing the handwritten Dogra script dataset, we have surveyed various sources like archival departments, libraries and museums in J&K. The collected handwritten document images are then pre-processed and segmented in order to obtain final handwritten character dataset.

From the literature, it is evident that no research work has ever been reported on Dogra script till date. So, this is the first attempt made in this direction. Since

Dogri language is one of the scheduled languages of India, so this research work will lead to significant movement for the revival of ancient Dogra literature available in libraries, temple inscriptions and museums. In the future, the handwritten Dogra character dataset can be used to obtain benchmark results for further research in this field.

## References

1. Boulid, Y., Souhar, A., Ouagague, M.M.: Spatial and textural aspects for Arabic handwritten characters recognition. *Int. J. Interact. Multimedia Artif. Intell.* **5**(1), 86–91 (2018)
2. Shi, C.Z., Gao, S., Liu, M.T., Qi, C.Z., Wang, C.H., Xiao, B.H.: Stroke detector and structure based models for character recognition: a comparative study. *IEEE Trans. Image Process.* **24**, 4952–4964 (2015)
3. Sharma, R., Kaushik, B., Gondhi, N.: Character recognition using machine learning and deep learning—a survey. In: International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 341–345. IEEE (2020)
4. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015)
5. Zhang, Q., Zhang, M., Chen, T., Sun, Z., Y. Ma, Y., Yu, B.: Recent advances in convolutional neural network acceleration. *Neurocomputing* **323**, 37–51 (2019)
6. Desai, A.A.: Gujarati handwritten numeral optical character reorganization through neural network. *Pattern Recogn.* **43**(7), 2582–2589 (2010)
7. Garg, N.: Handwritten Gurumukhi character recognition using neural networks. Master's thesis. Thapar University, Patiala (2009)
8. Sharma, D., Jhajj, P.: Recognition of isolated handwritten characters in Gurmukhi script. *Int. J. Comput. Appl.* **4**(8), 9–17 (2010)
9. Roy, S., Das, N., Kundu, M., Nasipuri, M.: Handwritten isolated Bangla compound character recognition: a new benchmark using a novel deep learning approach. *Pattern Recogn. Lett.* **90**, 15–21 (2017)
10. Shelke, S., Apte, S.: A fuzzy based classification scheme for unconstrained handwritten Devanagari character recognition. In: International Conference on Communication, Information & Computing Technology, pp. 1–6. IEEE (2015)
11. Sahare, P., Dhok, S.B.: Multilingual character segmentation and recognition schemes for Indian document images. *IEEE Access* **6**, 10603–10617 (2018)
12. Raj, R., Antony, M., Abirami, S.: Offline tamil handwritten character recognition using statistical based quad tree. *Australian J. Basic Appl. Sci.* **10**(2), 103–109 (2016)
13. Pandey, A.: Preliminary proposal to encode the Dogra Script in Unicode. Vol. 2. L2/15-213. <http://www.unicode.org> (2015)
14. Yadav, M., Purwar, R.K., Mittal, M.: Handwritten Hindi character recognition: a review. *IET Image Proc.* **12**(11), 1919–1933 (2018)
15. Otsu, N.: A thresholds election method from gray level histograms. *IEEE Trans. Syst. Man Cybern.* **9**(1), 62–66 (1979)
16. Sharma, R., Kaushik, B.: Offline recognition of handwritten Indic scripts: a state-of-the-art survey and future perspectives. *Comput. Sci. Rev.* **38**, 100302 (2020)
17. Singh, N.: An Efficient Approach for handwritten devanagari character recognition based on artificial neural network. In: 5th International Conference on Signal Processing and Integrated Networks, pp. 894–897. IEEE (2018)
18. Ha, J., Haralick, R.M., Phillips, I.T.: Document page decomposition by the bounding-box project. In: Proceedings of 3rd International Conference on Document Analysis and Recognition, pp. 1119–1122, IEEE (1995)

19. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks, In: European Conference on Computer Vision, pp. 813–833, Springer, Cham (2013)
20. Schmidhuber, J.: Deep learning in neural networks: an overview. *Neural Netw.* **61**, 85–117 (2015)

# “Device Design of 30 and 10 nm Triple Gate Single Finger Fin-FET for on Current ( $I_{ON}$ ) and off Current ( $I_{OFF}$ ) Measurement”



Sarika M. Jagtap and Vitthal J. Gond

**Abstract** Nowadays, users need portable gadgets like laptops and cellular phones with small in size which occupies less area, consumes low power and having low cost. Justifying Moore’s law by designing the smaller size transistors on the silicon wafer, more numbers of transistors available on a single wafer help to design complicated circuits with very low cost. Scaling plays vital role to decide the size of transistor with high performance. Most attracted multi-gate technology for researchers as well for industry is Fin-FET for nano-scale design. The nano-scale Fin-FET technology provides best solution for Moore’s law. This paper focuses on how Fin-FET helps to reduce short channel effect and also presents design of 30 nm and 10 nm single Fin-FET with Triple Gate. Leakage current, threshold voltage and drain drive current evaluated from device design by using high K of dielectric material. Simulation carried out using COMSOL MULTIPHYSICS Version 5.3

## 1 Introduction

Scaling has been expected toward smaller size, advanced speed, low stimulus and higher density of the semiconductor devices. As MOSFET channel thickness is climbed to the nano-meter regime, the gate cannot handle the position activity so damage the control action on the channel, and some critical belongings are occurred like hot carrier things, punch through, mobility degradation and short channel properties are occurring.

CMOS skill introduces the size of the device reduced to 10 micrometer in 1971 near 90 nanometer, and now in 2020 we avoid these possibly, we will shrink the to below 10 nm (7 nm, 5 nm). Size reductions of transistor cause unwanted technological effect that is short channel effect. Existing CMOS technology has a big problem of

---

S. M. Jagtap (✉)  
MVP Samaj's KBTCOE, Nashik, India

V. J. Gond  
MET's Bhujbal Knowledge City, COE, Nashik, India

short channel effect, and second most unwanted effect is leakage current; it introduces when gate length is reduced. Due to this limitation of CMOS, it cannot use beyond 22 nm reduction.

As MOSFET has planer device technology, it has very grave preventative factors of high device design cost, power restraints and very sensitive to process variations [1]. By providing alternative solution to the current technology, Fin-FET constructed multiple gate quasi-planer devices. To change nano-scale CMOS below 45 nm MOSFET 3-dimensional double gate (DG), triple gate (TG) device called Fin-FET is good choice. As Fin-FET may be double, triple or more than triple gate, i.e., multi-finger device, it has well gate action of blocking leakage over the perfect semiconductor channel, which diminishes short channel result (SCE) [2].

In this research work, we present how the scaling helps to improve device performance, and it is observed by designing a sample of 30 nm and 10 nm single Fin-FET with Triple Gate. From this model design, we observe the long channel and short channel device output in terms on threshold voltage, leakage current, subthreshold slope and drain-induced barrier lowering. This paper presents how Fin-FET benefits to reduce short channel effect by decreasing channel length and also how it helps to improve the device performance by using multi-gate. Modeling and simulation of 30 nm and 10 nm TG Fin-FET device design based on geometry is presented. The conventional MOSFET is in 2D planar structure, and the gate poses flat over the source and drain, whereas Fin-FET nonplanar is a 3D structure that covers the gate with gate oxide around the fin-shaped source and drain.

## 1.1 *Review of Literature*

Fin-FET technology has been born as a result of the increase in the levels of integration. The primary multi-gate transistor was distributed by Heida in 1987. As CMOS-based gadgets are scrambled to nano-meter region, due to scaling, several complications occur like drain-induced barrier lowering, punch through, mobility degradation and off current which extremely disturb the device concert.

Narendar and Mishra [1] In this work, researcher group has been analyzing independent gate Fin-FET for threshold voltage discrepancies by using work function. Effective energy of MOSFET count depends on how much area of depletion is fully charge. For lengthy, i.e., more distance between drain to source called as long channel devices, silicon material is effectively used as gate terminal supplies. Metal work function plays important role to change doping. High k dielectrics like  $HFO_2$  is better to improve the performance in terms of leakage current.

In paper, [2] main problem associated with the Fin-FET device is higher leakage and lower threshold voltage. By considering the geometry of Fin-FET like fin thickness, height, width on current and off current is affected. By suitable use of metal work function and low or high-k dielectrics, entire current of device affected. Advanced the dielectric asset of gate part, reduced the leakage, and developed the beginning voltage can be found a improved device appearance.

The blade, i.e., fins breadth of Fin-FET thickness must remain keep a smaller amount than one-third of channel dimension to reduce effects of short channel. In planar CMOS, due to long channel length, some harmful effects are occurred which are completely avoided in bulk Fin-FET. [3] Reduction in fin size remains toward reduction in leakage owing to short channel effects. Variation in work function and high-k dielectrics centrals to difference in beginning voltage.

The key factor of VLSI-based project is the power requirements, leakage current and size. Here, [4] author explores how circuits based on Fin-FETs expertise that is probable to complement bulk CMOS at 22-nano-meter, offer stimulating delay-power adjustments.

Due to scrambling of metal oxide semiconductor field effect transistor, nonstop reduction in turn on or operating voltage is detected, and off current or leakage current, mobility degradation, punch through effects are produced [5]. By suitable use of metal work function, we avoid these effects. Work function is the lowest energy desired to eliminate an electron from a solid to a point nearly separated from the solid surface. By affecting the work function in the Fin-FET modeling, it is possible to set appropriate operating voltage at same supply voltage and possibly decrease leakage currents. By varying in metal gate work function from 4.1 eV to 4.4 eV, author offers direct variation in threshold voltage.

Chopade [6] with silicon dioxide, the thickness of gate oxide is so small that it can barely produce a gate current. So the use of high-K material as a gate oxide is preferred. In this study, author proposes a device structure for leakage suppression using a pile gate with a different work function in the bottom gate electrode. The core gain of Fin-FET construction is there is no necessity for an extra substrate doping.

By introducing assets of MOSFET like fin thickness, doping concentration, gate material, and oxide thickness, author represents performance improvements. The depth of oxide layer is related to square root of oxidation period. From I-V characteristics, it originates that the threshold voltages ( $V_t$ ) for MOSFETs with changed oxide layer widths are related to the square root of the ( $V_{gs}$ ) [7].

In [8], author concentrates on how to improve the performance parameters like drain-induced barrier lowering, gate-induced barrier lowering, subthreshold slope, on current, threshold voltage of CMOS device which is observed by changing doping concentration of drain and source regime. By designing, author found that the doping differences releases the threshold voltage of Fin-FET and improves other parameters.

Mohd Radzi and Sanudin [9] Author presented the transistor performance by changing oxide thickness. The imitation grades that the threshold voltage is directly proportional to oxide thickness. Drain current is mainly related to operating drain voltage for smallest variation in gate voltage up to the border of its saturation level. The consequences prove that the leakage current, i.e., off current ( $I_{OFF}$ ) is improved as oxide thickness grow into stripper.

Multi-finger Fin-FETs are presented; here [10] fins are parallel to each other as the device is triple gate. Increase in quantity of fins peaks to upgrading the current over the device. The breadth of the fin interprets the operative channel length of the device [10]

By reducing threshold voltage, power gets affected. However, as drain voltage is reduced, the transistor takes time to change the transition from off to on with the partial voltage swing offered. Subthreshold leak current is most important issue of recent high performance [11]. The double, triple or multi-gate MOSFETs are playing most important role to improve the performance having two gates, triple, gate all around on any side of the channel. In this work, writer learning the device sizes things on the presentation of a double gate silicon MOSFET via energy stability typical with Bohm quantum potential.

In chapter, author does the MOSFET modeling to present gate tunneling for current measurement. The greater the gate voltage, the bigger the electric fields created in the substrate, thus growing the probabilities of tunneling.

By using Poisson and quantum transport equations, researcher simulates current voltage characteristics [13]. Reduction of short channel properties is found in double gate two-dimensional metal oxide semiconductor field effect transistor with changed in metal gates.

Threshold voltage measurement done by researcher for the lower power operating devices. Author designs 10 nano-meter NMOS Fin-FET to achieve low power requirement. By changing work function, author found off current of devices and threshold voltage.

Double gate means front and back gate transistors are modeled with completely depleted silicon layer have a spreading of the electrical potential inside the layer, where the potential at the center being different from zero [15].

In this effort [16] scholars' effort on assessment of Fin-FET strategies with bulk CMOS skill in nano-meter regions. As the device is triple gate (front, back, top), the channel is bounded in three extents in Fin-FET. High on current and output resistance are measured from the design.

In this paper, the variation of device characteristics of double gate Fin-FET with respect to the thickness of oxide and different dielectrics at 14 nm gate length is presented [17]. Gates wrapped around the vertical channel (or) fin allows ease in fabrication and compatibility with CMOS fabrication process. This study explains the sensitivity of Fin-FET performance metrics and short channel metrics to variation in dielectric constant and thickness of  $T_{ox}$ . The oxide layer wrapping the sides of the fin improves the control over the gate that makes them superior to planar MOSFETs in reduction of short channel metrics such as operating or device active voltage, pinch off and drain-induced barrier lowering.

The gate terminal dielectric  $SiO_2$  is changed with several high-k dielectric materials. The author planned gate dielectric k values are varied from 3.9, 7 and 22. As the k values are varied from 3.9, 7 and 22, the drain current also growths. The leakage or off current and the drain-induced barrier lowering parameters are start to decline exponentially which implies that the device has superior short channel effect suppression capability. The lateral electric field also increases with increase in gate dielectric value.

Author presents how high-k dielectrics help to control the leakage current and maintaining the threshold voltage [19].

Researcher investigates the [20] result of collective improving the quantity of fins affects on drain on-off current methods. Here, the drain current calculation is demonstrated by seeing the belongings of the depletion charges in the gate region, i.e., the channel close to the source/drain and substrate boundary.

Here, in [21], author settled a triple gate Fin-FET device for improved threshold voltage ( $V_{th}$ ), subthreshold slope (SS), high on current ( $I_{on}$ ) and high  $I_{on}/I_{off}$  ratio. By scrambling down the oxide thickness, author found intensification in on current and threshold voltage.

Using high-k dielectric, i.e.,  $HfO_2$  and  $ZrO_2$  demonstrates comparatively high threshold voltage levels due to imperfections at the gate electrode edge. This decreases drive current [22].

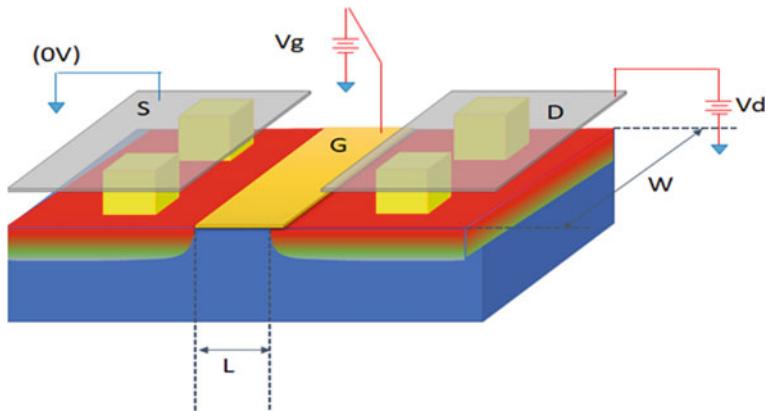
The drawback of the planar transistors the short channel effect when decreasing the channel length of the device. The idea of Multi-gate FETs became the ideal solution to the negative effects of channel decreasing of planar MOSFET. The different structures and their fabrication techniques of Multi-gate devices provide chip makers with devices with small channel length, and faster switching efficient speed operation and lower power consumption, due to the much-improved control over the device channel [23].

Carusone et al. [24] in this study, researcher simulates the changed types of electrical characteristics for dissimilar operating regions and changed gate dimensions and also for different oxide thickness. Fin-FET routes can reach lesser efficient voltage supply and optimal energy consumption associated to CMOS circuits.

## 2 MOSFET and FIN-FET Structure

Metal oxide semiconductor field effect transistor popularly referred to as MOSFET is most commonly used device for semiconductor circuits. Basic structure of the device is shown through its cross-section. Device formed in semiconductor material is referred to as frontend while metallization, contacts and poly-silicon are referred to as backend [3].

In a typical process, semiconductor may be lightly doped p-type ( $p^-$ ). NMOS device has four terminals namely drain, gate source and body. Drain and source are formed by  $n^+$  embedded in  $p^-$  substrate and separated physically by distance called gate length (L). The other dimensions of gate are gate width (W). Gate area (WL) is covered with nonconductive polymorphous silicon (poly-silicon). Metal ion implant on top surface of poly-silicon makes it conductive. This constitutes gate contact using conductive poly-silicon. Contacts to drain and source are made by process referred to as metal lift-off. A MOSFET structure is shown in Fig. 1 which describes the gate and channel arrangement between source and drain. Different types of dielectric materials can be used to prepare substrate and channel of device [4]. Trade off in MOSFET design is that if channel length decreases, leakage current increases which consumes more power. To reduce this effect, designers use low K dielectric material but this causes to reduce oxide capacitance which degrades performance of the device.



**Fig. 1** MOSFET construction

## 2.1 Threshold Voltage

Threshold voltage of a device depends on silicon–silicon dioxide interface work function, also known as metal work function ( $\phi_{SM}$ ), inversion potential ( $2\phi_f$ ), stored gate charge and gate oxide capacitance ( $\frac{Q_b}{C_{ox}}$ ) and ion implant under the gate at silicon–silicon dioxide interface ( $\frac{Q_{ss}}{C_{ox}}$ ). Threshold voltage can be assumed as voltage drop across gate and substrate [5]. All the derivations in this section are for unit gate part. Gate oxide capacitance ( $C_{ox}$ ) per unit gate area is given by.

$$C_{ox} = \frac{\epsilon_{ox}}{t_{ox}} \quad (1)$$

Current flows from drain to source and is modulated by gate voltage (through control over gate channel depth). For this analysis, we assume that body terminal is tied to the source and there is no body effect.

$$I_D = \frac{\mu_n C_{ox}}{2} \frac{W}{L} [2(V_{GS} - V_t)V_{DS} - V_{DS}^2] \quad (2)$$

## 2.2 Overdrive Voltage

Voltage required to create gate channel in an enhancement mode MOS device is called threshold voltage. Gate voltage above the threshold voltage is called overdrive voltage.

$$V_{OV} = V_{GS} - V_t \quad (3)$$

$$I_D = \mu_n C_{\text{ox}} \frac{W}{L} V_{\text{OV}} V_{\text{DS}} \quad (4)$$

This is equivalent to resistance.

$$R_{\text{DS}} = 1 / \left( \mu_n C_{\text{ox}} \frac{W}{L} V_{\text{OV}} \right) \quad (5)$$

Gate to source and gate to drain capacitance are equal.

$$C_{\text{gs}} = C_{\text{gd}} = \frac{W L C_{\text{ox}}}{2} \quad (6)$$

$C_{\text{ox}}$  Gate oxide capacitance.

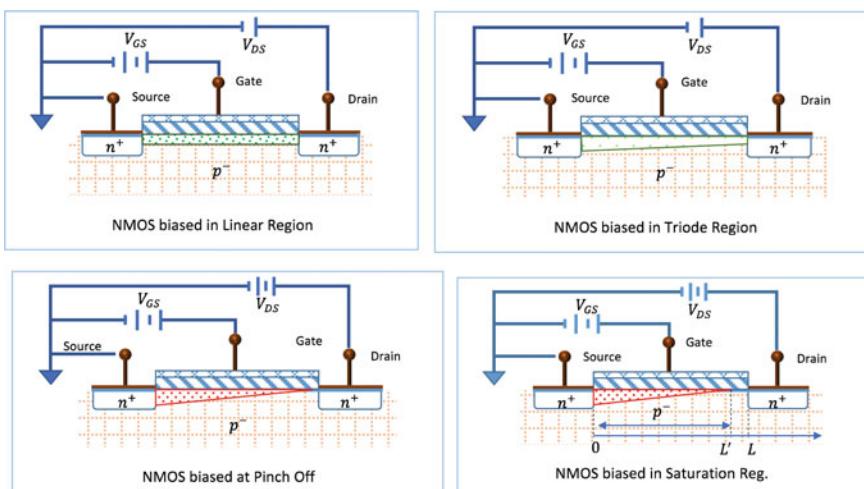
$W$  Width of the channel.

$L$  Gate length.

$V_{\text{OV}}$  Overdrive voltage.

$V_s$  Drain source voltage.

The planar MOSFET device has flat channel, i.e., MOSFET is parallel work whereas, and the Fin-FET channel is a thin steeply effort. Fins are entirely covered around the gate and channel is shaped between the source and the drain stations as shown in Fig. 2. The current flows in corresponding to the plane in Fin-FET, whereas the designed conducting channel is molded everywhere the fin boundaries [4]. With the similar structure of Fin-FET, the gate is able to abundantly depleted in the channel, and this formed conducting channel having much-improved electrostatic control over the channel. Fin-FETs can be classified by gate structure or type of substrate [6].



**Fig. 2** MOSFET working in different region

Basically, Fin-FET is classified in two types:

1. Shorted-gate (SG) Fin-FETs
2. Independent-gate (IG) Fin-FETs.

The feature of the Fin-FET is that the leading channel is enfolded through a narrow semiconductor “fin” that procedures the gate the device and the fin thickness control the effective channel distance of the device [3].

MOSFET is a planar device, whereas Fin-FET is work as nonplanar device, and there is a nitride spacer among the gate and the source and drain, which are usually higher [7].

The Fin-FET structure is built on a perpendicular silicon Fin considered by the gate length, fin height ( $H_{\text{fin}}$ ) and the silicon thickness as shown in Fig. 2.

### 3 Device Design of FIN-FET

The fin field effect transistor performance depends on the process-induced variation under systematic value of,

1. Gate length [ $L_g$ ],
2. Fin height [ $H_{\text{fin}}$ ],
3. Fin width [ $W_{\text{fin}}$ ],
4. Gate oxide thickness [ $t_{\text{ox}}$ ],
5. Channel doping.

$$\lambda = \sqrt{\frac{e_{\text{ox}}}{e_{\text{si}}} \left( 1 + \frac{e_{\text{ox}} * t_{\text{si}}}{4 * e_{\text{si}} * t_{\text{ox}}} \right) t_{\text{si}} * t_{\text{ox}}} \quad (7)$$

$$\lambda H_{\text{fin}} = \sqrt{\frac{e_{\text{si}}}{e_{\text{ox}}^4} \left( 1 + \frac{e_{\text{ox}} T_{\text{si}}}{4 e_{\text{si}} T_{\text{ox}}} \right) H_{\text{fin}} T_{\text{ox}}} \quad (8)$$

$$\lambda c = \frac{1}{\sqrt{\left(\frac{1}{\lambda}\right)^2 + \left(\frac{\Lambda}{\lambda H_{\text{fin}}}\right)^2}} \quad (9)$$

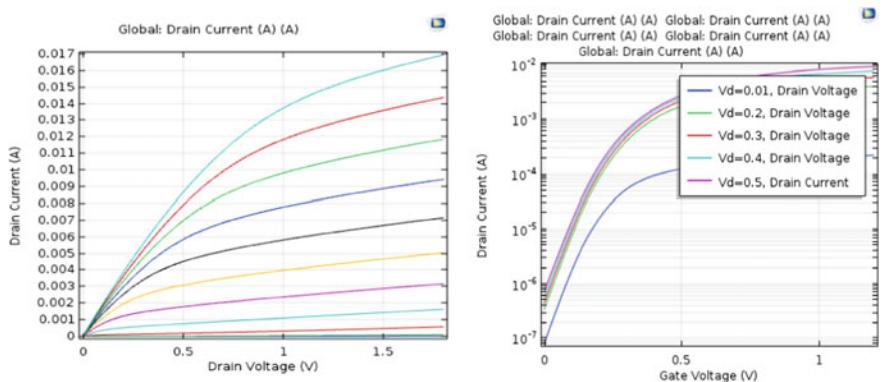
Nis operative amount of gates,  $e_{\text{si}}$  is permittivity of silicon,  $e_{\text{ox}}$  is permittivity of gate oxide [8, 9] (Table 1).

$t_{\text{ox}}$ is thickness of gate oxide,  $t_{\text{si}}$  is thickness of fin,  $H_{\text{fin}}$  is height of fin.

Front view and top view of single Fin-FET structure are shown in Figs. 3 and 4. Figure 5 shows geometry of 30 nm single Fin-FET (Figs. 6 and 7).

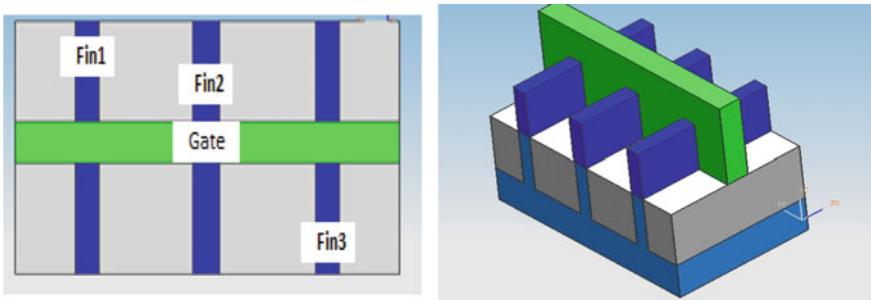
**Table 1** Physical parameters used in the design of single Fin-FET with 30 nm Fin width and 30 nm gate length

S. No	Device parameter	Symbol	N-Fin-FET
1	Height of fin	$H_{\text{fin}}$	90 (nm)
2	Width of fin	$W_{\text{fin}}$	100 (nm)
3	Fin thickness (depth)	$L_{\text{fin}}$	30 (nm)
4	Gate length	$L_G$	30 (nm)
5	Gate height	$H_{\text{gate}}$	60 (nm)
6	Gate work function	$\phi_g$	4.1
7	Oxide thickness	$T_{\text{ox}}$	1 (nm)
8	Drain voltage	$V_d$	10mv
9	Gate voltage	$V_g$	1 V
10	Gate material	$Si$	Si
11	Oxide relative permittivity	$\epsilon_{\text{ins}}$	6.9
11	Channel doping concentration	$N_{A0}$	1e19 (1/cm <sup>3</sup> )
12	Source/drain doping concentration for NMOS	$N_{D0}$	1e21 (1/cm <sup>3</sup> )
13	Junction depth	$D_j$	1 (nm)

**Fig. 3** I-V characteristics of MOSFET—NMOS output characteristics

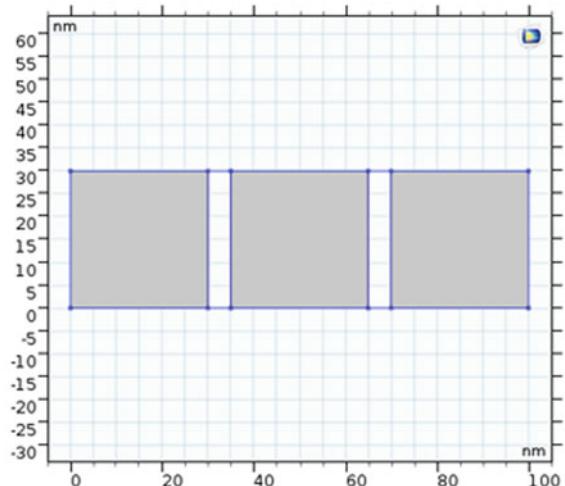
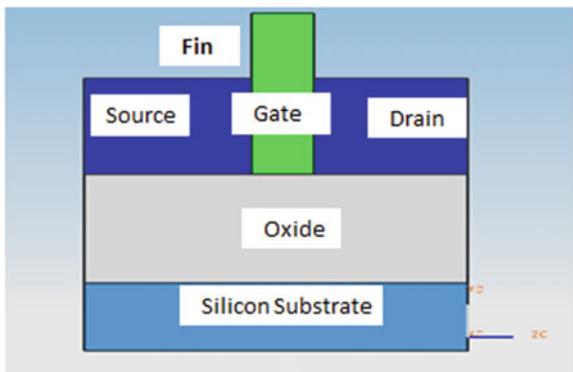
## 4 Conclusion

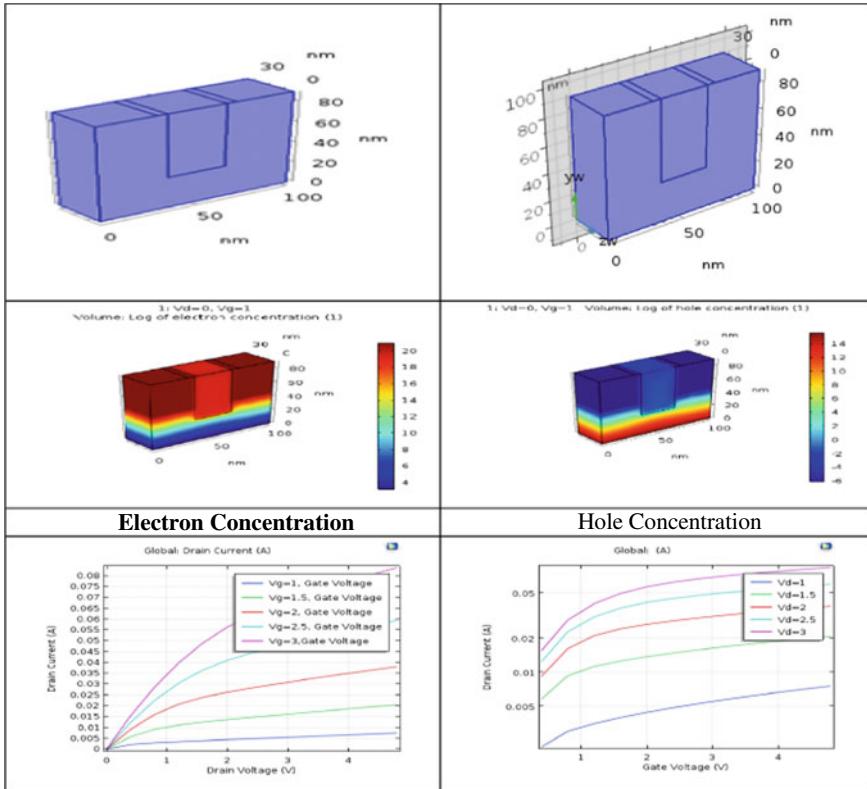
In this paper, we introduced the basic device design of 30 nm and 10 nm fin thickness for 30 nm gate length single finger Fin-FET using triple gate (front, back and top gate). Fin thickness and gate length improve threshold voltage to enhance speed of device by using geometry parameters. Using changed physical geometry parameters, we approach the single Fin-FET, and by varying gate length and fin, thickness



**Fig. 4** Fin-FET structure

**Fig. 5** Front view of single Fin-FET and plane geometry of single Fin-FET



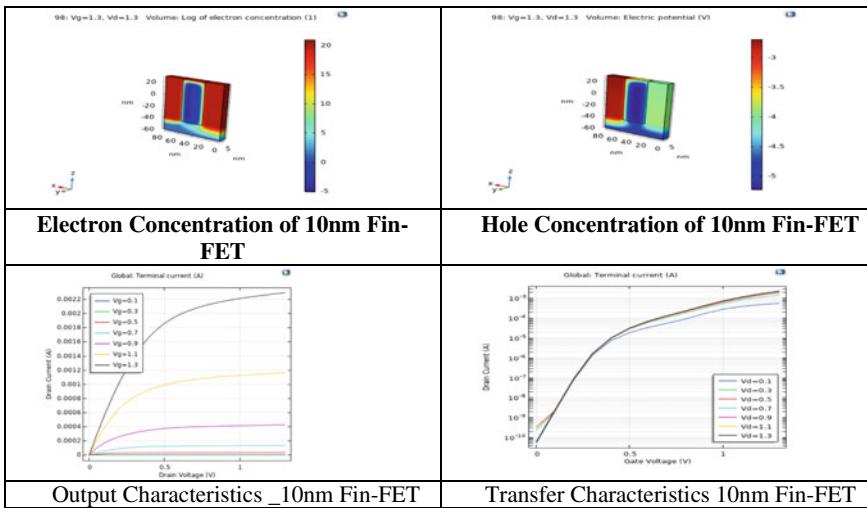


**Fig. 6** I-V characteristics of 30 nm Fin-FET

detects effect on  $I_{off}$ ,  $V_t$ , SS and DIBL. Their presentation studied in terms of I-V characteristics and transfer characteristics.

For long channel effect, pinch off rises at overdrive voltage ( $V_{gs}-V_t$ ). By falling the fin thickness, the leakage current decreases which helps to expand the presentation of the device.

In this , we introduced the basic device design of 30 nm and 10 nm fin thickness for 30 nm gate length on single finger using trigates (top, front, back). Fin thickness and gate length improve threshold voltage to enhance speed of device by using geometry parameters.



**Fig. 7** I-V characteristics of 10 nm Fin-FET

**Acknowledgements** I thankfully acknowledge Ni<sub>2</sub> Logic Design, Pune, for providing the licensed tool for implementation of Fin-FET fdesigns.

## References

1. Narendar, Mishra, R.: Threshold voltage control schemes in FIN-FETS. Int. J. VLSI Des. Commun. Syst. (VLSICS) 3(2):175–191 (2012). <https://doi.org/10.5121/vlsic.2012.3215>
2. Anju, C.: Performance analysis of wavy Fin-FET and optimization for leakage reduction. In: 2016 IEEE International Symposium on Nano electronic and Information Systems, pp 83–85. <https://doi.org/10.1109/iNIS.2016.43>
3. Shukla, S., Gill S.S.: Comparative simulation analysis of process parameter variations in 20 nm triangular Fin-FET. Act Passive Electron Compon 2017, 8 pp. Article ID 5947819. <https://doi.org/10.1155/2017/5947819>.
4. Mishra, P., Anish, M., Jha, N.K.: Fin-FET circuit design. Nanoelectronic Circuit Design. Springer Science New York, pp 23–54 (2011)
5. Ranka, D., Rana, A.K.: Performance evaluation of FD-SOI MOSFETs for different metal gate work function. Int. J. VLSI Des. Commun. Syst. (VLSICS) 2(1), 11–24 (2011)
6. Chopade, S.S., Padole, D.V.: Dual material gate approach for low leakage FIN-FET. Int. J. Technol. (2017). <https://doi.org/10.14716/ijtech.v8i1.3699>
7. Fan, J.-C., Lee, S.-F.: Effect of oxide layer in metal-oxide-semiconductor systems. In: MATEC Web of Conferences SMAE 2016, 5 pp. <https://doi.org/10.1051/06103> (2016). [matecconf/2016MATEC Web of Conferences 6SMAE 2016706103]
8. Keerti Kumar, K., Anil, P., Bheema, R.N.: Parametric variation with doping concentration in a Fin-FET using 3D TCAD. Int. J. Comput. Appl. 3, 21–23. [International Conference on Microelectronics, Circuits and Systems (MICRO-2014)] 0975 – 8887
9. Mohd Radzi, N., Sanudin, R.: Effect of oxide thickness variation in sub-micron NMOS transistor. Int. Res. Innov. Summit (IRIS2017) 10. IOP Publishing. <https://doi.org/10.1088/175-899X/226/1/012145>

10. Somra, N., Sawhney, R.S.: 32 nm Gate Length Fin-FET: impact of doping. *Research Gate* (2015). *Int. J. Comput. Appl.* **122**(6), 11–14 (2015). 0975 – 8887
11. Hasan, M., Hassan, E.: Study of scaling effects of a double gate silicon MOSFET. In: 10th International Conference on Electrical and Computer Engineering, 20–22 Dec 2018, pp. 169–172
12. Chaudhry, A.: Fundamentals of nano-scaled field effect transistors. *Nanoscale Effects: Gate Oxide Leakage Currents*. Springer Science New York (2013)
13. George James, T. Joseph, S.: The influence of metal gate work function on short channel effects in atomic-layer doped DG MOSFET. *J. Electron Devices* **8**, 310–319 (2010)
14. Walke, A.M.: Design strategies for ultralow power 10 nm Fin-FETs. In: 2017 Rochester Institute of Technology RIT Scholar Works
15. Cerdeira, A., Estrada, M., Alvarado, J.: Review on double-gate Mosfets and Fin-Fets modeling. *Facta Univ. Ser. Electron. Energetics* **26**(3), 197–213 (2013). <https://doi.org/10.2298/FUEE1303197C>
16. Farkhani, H., Peiravi, A., Kargaard, J.M., Moradi, F.: Comparative study of Fin-FETs versus 22 nm bulk CMOS technologies: SRAM design perspective. In: 2014 27th IEEE International System-on-Chip Conference (SOCC) 2–5 Sept 2014, pp. 449–454
17. Mushahhid Majeed, M.A., Rao, S.: Influence of thickness of oxide and dielectric constant on short channel metrics in Fin-FETs. *J. Adv. Res. Dyn. Control Syst.* **9**(4), 57–64 (2017)
18. Nirmal, D., Thomas, D.M.: Impact of channel engineering on Fin-Fets using high-K dielectrics. *Int. J. Micro Nano Electron. Cir. Syst.* **3**(1), 6 (2011)
19. Yin, H., Yao, J.: Advanced transistor process technology from 22- to 14-nm node (2018)
20. Sivasankaran, K., Mallick, P.S.: Impact of device geometry and doping concentration variation on electrical characteristics of 22 nm Fin-FET. In: 2013 (ICECCN 2013), pp. 528–531
21. Gupta, T.K.: Copper interconnect technology. *Dielectric Materials*. Springer Science (2009)
22. Shehata, N., Gaber, A.-R.: 3D multi-gate transistors: concept, operation, and fabrication. *J. Electr. Eng.* (2015)
23. Hossain, M.Z., Hossain, M.A.: Electrical characteristics of trigate Fin-FET. *Glob. J. Researches Eng. Electr. Electron. Eng.* (2011)
24. Carusone, T.C., Johns, D.A., Martin, K.W.: *Analog Integrated Circuit Design*, 2nd edn. John Wiley & Sons, Inc. (2012). ISBN 978-0-470-77010-8

# Fact Check Using Multinomial Naive Bayes



**Madhavi Ajay Pradhan, Ankita Shinde, Rohan Dhiman, Shreyas Ghorpade, and Swapnil Jawale**

**Abstract** An easy access to social media platforms has made information available effortlessly and thus has increased the intricacies to distinguish between true and falsified information. The credibility or reliability on social media platforms is also at stake. It is of utmost necessary to address this as a severe issue and act on it promptly. The extensive spread of counterfeit news has the potential for creating negative impacts on vast audience. Therefore, fake news detection on social media has become a very critical agenda in today's world. This paper proposes a prototype to detect whether a news is fake or real using the multinomial Naive Bayes algorithm and its various architectures. Furthermore, the proposed prototype is capable of handling the unstructured data as the news can be in the form of images. In addition to this, the use of Django which is a high-level Python framework that allows the development of UI very easily with multiple designing options. As there was a high need of a 24/7 working server, the system has been deployed on Amazon Web Services EC2 Server as it gave less downtime and is highly reliable. Experimentation was done on the synthetic COVID news dataset created by collecting COVID news on social platforms.

## 1 Introduction

Grouping of any news thing into Fake or Genuine one has offered ascend to an incredible enthusiasm from scientists around the globe. Different examination contemplations have been done to watch the impact of distorted or created news content on masses and reaction of individuals after running over such news things. Counterfeit news and fabricated news can be in either textual or in image format. This paper addresses the classification of both types. In the twenty-first century, individuals are more into Web-based media as opposed to investing important energy in perusing the news on approved news sites, they contribute additional time via online

---

M. A. Pradhan · A. Shinde · R. Dhiman (✉) · S. Ghorpade · S. Jawale

Department of Computer Engineering, AISSMS College of Engineering, University of Pune, Pune, Maharashtra, India

media, and they doubtlessly accept whatever news going ahead social media will be valid; however, it is phony news just to conceal reality or misdirect perusers and have negative effect about the genuine news [1]. Initially, counterfeit news can break the creativity equilibrium of the entire news environment. Furthermore, counterfeit news deliberately wins buyers to acknowledge one-sided or false convictions. Counterfeit news is typically controlled by the advertiser to pass on political messages or impact. Thirdly, counterfeit news changes the manner in which individuals decipher and react to genuine news [2]. For example, the recent COVID situation has caused a havoc around the globe. People in order to immune themselves may believe in a news that talks about a medicine which cures COVID-19 and might consume them and end up harming themselves, which in reality is just a falsified statement. This paper explores the use of classification algorithms that can be used to segregate fake news from real news.

## 2 Literature Survey

Gilda [3] made the use of TF-IDF along with bigram, syntactical structure and also various other features to find out the best fit factor for fake news detection. In the paper, it was found out that bigram and TF-IDF give the models that have a very high effect in the classification of the articles.

Kim et al. [4]—proposed two different ways: per-document text normalization and feature weighting method. While these are fairly impromptu strategies, the creator's proposed gullible Bayes text classifier performs very well in the standard benchmark assortments, rivaling cutting edge text classifiers dependent on an exceptionally unpredictable learning strategy, for example, SVM. Exploratory outcomes on the two assortments show that the proposed model is very valuable to construct probabilistic content classifiers with minimal additional expense as far as existence, compared to the conventional multinomial classifiers. Relative recurrence or smoothed relative recurrence is sufficient to standardize term frequencies.

Conroy et al. [5] proposed a hybrid approach that consolidates linguistic cue and AI, with network-based conduct information for counterfeit news recognition. Despite the fact that the methodology they utilized gives high exactness arrangement, they are restricted uniquely to printed information and do not investigate the unstructured information.

Tijare [6] had proposed models for fake news detection utilizing Naive Bayes, SVM, and long-short term memory (LSTM) which is an augmentation of RNN and Keras-based neural organization. LSTM-based model gives the most elevated precision when the content is naturally a serialized object.

Kim et al. [7] used a different approach by using the unified word vector for the various key sentences of article.

Granik et al. [8] proposed fake news detection model using Naive Bayes classifier and for their explanation they made the use of Facebook news posts dataset.

A similar kind of a problem where there are two labels and a classifier was used to predict the outcome by Pradhan and Bamnote [9] in their research. They proposed a classifier to find out whether a particular patient is diagnosed with diabetes or not using support vector machine.

Analyzing all the research done and after a comparison, a suitable algorithm was selected was moving further with the research.

### 3 Proposed System

The proposed plan is divided into three main modules

1. Basic textual fact check.
2. OCR functionality for news article in image format.
3. Combining both functionalities 1 & 2 and deploying it to a 24/7 working server to be accessed by the user anytime.

#### 3.1 Basic Textual Fact Check

A dataset will be first scraped from various news Web site. The dataset will consist articles divided into two categories fake and real. This dataset will be made to be used for the training purpose of the model. A Python code will then be developed in which the use of machine learning model, i.e., multinomial Naïve Bayes will be done. The dataset will be provided as an input to the algorithm for the training purpose, and then the trained model will be prepared and saved on the local server. A separate Python code would be developed in which input would be taken from the user for the news he or she wants to check the credibility for. The code would take in the input and predict the value by analyzing it from the model saved by the algorithm. The output would be displayed in accordance with the reply received from the algorithm.

#### 3.2 OCR Functionalities

Many news articles nowadays that are circulated over the Internet are in the form of images. So providing the user an option where he or she can upload the news image consisting of some written news, so that the system predicts and lets them know what it feels about the news. So the program would make the use of Pytesseract for the purpose of text extraction, and then providing the extracted text to the algorithm that would again predict the output from the model in the similar manner of the simple textual news.

### ***3.3 Combining and Deploying the Program Over a 24/7 Working Server***

Both the programs for simple text and OCR will be combined into a single program. For the users to interact with the software, a UI is needed. So the UI will be built using Django with the use of the language HTML. The user interface would be kept simple so that anyone can use the software without any extra knowledge or training. The whole software will be deployed on Amazon Web Server EC2 instance, with the aim that the software is accessible to the public 24/7.

### ***3.4 System Goals***

1. Provide a simple user interface.
2. Long-term goal for the software is to help people input a particular news and check its credibility.
3. We are trying to solve the problem of fake news that is being spread all over the social media.

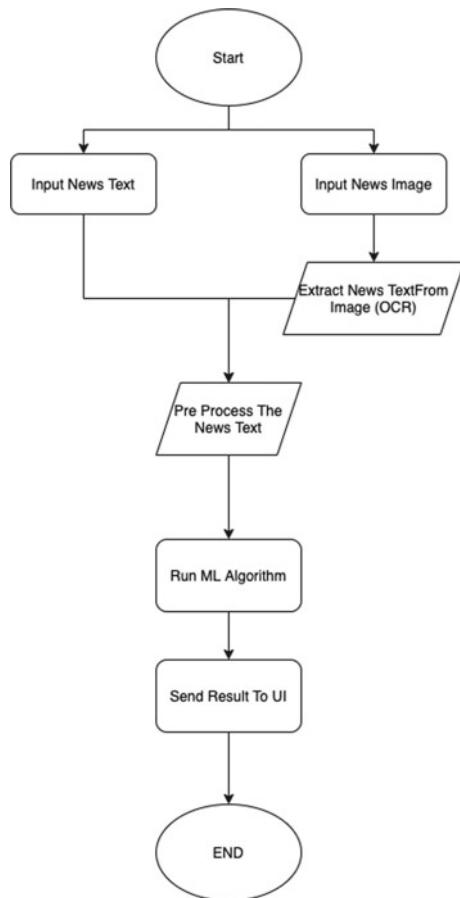
### ***3.5 Overall Design***

The design of the software was kept very simple so that any user can use the software without any prior training. First, the whole software was deployed over a Web site. The Web site has a very simple and appealing design. For taking the input from the user for the textual news, a simple text box is provided where the news article can be written or pasted and then on a button click the Web site sends the request to the server, and the algorithm is run in the background and the output is displayed to the user on the same page of the Web site itself. The second section contains an option to upload a news article in the image format. After the image is uploaded a button is made available, on click of which similar to the previous section the request is sent to the server, the algorithm runs on the server, and the output is displayed on the same page to the user (Fig. 1).

## **4 Proposed Algorithm**

The algorithms that gave us the best accuracy and an amazing result were multinomial Naïve Bayes. Multinomial Naive Bayes is one of the specialized versions of Naïve Bayes family that is designed text documents. Multinomial Naive Bayes explicitly models the word counts and adjusts the underlying calculations to deal with in. The

**Fig. 1** Working of the system



Web site that we built gives the users an option to upload a news image. In order to verify the truthfulness of this image as a news, the text has to be extracted from the image. This is done by using optical character recognition. OCR is implemented by using Python tesseract to “read” the text off an image. The use of Django has been made as it is a high-level Python Web framework that allows rapid development of a secure and maintainable Web site. Django mostly takes care of much of the hassle of Web development, so its easier for us to focus on the accuracy of the system. Ultimately the system has been deployed on Amazon Web Services EC2 Server to give us a Web site that runs 24/7 with minimum downtime and maximum reliability.

## 4.1 Multinomial Naive Bayes

Multinomial Naive Bayes is used to classify discrete features, for example, word count for text classification. Multinomial Naive Bayes is designed for text documents. Multinomial Naive Bayes expressly models the word tallies and modifies the fundamental estimations to manage it. The probability of the input sentence with the label is calculated and according to the probability attained the result is formulated.

**Practical Example** Considering a practical example. The user input's a news "Thorough hand-washing with an ordinary soap is effective in killing coronavirus (COVID-19) Soap and water alone, when used as per the WHO handwashing guidelines, are effective and easy for killing coronavirus" and wants to check whether the news is "REAL" or "FAKE".

The following example dataset is used for the training purpose (Table 1):

**Feature Engineering** In the initial step, we will extricate the highlights. We need mathematical highlights as contribution for our classifier. So we can consider word frequencies, for example, tallying the event of each word in the record. Thus, we need to calculate the probability.

Bayes' Theorem is helpful for managing restrictive probabilities, since it gives an approach to us to invert them.

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

NewsInput = Thorough hand-washing with an ordinary soap is effective in killing coronavirus (COVID-19) Soap and water alone, when used as per the WHO handwashing guidelines, are effective and easy for killing coronavirus.

1. **Simple Probability Approach** So, in our case the probabilities we will need to calculate are

$$P(\text{REAL}|\text{Newsinput}) = \frac{P(\text{Newsinput}|\text{REAL}) \times P(\text{REAL})}{P(\text{NEWS})}$$

$$P(\text{FAKE}|\text{Newsinput}) = \frac{P(\text{Newsinput}|\text{FAKE}) \times P(\text{FAKE})}{P(\text{NEWS})}$$

**Table 1** Sample dataset

News	Classification
US and UK have selected PM Modi to lead a coronavirus task force	FAKE
14-hour 'Janta curfew' will not break the cycle of infection	REAL
WhatsApp puts new limits on forwarding of viral messages	REAL
No lockdown after 15th April says Maharashtra CM	FAKE
Eating stuff like garlic, turmeric can save you from getting corona	FAKE

But, here, the problem is that the NewsInput is not present in the dataset, so the probability for both would come out to be zero. So, a simple probability approach cannot be followed.

2. **Being Naive** In the non-Naive Bayes way, we take a look at sentences in totality, in this manner once the sentence does not appear in the preparation set, we will get a zero likelihood, making it hard for additional figurings. While for Naive Bayes, there is a supposition that each word is autonomous of each other. Presently, we see singular words in a sentence, rather than the whole sentence.  
So, now, we can calculate the probabilities as:

$$\begin{aligned} P(\text{Newsinput}) = & \quad P(\text{Through}) \times P(\text{hand}) \times P(\text{washing}) \times P(\text{with}) \times P(\text{an}) \times \dots \\ & \times P(\text{easy}) \times P(\text{for}) \times P(\text{killing}) \times P(\text{coronavirus}) \end{aligned}$$

3. **Calculating Final Probabilities** So, now, we have to calculate the probabilities for  $P(\text{NewsInput}=\text{REAL})$  &  $P(\text{NewsInput}=\text{FAKE})$  and comparing both to find the one with the highest probability. But again a problem occurs, i.e., if a certain word is not present in the dataset so its probability again rounds up to zero, again giving us probability as zero for both REAL & FAKE.
4. **Overcoming The Problem Using Laplace Smoothing** Laplace smoothing is a strategy for smoothing clear cut information. A little example revision, or pseudotally, will be consolidated in each likelihood gauge. Subsequently, no likelihood will be zero. This is a method of regularizing Naive Bayes, and when the pseudotally is zero, it is called Laplace smoothing.

Estimator is calculated as follows:

$$\begin{aligned} \hat{\theta}_i &= \frac{x_i + \alpha}{N + \alpha d} \quad (i = 1, \dots, d), \\ \alpha &= \text{smoothing parameter} \\ d &= \text{number of unique words in dataset} \\ N &= \text{number of unique words in particular category} \end{aligned}$$

Using Laplace smoothing, it eliminates the problem of 0 probability, and so now the probabilities of individual words can be found out using the formula and then by calculating  $P(\text{NewsInput}=\text{REAL})$  &  $P(\text{NewsInput}=\text{FAKE})$  and comparing them and selecting the outcome with the higher probability will give the answer to which category the news belongs.

After calculating for the NewsInput, we get  $P(\text{NewsInput}=\text{REAL})$  higher than  $P(\text{NewsInput}=\text{FAKE})$ , so we can come to the conclusion from this that the NewsInput is REAL.

## 4.2 Improvising the Bayes Classifier

Using few techniques, we can improvise the Bayes classifier, for getting more accurate, precise, and faster results. The methods proposed in this paper are:

1. Removal of stopwords with NLTK
2. Stemming words with NLTK
3. Tokenize.

**Removal Of Stopwords With NLTK** Stopwords are the words that are used between sentences, words like “a, an, the, for”. These words can be ignored by the system or machine for further analysis of the input given to a particular algorithm. So the stopwords were first removed from the dataset of the news articles created for further training purpose.

**Stemming words with NLTK** Stemming of words means to find out a root word for set of words just in different tenses but having the same meaning. So the stemming of the words was done then and stored.

**Tokenize** The rest of the data after the removal of the stopwords and stemming of the words was then stored as tokens. These tokens were then used as the final dataset for the training of the model.

## 5 Experimental Results

The algorithm was tested on Jupyter Notebook v6.0.3 and the dataset used for training and testing was a self-created dataset named “Train COVID” (Tables 2 and 3).

We were able to obtain a classification accuracy of 76.37% with a maximum accuracy of 85% (Fig. 2).

## 6 Advantages

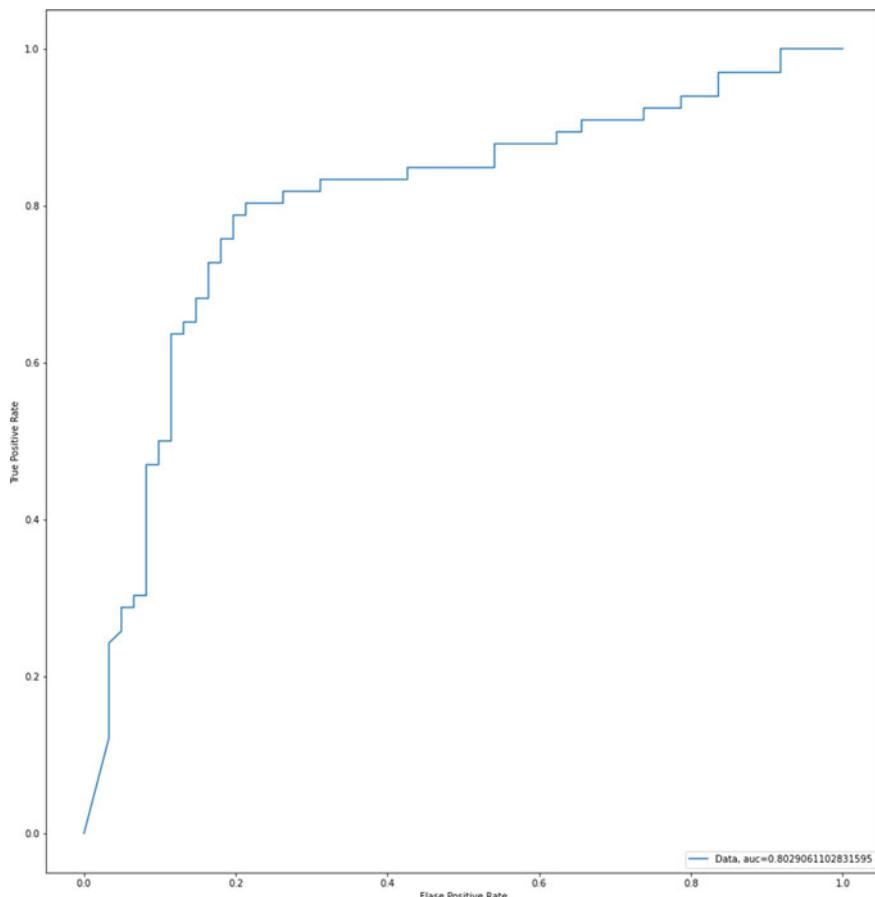
- Naive Bayes algorithm is easy to use.

**Table 2** Characteristics of the system

S. No.	Parameters	Value
1	N-grams range	3
2	Population size	640
3	Terminal set	2 attributes of COVID news dataset
4	Threads in parallel	3

**Table 3** Confusion matrix

Actual Value	Predicted Values	
	REAL	FAKE
REAL	True Positive 52	False Negative 9
FAKE	False Positive 21	True Negative 45
Sensitivity = 71.23% Specificity = 83.33%		

**Fig. 2** Receiver operating characteristics (ROC) plot

- The end product would be able to access from any device.
- Easy to use UI.
- No extra training or knowledge required
- Fast result prediction.
- Model training can be done daily and the software would be kept up to date.

## 7 Disadvantages

- Prediction for news containing quantitative data is not possible.
- The result cannot be 100% accurate as the dataset used to train the model cannot contain all the news articles around the world.

## 8 Conclusion

Maximum news that we find over the social media nowadays are fake. According to a daily research, more than 50% of the news that get's forwarded on WhatsApp is fake. Many tech giants like Google, WhatsApp, and Facebook are coming forward with various ways to stop the spread of the viral fake news.

But as in today's world, people won't stop creating such fake news and forwarding them, the system proposed in this paper comes out as a solution to all these problems. The system proposed here can help people identify which news is trusted and which is not very accurately, stopping the whole cycle of forwarding of viral fake news at a greater extent.

## References

1. Bedi, A., Pandey, N., Khatri, S.K.: A framework to identify and secure the issues of fake news and rumours in social networking. In: 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), Greater Noida, India, pp. 70–73. <https://doi.org/10.1109/PEEIC47157.2019.8976800>
2. Pradhan, M.A., Shinde, A., Dhiman, R., Ghorpade, S., Jawale, S.: Fake news detection methods: machine learning approach. Int. J. Res. Appl. Sci. Eng. Technol. **8**(VII) (IJRASET), 971–975. ISSN: 2321-9653. <https://doi.org/10.22214/ijraset.2020.29630>
3. Gilda, S.: Notice of violation of IEEE publication principles: evaluating machine learning algorithms for fake news detection. In: 2017 IEEE 15th Student Conference on Research and Development (SCoReD), Putrajaya, 2017, pp. 110–115, <https://doi.org/10.1109/SCORED.2017.8305411>
4. Kim, S.-B., Han, K.-S., Rim, H.-C., Hyon Myaeng, S.: Some effective techniques for Naive Bayes text classification. IEEE Trans. Knowl. Data Eng. **18**(11), 1457–1466 (2006). <https://doi.org/10.1109/TKDE.2006.180>
5. Conroy, N.J., Rubin, V.L., Chen, Y.: Automatic deception detection: methods for finding fake news. Proc. Assoc. Inf. Sci. Technol. **52**(1), 14 (2015)

6. Tijare, P.: A study on fake news detection using Naïve Bayes, SVM, Neural Networks and LSTM (2019)
7. Kim, N., Seo, D., Jeong, C.-S.: FAMOUS: fake news detection model based on unified key sentence information. In: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS) (2018)
8. Granik, M., Mesyura, V.: Fake news detection using Naïve Bayes classifier. In: 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON) (2017)
9. Pradhan, M., Bamnote, G.R.: Efficient binary classifier for prediction of diabetes using data preprocessing and support vector machine. In: Satapathy, S., Biswal, B., Udgata, S., Mandal, J. (eds) Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014. Advances in Intelligent Systems and Computing, vol 327. Springer, Cham (2015). [https://doi.org/10.1007/978-3-319-11933-5\\_15](https://doi.org/10.1007/978-3-319-11933-5_15)

# Author Index

## A

- Aali Al, Mansoor, 9  
Abhinith, Bakshi, 505  
Abhishek, P. S. R., 259  
Adilakshmi, T., 67, 241, 259, 269, 505  
Ajay, R., 617  
Ali, Mir Zahed, 269  
Anisha, P. R., 673, 683, 693  
Anitha, A., 545  
Anjana Gowri, G., 617  
Anuradha, T., 617  
Anusha, V., 573  
Aparna, R., 129  
Apoorva, K., 673  
Arya, Rakesh, 383  
Avanthi, M., 193  
Ayyappa Reddy, K., 605

## B

- Balaraju, J., 75  
Bal, Jyotisagar, 515  
Bandu, Sasidhar, 487  
Bansal, Bijender, 107  
Baranidharan, V., 585  
Barge, Yamini, 461  
Basheer, Abdul Adil, 415  
Beena Bethel, G. N., 121  
Bein, Franklin, 735  
Bethu, Srikanth, 573  
Bharamagoudar, Geeta, 201  
Bhasgi, Shivani S., 711  
Bhaskaran, Subhashini Sailesh, 9  
Bhateja, Vikrant, 627  
Bhattacharyya, Akash, 637

Bisht, Shubhangi, 95

Bitra, Surendra Kumar, 25

## C

- Catherine Joy, R., 527  
Chalapati Rao, K. V., 373  
Chandra Sekhar Reddy, N., 141  
Chandra Sekhar Reddy, P., 193  
Charles Babu, G., 573  
Chaudhari, Shilpa Shashikant, 129  
Chordia, Aishwarya, 425  
Chore, Amruta, 765

## D

- Damera, Vijay Kumar, 473  
Dansana, Jayanti, 645  
Datta, Suhrid, 395  
Dattatreya, V., 373  
Deepak Varma, V., 495  
Deshpande, Rashmi, 781  
Devaraj, Sujitha Juliet, 527  
Devaraju, R., 327  
Dhiman, Rohan, 813  
Divya, 107  
Dumala, Anveshini, 347  
Dutta, Arijit, 637, 655

## E

- Edwin Prem Kumar, G., 15

## G

- Garg, Ojas, 41

Gayathri, M., 395  
 Ghorpade, Shreyas, 813  
 Ghose, Smaranjit, 395  
 Giraddi, Shantala, 201  
 Gondhi, Naveen Kumar, 789  
 Gond, Vitthal J., 799  
 Goyal, Deepak Kr., 107  
 Gujarathi, Priyanka Vishwas, 747  
 Gupta, Roopam, 383  
 Gupta, Shelley, 41, 95  
 Gupta, Shirin, 95

**H**

Halkai, Abhijit, 719

**I**

Itkar, Suhasini, 425

**J**

Jagtap, Sarika M., 799  
 Jain, Amita, 87  
 Jain, Minni, 87  
 Jain, Niyati, 107  
 Jalaja, T., 259  
 Janardana Naidu, G., 253  
 Jawale, Deepali, 757  
 Jawale, Swapnil, 813  
 Jeba Priya, S., 285  
 Johri, Archita, 627  
 Joshi, Maulin M., 403  
 Joshua Jaistein, S., 285  
 Jyothirmai, J., 555

**K**

Kale, Mihir, 425  
 Kanade, Vijay A., 307  
 Kanakala, Srinivas, 31  
 Karthik, V., 585  
 Kaushik, Baijnath, 789  
 Kevin, I. C., 527  
 Kishore, Pinninti, 545  
 Kishor Kumar Reddy, C., 673, 683, 693  
 Kumar, Anubhav, 219

**L**

Lakshmi, Boggula, 141  
 Lakshmi Prasuna, A. V., 149  
 Lakshmi, Soanpet Sree, 505  
 Latheef, Jesmi, 363

Lukose, Jibin, 415  
 Lydia, M., 15

**M**

Madana Mohana, R., 693  
 Madan Kumar, C., 605  
 Madgi, Manohar, 201  
 Madhavi Latha, Makkena, 545  
 Madhuravani, B., 141  
 Madhur, M. S., 201  
 Malathy, C., 395  
 Malik, Sandeep, 757  
 Manjulatha, B., 337  
 Mayee, Mukta, 425  
 Mehrotra, Radhika, 41  
 Mishra, Swagatika, 595  
 Misra, Chinmaya, 637, 655  
 Mittal, Harish, 107  
 Moghe, Asmita A., 383  
 Moorthy, Priyanka, 353, 495  
 Moulieshwaran, B., 585

**N**

Nabi, Shaik Abdul, 159  
 Nagaratna, M., 473  
 Nagesh, A., 473  
 Naisal, S. A., 663  
 Nakrani, Naitik M., 403  
 Naveen Sundar, G., 285  
 Navya, R., 327  
 Nikitha, M., 487  
 Nukala, Chaitanya, 149

**P**

Pabboju, Shyam Sunder, 241  
 Pabboju, Suresh, 337  
 Padmaja, S., 487  
 Pal, Babita, 627  
 Pal, Deepika, 627  
 Panda, Ganapati, 595  
 Pandey, Manjusha, 451  
 Papasani, Anusha, 347  
 Patidar, Hemant, 773, 781  
 Patil, G. A., 51  
 Patil, Sandip Raosaheb, 747  
 Patra, Sudhangshu Sekhar, 637  
 Paul, Varghese, 317, 415  
 Pavan Kumar, C. S., 617  
 Pavan Kumar, N., 353  
 Pradhan, Madhavi Ajay, 813  
 Prasada Rao, P. V. R. D., 75

Prasad, C. V. P. R., 555  
 Prashanthi, Vempaty, 31  
 Praveen Kumar Reddy, A., 495  
 Preethi Lalithya, G., 617  
 Priyadarshan, Pradosh, 595  
 Priyadarshini, Aishwarya, 227  
 Purohit, Lalit, 209, 461  
 Purushotham Reddy, M., 495  
 Puttamadappa, C., 535

**R**

Raghava, M., 373, 555  
 Raja Sundrapandiyanleebanon, T., 285  
 Rajesh, Sreeja, 415  
 Ramadevi, Y., 605  
 Rama Narasimham, K. B. V., 555  
 Rao, Mekala Srinivasa, 169  
 Rath, Adyasha, 595  
 Ratheesh, T. K., 317  
 Rath, Manas Kumar, 515  
 Rautray, Siddharth Swarup, 451  
 Ravi, Kumar, 227  
 Ravinder Reddy, B., 67  
 Ravinder Reddy, R., 605  
 Ravi Teja, P., 617  
 Ravi, Vadlamani, 227  
 Reji Kumar, K., 663, 729  
 Roopa, V., 1  
 Roy, Chandrima, 451  
 Roy, Ruben, 655

**S**

Saha, Soma, 461  
 Sameen Fatima, S., 487, 565  
 Sanjay, R., 585  
 Sankara Babu, B., 573  
 Santhosh, I., 527  
 Sarangi, Prateek, 595  
 Sarraju, Sridevi, 735  
 Sashi Kumar, M. S. V., 441  
 Satapathy, Suresh Chandra, 627  
 Sateesh Kumar, R., 565  
 Saxena, Divya, 219  
 Seshashayee, M., 253  
 Shailaja, Varagiri, 149  
 Sharma, Gaurav, 87  
 Sharma, Reya, 789  
 Shinde, Ankita, 813  
 Shinde, Sandhya, 773  
 Shiva Krishna, R. M., 441

Shrivatsava, Poornima, 383  
 Shrivastava, Sonika, 209  
 Shyamala Devi, M., 353, 495  
 Sindhu, S., 1  
 Singh, Adarsh, 645  
 Singh, Archana, 41  
 Singh, Kamakhya, 655  
 Singh, Manoj Kumar, 535  
 Sireesha, V., 441  
 Sirisati, Ranga Swamy, 169, 183  
 Sree Sandhya, N., 121  
 Sridhar, M., 25  
 Srinivasa Kumar, C., 183  
 Srinivasa Rao, P. S. V., 169  
 Srinivas, J., 295  
 Sriram Karthik, M., 1  
 Sunitha, M., 269  
 Sunny Deol, G. J., 295  
 Suresh, Anitha, 535  
 Swain, Prasanta Kumar, 515  
 Swathi, P., 353, 495  
 Swetha, B., 149

**T**

Terdal, Sujatha, 711, 719  
 Thangabalaji, V., 585  
 Thankachan, Dolly, 765  
 Thonukunuri, Srinivasulu, 183  
 Triveni, B., 545

**V**

Vaishnavi, N., 1  
 Vankar, Madhura D., 51  
 VaraPrasada Rao, P., 295  
 Vardhan, Mettu Krishna, 159  
 Varma Tungala, Ravi, 353  
 Vasikaran, R., 1  
 Venkata Subba Reddy, K., 295  
 Vikkurty, Sireesha, 347  
 Vinay Kumar, S., 441  
 Vineetha, S., 363  
 Vivekanandan, Saranya, 353, 495

**Y**

Yadav, Preksha, 425  
 Yasmeen, Nuzhat, 683  
 Yindumathi, K. M., 129