

Vinit Kumar Gunjan
Jacek M. Zurada *Editors*

Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications

ICMISC 2020

Advances in Intelligent Systems and Computing

Volume 1245

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,
Warsaw, Poland

Advisory Editors

Nikhil R. Pal, Indian Statistical Institute, Kolkata, India

Rafael Bello Perez, Faculty of Mathematics, Physics and Computing,
Universidad Central de Las Villas, Santa Clara, Cuba

Emilio S. Corchado, University of Salamanca, Salamanca, Spain

Hani Hagras, School of Computer Science and Electronic Engineering,
University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University,
Gyor, Hungary

Vladik Kreinovich, Department of Computer Science, University of Texas
at El Paso, El Paso, TX, USA

Chin-Teng Lin, Department of Electrical Engineering, National Chiao
Tung University, Hsinchu, Taiwan

Jie Lu, Faculty of Engineering and Information Technology,
University of Technology Sydney, Sydney, NSW, Australia

Patricia Melin, Graduate Program of Computer Science, Tijuana Institute
of Technology, Tijuana, Mexico

Nadia Nedjah, Department of Electronics Engineering, University of Rio de Janeiro,
Rio de Janeiro, Brazil

Ngoc Thanh Nguyen, Faculty of Computer Science and Management,
Wrocław University of Technology, Wrocław, Poland

Jun Wang, Department of Mechanical and Automation Engineering,
The Chinese University of Hong Kong, Shatin, Hong Kong

The series “Advances in Intelligent Systems and Computing” contains publications on theory, applications, and design methods of Intelligent Systems and Intelligent Computing. Virtually all disciplines such as engineering, natural sciences, computer and information science, ICT, economics, business, e-commerce, environment, healthcare, life science are covered. The list of topics spans all the areas of modern intelligent systems and computing such as: computational intelligence, soft computing including neural networks, fuzzy systems, evolutionary computing and the fusion of these paradigms, social intelligence, ambient intelligence, computational neuroscience, artificial life, virtual worlds and society, cognitive science and systems, Perception and Vision, DNA and immune based systems, self-organizing and adaptive systems, e-Learning and teaching, human-centered and human-centric computing, recommender systems, intelligent control, robotics and mechatronics including human-machine teaming, knowledge-based paradigms, learning paradigms, machine ethics, intelligent data analysis, knowledge management, intelligent agents, intelligent decision making and support, intelligent network security, trust management, interactive entertainment, Web intelligence and multimedia.

The publications within “Advances in Intelligent Systems and Computing” are primarily proceedings of important conferences, symposia and congresses. They cover significant recent developments in the field, both of a foundational and applicable character. An important characteristic feature of the series is the short publication time and world-wide distribution. This permits a rapid and broad dissemination of research results.

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

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

Vinit Kumar Gunjan · Jacek M. Zurada
Editors

Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications

ICMISC 2020



Springer

Editors

Vinit Kumar Gunjan
Department of Computer Science
and Engineering
CMR Institute of Technology
Hyderabad, India

Jacek M. Zurada
Department of Electrical
and Computer Engineering
University of Louisville
Louisville, KY, USA

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-15-7233-3

ISBN 978-981-15-7234-0 (eBook)

<https://doi.org/10.1007/978-981-15-7234-0>

© Springer Nature Singapore Pte Ltd. 2021

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

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

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

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

Preface

Artificial intelligence (AI) and machine learning (ML) are the emerging technologies that are moving organizations faster than ever before. In this era of digital transformation, the success is based on using analytics to find huge amount of data with close insights. Historically, these insights were discovered manually through in-depth data analytics—but data complexity continues to increase, as does the complexity of data. AI and ML are the latest tools for data scientists, enabling them to rapidly refine the data to ensure its value.

The term ‘smart cities’ first gained traction back in the 1990s, when it was adopted as a way to illustrate the use of technology and innovation in urban development. Since then, the world has become increasingly urbanized. By 2020, 55% of the world’s population lives in urban areas, and the number is expected to increase to around 70% by 2050. This rapid urbanization will put increasing pressure on resources and address the demand for intelligent and sustainable environments. The requirements of citizens providing a higher quality of life will increase.

We are entering into a new era of computing technology that many are calling as the Internet of Things (IoT). Machine to Machine, Machine to Infrastructure, Machine to Environment, Internet of Everything, Internet of Intelligent Things, Intelligent System—call it what you want, but it is happening, and its potential is huge.

This book is comprised of selected and presented papers of the International conference on Recent Trends in Machine Learning, IOT, Smart Cities & Applications, 2020. It consists of selected manuscripts arranged on the basis of their approaches and contributions to the scope of the conference. The chapters of this book present key algorithms and theories that form the core of the technologies and applications concerned, consisting mainly of face recognition, evolutionary algorithms such as genetic algorithms, automotive applications, automation devices with artificial neural networks, business management systems, IoT, machine learning, data science and modern speech processing systems. This book also

covers recent advances in medical diagnostic systems, sensor networks and systems of VLSI domain. Discussion of learning and software modules in deep learning algorithms is added wherever suitable. In a nut shell the book will bring insights to the modern advancements involved in smart cities with an IoT and ML approach.

Hyderabad, India
Louisville, USA

Vinit Kumar Gunjan
Jacek M. Zurada

Contents

Automated Identification of Interictal Activity from EEG Signal Using Non-linear Features	1
Arshpreet Kaur, Karan Verma, Amol P. Bhondekar, and Kumar Shashvat	
A Survey on Phishing URL Detection Using Artificial Intelligence	9
Arpita Vadariya and Nilesh Kumar Jaday	
Prognosticating Liver Debility Using Classification Approaches of Machine Learning	21
Revelly Akshara and Sandhi Kranthi Reddy	
Robust UI Automation Using Deep Learning and Optical Character Recognition (OCR)	33
Mithilesh Kumar Singh, Warren Mark Fernandes, and Mohammad Saad Rashid	
Load Scheduling with Combinations of Existing Tariff Structure for Residential Consumers in Maharashtra, India—Case Study	45
Archana Talhar and Sanjay Bodkhe	
Mammogram Classification Using Rotation-Invariant Local Frequency Features	55
Spandana Paramkusham and C. Venkata Narasimhulu	
Internet of Things (IOT) Architecture—A Review	67
Achinta K. Palit	
Robust and Lightweight Control System for IoT Networks: Enabling IoT for the Developing World	73
Jithu G. Panicker and Mohamed Azman	
Real-Time Detection and Prediction of Heart Diseases from ECG Data Using Neural Networks	93
K. V. Sai Kiran, Mohamed Azman, Eslavath Nandu, and S. K. L. V. Sai Prakash	

Gender Classification Based on Fingerprint Database Using Association Rule Mining	121
Ashish Mishra, Shivendu Dubey, and Amit Sahu	
Cyber Terrorism-Related Multimedia Detection Using Deep Learning—A Survey	135
Himani Mandaviya and Snehal Sathwara	
The Role of Technologies on Banking and Insurance Sectors in the Digitalization and Globalization Era—A Select Study	145
Venkamaraju Chakravaram, Sunitha Ratnakaram, Nitin Simha Vihari, and Neelakantam Tatikonda	
Review of Recent Plagiarism Detection Techniques and Their Performance Comparison	157
Manpreet Kaur, Vishal Gupta, and Ravreet Kaur	
A Combination of 2DLDA and LDA Approach for Fruit-Grade Classification with KSVM	171
Yogeswararao Gurubelli, Malmathanraj Ramanathan, and Palanisamy Ponnusamy	
Facial Expression Extraction and Human Emotion Classification Using Convolutional Neural Network.....	179
Amruta Khot and Anmol Magdum	
Exploring Opportunities in Hydro Electric Power Plant with Heron's Fountain	189
Khude Anupam Tanaji and Patil Manoj Dhondiram	
A Comprehensive Survey on Application Layer Protocols in the Internet of Things	203
Hemant Sharma, Ankur Gupta, and Madhavi Latha Challa	
Empirical Laws of Natural Language Processing for Hindi Language	217
Arun Babhulgaonkar, Mahesh Shirasath, Atharv Kurdukar, Hrishikesh Khandare, Adwait Tekale, and Manali Musale	
Classification Method to Predict Chances of Students' Admission in a Particular College	225
Danny Joel Devarapalli	
Doppler Shift Based Sampling Rate Conversion for GFDM Underwater Acoustic Communication	239
V. S. Kumar, A. Chakradhar, M. Shiva Prasad, and Y. Shekar	

A Novel Recommendation System for Housing Search: An MCDM Approach	251
Shreyas Das, Swastik Ghosh, Bhabani Shankar Prasad Mishra, and Manoj Kumar Mishra	
Health-Related Tweets Classification: A Survey	259
Kothuru Srinivasulu	
Smart Farming	269
Nihar Sardal, Ankit Patel, and Vinaya Sawant	
Solar-Powered Smart Agriculture and Irrigation Monitoring/Control System over Cloud—An Efficient and Eco-friendly Method for Effective Crop Production by Farmers in Rural India	279
Syed Musthak Ahmed, B. Kovela, and Vinit Kumar Gunjan	
Exploration of Classification Algorithms for Divorce Prediction	291
Danussvar Jayanthi Narendran, R. Abilash, and B. S. Charulatha	
Pronunciation Similarity Matching Using Deep Learning	305
Ashish Upadhyay, Bhupendra Kumar Sonwani, Vimal Anand Baghel, Yash Kirti Sinha, Ashish Singh Patel, and Muneendra Ojha	
Noise Reduction in SAR Images with Variable Mode CT	315
R. Durga Bhavani, A. Ravi, and N. Mounika	
Modeling IoT Based Automotive Collision Detection System Using Support Vector Machine	323
Nikhil Kumar, Debopam Acharya, and Divya Lohani	
Optimization-based Resource Allocation for Cloud Computing Environment	333
M. Chidambaram and R. Shanmugam	
Hardware Trojan Detection Using Deep Learning-Deep Stacked Auto Encoder	345
R. Vishnupriya and M. Nirmala Devi	
IOT and Intelligent Asthmatics Monitoring Sensors—A Literature Survey	355
Aditya Bothra, Saumya Bansal, and Surender Dhiman	
Exploring in the Context of Development of Smart Cities in India	365
Smita Bharne and Suryakant Patil	
Microservices and DevOps for Optimal Benefits from IoT in Manufacturing	375
Anurag Choudhry and Anshu Premchand	

Semantic Interoperability for IoT Agriculture Framework with Heterogeneous Devices	385
P. Salma Khatoon and Muqeem Ahmed	
Boosting Approach for Multiclass Fake News Detection	397
Rajkamal Karedula and Pradeep Singh	
A Sentiment-Based Recommender System Framework for Social Media Big Data Using Open-Source Tech Stack	407
Shini Renjith, Mable Biju, and Monica Merin Mathew	
Hardware Trojan Detection Using Machine Learning Technique	415
Nikhila Shri Chockaiah, S. K. Swetha Kayal, J. Kavin Malar, P. Kirithika, and M. Nirmala Devi	
An Application Suite: Effectiveness in Tracking and Monitoring of Skill Training Programs	425
Balu M. Menon, P. Aswathi, and Shekar Lekha	
Automated Water Management System Using Internet of Things	435
Ritik Gupta, B. Shivalal Patro, and Manas Chandan Behera	
Model-Based Observer Performance Study for Speed Estimation of Brushed DC Motor with Uncertain Contact Resistance	441
Sayantan Moulik and Biswajit Halder	
Severity Prediction of Software Vulnerabilities Using Textual Data	453
Ruchika Malhotra and Vidushi	
Data Analysis of Cricket Score Prediction	465
Suyoga Srinivas, Naveen N. Bhat, and M. Revanasiddappa	
Anchor-Based Effective Node Localization Algorithm for Wireless Sensor Networks	473
Basavaraj M. Angadi and Mahabaleshwar S. Kakkasageri	
A Survey on Security on Medical Data and Images in Healthcare Systems	481
Swarnali Sadhukhan, Mihir Sing, Koushik Majumder, Santanu Chatterjee, and Subhanjan Sarkar	
Pothole and Speed Bump Classification Using a Five-Layer Simple Convolutional Neural Network	491
Anju Thomas, P. M. Harikrishnan, J. S. Nisha, Varun P. Gopi, and P. Palanisamy	
Smart Ecosystem to Facilitate the Elderly in Ambient Assisted Living	501
Ashish Patel and Jigarkumar Shah	

Data-Driven Stillbirth Prediction and Analysis of Risk Factors in Pregnancy	511
Aravind Unnikrishnan, K. Chandrasekaran, and Anupam Shukla	
A Comparative Study of Heuristics and Meta-heuristic Algorithms with Issues in WSN	525
Prince Rajpoot, Kumkum Dubey, Nisha Pal, Ritika Yaduvanshi, Shivendu Mishra, and Neetu Verma	
A Systematic Review on an Embedded Web Server Architecture	533
Suman Kumar Panday, R. V. V. Krishna, and Durgesh Nandan	
Water Sharing Marketplace Using IoT	543
Dendukuri Ravi Kiran, Aki Rohith, Kothur Dinesh Reddy, and G Pradeep Reddy	
Augmenting the Existing CBPM Maintenance Philosophy Currently in Use in the Marine Sector with Intelligent Predictive—CBPM Philosophy	553
Minakshi Gautam, Vaishnavi S. Ramu, Sachin Sinha, Pranay Kumar Reddy, Monica Kondur, and S. Suresh Kumar	
An Investigation on Rolling Element Bearing Fault and Real-Time Spectrum Analysis by Using Short-Time Fourier Transform	561
M. Siva Santhoshi, K. Sharath Babu, Sanjeev Kumar, and Durgesh Nandan	
A Review of 4-2 Compressors: Based on Accuracy and Performance Analysis	569
P. Venkata Ganesh, E. Jagadeeswara Rao, and Durgesh Nandan	
Emotion Recognition Using Chatbot System	579
Shraddha Pophale, Hetal Gandhi, and Anil Kumar Gupta	
A Hybrid and Improved Isolation Forest Algorithm for Anomaly Detection	589
G. Madhukar Rao and Dharavath Ramesh	
Ranger Random Forest-Based Efficient Ensemble Learning Approach for Detecting Malicious URLs	599
G. Madhukar Rao and Dharavath Ramesh	
Smart Camera for Traffic Control by Sing Machine Learning	609
Priya Tiwari, Santosh Jagtap, and Dattatray Bade	
Systematic Review on Full-Subtractor Using Quantum-Dot Cellular Automata (QCA)	619
Sri Sai Surya, A. Arun Kumar Gudivada, and Durgesh Nandan	
Intelligent Resource Identification Scheme for Wireless Sensor Networks	627
Gururaj S. Kori and Mahabaleshwar S. Kakkasageri	

A Systematic Review on Various Types of Full Adders	635
D. Dhathri, E. Jagadeeswara Rao, and Durgesh Nandan	
Design and Implementation of Smart Real-Time Billing, GSM, and GPS-Based Theft Monitoring and Accident	
Notification Systems	647
B. Jyothi Priya, Parvateesam Kunda, and Sanjeev Kumar	
Authentication of Vehicles in Vehicular Clouds:	
An Agent-Based Approach	663
Shailaja S. Mudengudi and Mahabaleshwar S. Kakkasageri	
ANN-Based Model to Predict Reference Evapotranspiration for Irrigation Estimation	671
Neha K. Nawandar, Naveen Cheggoju, and Vishal Satpute	
Entropy: A New Parameter for Image Deciphering	681
Naveen Cheggoju, Neha K. Nawandar, and Vishal R. Satpute	
A Systematic Review of Approximate Adders: Accuracy and Performance Analysis	689
M. Lakshmi Akhila, E. Jagadeeswara Rao, R. V. V. Krishna, and Durgesh Nandan	
A Review Paper Based on Image Security Using Watermarking	697
V. Ch. S. Ravi Shankar, R. U. S. D. Vara Prasad, Rama Vasantha Adiraju, R. V. V. Krishna, and Durgesh Nandan	
Smart Healthcare Analytics Solutions Using Deep Learning AI	707
K. P. Subiksha and M. Ramakrishnan	
Design of 32—Bit MAC Unit Using Vedic Multiplier and XOR Logic	715
Aki Vamsi Krishna, S. Deepthi, and M. Nirmala Devi	
Deep Learning Model for Detection of Attacks in the Internet of Things Based Smart Home Environment	725
Raveendranadh Bokka and Tamilselvan Sadasivam	
Smart Irrigation Using Decision Tree	737
Chinmay Patil, Shubham Aghav, Sagar Sangale, Shubham Patil, and Jayshree Aher	
Contextually Aware Multimodal Emotion Recognition	745
Preet Shah, Patnala Prudhvi Raj, Pragnya Suresh, and Bhaskarjyoti Das	
A Machine Learning Based Approach for Prediction of Actual Landing Time of Scheduled Flights	755
S. Deepudev, P. Palanisamy, Varun P. Gopi, and Manjunath K. Nelli	

Data Integrity and Security in Distributed Cloud Computing—A Review	767
Abdullatif Ghallab, Mohammed H. Saif, and Abdulqader Mohsen	
Independent Learning of Motion Parameters for Deep Visual Odometry	785
Rahul Kottath, Rishab Kaw, Shashi Poddar, Amol P. Bhondekar, and Vinod Karar	
Smart Street Lights to Reduce Death Rates from Road Accidents	795
Rajat, Naresh Kumar, and Manoj Sharma	
An Improved Approach for Face Detection	811
C. A. Rishikeshan, C. Rajesh Kumar Reddy, and Mohan Krishna Varma Nandimandalam	
Efficient Band Offset Calculation Method for HEVC and Its VLSI Implementation	817
I. Manju, K. S. Srinivasan, E. Rohith Kumar, and R. Haresh	
Clinical Skin Disease Detection and Classification: Ensembled VGG	827
Gogineni Saikiran, G. Surya Narayana, Dhanrajnath Porika, and Gunjan Vinit Kumar	
A Review of Smart Greenhouse Farming by Using Sensor Network Technology	849
D. Chaitanya Kumar, Rama Vasantha Adiraju, Swarnalatha Pasupuleti, and Durgesh Nandan	
A Study on Low-Frequency Signal Processing with Improved Signal-to-Noise Ratio	857
G. Pavan Avinash, P. Ramesh Kumar, Rama Vasantha Adiraju, and Durgesh Nandan	
Factors that Determine Advertising Evasion in Social Networks	865
Jesús Silva, Yisel Pinillos-Patiño, Harold Sukier, Jesús Vargas, Patricio Corrales, Omar Bonerge Pineda Lezama, and Benjamín Quintero	
Classification of Academic Events from Their Textual Description	875
Jesús Silva, Nicolas Elias María Santodomingo, Ligia Romero, Marisol Jorge, Maritza Herrera, Omar Bonerge Pineda Lezama, and Francisco Javier Echeverry	
Geosimulation as a Tool for the Prevention of Traffic Accidents	883
Amelec Viloria, Noel Varela, Luis Ortiz-Ospino, and Omar Bonerge Pineda Lezama	

Identification of Author Profiles Through Social Networks	893
Jesús Silva, Nicolas Elias Maria Santodomingo, Ligia Romero, Marisol Jorge, Maritza Herrera, Omar Bonerge Pineda Lezama, and Francisco Javier Echeverry	
Real Road Networks on Digital Maps with Applications in the Search for Optimal Routes	901
Amelec Viloria, Noel Varela, David Ovallos-Gazabon, Omar Bonerge Pineda Lezama, Alberto Roncallo, and Jairo Martinez Ventura	
Design of a Network with Sensor-Cloud Technology Applied to Traffic Accident Prevention	911
Amelec Viloria, Noel Varela, Yaneth Herazo-Beltran, and Omar Bonerge Pineda Lezama	
Comparison of Bioinspired Algorithms Applied to Cancer Database	921
Jesús Silva, Reynaldo Villareal-González, Noel Varela, José Maco, Martín Villón, Freddy Marín-González, and Omar Bonerge Pineda Lezama	
Indicators for Smart Cities: Tax Illicit Analysis Through Data Mining	929
Jesús Silva, Darwin Solano, Claudia Fernández, Lainet Nieto Ramos, Rosella Urdanegui, Jeannette Herz, Alberto Mercado, and David Ovallos-Gazabon	
Classification, Identification, and Analysis of Events on Twitter Through Data Mining	939
Jesús Silva, Pedro Berdejo, Yuki Higa, Juan Manuel Cera Visbal, Danelys Cabrera, Alexa Senior Naveda, Yasmin Flores, and Omar Bonerge Pineda Lezama	
Algorithm for Detecting Polarity of Opinions in University Students Comments on Their Teachers Performance	949
Jesús Silva, Edgardo Rafael Sanchez Montero, Danelys Cabrera, Ramon Chacon, Martin Vargas, Omar Bonerge Pineda Lezama, and Nataly Orellano	
Prediction of the Efficiency for Decision Making in the Agricultural Sector Through Artificial Intelligence	959
Amelec Viloria, Alex Ruiz-Lazaro, Ana Maria Echeverría González, Omar Bonerge Pineda Lezama, Juan Lamby, and Nadia Leon Castro	
Model for Predicting Academic Performance in Virtual Courses Through Supervised Learning	967
Jesús Silva, Evereldys Garcia Cervantes, Danelys Cabrera, Silvia García, María Alejandra Binda, Omar Bonerge Pineda Lezama, Juan Lamby, and Carlos Vargas Mercado	

Contents	xv
Genetic System for Project Support with the Sequencing Problem	977
Amelec Viloria, Noel Varela, Carlos Herazo-Beltran, Omar Bonerge Pineda Lezama, Alberto Mercado, Jairo Martinez Ventura, and Hugo Hernandez Palma	
Method for the Recovery of Images in Databases of Skin Cancer	985
Amelec Viloria, Noel Varela, Narledys Nuñez-Bravo, and Omar Bonerge Pineda Lezama	
Author Index	995

About the Editors

Vinit Kumar Gunjan is an Associate Professor of Computer Science & Engineering at CMR Institute of Technology Hyderabad (Affiliated to Jawaharlal Nehru Technological University, Hyderabad). He is an active researcher, and has published papers at IEEE, Elsevier & Springer conferences, authored several books, and edited volumes of Springer series, most of which are indexed in the SCOPUS database. In 2016, he received the prestigious Early Career Research Award from the Science Engineering Research Board, Department of Science & Technology, Government of India. He is a senior member of IEEE, and is active in the IEEE Hyderabad section. He is currently the Secretary of the Computational Intelligence Society, and has served as the Treasurer, Secretary & Chairman of IEEE Young Professionals Affinity Group & IEEE Computer Society.

Jacek M. Zurada (SM'85, F'96, LF'14) is a Professor of Electrical and Computer Engineering and Director of the Computational Intelligence Laboratory at the University of Louisville, Kentucky, USA, where he served as Department Chair and a Distinguished University Scholar. He was a Professor at Princeton University; Northeastern University; Auburn University; the National University of Singapore; Nanyang Technological University in Singapore; the Chinese University of Hong Kong; the University of Chile, Santiago; Toyohashi University of Technology, Japan; the University of Stellenbosch, South Africa; and the University of Marie-Curie, Paris, France, and was a Postdoctoral Fellow at the Swiss Federal Institute of Technology, Zurich, Switzerland.

Automated Identification of Interictal Activity from EEG Signal Using Non-linear Features



Arshpreet Kaur, Karan Verma, Amol P. Bhondekar, and Kumar Shashvat

Abstract Analysis of EEG (Electroencephalography) for the detection of interictal activity amidst of artefacts for supporting the diagnosis of epilepsy is a time-consuming process that requires high expertise and experienced neurologist. The objective of the work is automated identification interictal activity in order to assist neurologists and also reduce the time consumed in visual inspection. For this work, four cases distinguishing interictal from a controlled activity are considered from the Bonn database. Non-linear properties from the complete signal such as correlation dimensions and properties such as approximate entropy, sample entropy and fuzzy approximate entropy from the specific sub-bands of frequency range A5 (0–2.7 Hz), D5 (2.71–5.4 Hz), D4 (5.4–10.8 Hz), D3 (10.85–21.7 Hz), and D2 (21.7–43.4 Hz) are used. Backpropagation neural network is used as a classifier. Performance parameters accuracy, sensitivity, and specificity are calculated. Ten-fold cross method is used as a validation method. For case 1, case 2, case 3, and case 4, the highest accuracy of 99.9%, 99%, 98.5%, and 99% has been achieved, respectively. The problem of interictal identification activity still remains unmapped to some extent this work focuses on the same and in this work we have achieved good performance measures in context to the same.

A. Kaur (✉) · K. Verma
National Institute of Technology, New Delhi, Delhi, India
e-mail: arshpreet@nitdelhi.ac.in

K. Verma
e-mail: karanverma@nitdelhi.ac.in

A. P. Bhondekar
CSIR-Central Scientific Instruments Organization, Chandigarh, India
e-mail: amolbhondekar@gmail.com

K. Shashvat
Bharti Vidya Peeth's College of Engineering, New Delhi, Delhi, India
e-mail: shashvat.sharma13@gmail.com

Keywords Epilepsy · Discrete wavelet transform · Back propagation neural network · Sample entropy · Approximate entropy · Fuzzy approximate entropy · Correlation dimension

1 Introduction

Epilepsy is a neurological condition affecting the nervous system of a person and is prevalent in world's 1% population [1]. It is the second-most common neurological disorder. EEG is a tool used by neurologists to support the diagnosis of this disease. EEG examination last from twenty minutes to hour or even longer for patients admitted for continuous monitoring. Generally, abnormalities (interictal activity) are looked for in an EEG admits of various physiological and non-physiological artefacts which may include a bad electrode, eye-flutter, eye blink artefact, etc. Researchers have been developing methodologies to understand and interpret EEG brain waves form many decades not only to support a diagnosis of various diseases but to understand sleep patterns [2] and syndromes [3], correlate them with a mood of people and even understand effect of different types of music on people [4]. Among features used to study EEG, nonlinear features are popular among researchers which can be attributed to the non-stationary nature of EEG. In [5] Babloyantz et al. (1985) used LLE (Lyapunov exponents) and CD(correlation dimension) to study sleep wave signals of sleep stage two and stage four and presented that some stages of sleep are characterized by chaotic attractors. In [6] Acharya et al. (2012) used approximate entropy, sample entropy phase entropy 1 and phase entropy 2 as features and fed them into seven classifiers out which fuzzy classifier performed the best classifying between (normal)-(pre-ictal)-(ictal) class with 98.1% classification accuracy. In [7] Acharya et al. (2015) studied ten entropy features and RQA and compared these for various two-class problems in which controlled and ictal were compared and interictal and ictal were considered for three-class problem controlled, interictal and ictal were considered the researcher concluded that entropies are a state of art features in understanding EEG states. In [8] Tawfik et al. (2016) used WPE (Weighted Permutation Entropy) and Support vector machine classifier for three cases (Z-S), (Z-F-S) and (Z-O-N-F-S) for which classification accuracy of 99%, 97.5%, 93.75% was achieved. In [9] Acharya et al. (2019) studied use of different types of nonlinear features which are entropy-based such as fractal dimension, Hjorth, LLE, Reoccurrence qualitative analysis, Lempel-ziv complexity are discussed; the study concluded that the nonlinear features hold the capability to detect minute changes in EEG from focal and non-focal classes and have the capability to be used in clinical settings. Though much work has been done understanding and classifying between various EEG states but identifying interictal activity from the activity of controlled patients has the potential to be explored fully. In this work, we have used Bonn data from [10], which is a publically available data set. Figure 1 depicts various sets in the available data. The data set has five sets corresponding to five states; set A and set B are taken from a controlled group. Set A with eyes open and Set B with eyes closed.

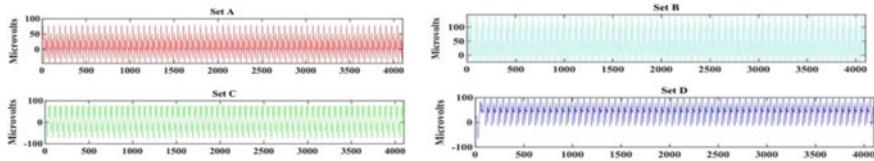


Fig. 1 Plots of different sets used

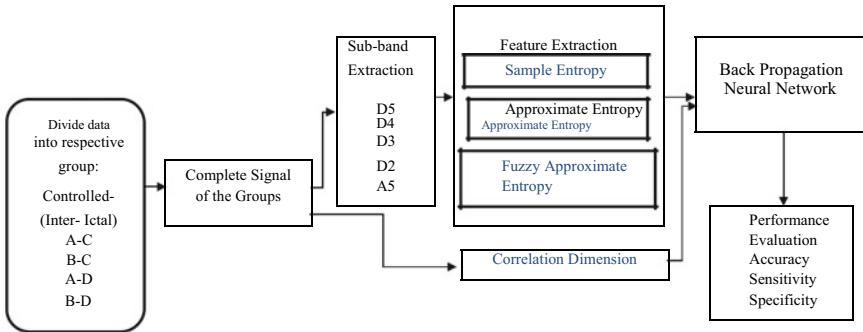


Fig. 2 Work methodology

Set C and Set D hold interictal discharges from five patients from the epileptic zone in set C and opposite to epileptic zone in set D. Set E contains ictal activity. In this work, our focus is to distinguish controlled activity from interictal activity. For this work, we have considered four cases under group controlled and interictal; which are A-C, B-C, A-D, and B-D (Fig. 2).

2 Methodology

2.1 Data Details and Group Division

For this work the objective was to classify between data taken from controlled patients and data taken from people with interictal activity; therefore, a total of four cases were considered. Total of 100 files was present in each set. Each signal was recorded for 23.6 s at a sampling frequency of 173.6 Hz and hence had 4096 data points.

Table 1 EEG sub-band division

Sub-band	Preprocessing (Hz)
D2	21.7–43.4
D3	10.85–21.7
D4	5.4–10.8
D5	2.71–5.4
A5	0–2.71

2.2 Data Preprocessing

All the datasets considered in this work are divided into five sub-bands; ‘db4’ wavelet is used for the sub-band division. Table 1 depicts the specific frequency range.

Feature Extraction:

For this work, three entropy-based features are Approximate Entropy, Sample Entropy, and fuzzy approximate entropy are extracted out of the five sub-bands used mentioned in Table 1. Correlation dimension is extracted out of complete signal. Total 16 features are used to train the model. These entropy-based features have been also used by previous researches either independently or with combination with other features and have been elaborately explained in [11–15]. For this work, the value of parameters used are as follows N is taken to be 4097, m is embedding dimension; it is set to two and r is vector comparison distance. It is set to 0.2 times the standard deviation of the signal. Correlation dimension measure chaotic signal complexity for time-domain signal which is uniformly sampled is used in this work is specified more clearly in [16].

2.3 Classification

Backpropagation neural network was used in this work. The number of hidden layers used in this algorithm was set to 10. The maximum epoch was set to 1000. Tan-Sigmoid was used as a transfer function. Levenberg Marquardt optimization is used for this work. The validation method opted is k -fold where the value of k is set to 10. Three parameters Accuracy, Sensitivity, and Specificity are calculated. Following Eqs. (1–3) specifies the formulas. For the equations considered true positive cases are where the interictal activity is correctly classified, true negative is one where there is no actual activity and it is actually classified as such.

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}} \quad (1)$$

$$\text{Sensitivity: } = \frac{\text{TruePositive}}{\text{TruePositive} + \text{FalseNegative}} \quad (2)$$

$$\text{Specificity: } = \frac{\text{TrueNegative}}{\text{TrueNegative} + \text{FalsePositive}} \quad (3)$$

3 Results

Table 2 holds the results obtained using the methodology proposed in this work. For this work, four cases have been evaluated for all of which accuracy, sensitivity, and specificity have been calculated. Table 2 contains the results obtained. Highest accuracy of 99% has been achieved for case 2 (B–C) when comparing controlled group and interictal activity recorded from epileptic zone; while comparing controlled activity from and interictal activity from opposite from epileptic zone highest accuracy of 98.98% has been achieved for case 4 (B–D). This work indents on comparing interictal activity and activity from the controlled patients. For this work Backpropagation neural is used to classify between them. Nonlinear entropy-based features approximate entropy, sample entropy, and fuzzy approximate entropy were extracted from five sub-bands of both types of signals and correlation dimensions from a complete signal.

Figures 3 and 4 compares correlation dimension of two sets used in this study from which it can be observed that interictal activity which is represented by Set C and Set D has higher correlation dimension than activity from controlled patient represented by Set A and Set B. Table 3 compares different studies on a similar database.

Table 2 Results obtained from proposed method

	Accuracy	Sensitivity	Specificity
A–C	99.5	99	100
B–C	99	99	99
A–D	98.5	99	98
B–D	99	98	100



Fig. 3 Correlation dimension plot of set A and set C

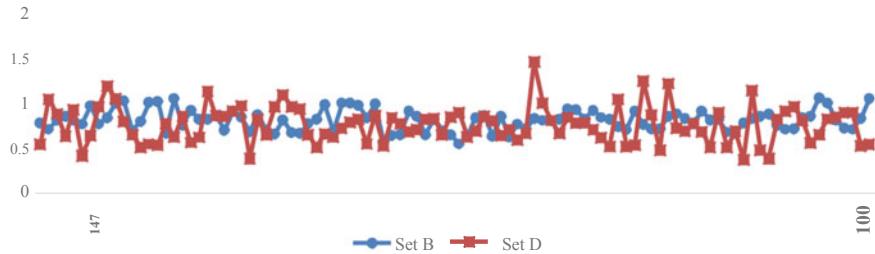


Fig. 4 Correlation dimension of set B and set D

Table 3 Comparison of results with previous research work

Researcher	Year	Feature	Classifier	Classification	Accuracy of different cases (%)
Sriraam [20]	2017	WPLogEn	Recurrent Elman neural network	A–C	99.7
				A–E	99.7
				C–E	99.85
Manish et al. [17]	2017	LS-SVM	ATFFWT and FD	AB–CD	92.5
Wang et al. [19]	2017	LDWT-based feature extraction	SVM	CD–AB–E	93.9
Jaiswalet al. [21]	2016	1D–LBP	Bayes Net	A–D	99.50
Mahmut HEK'IM [18]	2016	Discretization-based entropy	Neuro-fuzzy inference system (ANFIS)	AB–CD	92
Proposed work	2019	Approximate entropy, sample entropy, fuzzy approximate entropy, correlation dimension	Backpropagation neural network backpropagation neural network	A–C	99.5
				B–C	99
				A–D	98.5
				B–D	99

Previous researchers have worked on compared (a) activity from controlled patients with epileptic, (b) activity from controlled patients with interictal, (c) interictal with epileptic activity taking different cases. However, not much work is available to compare activity from controlled patients with interictal activity. This study has considered four such cases. From Table 3 it can be seen that different researchers have worked on different cases on similar problems and the proposed methodology outperforms [17–19] to the good margin where a different combination of sets for the same problem is considered. Though, work done in [20, 21] have marginally outperformed the proposed method by 0.2% and 1% for case 1 and case 3 considered in

our work, respectively. However, not all cases have been considered in any of the recent and previous studies.

4 Conclusion

To identify interictal discharges from routine EEG admits of artefacts through visual inspection is a crucial but time-consuming task. To support the diagnosis of epilepsy it is an important process. The automation of this process will ensure will supplement the neurologist. This work is done to conquer the same problem. For validation k-fold validation method is used in this work. We have considered four cases from the publically available dataset and have achieved good performance measures.

Acknowledgements The authors express appreciation to R. G. Andrzejak et al. for their public accessible database [10].

References

1. Fisher RS, van Emde BW, Blume W et al (2005) Epileptic seizures and epilepsy: definitions proposed by the International League against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* 46:470–472
2. Kobayashi T, Misaki K, Nakagawa H, Madokoro S, Ihara H, Tsuda K et al (1999) Non-linear analysis of the sleep EEG. *Psychiatry Clin Neurosci* 53(2):159–161. <https://doi.org/10.1046/j.1440-1819.1999.00540>
3. Lin R, Lee RG, Tseng CL, Zhou HK, Chao CF, Jiang JA (2006) A new approach for identifying sleep apnea syndrome using wavelet transform and neural networks. *Biomed Eng Appl Basis Commun* 18(3):138–144
4. Lin W-C, Chiu H-W, Hsu C-Y (2005) Discovering EEG signals response to musical signal stimuli by time-frequency analysis and independent component analysis. In: Proceedings of the 27th annual IEEE engineering in medicine and biology conference. Shanghai, China, pp 2765–2768
5. Babloyantz A, Nicolis C, Salazar JM (1985) Evidence of chaotic dynamics of brain activity during the sleep cycle. *Phys Lett* 111A:152–157
6. Acharya UR et al (2012) Automated diagnosis of epileptic EEG using entropies. *Biomed Signal Process Control* 7(4):401–408
7. Acharya UR, Fujita H, Sudarshan VK, Bhat S, Koh JEW (2015) Application of entropies for automated diagnosis of epilepsy using EEG signals: a review. *Knowledge-Based Syst* 88:85–96
8. Tawfik NS, Youssef SM, Kholief M (2016) A hybrid automated detection of epileptic seizures in EEG records. *Comput Electr Eng* 1(53):177–190
9. Acharya UR et al (2019) Characterization of focal EEG signals: a review. *Future Gener Comput Syst* 91:290–299
10. Andrzejak RG, Lehnertz K, Mormann F, Rieke C, David P, Elger CE (2001) Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: dependence on recording region and brain state. *Phys Rev E* 64(6):061907
11. Kumar Y, Anand MLDRS (2014) Epileptic seizures detection in EEG using DWT-based ApEn and artificial neural network, pp 1323–1334

12. Chen LL, Zhang J, Zou JZ, Zhao CJ, Wang GS (2014) A framework on wavelet-based nonlinear features and extreme learning machine for epileptic seizure detection. *Biomed Signal Process Control* 10(1):1–10
13. Xiang J, Li C, Li H, Cao R, Wang B, Han X, Chen J (2015) The detection of epileptic seizure signals based on fuzzy entropy. *J Neurosci Methods* 243:18–25
14. Song Y, Liò P (2010) A new approach for epileptic seizure detection: sample entropy based feature extraction and extreme learning machine. *J Biomed Sci Eng* 3(6):556–567
15. Bai D, Li X, Qiu T (2007) The sample entropy and its application in EEG based epilepsy detection. *J Biomed Eng* 24(1):200–205
16. Theiler J (1987) Efficient algorithm for estimating the correlation dimension from a set of discrete points. *Phys Rev A* 36(9):4456 American Physical Society
17. Sharma M, Pachori RB, Acharya UR (2017) A new approach to characterize epileptic seizures using analytic time-frequency flexible wavelet transform and fractal dimension. *Pattern Recogn Lett* 15(94):172–179
18. Hekim M (2016) The classification of EEG signals using discretization-based entropy and the adaptive neuro-fuzzy inference system. *Turkish J Electr Eng Comput Sci* 24(1):285–297
19. Wang Y, Li Z, Feng L, Bai H, Wang C (2018) Hardware design of multiclass SVM classification for epilepsy and epileptic seizure detection. *IET Circ Dev Syst* 12:108–115
20. Sriraam SRN (2017) Classification of epileptic seizures using wavelet packet log energy and norm entropies with recurrent Elman neural network classifier. *Cogn Neurodyn* 11(1):51–66
21. Jaiswal AK, Banka H (2017) Local pattern transformation based feature extraction techniques for classification of epileptic EEG signals. *Biomed Signal Process Control* 34:81–92. <https://doi.org/10.1016/j.bspc.2017.01.005>

A Survey on Phishing URL Detection Using Artificial Intelligence



Arpita Vadariya and Nilesh Kumar Jadav

Abstract Nowadays, a cyberattack is one of the most common security threats. The attackers generally try to steal confidential information using social engineering platforms. With the evolution of the internet in the last few years, phishing has also rapidly grown on the internet. By phishing, Internet users have a financial loss of billions of dollars per year. Phishing is a social engineering attack used to identify and steal the victim's personal information; it is generally done by e-mail, text messages, website spoofing, etc. Attackers send some attractive messages to the victims to misguide him/her and perform an attack, by using traditional methods; we cannot mitigate the ratio of the phishing attack. So, the researcher's focus on moving from the conventional approach to machine learning techniques to secure the use of the internet.

Keywords Cybersecurity · Phishing website detection · Social engineering · Machine learning · Classification algorithms

1 Introduction

Cybersecurity used to protect computer-based and internet-connected applications from cyber attackers to secure confidential information. Generally, we are using antivirus software to protect our personal computers. However, for industrial purpose, we adopt an Industrial Control System (ICS) use as the combination of the control systems in the industry such as the SCADA system, DCS, PLC and RTU [1]. ICS monitor complex industrial mechanism and critical infrastructures such as healthcare, chemical processes and many large-scale areas [2]. SCADA is a type of ICS which use as a honeypot to detect outside attacks and malicious requests within its network

A. Vadariya · N. K. Jadav (✉)
Marwadi University, Rajkot, Gujarat, India
e-mail: nileshjadav991@gmail.com

A. Vadariya
e-mail: arpitavadariya8@gmail.com

[3]. The newest technologies come along with the use of the highest-requirements of the internet. According to [4], in the future, we will have a network of about 6 billion connected devices and we should have a reliable solution to protect them. The number of networks we have, the complexity of cybersecurity threats will be increased. So, we need new security solutions every time to protect our devices and systems [5]. Computing and communication have experienced exceptional changes in ongoing decades. Computation favour on the go with a large demand for mobility support in communicating [6, 7]. The Internet can access using three technologies such as Broadband which is expensive, Wi-Fi having short-range, Dial-up which is slow and outdated and major developing countries such as India, China, Brazil and Mexico are in more potential in demanding Broadband access due to the recent austerity of Internet requirements. Wireless communication can provide such services to a great extent to a larger audience, but the wireless system has many glitches that make it uncomfortable to provide reliable services [8]. Due to a huge number of users in a wireless environment communication model also have shifted to the concept of Cognitive Radio Networks [9, 10] for greater utilization of wireless spectrum. Cyber-attackers always try to breach these security and control systems to get the confidential information of users. Phishing is the most common method to exploit user accounts and get the information. It can be done by either using the URL or Web Page method. Traditional methods to detect a phishing attack were not given the proper output so we have been shifted to Artificial Intelligence based techniques for accurate detection.

In the past two decades, different AI techniques we have been used to detect phishing website client-side on random forest, SVM and Naive Bayes do best respecting the highest true positive rate and accuracy [11]. In [12], the researchers present an Adaptive Neuro-Fuzzy Interface System (ANFIS) based powerful system using the integrated features of the text, images and frames for phishing web detection and protection after they provide the best solution for phishing web detection system. In [13], the paper aims to give a comprehensive review and a structural knowledge of malicious URL detection methods using machine learning. In real-time anti-phishing system, which uses different types of classification algorithm with NLP based feature is proposed, such as Decision tree, adaboost, Kstar¹, KNN, random forest, SMO, Naive Bayes and after the finalized result base on algorithms, only random forest with NLP-based features are given best accuracy rate [14].

1.1 *Cybersecurity*

Cybersecurity or IT security is the technique that is used to protect a network, and Programs, data from the person who has no legal access. Due to the improvement of education, agriculture, healthcare, the use of the internet in every sector is widely increasing. We can use the internet for various time security purposes. Figure 1 divides it into a few common categories.

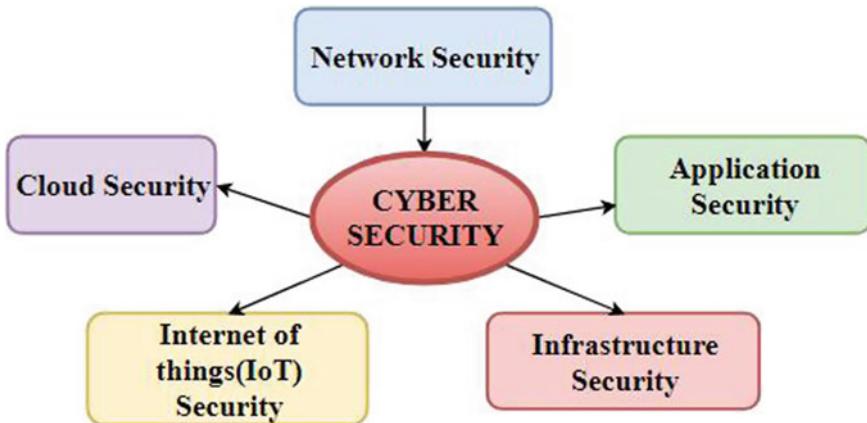


Fig. 1 Different area of cybersecurity

1. **Cloud Security:** It is secure of data storage in online like stealing data, leakage data.
2. **Network Security:** Network Security means any criminals or hackers are do not access to the file or directory in a computer system or network.
3. **Application Security:** It secures the data and code within the application against the attacker or unauthorized person.
4. **Internet of Things (IoT) Security:** It is the technological field protection to the devices and networks in the internet of things (IoT).
5. **Infrastructure Security:** It is the protection of infrastructures like financial and common area infrastructures in banking, hospitals, hotels, airports, railway transportation, etc.

1.2 Cyberattack

A cyberattack is an attack done on the computer from different one or more networks. In which, the aim is to make the target computer disable to the system. In which intention is the get access to the data of the target computer. Figure 2 shows the different types of a cybersecurity attack.

1. **Phishing Attack:** Phishing is one type of social engineering attack, in which attackers or users are connected via email, telephone, or text message. The attacker sends any fake message to the user and steals the user's personal information.
2. **Password Attack:** It is a process of retrieving a password from the data store in a computer system. A common method to retrieving a password used in a brute-force attack to try guess password and check them against as possible.

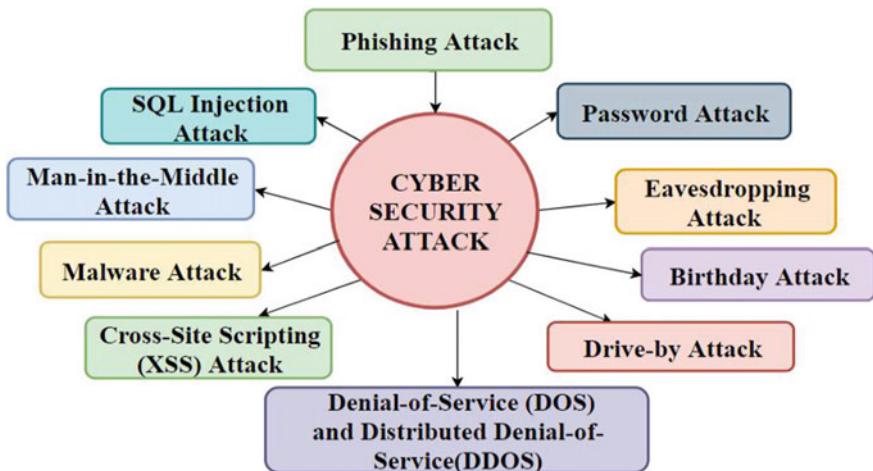


Fig. 2 Different types of cybersecurity attack

3. **Eavesdropping Attack:** It happens in the interference of network traffic. With the use of eavesdropping, an attacker can obtain the password, credit card number and other personal information users are sending to the network.
4. **Birthday Attack:** It is one kind of cryptographic attack which uses maths after the birth-day query in the probability method. However, this attack can be used to communicate between two or more people.
5. **Drive-by Attack:** It attacks to focus on the user through their browser, and it is download malware on their computer system and after they visit the malicious website. It can also hap-pen users are visiting the real web site.
6. **Denial-of-Service Attack** Hackers use this attack to make a network or machine unavailable to the users accessing it. The main purpose of this attack is to prevent users from accessing a service such as the Internet.
7. **Distributed denial-of-service Attack:** It is similar to a Denial-of-Service attack, but the result is much different, it allies many computers and many connections. The computer behind such attacks and distributed around the whole world.
8. **Cross-site scripting Attack:** cross-site scripting means the malicious script is inserted into helpful or trusted websites and webpage.
9. **Malware Attack:** Malware is malicious software used and damage to the computer system and harm to the computer user. It is a typically steal data try something on the computer system.
10. **Man in the Middle Attack:** It occurs an attacker enters between the communication to the client and a server.

11. **Structure Query Language Injection Attack:** The attacker is inserting unwanted data into the program, which process editor command or query which changes to the program. Injection attacks are a risky attack on web application attacks.

1.3 Cybersecurity and Its Domain

The most difficult tasks are to protect cyberspace and social engineering from non-ethical hacker's and criminals, in the current scenario, technologies are grown up and is used in various field like network, cloud, application, etc. [14]. However, for everything, there are challenges to like stealing and manipulating a user's personal information. Which are the main research area for the security researcher's how to protect victim's personal information in social media, every people spend lots of hours and send some documents if there is no security protection then there will be a great issue of privacy. Social engineering is the main threat for attacks like phishing attacks are to gain or steal victim's personal information. In this paper, the researchers will try to deliver that what kind of phishing attack is happening now a day's, how attacker gain information to attack there are so many security issues we are facing when we are using the internet, but we have to reduce it. If any unwanted message or mail from, the attacker sends to, the users after the users open the mail or message and they are filled up all the sensitive, information, mostly these types of messages are coming to the fake company, other social sites. Regarding security purposes, we can apply some security settings from users then the people can protect our data at some level and after that, we require some devices to secure our data. However, many methods to perform such an attack, such as Social engineering, Phishing Attack, DoS, DDoS, Main-in Middle, SQL Injection, etc. [13]. We have to use some tools or methods to detect malicious websites Social engineering creates based on the scam website from the attackers to the users. It is the term used for a broad area of dangerous activities performed through human Interaction via email, website, phone, USB drives, etc. It also uses emotional manipulation to trick users into submitting security mistakes or giving personal information [15]. Attackers first gather information about a victim and the victim cannot aware of the scam the website, after that, the attacker gains the victim's trust after giving unwanted information to break the security systems, such as sharing personal information or providing access to critical resources. In [16] social engineering happens in different sectors of business and services as well as the percentage of risk it contains. It is used in different types of techniques to prevent the system from stealing confidential information shown in Fig. 3. To protect the users from the attacker's same awareness of protecting your inbox, don't open in any suspicious like those are not aware of it, and don't share secret information with anyone like username, password, credential information, etc. The researchers are implemented based on an efficient algorithm to detect a phishing website and protect users in real-time. If any suspicious activity is occurring, it will automatically detect the website with the help of host-based features. Among the use of conventional

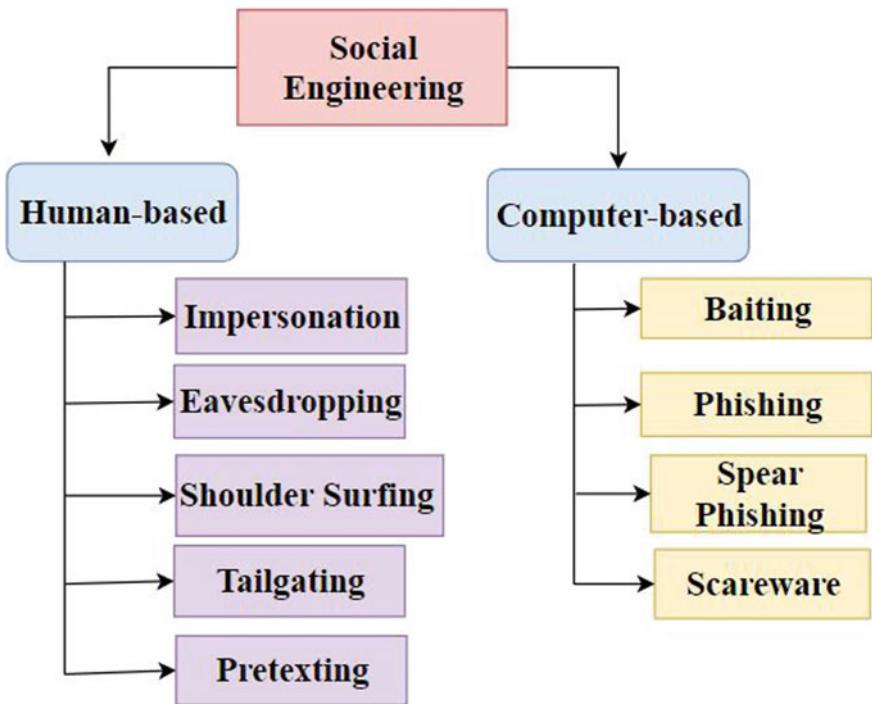


Fig. 3 Different techniques based on social engineering

techniques, in the world, the ratio of social engineering attacks is increasing through the Internet. To reduce the social engineering attack rate now, we are moving from the traditional way to Artificial Intelligence, Machine Learning and Deep Learning. It provides a variety of datasets from which computers can learn from the help of previous data. Accordingly, computers train themselves continuously without the help of a specific program. Including the way of Artificial Intelligence, we can also decrease the computer attack [12].

2 Social Engineering

Social engineering is one of the simplest methods to gather information about a targeted person through the process of using a human weakness that is inherent to each organization and this attack is avoided to cybersecurity systems through fraud, using the most vulnerable link, the people are affected [17]. It is the planning of people into doing unwanted actions or disclosing private information related to the victim. The term typically refers to fraud for information gathering, identity theft

and computer system access, the social engineering attacks include direct interaction between the users and attackers by electronic devices like electronic communication via mobile, email and other social sites. It exploits the trusted person that the user unconsciously placed in the attackers. However, they often act as company employees, colleagues and friends, they are gaining illegal access to the system the following appearance of protecting the user's credential information [18]. It is a non-technical approach that relies on human interaction and frequently Involves deceiving people into breaking the traditional security methods. Social engineering attacks are more challenging to manage, and they depend upon human behaviour and take advantage of the vulnerable person. Nowadays, technology solutions, it is defined in social engineering, it helps to user awareness, and also helps to protect confidential information.

2.1 Different Techniques Based on Social Engineering

Social engineering is attackers gathering information about victims and after they are stealing the confidential information. Social engineering is classified, into two types of techniques is Human-based techniques, and the other one is digital-based techniques. Figure 3 shows Different Techniques based on social engineering.

2.1.1 Human-Based Techniques

- A. **Impersonation:** Taking up the role of another person in order to entertain someone or for the purpose of fraud.
- B. **Eavesdropping:** It securely listens to the confidential conversations of others without their approval.
- C. **Shoulder Surfing:** It is used to obtain a Victim's personal details such as identification number and password.
- D. **Tailgating:** It is when a person obtains illegal access into convenience by using methods and tactics to fool the Users of that organization.
- E. **Pretexting:** Attacker tries to show that he is school, colleagues, police, or someone who has right-to-know authority and tries to get confidential information.

2.1.2 Computer-Based Techniques

- A. **Baiting:** Baiting attacks use a fake promise to the victim's desire or interest. The attacker steals their credential information or infected system with malware.
- B. **Phishing:** Phishing is a fraudulent activity to designed and collect sensitive information like username, password and credit card details. The attacker sends

millions of fraud email messages to the victims from the fake website and requests that you give personal details.

- C. **Spear Phishing:** Spear phishing is a direct attack on a specific person or company.
- D. **Scareware:** It involves the victim's being attacked with fake calls, messages and invented threats, giving the user to believe that their machine has some threat.

2.2 *Phishing*

Phishing is one type of social engineering attack, and every person should be read and protect themselves. That used to send an email and fake websites. Attackers are trying to get a victim's personal information such as username, password, bank account information and credit card details in electronic communication [19]. The attacker sends to the email or message to the victims, to enter secret information on the scam website which looks and feels like a legitimate website. Phishing attacks are constantly to affectation a major threat for computer system protectors, repeatedly creating the initial step too many stage attacks [20]. Attack method to also classify into three classes: attack initialization, data collection and system penetration [21]. The main goal of, phishing is to steal the credential information, or sometimes malware is installed on a victim's system.

2.3 *Different Types of Phishing*

- I. **Deceptive Phishing:** The attacker is collected personal information from victims. It has used the information to steal money or start other attacks.
- II. **Man-in-Middle Attack:** The attacker enters between the communication of the client and the server.
- III. **Malware Phishing:** Malware is, malicious software it is, harm to the user's computer. It is running malicious software to the user's computer. The various form of malware based on phishing is Key loggers, session Hi-jacker, web trojans and Data Theft.
- IV. **Domain Name System-Based:** Phishing is to avert the trustworthiness of the query strategy for a domain name. Types of Domain Name System-Based phishing host document harming, dirtying client's Domain Name System reserve, intermediary server bargain.

3 Artificial Intelligent and Its Different Area

Artificial intelligence (AI) is a study of how to make computers do the things which at present human can do batter. Artificial intelligence utilizes techniques from several

fundamental disciplines like a statistic, mathematics, engineering, neural science, computer science and, which then underpin numbers of technological capability are computer vision, natural language processing, information retrieval, information filtering, predictive analytics, decision analysis, robotics. Artificial intelligence covered is subset such as Machine Learning, Deep Learning, etc. It can be classified into two types weak or strong. Weak Artificial Intelligence—It is designed and trained for a particular task. Strong Artificial Intelligence—It is the generalized human abilities.

3.1 Machine Learning

Machine Learning (ML) is a sub-branch of Artificial Intelligence. It has, the capability to learn without being explicitly programmed. Machine learning is many algorithms used to classified problems like speech recognition, and document classification, etc. It can be divided into four types.

- **Supervised:** data has known labels or output. It is also used for classification and regression.
- **Unsupervised:** Labels or output unknown and it is focused on finding patterns and gaining insight from the data. It is also used for clustering.
- **Semi-Supervised:** Labels or output known for a subset of data is a mixture of supervised learning and unsupervised learning.
- **Reinforcement:** Focus on marketing decisions based on previous experience.

3.2 Deep Learning

Deep learning (DL) is a subset of machine learning. It is used in multi-layered neural networks are connected to each other and build algorithms that find the best way to perform task on a large set of data. Basic applications of Deep learning in today's world are Image Recognition, Speech Recognition and Natural Language Processing, Robots and Self Driving car, Drug discovery and better diagnostics of diseases in Healthcare (Tables 1 and 2).

4 Conclusion

Phishing URL detection is a very crucial part of cybersecurity. In this paper, we concluded the survey of various Phishing URL Detection using machine learning techniques. Phishing, the main task is to steal the victim's personal information such as username, password, bank information and credit card number. A phishing attack can be done by either human-based feature or computer-based feature. With

Table 1 Advantages and disadvantages of different machine learning and deep learning algorithms

Name of algorithm	Advantages	Disadvantages
Decision Tree	Easy to understand Efficient algorithm Order of instances has no effect	Classes must be mutually exclusive Decision tree depends on selection order
Naive Bayes	Based on statical model Easy to understand Efficient training algorithm Order of instances has no effect Useful across domain	Missing Values of Attribute. Attributes to be independent. Normal distribution on numeric. Classes must be exclusive. Attribute mislead classification. Attribute and class Frequency are accuracy
Neural network	Used classification/regression Based on nodes Tolerate noisy inputs	Difficult to understand structure. Too many attributes can result in overfit Network structure can be determined by expert
Support vector machine	Model nonlinear boundaries Easy to control complexity	Training is slow compare to Naïve Bayes and Decision Tree Difficult to understand structure of algorithm
Linear regression	It is a simple method It is easy to use and understand	Relation between dependent and independent variables. Independent observation
Logistic regression	It is simple and implemented is fast and easy	Highly interpretable
Kernel-means	Kernel-Means times computationally faster than hierarchical clustering	Difficult to predict Kernel-value. Different initial partitions can result in different final clusters
Random forest	Random forests are extremely flexible High accuracy	Random forests are their complexity. Harder and time-consuming to construct than decision trees
Kernel nearest neighbour	Kernel nearest neighbour is very easy to implement Kernel nearest neighbour algorithm much faster than other algorithms	Does not work well with large dataset And high dimensions Need feature scaling

the help of machine learning techniques, we can detect the phishing website on the internet. This paper surveyed the different types of phishing URL detection techniques using Machine Learning and Deep Learning techniques, their advantages and disadvantages over traditional methods and many more.

Table 2 Different types of tools used in machine learning and deep learning techniques

Tool name	Description
1. Matlab	It is a programming language developed for Mathematics work. It has in-built mathematical functions. It starts with the matrix programming language. It is a simple and easy way to get a fast and effective result
2. Weka	Weka is the open-source tool used to implement data mining algorithms. It is collected the machine learning methods for the task of data mining. The techniques to apply directly to the dataset or collect from your java code and weka contains preparation, classification, clustering, regression, association rules and visualization
3. Rapid miner	Rapid miner is an open-source software in an environment of machine learning and data mining. It is similar to the weka tool. Rapid miners used to solve for research work and real-world data mining and machine learning task
4. Keras	Keras is open-source software. It provides high-level neural network
5. Scikit learn	Python this tool used for Deep learning. Keras is written in python and it is running on popular neural network framework like Tensor Flow. Scikit learn tool is open-source tool it is used for data analysis. Scikit learn is used many features such as classification, regression and clustering model processing. It is easy to use

References

- Dutta N, Tanchak K, Delvadia K (2019) Modern methods for analyzing malware targeting control systems. In: Recent developments on industrial control systems resilience, pp 135–150
- Ibrahim M, Alsheikh A (2019) Determining resiliency using attack graphs. In: Recent developments on industrial control systems resilience, pp 117–133. https://doi.org/10.1007/978-3-030-31328-9_6. Accessed 24 Jan 2020
- Dutta N, Jadav N, Dutiya N, Joshi D (2019) Using honeypots for ICS threats evaluation. In: Recent developments on industrial control systems resilience, pp 175–196
- Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025 (inbillions). <https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/>
- Ionescu O, Dumitru V, Pricop E, Pircalabu S (2019) Innovative hardware-based cybersecurity solutions. In: Recent developments on industrial control systems resilience, pp 283–299. https://doi.org/10.1007/978-3-030-31328-9_12
- Dutta N, Misra IS (2014) Multilayer hierarchical model for mobility management in IPv6: a mathematical exploration. *Wirel Pers Commun* 78(2):1413–1439
- Dutta N, Misra IS (2007) Mathematical modelling of HMIPv6 based network architecture in search of an optimal performance. In: IEEE 15 th ADCOM, Guwahati, India, pp 599–605
- Jadav N (2019) Improving BER of WiMAX using interference mitigation techniques. In: 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)
- Dutta N, Sarma HK, Polkowski Z (2018) Cluster based routing in cognitive radio adhoc networks: Reconnoitering SINR and ETT impact on clustering. *Comput Commun* 1(115):10–20
- Dutta N, Sarma HK (2017) A probability based stable routing for cognitive radio adhoc networks. *Wirel Netw* 23(1):65–78
- Jain AK, Gupta BB (2018) Towards detection of phishing websites on client-side using machine learning based approach. *Telecommun Syst* 68(4):687–700

12. Adebowale M, Lwin K, Sánchez E, Hossain M (2019) Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. *Expert Syst Appl* 115:300–313
13. Sahoo D, Liu C, Hoi SCH (2017) Malicious URL detection using machine learning: a survey
14. Sahingoz OK, Buber E, Demir O, Diri B (2019) Machine learning based phishing detection from URLs. *Expert Syst Appl* 1(117):345–357
15. Mao J, Bian J, Tian W, Zhu S, Wei T, Li A, Liang Z (2018) Detecting phishing websites via aggregation analysis of page layouts. *Procedia Comput Sci* 129:224–230
16. Yang P, Zhao G, Zeng P (2019) Phishing website detection based on multidimensional features driven by deep learning. *IEEE Access* 7:15196–15209
17. Conteh NY, Schmick PJ (2016) Cybersecurity: risks, vulnerabilities and countermeasures to prevent social engineering attacks. *Int J Adv Comput Res* 6:23–31
18. Breda F, Barbosa H, Morais T (2017) Social engineering and cyber security. In: Inted 2017 Proceedings. <https://doi.org/10.21125/inted.2017.1008>. Accessed 6 Sept 2019
19. Mao J, Tian W, Li P, Wei T, Liang Z (2017) Phishing-alarm: robust and efficient phishing detection via page component similarity. *IEEE Access* 5:17020–17030
20. Gutierrez C, Kim T, Corte R, Avery J, Goldwasser D, Cinque M, Bagchi S (2018) Learning from the ones that got away: detecting new forms of phishing attacks. *IEEE Trans Dependable Secure Comput* 15(6):988–1001
21. Aleroud A, Zhou L (2017) Phishing environments, techniques, and counter measures: a survey. *Comput Secur* 68:160–196

Prognosticating Liver Debility Using Classification Approaches of Machine Learning



Revelly Akshara and Sandhi Kranthi Reddy

Abstract In the field of Machine Learning in Healthcare, one of the most compelling factors is the prognosis of illness by examining the characteristics which have the most impact on its recognition. One of the deadliest diseases is a liver disease that accounts for about two million deaths annually worldwide. Prognosticating it in an early stage is helpful to get diagnosed at the right time as this may lead to a complete recovery in some patients. We can effectively prognosticate liver disease using supervised learning techniques specifically classification methods of machine learning. In this paper, different classification methods are applied to the Indian Liver Patient Records Dataset downloaded from kaggle.com to prognosticate liver disease.

Keywords Machine learning · Liver disease · Dataset

1 Introduction

Liver Disease is the major chronic disease that occurs worldwide irrespective of age, sex, or region that accounts for about 2 million deaths annually worldwide ranking as the 20th leading causes of death [1]. Prognosis of liver disease is made easy by the advancements in the use of machine learning and data mining in healthcare.

R. Akshara · S. K. Reddy (✉)

Assistant Professor, Department of Computer Science and Engineering, Vignan Institute of Technology and Science, Yadadri Bhuvanagiri District, Telangana, India
e-mail: kranthi.sandhi@gmail.com

R. Akshara
e-mail: akshara.revelly@gmail.com

2 Machine Learning in Health Care

Machine learning allows the system to study the data implicitly and improve its performance from experience with minimal human involvement [1]. The applying procedure of machine learning on different data are: (i) Accumulating the Data, (ii) Interpreting and Devising the Data, (iii) Training a model on the Data, (iv) Assessing model's Performance, (v) Improving model's Performance [2]. Based on learning criteria, Machine learning algorithms are of three types [2]. They are (1) Supervised Learning, (2) Unsupervised Learning, (3) Reinforcement Learning.

Application of Machine learning algorithms performs effectively in prognosis of several diseases due to its ability to combine; organize and aggregate huge amounts of data from various sources and its feature of minimizing dimensionality helps to increase algorithm's overall performance [3].

Liver is one of the major parts in humans; its debility may sometimes cause death. So, initially prognosticating it is essential which can be done effectively using machine learning.

3 Related Work

In 2019, Chieh-Chen Wu, Wen-Chun Yen, Wen-Ding Hsu, Md. Mohaimenul Islam, Phung Anh (Alex) Nguyen, Tahmina Nasrin Poly, Yao-Chin Wang, Hsuan-Chia Yang, Yu-Chuan (Jack) Li has used the classification models namely, Random Forest (RF), Naïve Bayes (NB), artificial neural networks (ANN) and logistic regression (LR) to predict Fatty Liver Disease and concluded that the Random Forest model accurately predicts Fatty Liver Disease patient using minimum clinical variables [4]. In 2018, Insha Arshad, Chiranjit Dutta, Tanupriya Choudhury, Abha Thakral has used machine learning techniques namely, SMO algorithm, Naïve Bayes algorithm and J48 algorithm to detect liver disease due to excessive alcoholism and concluded that SMO gives the best result [5].

4 Dataset Analysis

To prognosticate liver disease of a patient, we used data sets which are patient records collected from North East of Andhra Pradesh, India, download from kaggle.com. The data set contains the following attributes:

- Age of the patient
- Gender of the patient
- Total Bilirubin
- Direct Bilirubin
- Alkaline Phosphotase

- Alamine Aminotransferase
 - Aspartate Aminotransferase
 - Total Proteins
 - Albumin
 - Albumin and Globulin Ratio
 - Dataset: field used to split the data into two sets (patient with liver disease, or no disease).
- Class variable (1 or 2).

The dataset has 416 liver patient records and 167 nonliver patient records collected from 441 male patient records and 142 female patient records. The “Dataset” class label is used to divide groups into liver patients (liver disease) or not (no disease).

The dataset is divided into two parts, a training set that contains 80% of the dataset as a training set and 20% as a test set. In this proposed work we applied some classification models liver datasets using built-in packages of R Programming [6]. R Programming works efficiently for Statistical Data Analysis and Machine Learning. R is an open-source under the GNU public license, its environment is written primarily in C, FORTRAN and R. R is one of the best platforms and provides many built-in packages to run different machine learning algorithms on different datasets, predict() function is used to predict the output of new input. R is an implementation of an earlier programming language called S developed at Bell Laboratories by John Chambers, which can be downloaded from <https://cran.r-project.org/bin/windows/base/>. The following graphs are generated in R Programming which shows the correlation of training data set (Fig. 1) and test data set (Fig. 2). It is observed that the graph is generated for all attributes of the data set. One attribute to other attributes are strongly correlated because maximums graphs showing positive correlation/direction.

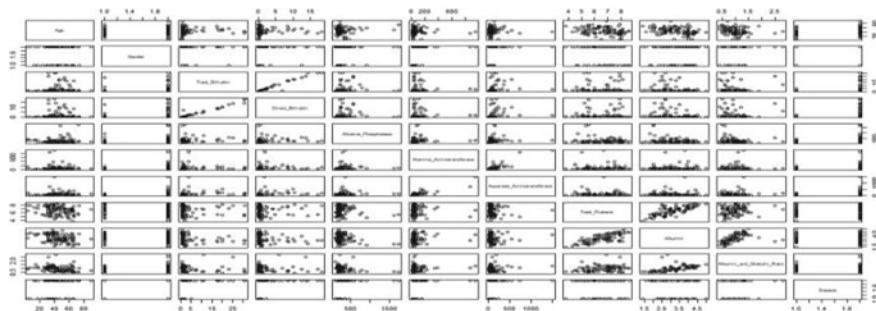


Fig. 1 Correlation among the attributes of training set

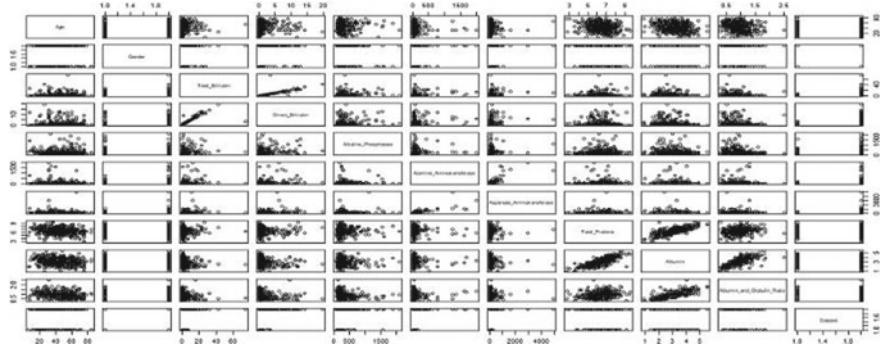


Fig. 2 Correlation among the attributes of test set

5 Classification Approach for Prognosticating Liver Disease

Classification approach which is a supervised learning technique of machine learning yields accurate results for chronic diseases like liver disease. In our proposed work, we have used techniques such as Decision Tree, Logistic Regression and Random Forest Tree. The Following Fig. 3 describes our proposed model for the classification.

6 Implementation

6.1 Decision Tree

Decision tree is one of the supervised learning algorithms that build the model in tree form, used for classification or regression problems. The main aim of using Decision Tree is to create a training model that is used to prognosticate class or classify the given data with a set of decision rules extracted from dataset. Each internal node represents an attribute and each leaf node represents class label. The first task in Decision Tree is to identify the root node from the set of attributes, the root node/parent node at each level can be identified based on the attribute selection measures, namely; information gain and Gini index.

Information Gain: The attribute with highest gain can be considered as a significant node or root node or critical node and it is calculated based on the entropy of class. The entropy of class is calculated as

$$E(C) = -p \log_2(p/(p+n)) - q \log_2(n/(n+p)).$$

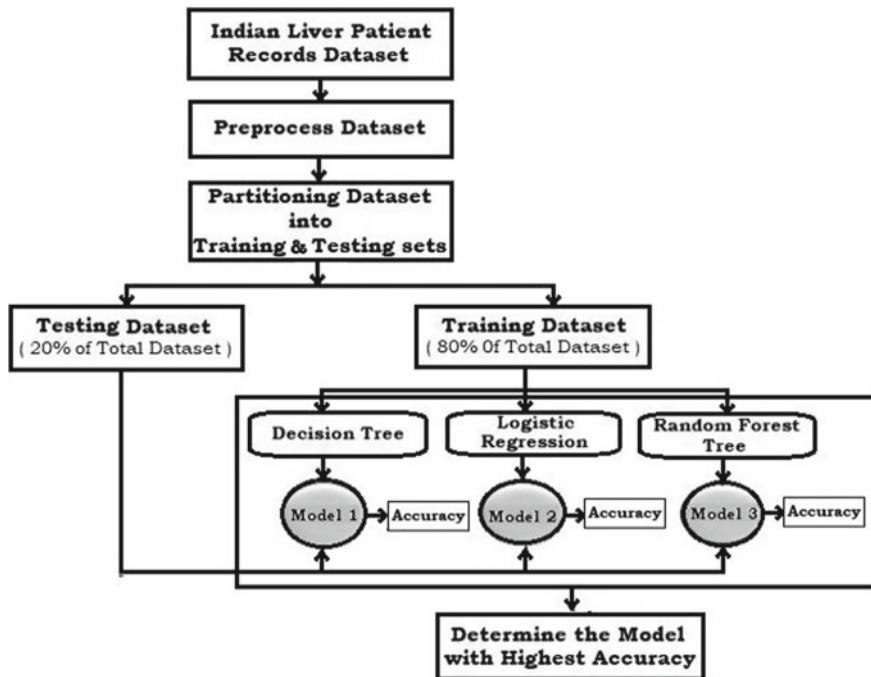


Fig. 3 Proposed model

where p is probability of success, q is probability of failure. After calculating entropy of class, calculate entropy of all attributes, finally we calculate information gain for all attributes as

$$\text{Information Gain (Attribute)} = \text{Entropy of class} - \text{Entropy of attribute}.$$

The attribute with highest information gain is considered as root node/critical node/significant node. This process continued at each level to identify the critical node.

Gini Index: Gini Index is another important measure used to identify the root node. The Gini index is calculated for all attributes and the attribute with lower Gini index is considered as root node/critical node.

The Decision Tree algorithm is applied to liver datasets in R Programming with the help of a party package. After training the model, the following tree (Fig. 4) is generated in R Programming.

The confusion matrices in Fig. 5 are generated for the training set and test set.

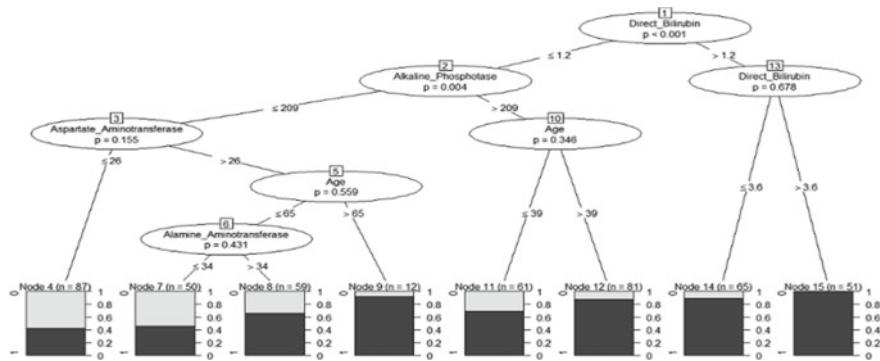


Fig. 4 Decision Tree for liver dataset

Decision Tree Algorithm				
Training Data Set				
Class	0	1	Total	Total Observations
0	77	60	137	466
1	57	272	329	
ACCURACY	74.9%	Miss Classification Error		0.251073

Test Data Set				
Class	0	1	Total	Total Observations
0	16	23	39	117
1	17	61	78	
ACCURACY	65.9%	Miss Classification Error		0.3418803

Fig. 5 Confusion matrix for decision tree

6.2 Logistic Regression

Logistic Regression is a classification algorithm, which is a special type of linear regression, used to predict a binary outcome (1/0, Yes/No, True/False) for a given set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to logit() function. Logistic regression equation is derived from a linear regression equation with the dependent variable as follows:

$$y = \beta_0 + \beta_1(\text{Age}) \quad (\text{a})$$

In logistic regression, we are only concerned about the probability of outcome dependent variable (success or failure). To make the probability of less than 1, we must divide p by a number greater than p as follows:

$$p = \exp(\beta_0 + \beta(\text{Age}) / \exp(\beta_0 + \beta(\text{Age})) + 1 \quad (\text{b})$$

$$p = e(\beta_0 + \beta(\text{Age})) / (e(\beta_0 + \beta(\text{Age})) + 1) \quad (\text{c})$$

$$p = e^y / (e^y + 1) \quad (\text{d})$$

where p is the probability of success and (d) is logit function. If p is the probability of success, $1 - p$ will be the probability of failure which can be written as:

$$q = 1 - p = 1 - (e^y / (1 + e^y)) \quad (\text{e})$$

Odd ratio is defined as the probability of success divided by probability of failure. On dividing, (d)/(e), we get,

$$p / (1 - p) = e^y$$

After taking log on both sides:

$$\log(p / (1 - p)) = y$$

Substituting the value of y in the above equation we get

$$\log(p / (1 - p)) = \beta_0 + \beta(\text{Age})$$

The Logistic Regression algorithm is applied to liver datasets in R Programming with the help of a fundamental package. After training the model, the following graph (Fig. 6) is generated in R Programming:

The following confusion matrices (Fig. 7) are generated for the training set and test set.

6.3 Random Forest Tree

The Random Forest Tree is also a type of supervised learning algorithm which is used for classification or regression problems [7]. It is an ensemble method based on divide and conquers strategy which generates multiple decision trees [8] on randomly split datasets based on information gain, gain ratio and Gini index [9]. The group of generated decision trees is known as forest. The Random Forest Tree works in three steps: (i) split datasets into random samples. (ii) Construct Decision Tree for each sample and group them by making it a forest. (iii) Gather a prediction result from each Decision Tree and finalize the prediction results where more trees are generated in forest [9].

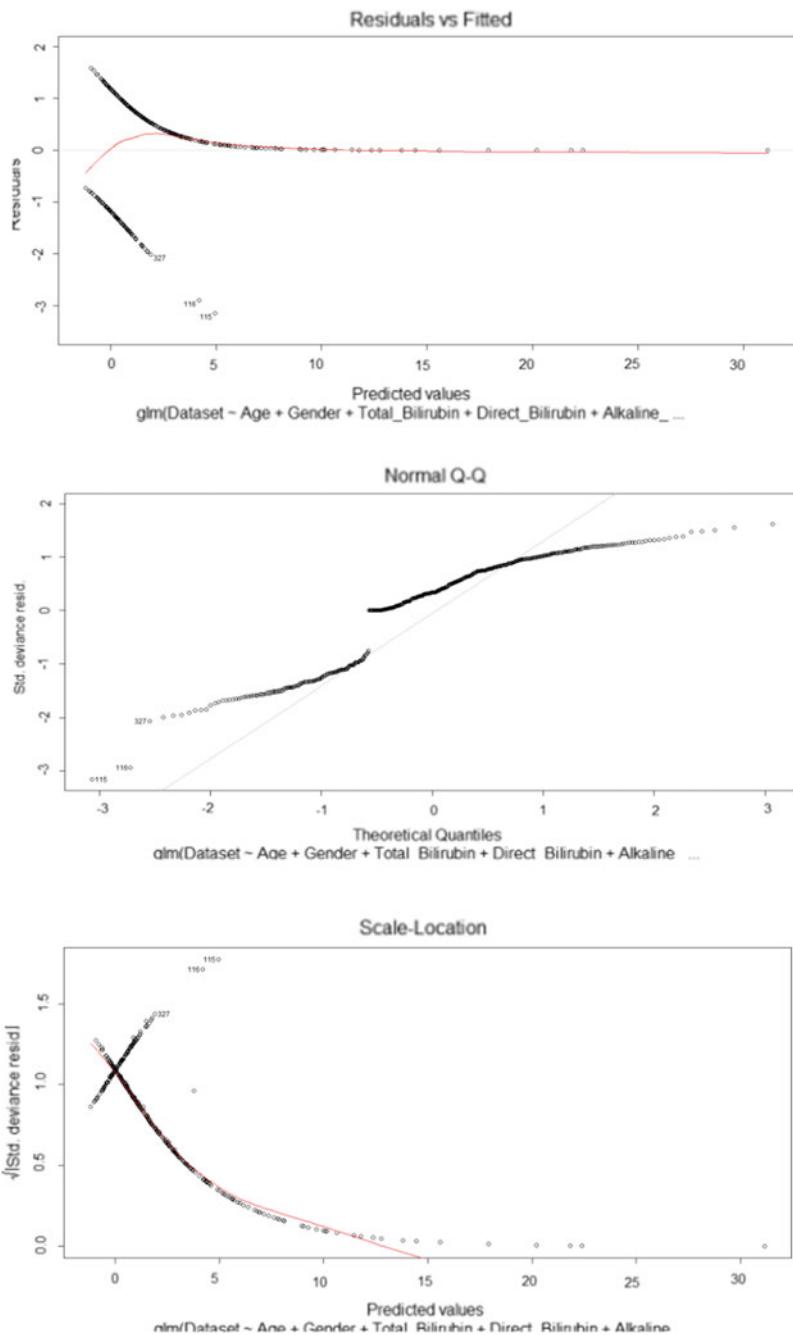
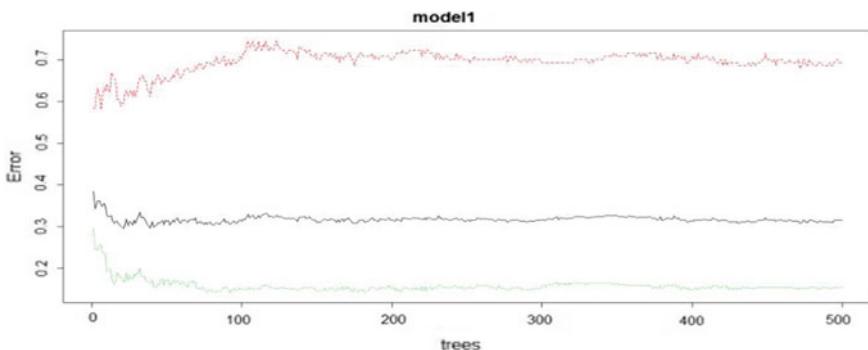


Fig. 6 Logistic regression graphs for liver dataset

Logistic Regression Algorithm				
Training Data Set				
Class	0	1	Total	Total Observations
0	25	14	39	403
1	85	279	364	
ACCURACY	75.5%	Miss Classification Error		0.2456576
Test Data Set				
Class	0	1	Total	Total Observations
0	10	14	24	180
1	45	111	156	
ACCURACY	68%	Miss Classification Error		0.3125

Fig. 7 Confusion matrix for logistic regression**Fig. 8** Random Forest Tree for liver dataset

The Random Forest Tree algorithm is applied to liver datasets in R Programming with the help of the RandomForest package. After training the model, the following tree (Fig. 8) is generated.

The following confusion matrices (Fig. 9) are generated for the training set and test set.

7 Results

After applying Decision Tree, Logistic Regression and Random Forest Tree Liver datasets, the following results [Table 1 and graphs (Fig. 10)] are generated.

Random Forest Tree Algorithm				
Training Data Set				
Class	0	1	Total	Total Observations
0	41	49	90	452
1	93	269	362	
ACCURACY	68.6%	<i>Miss Classification Error</i>		0.3141593
Test Data Set				
Class	0	1	Total	Total Observations
0	14	16	30	131
1	20	81	101	
ACCURACY	76.7%	<i>Miss Classification Error</i>		0.2635659

Fig. 9 Confusion matrix for random forest tree

Table 1 Results after applying Decision Tree, Random Forest Tree and logistic regression on liver dataset

Algorithm	Training dataset accuracy	Test dataset accuracy	Training misclassification error	Test misclassification error
Decision tree	74.9	65.9	0.251073	0.342
Logistic regression	75.5	68	0.2456576	0.312
Random Forest Tree	68.6	76.7	0.3141593	0.263

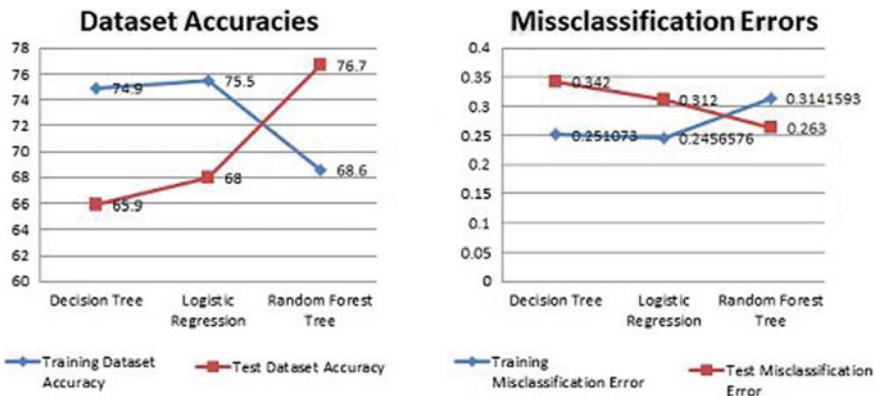


Fig. 10 Comparison graphs for obtained results of Decision Tree, Logistic Regression and Random Forest Tree applied on liver dataset

8 Conclusion

Debility of liver can be predicted efficiently using machine learning classification approach. In the proposed model, we have applied three classification techniques to prognosticate liver debility, which are Decision Tree, Logistic Regression and Random Forest Tree on Liver Dataset of North East of Andhra Pradesh, India. On Testing Dataset, Decision Tree has generated an accuracy of 65.9% Logistic Regression has generated accuracy of 68% and Random Forest Tree has generated the highest accuracy of 76.7%. By this, we can conclude that Random Forest classifies and Prognosticates Liver Debility better than the other models that we have used.

References

1. Asrani SK, Devarbhavi H, Eaton J, Kamath PS (2019) Burden of liver diseases in the world. *J Hepatol* 70(1):151–171
2. Nithya B, Ilango V (2017) Predictive analytics in health care using machine learning tools and techniques. In: International conference on intelligent computing and control systems. 978-1-5386-2745-7/17/\$31.00 ©2017 IEEE
3. Jain D, Singh V (2018) Feature selection and classification systems for chronic disease prediction: a review. *Egyptian Inf J.* <https://doi.org/10.1016/j.eij.2018.03.002>
4. Wu C-C, Yen W-C, Hsu W-D, Islam MM, Nguyen PAA, Poly TN, Wang Y-C, Yang H-C, Li Y-CJ (2019) Prediction of fatty liver disease using machine learning algorithms. In: Computer methods and programs in biomedicine. <https://doi.org/10.1016/j.cmpb.2018.12.032>
5. Arshad I, Dutta C, Choudhury T, Thakral A (2018) Liver disease detection due to excessive alcoholism using data mining techniques. In: International conference on advances in computing and communication engineering (ICACCE-2018)
6. CRAN Packages available at https://cran.r-project.org/web/packages/available_packages_by_name.html
7. Random Forest Algorithm available at <https://towardsdatascience.com/the-random-forest-algorithm-d457d499ffcd>
8. Random Forest Classifier available at <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>
9. Random Forest Algorithm available at <https://www.datacamp.com/community/tutorials/random-forests-classifier-python#algorithm>

Robust UI Automation Using Deep Learning and Optical Character Recognition (OCR)



Mithilesh Kumar Singh, Warren Mark Fernandes,
and Mohammad Saad Rashid

Abstract In the complete software development life cycle (**SDLC**), testing and maintenance stand slowest, and the reason being entire process is manual or manually maintained automation. Over time, companies have realized that it is very costly and time consuming. The biggest reason for manual maintenance of automation is its dependency on dynamic properties of the UI elements (xpath, class, ID) which directly depend on document object model (**DOM**). Cloud releases are very frequent, and changes in the UI properties are expected. In this paper, we are presenting a UI technology agnostic approach, which does not depend on the DOM and UI properties of the application rather behaves like a human and visually parses the screen. We use artificial intelligence to detect the UI elements, and even, just a mock-up of a screen can be parsed this way. This also enables test-driven development (**TDD**) of UI, among many other use cases. We used annotated UI image data for training our deep learning model, and the method has been validated with 94% accuracy on new UI elements. This approach does not require manual maintenance of the generated automates unless there is a functional change in the application. We propose UI technology agnostic, zero touch, self-healing UI automation.

Keywords Deep learning · Computer vision · Object detection · OCR · Visual automation · Software testing · UI automation

M. K. Singh (✉)

Intelligent RPA SAP Labs, Bangalore, India
e-mail: 1mithilesh194@gmail.com

W. M. Fernandes

STE SAP Labs, Bangalore, India
e-mail: f.warren210@gmail.com

M. S. Rashid

PSCC SAP Labs, Bangalore, India
e-mail: msaadr94@gmail.com

1 Introduction

Before today, implementing a visual-based automation approach was not easy, and some obstacles could have been non-existence of technology or deep learning-powered solution like proposed by Joseph et al. (2016) that took a new direction in computer vision and helped detect different attributes/objects on a single image/screenshot [1–3]. Owing to the increased user engagement with graphical interfaces, the demand for UI (user interface) applications has increased dramatically [4, 5] causing every software company to implement it. To increase availability and connectivity, companies are moving their applications to the cloud. Cloud releases are very frequent (at least once a quarter), and hence, companies are under increasing pressure to raise the quality of the application/product, reduce the time to market, and lower the costs involved [6]. In order to achieve the aforementioned goals, very effective UI automation methods are required. The limitations of existing UI automation tools/solutions are their dependency on UI components, APIs, layouts, and DOM properties which are very fragile (tend to change) [6–8]. Testing processes have many aspects of validation such as reliability, usability, integrity, maintainability, and compatibility [9]. Testing and maintenance are one of the important aspects of software engineering [10] and have become the slowest phase in the complete **SDLC** [9, 11], reason being the entire process is manual or manually maintained automation [12–14]. Software testing is already very labor and resource intensive, accounting for up to 50–60% of total software development costs [7, 8, 15] and 50–75% of the total cost of a product release [11, 16]. Testing a graphical user interface (GUI) requires human testers to validate the application’s visual behavior with the intended results [17]. Over time, companies have taken cognizance of the impact to (total cost of ownership (TCO) this causes. Automation UI testing was intended to lower the cost, increase the product quality, and reduce the maintenance effort, but the involvement of cost for automation cycle is still very high and regularly overshoots the planned budget; in [15], the authors talk about the huge costs still associated with automation testing [18]. In [11], the authors have mentioned seven down sides of automation testing ranging from manual maintenance of tests to the slow and unreliable nature of automation scripts. Dependency on manual test cases and manual maintenance of automation has instilled fears with respect to the unanticipated costs of automation. Manual test case generation depends on: 1. Testers who write it based on the user guideline documents and 2. Generating the test cases based on the availability of the meta document. Both the above approaches involve the human to understand the automation framework and then generate test cases based on the requirements. In this proposed solution, we introduce a concept which requires very minimal effort on the part of testers to generate test cases. We propose a three-part syntax to define the actions to be performed that we call the triplet approach: <action> <label> <value> (e.g., fill username testuser). The complete test case is composed of a set of triplets; testers can write a test cases just by looking at an application’s UI unlike existing solutions that require testers to locate technical.

UI properties using browser developer tools or source code [19]. Maintenance of automation is the biggest challenge, and as mentioned above, the current automation approaches depend on the document object model (DOM) and UI elements properties of the application [15]. DOM/UI elements mainly consist of combination of hypertext markup language (HTML) tags and their attributes such as ID, class, xpath (relative path to the UI elements). In addition, finding the DOM properties from the application is very time consuming. The properties of the UI elements are very fragile, changing every product release breaking the automated test cases, leading to a large number of false negatives. To fix these broken automation scripts, manual effort is needed; testers must analyze the automates to find the reason of failure and then fix it which is very costly and time consuming. In this paper, we present a deep learning-powered UI technology agnostic approach, which does not depend on the underlying technology or DOM/properties of UI elements of the application. This approach tackles the aforementioned fragile behavior that can no longer affect the automated test cases unless there is a functional change. Our approach is inspired by the testing behavior of the manual testers; it parses the screen the same way a manual tester does and identifies the UI elements based on their visual appearance; for example, a button's visual appearance will always be like a button on any UI application, and it cannot appear to be a text field or checkbox, if that is the case then the script should fail as an ideal fail because even a human tester would not identify that. Our approach is visual, which generalizes on different UI applications unlike [17, 20], which restrict to a fixed UI element. The proposed solution will overcome the most challenging, time consuming, and costliest problem while enabling UI test-driven development (TDD) among other use cases. There is a huge potential for improvement in the approached solution domain (UI automation).

2 Related Work

There are two phases to UI automation: First is the test case generation, and second is the execution of the generated test cases.

Test case generation can be difficult based on the requirement of the automation framework; the simplest test case generation is writing a test case by looking at the UI application; in the paper [17], Tsung et al. (2010) have proposed visual test and execution, which uses the images of UI elements directly inside the test cases (taken while creating it); this method is very effective when the application is static, and the UI elements used in the test creation remain the same on the UI application. When there is a change in the UI application, testers have to take a snapshot of the UI element and then create the new test cases or update the old image with the new one. Unlike this approach, we have flexibility in UI component identification, which gives us an advantage, as it does not rely on the UI element snapshots while creating test cases and generalizes for non-functional changes (change in element properties), e.g., UI element's color and size.

Amant et al. (2001) talk about the different visual properties to consider, such as size, shape, color, appearance. Recent increases in computing power have opened vast opportunities for image processing and have augmented the possibilities of real-time analysis of the images. The approach [21] records the static position of the objects on the screen which may change for different screen size's images [21]; the proposed solution has managed the resolution change drawback by providing a resolution manager, which uses the current size of the image to predict the UI element's location based on the trained neural network, which generalizes on different aspect ratio images.

3 Methodology

Our method is divided into two phases; the first phase is the test case (we call it **instruction set**) generation; the most popularly used method in test automation industries is “record and play” [22, 23] which retrieves and saves the UI properties while recording and then performs the test on the UI application. We have used a different approach because our solution does not require properties of the UI elements of the application. The second phase is the execution engine which takes the instruction set as input and creates an automation script without any further human intervention, which will be used as many times as required.

3.1 Instruction Set

The first required step to create the automation is the instruction set, which describes the stepwise functionality to be followed by the automation script. Existing methods of creating an instruction set involves a human reading a PDF/text document file and using that to record the UI application which generates a JSON/XML file which is used by the automation engine to replay the same steps. We propose a triplet-based instruction set, which holds the minimal required information to test and automate the UI application. The proposed triplet holds the three-key information: **(a)** action, **(b)** label, and **(c)** value. The “action” of the instruction specifies the type of UI element on the screen; typical UI actions are fill, click, select, drag, etc. This approach handles all kinds of label-based action types. The “label” in this instruction set is the visible human readable text which is present on the UI application screen.

The “value” in the instruction set is the actual data which will be applied to the UI element. In Fig. 1, the instruction set consists of four steps: The first step will **navigate** to the application, and the second step is <fill user testuser>, which means the type of action is **fill**, the label on the screen is “user,” and the value to be filled is “testuser.” Similarly, the third step is <fill password testpass> which means the type of action is **fill**, the label is “password,” and the value to be filled is “testuser,” In the fourth step, <click log on>, the action type is **click**, which means this is a

1. **navigate https://my_app.com**
2. **fill user testuser**
3. **fill password testpass**
4. **click log on**

Fig. 1 Instruction set

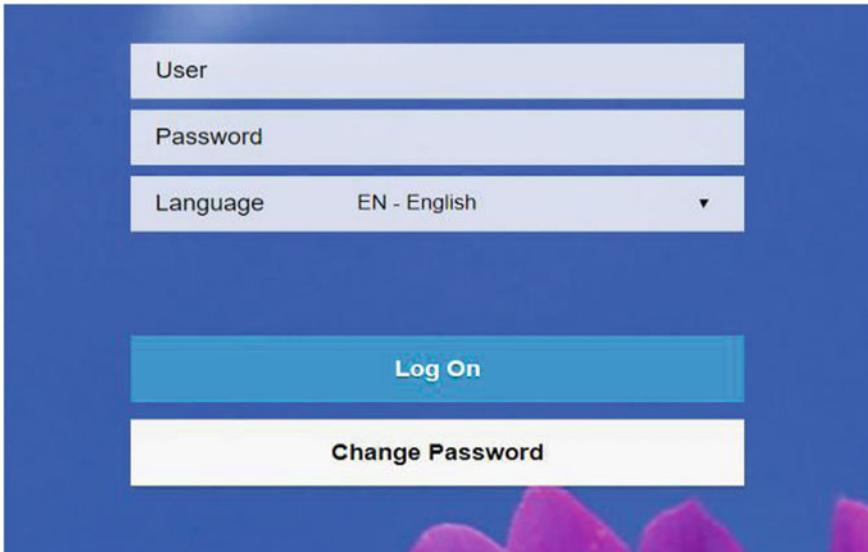


Fig. 2 UI application snapshot

clickable element, and second item (“log on”) is the label where it expects a click event. Different entities in a step are separated by double space, and there are double space among action type fill, label “user,” and value “testuser.” Single space between entities will be treated as single entity; for example, in Step 3, entity “log” and “on” is a single space separated entity, and it is treated as “log on,” a single label. Figure 2 is a snapshot of the SAP S/4 HANA cloud log on screen. The instruction set in Fig. 1 was written just by looking at the snapshot.

3.2 Execution Engine

In simple terms, “*it uses deep learning to detect the UI elements/controls and uses OCR to read the texts on the detected UI elements/controls if any.*”

User interface consists of multiple UI elements on a single page which makes the detection process very complex. Visually detecting and understanding the UI elements/controls undergoes two major processes.

I. Object Detection Using Deep Learning (AI engine)

Object detection means detecting the various UI elements/controls/objects on the image. Automatically detecting objects/icons on the screen requires understanding of each object which demands a large annotated dataset. The required dataset consists of two different files. First is the image file which holds the pixel information of every object, and second is the image annotation XML file, which holds the pixel location of every intended object. There will be a possibility wherein the image contains many objects and you want to detect only few of them; in this case, annotations help the neural network to only focus on those objects which are required. Figure 3 shows a snap view of annotated file which holds the information about the input image, folder containing, its file location, file name, its size, and the intended object locations on the input image file. While training after every iteration, network generates multiple

```

<annotation>
    <folder>folder_name</folder>
    <filename>image_name.jpg</filename>
    <path>image_file_location</path>
    <size>
        <width>1920</width>
        <height>922</height>
        <depth>3</depth>
    </size>
    <object>
        <name>textfield</name>
        <bndbox>
            <xmin>222</xmin>
            <ymin>155</ymin>
            <xmax>617</xmax>
            <ymax>197</ymax>
        </bndbox>
    </object>
</annotation>
```

Fig. 3 Annotated file snapshot [PASCAL VOC [24] XML format]

files including weights and checkpoints. The weight files are used by neural network at the start of every iteration to improve the training, and at the end of the iteration, the weight files will be updated for the next iteration. In case of training failure, checkpoint will be used by network to recover from the failure. After successful training, network generates a single file, containing only the required trained information capable of predicting on new testing data. In a production environment, the same file will be used by serving model for the prediction requested by the execution engine.

II. Optical character recognition (OCR) [OCR engine]

AI engine predicts the objects location on the image, and to differentiate different labeled same UI elements, we used optical character recognition (OCR); it reads the text in images character by character. We used the similar approach as [25] while training the neural network. OCR engine takes the predicted UI elements as input and returns the textual data available inside the image. Training the OCR engine to detect textual data inside the image required a different set of images containing text against different backgrounds. We automated this process to generate the training dataset. The generated data contained all possible combinations of SAP's application's background color and text which gave us a diverse range for detecting the textual data inside any application's snapshots. We trained a convolutional neural network with dataset of size ~20 k on 2vCPU, 16 GB RAM virtual computer, and we got an accuracy that successfully passed all the required test cases. Execution engine takes the instruction sets as input and bucketizes each instruction, associating the label with the action and value of the instruction. Then, it navigates to the required application page and performs the instruction. It takes a screenshot of the required page (specified in the instruction set with the **navigate** action item) (Fig. 1). The screenshot is passed to the AI engine, which helps to detect all required UI elements inside the screenshot. The AI engine uses the trained model, and based on its learning, it locates all available objects with their type in the screenshot. The results are returned to execution engine, which will use the location of the objects and crop the screenshot at the predicted locations. The cropped images are passed to the OCR engine which will return the text containing if any, back to the execution engine. Based on the labels from the instruction set, it matches the results returned from OCR engine. Now, it will click on the location which is returned from the AI engine, and execution will proceed to the next instruction. The complete flow of execution is mentioned in Fig. 4, from parsing the instruction to detecting objects and recognizing text and then finally actual execution. After the execution of automation, the execution engine will generate a report which will hold the execution status of every step. For each failed step, there will be a description stating the reason of failure with a screenshot attached, while for success cases the screenshots will be attached to the generated report. There will be two scenarios while creating new automation test cases, and we have handled those separately:

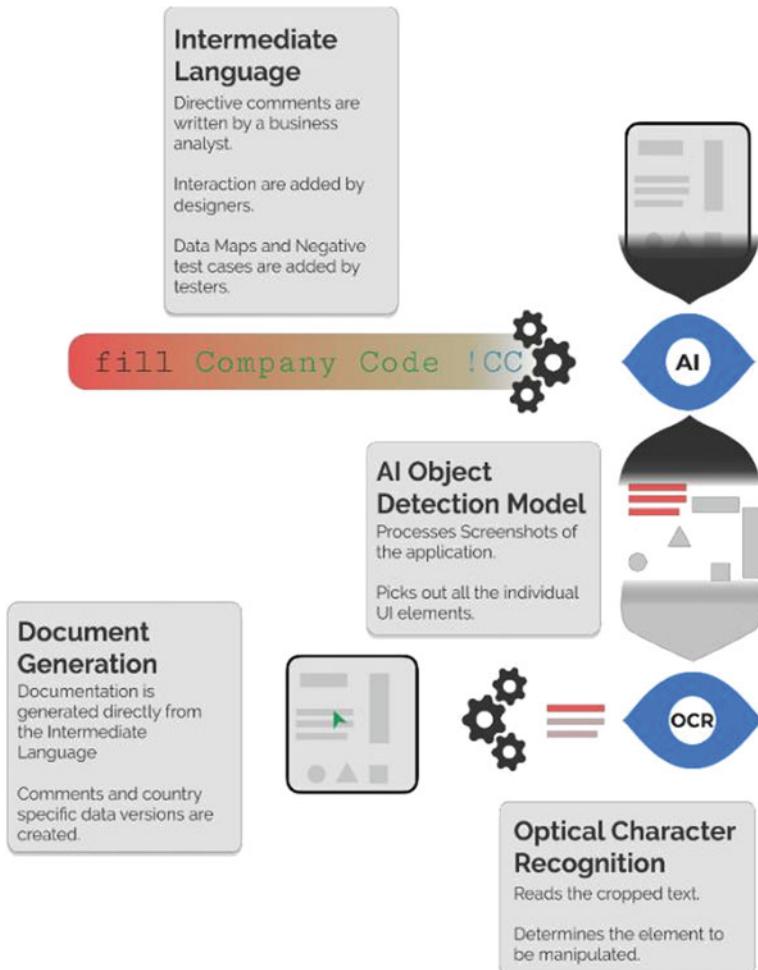


Fig. 4 End-to-end execution flow of the system

a. UI Elements with a Label

When a UI element contains a label, it becomes easy for OCR to read the text and differentiate it from other similar UI elements. In this case, the process follows two steps to detect the elements uniquely, which is using the AI engine to detect the elements and OCR to reads its text.

b. UI Elements with No Label

When the UI elements do not contain any text/label, there is no way to detect it uniquely other than its visual appearance (e.g., icons); we have trained the neural network just to predict the specific logos and icons. While training, we gave a name

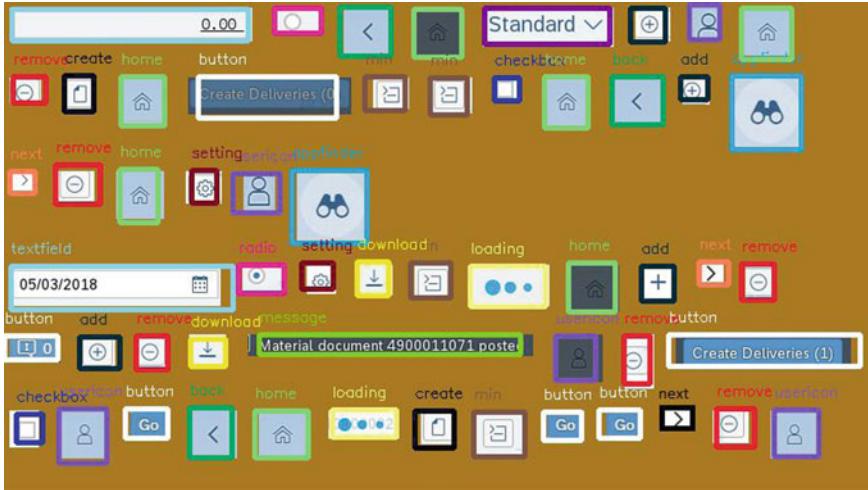


Fig. 5 AI model (20 classes) prediction on test data

to each logo/icon and that name will be predicted while creating the automation. In Fig. 5, there are prediction for the icons, e.g., home, create, back, remove. We will use these names while creating the automation scripts. For example—to click on the create icon, the automation instruction will be <Click create>.

At the time of running the automate, the execution engine takes the predicted results from the AI engine, and it will look for the prediction of type “create,” this is then interpreted as a non-standard UI action type, and the OCR engine will not be called. It will take the AI engine’s predicted (icon’s) location and will perform the specified action, “Click” in this case.

4 Results

We started the complete process with the aim of detecting only seven objects from a single image. We collected around 250 images for each object, and in total, we collected total around 2000 images. We have used YOLO architecture to train the model for seven classes.

After approximately 36 h of training, we stopped the process, where the loss was minimum at about 3.6 every epoch (Fig. 6) [26]. We tested the trained model on the new images, and we got a good accuracy.

This result motivated us to move forward, address the wider challenge, i.e., include more UI elements. We considered the most used SAP S/4 HANA cloud applications and their complete process flows, selecting 20 different classes that were required for their execution. We trained the neural network for more than 80 h on 2vCPU, 16 GB RAM virtual computer; this process could have been expedited greatly by the

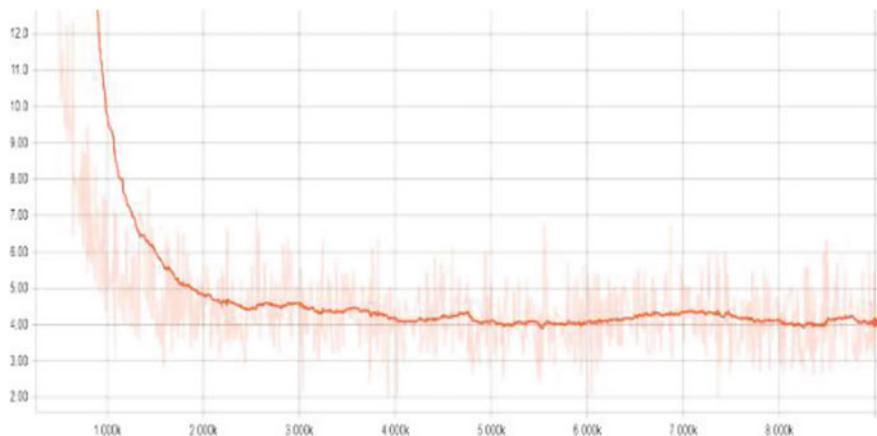


Fig. 6 AI model average loss for seven classes of around 96%, with which we were able to detect all the seven classes in any UI environment [Fig. 7]

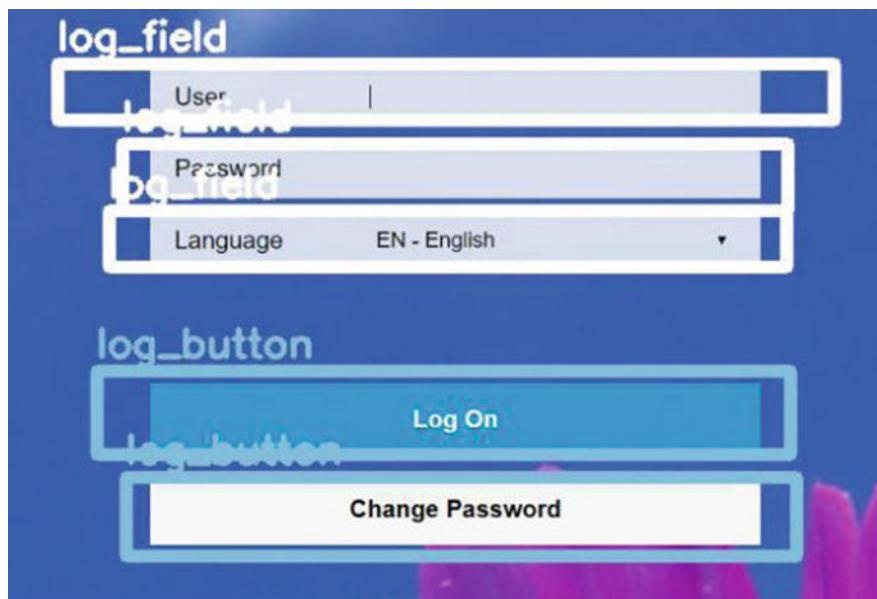


Fig. 7 Prediction of object by AI model

usage of GPUs. We achieved a phenomenal accuracy, and to validate it, we tested the network on different background images and what we found was totally unbelievable; we touched to golden 95% accuracy (Fig. 5).

To validate our OCR results, we ran the OCR model on all the test data and found that it was able to detect all the text with a 99.9% accuracy. Then, we validated the

Table 1 OCR accuracy on different input categories

Text type	Data size (k)	Accuracy (%)
Black text in white background	5	100
White text in black background	5	99.9
Black text in blue background	5	100
Black text in gray background	5	99.6
Text with same background color	2	0.1

extensibility of OCR model on generated samples with different background colors and different text transforms (Table 1). We additionally included the special case of same text color and background color. Here, its accuracy was exceptionally poor, and this behavior proved beneficial to our use case since in a UI application a human cannot see such texts, and hence, the automation should fail.

5 Conclusion

In this paper, we presented a very novel and urgently required method to solve the biggest UI automation problem using deep learning to detect UI elements on an image and followed by OCR to read the textual data from the UI application screenshots. This approach covered the end-to-end UI automation flow with no requirement for automation maintenance. Our very next target on the roadmap is to reduce the manual effort involved in the annotation of images, which will enable faster scaling up. Exceptionally, large training data could be the key to analyzing very complex images. We will also focus on increasing the accuracy of the AI model, which will lead to an ideal zero-maintenance system that will not require any human intervention once the automation script is created.

References

- Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
- Li J, Liang X, Shen S, Xu T, Feng J, Yan S (2018) Scale-aware fast R-CNN for pedestrian detection. *IEEE Trans Multimedia* 20(4):985–996
- Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
- Li P, Huynh T, Reformat M, Miller J (2007) A practical approach to testing GUI systems. *Empirical Softw Eng* 12(4):331–357
- Dallmeier V, Burger M, Orth T, Zeller A (2013) WebMate: generating test cases for web 2.0. In: International conference on software quality. Springer, Berlin, Heidelberg, pp 55–69

6. Börjesson E, Feldt R (2012) Automated system testing using visual gui testing tools: a comparative study in industry. In: 2012 IEEE fifth international conference on software testing, verification and validation. IEEE, pp 350–359
7. Memon AM (2002) GUI testing: pitfalls and process. Computer 8:87–88
8. Benedikt M, Freire J, Godefroid P (2002) VeriWeb: Automatically testing dynamic web sites. In: Proceedings of 11th international world wide web conference (WWW'2002)
9. Maheshwari S, Jain DC (2012) A comparative analysis of different types of models in software development life cycle. Int J Adv Res Comput Sci Softw Eng 2(5):285–290
10. Monier M, El-mahdy MM (2015) Evaluation of automated web testing tools. Int J Comput Appl Technol Res 4(5):405–408
11. Kazmi R, Afzal RM, Bajwa IS (2013) Teeter-totter in testing. In: 2013 Eighth international conference on digital information management (ICDIM). IEEE, pp 194–198
12. Memon AM (2007) An event-flow model of GUI-based applications for testing. Softw Test Verification Reliab 17(3):137–157
13. Sharma M, Angmo R (2014) Web based automation testing and tools. Int J Comput Sci Inf Technol 5(1):908–912
14. Jain A, Jain M, Dhankar S (2014) A Comparison of RANOREX and QTP Automated Testing Tools and their impact on Software Testing. IJEMS 1(1):8–12
15. Harrold MJ (2000) Testing: a roadmap. In: Proceedings of the conference on the future of software engineering. ACM, pp 61–72
16. Fuggetta A, Di Nitto E (2014) Software process. In: Proceedings of the on future of software engineering. ACM, pp 1–12
17. Chang TH, Yeh T, Miller RC (2010) GUI testing using computer vision. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 1535–1544
18. Sjösten-Andersson E, Pareto L (2006) Costs and benefits of structure-aware capture/replay tools. In: SERPS'06, p 3
19. Nagarani P, VenkataRamanaChary R (2012) A tool based approach for automation of GUI applications. In: 2012 Third international conference on computing communication and networking technologies (ICCCNT). IEEE, pp 1–6
20. Yeh T, Chang TH, Miller RC (2009) Sikuli: using GUI screenshots for search and automation. In: Proceedings of the 22nd annual ACM symposium on user interface software and technology. ACM, pp 183–192
21. Amant RS, Lieberman H, Potter R, Zettlemoyer L (2001) Visual generalization in programming by example. In: Your wish is my command, pp 371–XIX
22. Jovic M, Adamoli A, Zaparanuks D, Hauswirth M (2010) Automating performance testing of interactive java applications. In: Proceedings of the 5th workshop on automation of software test. ACM, pp 8–15
23. Andrica S, Candea G (2011) WaRR: a tool for high-fidelity web application record and replay
24. Everingham M, Winn J (2011) The PASCAL visual object classes challenge 2012 (VOC2012) development kit. Rep, Pattern Analysis, Statistical Modelling and Computational Learning, Tech
25. Yin XC, Yin X, Huang K, Hao HW (2014) Robust text detection in natural scene images. IEEE Trans Pattern Anal Mach Intell 36(5):970–983
26. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, Kudlur M (2016) Tensorflow: a system for large-scale machine learning. In: OSDI, vol 16, pp 265–283
27. Agarwal S, Awan A, Roth D (2004) Learning to detect objects in images via a sparse, part-based representation. IEEE Trans Pattern Anal Mach Intell 26(11):1475–1490

Load Scheduling with Combinations of Existing Tariff Structure for Residential Consumers in Maharashtra, India—Case Study



Archana Talhar and Sanjay Bodkhe

Abstract The growth of any country depends upon the availability, accessibility and growth of electricity. The need of energy rises due to the modern lifestyle and advanced technologies in residential sector and hence in industrial sector. The uneven demand in energy during different time intervals of a day leads to unbalanced load curve at power plant. To balance the load curve, demand during peak hours need to be reduced and this can be adjusted during off peak hours. Therefore, this paper presents the load scheduling scheme for residential consumers with the combinations of existing tariff structure. This will help in reducing the peak load thereby benefitting both consumers and utility. The case study of residential consumer with the combinations of tariff structures like (1) slab-wise tariff, (2) slab-wise tariff and Time of day tariff (ToD) with usual load pattern and (3) slab-wise tariff & ToD tariff with load scheduling is presented to validate the proposed load scheduling scheme. The same proposal with little modifications can be implemented for industrial sector also.

Keywords Combined tariff · Existing tariff · Demand · Load curve · Load scheduling · Residential sector · Time of day

1 Introduction

Due to modern lifestyle and advanced technology, the demand of energy is increasing day by day. The energy generating resources are limited and will be vanishing in coming years. Generating more electricity cannot be the only solution to meet the increasing demand. There are various resource constraints for generation such as limited land, fuel, water, etc. There are also social and environmental concerns in the

A. Talhar (✉) · S. Bodkhe
Ramdeobaba College of Engineering and Management, Nagpur, India
e-mail: archanabelge2@gmail.com

S. Bodkhe
e-mail: bodkheshb@rk nec.edu

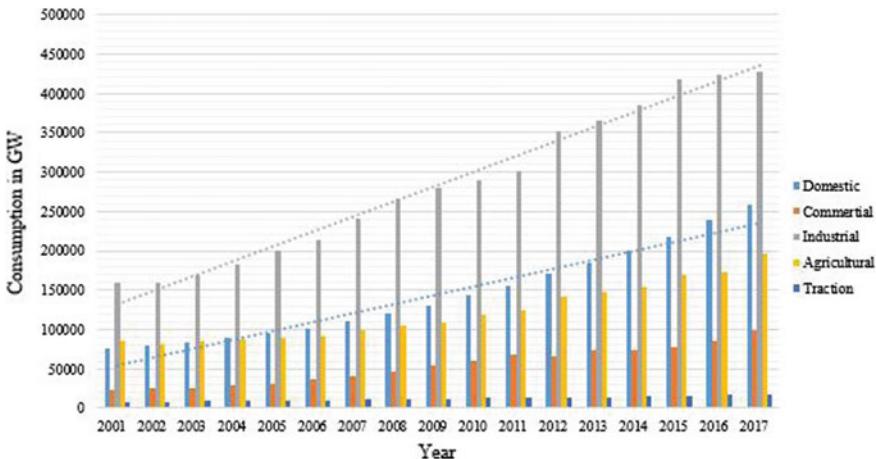


Fig. 1 Growth of electricity consumption in India. *Source* Central Electricity Authority Report, Government of India, 2017

siting of new power plants. Hence, it is necessary to use energy with responsibility [1, 2]. The major share in energy consumption is from industrial and commercial sector. Therefore, for balancing the load curve and for reducing the electricity, Government has implemented many policies and schemes. But now the scenario is changed. It is observed that after industrial sector, the growth of energy consumption is highest in domestic sector shown in Fig. 1 [3]. Therefore, for reducing energy consumption and energy bill in residential sector, many researches are trying to find the solutions. The main focus is on smart home energy management systems. A system which takes care by itself without human intervention depending upon the availability of supply is called as smart home or modern homes in which home appliances can be remotely monitored and controlled by the owner via a mobile app, Internet of Things, Wi-Fi, ZigBee, etc. The Major benefits of smart home energy management systems are:

- It facilitates owner to remotely control the home appliances.
- It configures time schedules for smart home-enabled devices to help in controlling costs.
- It is more energy-efficient (i.e., green homes).
- It provides convenience and potential time savings.

In paper [4], author has presented the hardware for smart home energy management system using communication and sensing technology which helps in reducing the cost of electricity. In paper [5, 6], author has presented energy and cost efficient scheduling algorithm for reducing electricity bill. In paper [7], author has presented home energy management system with radio frequency (RF) communication which considers customer load priorities and preferences. In paper [8], author has presented hardware design of smart meter which provides remote monitoring, energy management, feasible control. In paper [9], author has presented smart home renewable

energy management system results by saving in home energy bill. In paper [10], author has discussed the energy management programs like energy efficiency standards, labeling and policy instruments to tackle the standby power losses. In paper [11], author has presented design of low cost, low maintenance smart stick-on sensors for monitoring utility assets. In paper [12, 13], author has presented system architecture to lower electricity bills and smart energy distribution and management system for renewable energy distribution. In paper [14], author has presented comparison of the energy storage technologies with different parameters like durability in terms of time and cycles etc. In paper [15, 16], author has presented a detailed architecture of net meter. It keeps track of the difference between the electricity imported from grid and the surplus energy or exported electricity to grid. In paper [17, 18], author has presented comparison of the performance of flat price, time of use, real time pricing and enhanced time of use pricing scheme. In paper [19], author has presented methodology to systematic approach for utility engineers to reduce the peaks.

2 Need of Load Scheduling

A survey of load consumption pattern of Maharashtra state is carried out. To understand the load demand, data is collected for all seasons. The analysis for summer, winter and rainy seasons for the last four years is carried out on the sample basis and shown in Figs. 2, 3 and 4 [20].

It is observed that from 0900 h–1200 h and 1800 h–2200 h load demand is very high and are called as morning peak and evening peak respectively. From 2200 to 0600 h demand is very less and this period is called as off peak, whereas 0600–0900 h and 1200–1800 h demand is moderate and is called as base period. Because of this

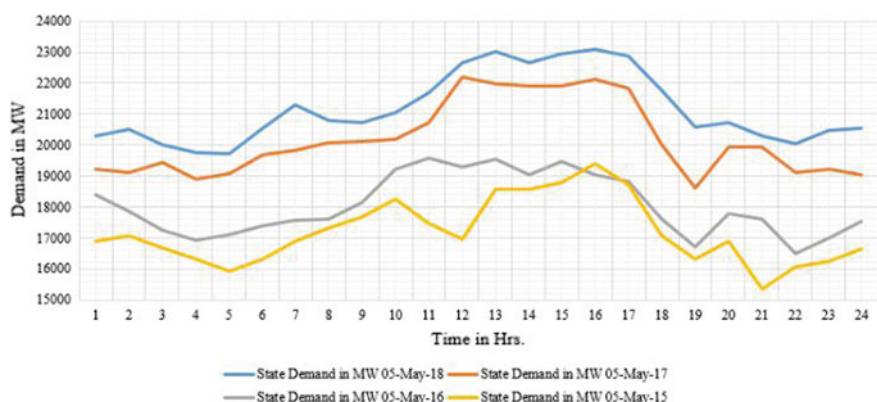


Fig. 2 State demand in MW for summer season. *Source* Maharashtra SLDC, 2018

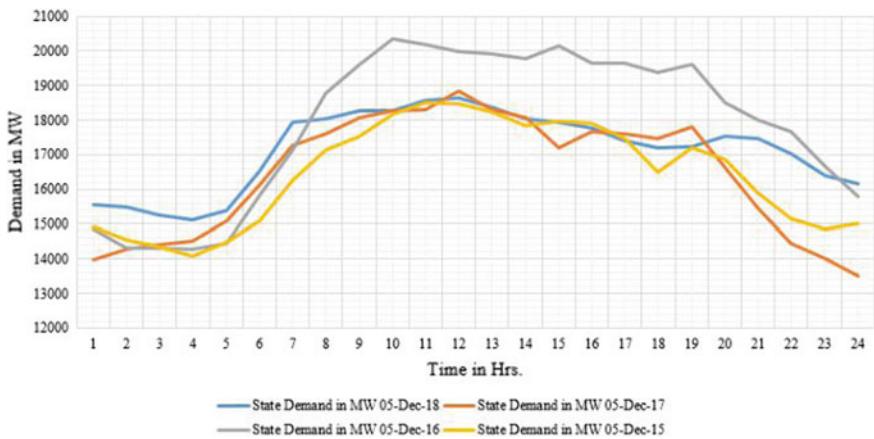


Fig. 3 State demand in MW for winter season. *Source* Maharashtra SLDC, 2018

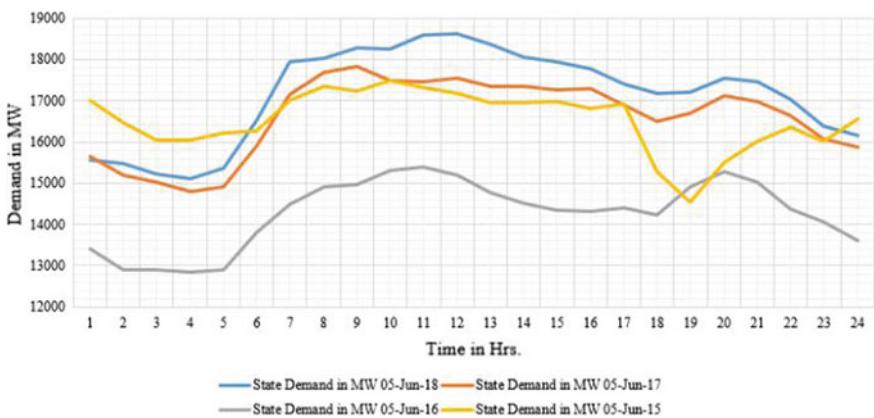


Fig. 4 State demand in MW for rainy season. *Source* Maharashtra SLDC, 2018

variation in demand of energy during different time intervals in a day, utility and consumers both face problems as discussed below.

2.1 Problem Faced by Utilities

Very high demand during peak hours. In peak hours, demand is very high. For fulfilling this demand, utility requires additional capacity of generators, thereby increasing capital cost. In case if this is not possible, then purchase power from other utilities like Adani, TATA etc. which again leads to increase in power purchase cost.

Low demand during off peak hours. For reducing generation during off peak hours, the only solution is to shut down few generators. This process of start and stop of generators as per load demand requires additional time and fuel which is very expensive and leads to power loss also.

Fuel consumption during peak hours increased. This ultimately leads to carbon emission.

2.2 Problem Faced by Consumers

Load-shedding. Due to unbalance load curve, i.e., high demand in peak periods and low demand in off peak periods, consumers staying at remote location are not getting continuous electricity supply.

Higher electricity bill. As the consumption increases, electricity bill of consumers also increases in slab-wise tariff structure. At present, ToD tariff is available only for non-residential consumers.

Electricity tariff variation due to multiple utilities. In Maharashtra, except Mumbai, consumers have no choice for utility selection. Whereas in Mumbai, consumers have many options like Adani, TATA, BEST and MSEDC.

3 Proposed Load Scheduling Scheme

The proposed scheme is based on the combination of existing tariff structure (slab-wise tariff and ToD tariff) and load scheduling with the sole objective of minimizing bill of residential consumer. So, in this work an assumption is made that system is running on ToD. In ToD tariff, the rates of electricity are high during the peak period, medium during the shoulder period and low during the off peak period. So if the proper load scheduling is carried out then it will help the consumer to minimize the electricity bill. For demonstration purpose, twelve loads are considered (1—LED light, 2—Fan, 3—Refrigerator, 4—Water purifier, 5—Television, 6—Mixer, 7—Washing machine, 8—Iron, 9—Water pump, 10—Water heater, 11—Microwave oven, 12—Air conditioner). The load is divided into three categories: Highly Essential (HE), Moderately Essential (ME) and Least Essential (LE) loads. Depending upon the essentiality, consumer preference and ToD the load will operate to minimize the electricity bill. The flow chart for the same is shown in Fig. 5.

The Algorithm is explained with following steps:

- **Step (1)** User defines Time of Day (ToD) slots for 24 h. These 24 h are divided into four intervals: Off Load Period (OLP), Base period (BSP), Moderate Load Period (MLP) and Extensive Load Period (ELP).
- **Step (2)** User defines load as per essentiality: Highly Essential (HE), Moderately Essential (ME) and Least Essential (LE).

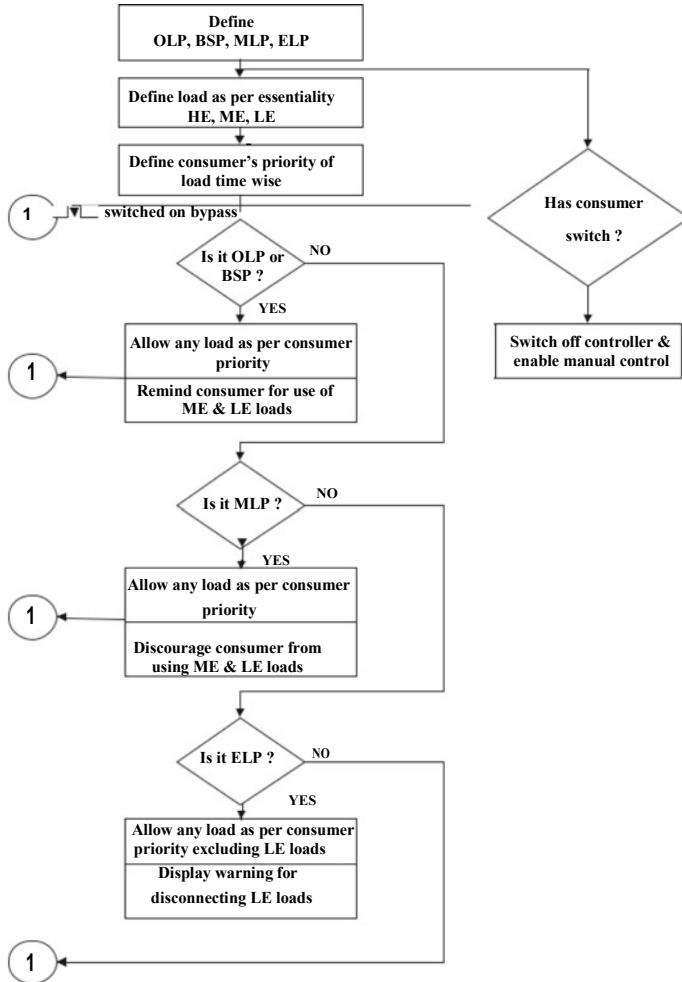


Fig. 5 Proposed load scheduling scheme

- **Step (3)** Define consumer's priority of load time wise.
- **Step (4)** During OLP: Allow any load as per consumer's priority and give reminder to consumer for use of ME and LE loads.
- **Step (5)** During BSP: Allow any load as per consumer's priority and give reminder to consumer for use of ME and LE loads.
- **Step (6)** During MLP: Allow any load as per consumer's priority and discourage consumer for use of ME and LE loads.
- **Step (7)** During ELP: Allow any load as per consumer's priority excluding LE loads and give warning to consumer for disconnecting LE loads.

- Step (8)** After step 1, If bypass switch is ‘ON’ then controller will switch off and enable manual control mode.

The tariff structure of one of largest utility in Maharashtra, i.e., Maharashtra State Electricity Distribution Company Limited (MSEDCL) is used for bill calculations. As initial assumption is made that real time tariff is implemented so, ToD tariff structure of MSEDCL is used for calculations. Tables 1 and 2 shows the tariff structure used for calculations.

The usual load consumption pattern of a sample case study is shown in Fig. 6. With this load consumption pattern, the monthly energy consumption is 829 units. According to slab-wise tariff structure, the monthly electricity bill is Rs. 8702.89/month. After implementation of time of day tariff, consumer is getting a

Table 1 Slab-wise tariff structure of MSEDCL for residential sector

Consumption slab (kWh)	Fixed/demand charge (Rs./month)	Wheeling charge (Rs./kWh)	Energy charge (Rs./kWh)
0–100 units	1-Phase: 65 per month 3-Phase: 185 per month	1.18	3.07
101–300 units		1.18	6.81
301–500 units		1.18	9.76
501–1000 units		1.18	11.25
Above 1000 units		1.18	12.53

Table 2 ToD tariff structure of MSEDCL for non-residential sector

Time slots in hours	Energy charge above base charge (Rs./kWh)
2200–0600	−1.50 (Rebate)
0600–0900 and 1200–1800	+0.00 (Base rate)
0900–1200	+0.80 (Penalty)
1800–2200	+1.10 (Penalty)

Household appliance	12AM	1AM	2AM	3AM	4AM	5AM	6AM	7AM	8AM	9AM	10AM	11AM	12PM	1PM	2PM	3PM	4PM	5PM	6PM	7AM	8AM	9AM	10AM	11AM	12AM	OLP	ELP	OLP
	1	HE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	HE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	HE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	HE																											
5	ME																											
6	ME																											
7	ME																											
8	LE																											
9	LE																											
10	LE																											
11	LE																											
12	LE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Fig. 6 Usual load pattern of residential consumer sample case, Whereas, 1—LED light, 2—Fan, 3—Refrigerator, 4—Water purifier, 5—Television, 6—Mixer, 7—Washing machine, 8—Iron, 9—Water pump, 10—Water heater, 11—Microwave oven, 12—Air conditioner

rebate of Rs. 380.28 over an actual electricity bill. After rescheduling the high energy consumption load, i.e., water pump shifted from MLP, ELP to OLP; water heater shifted from MLP, ELP to BSP and OLP; energy consumption remains the same and now consumer is getting a rebate of Rs. 512.37.

4 Result

From the case study, carried out on the combination of existing tariff structure for residential consumers in Maharashtra, it is observed that, implementation of ToD tariff for residential consumers will encourage to reduce the electricity bills. The final results are presented in Table 3.

It is observed that the electricity bill with slab-wise tariff is Rs. 8702.89 per month and with combined tariff is Rs. 8322.61 per month. After rescheduling the high energy consumption load electricity, bill is Rs. 8190.52. The number of unit's consumption in all three cases is same but the bill has reduced in the second case by Rs. 380.28 per month and third case by Rs. 512.37 as compared to slab-wise tariff.

Table 3 Comparison of electricity bill of sample case study

Case study	No. of units consumed in one day	No. of units consumed in one month	Electricity Bill in Rs. per month	Remark
(1) Based on the existing slab-wise tariff and the usual load pattern	28	829	8702.895	–
(2) Based on the proposed combined tariff and usual load pattern	28	829	8322.615	Rebate of Rs. 380.28 earned by the consumer
(3) Based on the proposed combined tariff and rescheduled load pattern	28	829	8190.525	Rebate of Rs. 512.37 earned by the consumer

5 Conclusion

From the case study discussed in this paper, it is found that, if the consumer reschedules the load pattern with an objective to reduce the energy bill without compromising much on the comfort level, there will be enough scope to save in electricity bill. However, if the consumer does not care for the global need of energy conservation, consumer will have to pay more for the same electricity consumed. The utilities shall be able to charge extra from such consumers and pay a part of it or more as a rebate to the former ones. The reduction in load on the grid during peak hours will ultimately reduce the cost incurred in enhancement of generation capacity also. Thus, the proposed combined tariff structure, i.e., slab plus ToD tariff structure is superior over the existing slab-wise tariff for all concerned stakeholders.

References

1. Demand side management in india: an overview of state level initiatives. Prayas Energy group. <http://www.prayaspune.org/peg/publications/item/281-demand-side-management-in-india-an-overview-of-state-level-initiatives.html>
2. Talhar AS, Bodkhe SB (2019) The global survey of electrical energy distribution system: a review. IJECE 19:2247–2255. <https://doi.org/10.11591/ijece.v9i4>
3. Central Electricity Authority Report (Government of India) (2017) <http://www.cea.nic.in/annualreports.html>
4. Hu Q, Li F (2013) Hardware design of smart home energy management system with dynamic price response. IEEE Trans Smart Grid 4(4):1878–1887
5. Chen X, Wei T (2013) Uncertainty-aware household appliance scheduling considering dynamic electricity pricing in smart home. IEEE Trans Smart Grid 4(2):932–941
6. Guo Y, Pan M, Fang Y (2012) Optimal power management of residential customers in the smart grid. IEEE Trans Parallel Distrib Syst 23(9):1593–1606
7. Rajan S, Thomas M (2015) An efficient home energy management algorithm for demand response analysis in Indian scenario. In: IEEE Annual India Conference (INDICON) 2015
8. Mathavi S, Vanitha D, Jeyanthi S, Kumaran PS (2012) The smart home: renewable energy management system for smart grid based on ISM band communications. Int J Sci Eng Res 3(3):1–8
9. Al-Ali AR, Hag AE, Bahadiri M, Harbaji M, Haj YA (2011) Smart home renewable energy management system. Energy Procedia 12:120–126 Elsevier
10. Solanki PS, Mallela VS, Zhou C (2013) An investigation of standby energy losses in residential sector: solutions and policies. Int J Energy Environ 4(1):117–126
11. Moghe R, Lambert FC, Divan D (2012) Smart stick-on sensors for the smart grid. IEEE Trans Smart Grid 3(1):241–252
12. Mishra A, Irwin D, Shenoy P, Kurose J, Zhu T (2013) Green charge: managing renewable energy in smart buildings. IEEE J Sel Areas Commun 31(7):1281–1293
13. Byun J, Hong I, Kang B, Park S (2011) A smart energy distribution and management system for renewable energy distribution and context-aware services based on user patterns and load forecasting. IEEE Trans Consum Electron 57(2):436–444
14. Ferreira HL, Garde R, Fulli G, Kling W, Lopes JP (2013) Characterization of electrical energy storage technologies. Elsevier J Energy 53:288–298
15. Bedi HS, Singh N, Singh N (2016) A technical review on solar-net metering. In: India international conference on power electronics (IICPE)

16. Maharaja K, Balaji PP, Sangeetha S, Elakkiya M (2016) Development of bidirectional net meter in grid connected solar PV system for domestic consumers. In: International conference on energy efficient technologies for sustainability (ICEETS)
17. Nazar N, Abdullah M, Hassan M, Hussin F (2012) Time-based electricity pricing for demand response implementation in monopolized electricity market. In: IEEE students conference on research and development (Malaysia)
18. Azman N, Abdulla M, Hassan M, Said D, Hussain F (2017) Enhanced time of use electricity pricing for industrial customers in Malaysia. Indonesian J Electr Eng Comput Sci 6(1):155–160
19. Shaikh S, Dharme A (2009) Time of use pricing-India, a case study. In: Third international conference on power systems, Kharagpur, INDIA
20. Maharashtra State Load Dispatch Center, mahasldc.co.in

Mammogram Classification Using Rotation-Invariant Local Frequency Features



Spandana Paramkusham and C. Venkata Narasimhulu

Abstract Breast cancer accounts for the highest mortality rate among women in the world. Mammograms play a prominent role in detecting abnormalities in the breast. Computer-aided diagnosis systems help the radiologist in detection abnormalities in less time. This work deals with the extraction of features from ROIs to reduce false positives in computer-aided diagnosis systems. In this paper, the rotational invariant local frequency technique is implemented using three methods for the extraction of features from mammogram region of interest (ROIs). Features obtained from ROIs are given to SVM for further classification of ROIs into normal-abnormal using SVM classifier via 10 fold cross-validation method. The proposed methods are validated using ROIs obtained from Image Retrieval in Medical Applications (IRMA) database for feature extraction

Keywords Breast · Mammograms · Feature extraction · Classification · SVM

1 Introduction

Cancer is caused due to the uncontrollable growth of abnormal cells. These cells grow proliferate because of damaged DNA (deoxyribonucleic acid). In breast, these cells grow abnormally locally and metastasize in lymph nodes or ducts. Breast cancer occurs about 14% of all cancers in women in India. According to Globacon 2018 survey [1], nearly 162,468 new cases are diagnosed with breast cancer. Thus, early detection is very important for the treatment of breast cancer. Early detection of breast cancer helps in decreasing mortality rates. The women who are at higher risk with no symptoms when clinically examined can be screened using mammography. This screening helps to predict cancer in early stages and decreases mortality rate.

S. Paramkusham (✉) · C. V. Narasimhulu

Geethanjali College of Engineering and Technology, Cheeryal Village, Keesara Mandal, Hyderabad, Telangana 501301, India
e-mail: spandanamadhav@gmail.com

However, the mammograms are error-prone due to the overlapping of abnormalities with dense tissue regions of mammogram and it is difficult for radiologists to detect breast abnormalities. This misinterpretation leads to an increase in false positives and false negatives by radiologists. Due to the huge amount of screening mammograms and fewer radiologists, double reading of mammograms can be carried out by computer-aided diagnosis systems. CAD system can act as double reader and can give a warning sign to radiologists to detect suspicious regions in mammograms such as masses and microcalcification. Computer-aided diagnosis system uses digital image processing, pattern recognition, and machine learning techniques to detect abnormalities. Among all the steps in the CAD feature extraction step plays vital as it extracts the texture properties of normal and abnormal ROIs. In this paper, we reduce false-positive cases by extracting features and classifying them into normal-abnormal.

In this paper, the features are extracted using rotational invariant local frequency (RILF) technique from mammograms using three different methods. RILF can be used effectively on mammograms in the presence of noise and has a comparatively small number of features when compared to other techniques [2]. Hence, we the RILF technique was first employed on mammograms for classification of mammogram regions.

2 Related Works

In literature, several feature extraction techniques have been proposed to detect abnormality in mammograms and classify them into normal/abnormal and benign/malignant. The method based on Haralick, correlogram function, and shape features were proposed to classify mammograms into normal/abnormal [3]. Zernike moments of different orders have been used to obtain features for the classification of malignant and normal regions in mammogram [4]. Curvelet transform method is used for feature extraction from approximation bands [5]. LBP features are computed from each block of subdivided image and given to classifier to differentiate mammogram regions [6]. Masses have been detected using LBP variance and shape descriptors [7]. The statistical features have been calculated from each map obtained from wavelet transform, and these features were submitted to Bayesian classifiers for abnormality detection [8]. Histogram of Oriented Gradients(HOG), Dense Scale Invariant Feature Transform (DSIFT), and Local Configuration Pattern(LCP) methods are concatenated to form feature vector and given to classifier to differentiate breast tissues into normal and abnormal [9]. Cross variogram and variogram functions have been utilized to detect asymmetric regions and distinguish the regions into mass/non-mass [10]. Local energy-based shape histogram (LESH) features have been calculated and fed to support vector machines to classify into benign and malignant abnormalities [11]. Spherical wavelet transform has been applied to mammograms to delineate breast abnormal regions into mass/non-mass and benign/malignant [12]. Phylogenetic trees have been constructed to compute the taxonomic diversity index (Δ) and

the taxonomic distinctness (Δn) from mammograms as features [13]. Asymmetric features based on fractals were computed to differentiate mammogram ROIs [14]. Many other feature extraction techniques were proposed in the literature. Some of the feature extraction techniques include grey level co-occurrence matrix, grey level run length matrix, statistical features, multiresolution methods for classification of mammograms [15, 16]. In this paper, features have been computed using the RILF technique with three different methods. In order, to obtain better accuracy, a classifier that increases the efficiency of CAD system has to be selected. Hence, SVM classifier via tenfold cross-validation has been utilized to evaluate the performance of RILF feature extraction technique.

3 Methodology

3.1 Rotational Invariant Local Frequency (RILF)

In RILF method [2], features are extracted from frequency components of local circular function ($LCF_{(N, R)}$). The value of the LCF at each pixel of the image for feature extraction is obtained by applying 1D Fourier transform. The $LCF_{(N,R)}(x, y) = (t_0, t_1, t_2, \dots, t_{N-1})$ have been applied on N circular neighbouring pixels at each pixel (x, y) with radius R as in local binary pattern (LBP) and is given in Eq. (1). The textural information is loosed when thresholding the neighbourhood pixels in LBP. To get rid of this problem Fourier transform is applied to the pixels of LCF function

$$f_n = \sum_{k=0}^{N-1} t_k e^{-\frac{2\pi i(n-1)}{N}}, \quad (n = 1 \text{ and } 2) \quad (1)$$

The low-frequency components do not contain noise and have 90% of the texture energy when compared to high-frequency components. Hence, we have considered low-frequency components f_1 and f_2 [2]. Then, the magnitude-based features are extracted from these two frequency channels (f_1 and f_2) by computing circularly shifted 2D Fourier transform. The 2D spectrum of each frequency channel CH_n , (where $n = 1$ and 2) is given in Eq. (2)

$$CH_n(k, l) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} |f_n(x, y)| \cdot e^{-2\pi i \left(\frac{xk}{W} + \frac{yl}{H} \right)} \quad (2)$$

The circular band-pass disk shape filters are applied on the spectrum of frequency channels to achieve rotational invariance property and these filters are defined in Eq. (3). W and H in the below equation give the width and height of the image.

$$D_{r1,r2}(x, y) = 1 \quad \text{if } r_1 \leq \sqrt{x^2 + y^2} \leq r_2$$

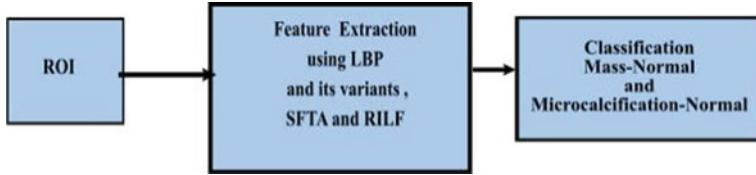


Fig. 1 Block diagram for the classification of ROIs using method 1

$$= 0 \quad \text{otherwise} \quad (3)$$

r_1 and r_2 are the inner and outer radius of disk shape filters. The magnitude descriptors of RILF are computed in Eq. (4) which is given below

$$\text{RILFMD}(r_1, r_2, n) = \frac{\sum_{k=-W/2}^{\frac{W}{2}-1} \sum_{l=-H/2}^{\frac{H}{2}-1} |CH_n(k, l)| \cdot D_{r_1, r_2}(k, l)}{\sum_{k=-W/2}^{\frac{W}{2}-1} \sum_{l=-H/2}^{\frac{H}{2}-1} D_{r_1, r_2}(k, l)} \quad (4)$$

4 Implementation of RILF Feature Extraction Technique

RILF feature extraction technique is applied to mammogram ROIs using three different methods. The features are extracted using three methods and classified using SVM classifier to validate the RILF technique.

4.1 Method 1

In this method, the RILF technique is applied directly to mammogram ROIs to compute texture features. Figure 1 shows a block diagram of method 1. The magnitude descriptors are computed from the RILF technique from channel 1 (ch1) and channel 2 (ch2) and given to SVM classifier for performance evaluation of the RILF technique. This method is implemented in our previous paper [17]. In this paper, we have implemented the RILF technique using method 2 and method 3 described below. Further, we have compared all the three methods.

4.2 Method 2

ROI is subdivided into $N \times N$ blocks in method 2. Each divided block undergoes a feature extraction process using the RILF technique. The feature vector obtained

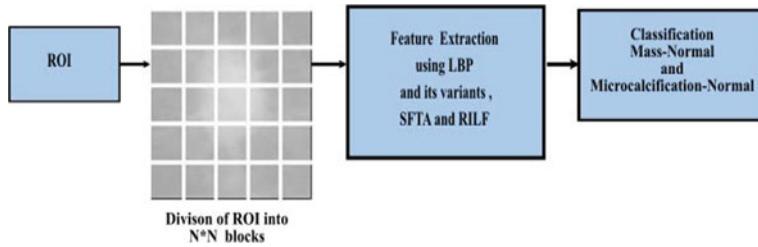


Fig. 2 Block diagram for the classification of ROIs using method 2

from all blocks is given SVM classifier for classification. The block diagram of this method is shown in Fig. 2. This type of feature extraction technique is named as block-wise-RILF (BRILF) technique.

4.3 *Method 3*

ROI is decomposed and represented in the scale space Gaussian pyramid in method 3 [18]. The scale-space equation of an image is given in Eq. (5)

$$G_k(x, y) = mnW(m, n)G_{k-1}(2x + m, 2y + n) \quad \text{for } k > 0$$

$$= I \quad \text{for } k = 0 \quad (5)$$

$G_k(x, y)$ represents k th level image pyramid and $W(m, n)$ is a pyramid filter (Gaussian) of size $m \times n$ and x, y are spatial coordinates of the image.

RILF technique is used to extract features from two levels of images obtained by scale-space equation and further, these features are combined and given to SVM classifier. Figure 3 gives a block diagram of method 3 and this technique is named as pyramidal-RILF (PRILF).

5 Classification

The evaluation of the above three methods using RILF technique is carried out using the support vector machine (SVM). Evaluation parameters such as accuracy, sensitivity and specificity are calculated to validate RILF technique.

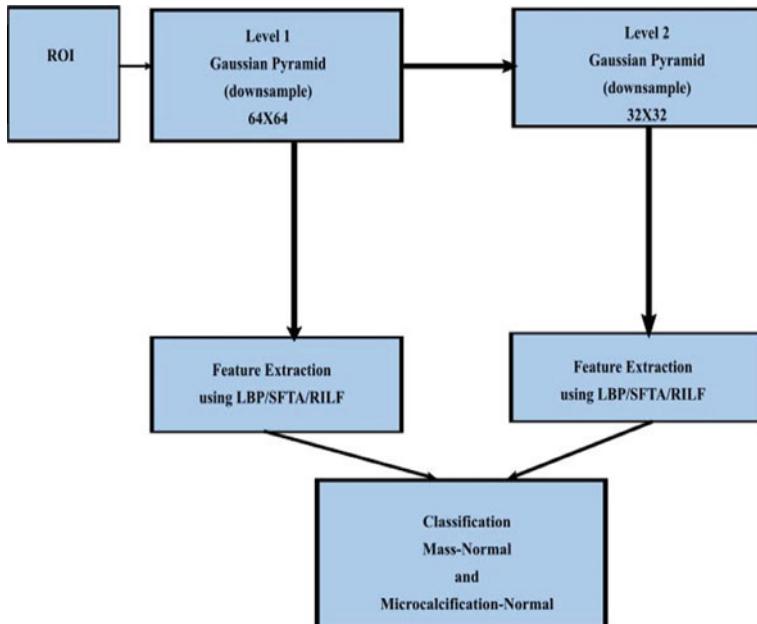


Fig. 3 Block diagram for the classification of ROIs using method 3

6 Results and Discussion

RILF feature extraction technique is utilized to classify mammogram using the three methods mentioned in Sect. 4 and the results obtained using RILF and its extended versions are presented in this section. The features extracted from three methods are given to SVM classifier (linear kernel) via the ten-fold cross-validation method for classification.

6.1 Database

The work has been carried out by considering two datasets from Image retrieval in medical applications (IRMA). IRMA database consists of normal, benign and malignant ROIs. For validation of the RILF technique, two datasets have been used from IRMA database. The first dataset consists of 1157 mass and 932 normal ROIs. The second dataset contains 688 microcalcifications and 932 normal ROIs. Testing and validation have been carried out using SVM classification with two datasets for the classification of normal-abnormal ROIs.

6.2 Results of Method 1

In this method, RILF achieved (accuracy of 93.53%, the sensitivity of 99.52% and specificity of 88.71%) with neighbourhood (8, 1) and ch1 and ch2 for classifying mass-normal ROIs as shown in Figs. 4 and 5 show that RILF gives the best values with accuracy of 91.11%, sensitivity of 97.45% and specificity of 82.50% with neighbourhood (8, 1) for microcalcification-normal classification as in [17].

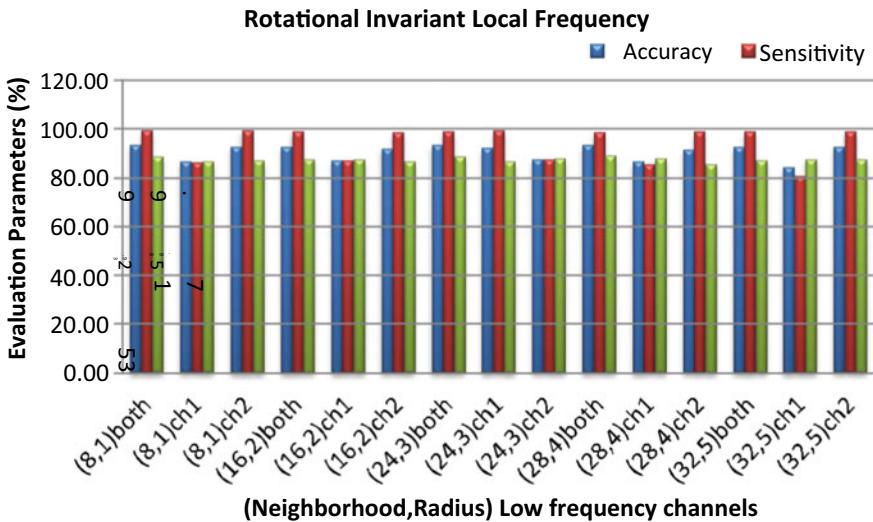


Fig. 4 Results of RILF (mass-normal)

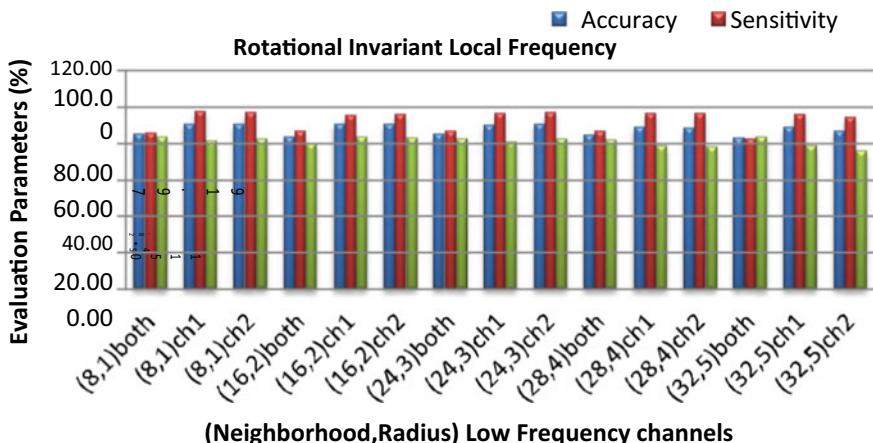


Fig. 5 Results of RILF (microcalcification-normal)

6.3 Results of Method 2

RILF is used to extract features using method 2 with $N = 2$ and $N = 4$. BRILF method has achieved an accuracy of 92.49%, the sensitivity of 98.96% and specificity of 87.26% using channel 2(ch1), $N = 2$ and with neighbourhood of (16, 2) as shown in Fig. 6 for mass-normal classification. Figure 7 shows that BRILF method gave an accuracy of 90.42%, sensitivity of 96.55% and specificity of 82.15% with $N = 4$, and used channel 2(ch2) with (8, 1) neighbourhood for microcalcification-normal classification. The features extracted using this method are given to the classifier for further validation.

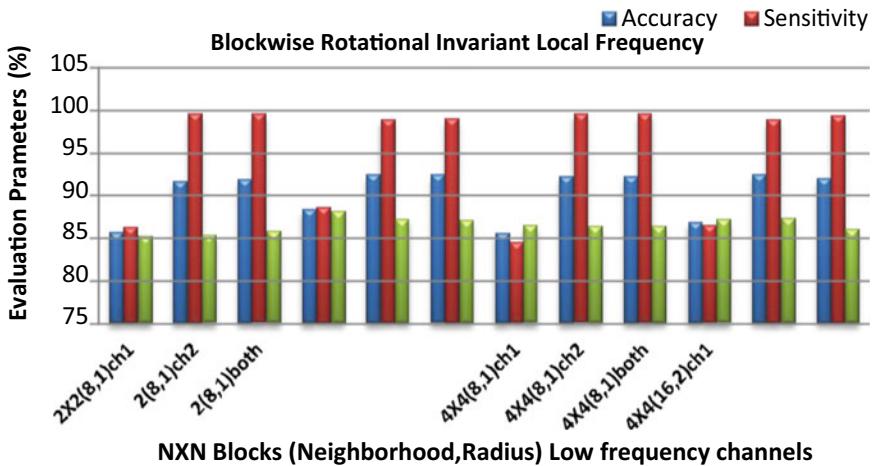


Fig. 6 Results of BRILF (mass-normal)

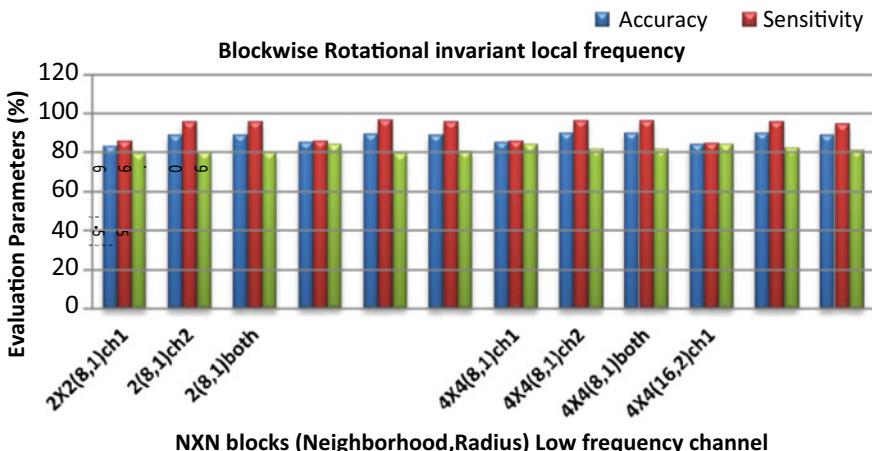


Fig. 7 Results of BRILF (microcalcification-normal)

6.4 Results of Method 3

RILF is applied to three levels of the image pyramid. PRILF gives the best accuracy of 94.36%, sensitivity of 99.45%, and specificity of 90.26% with channel 2 and neighbourhood (16, 2) for mass-normal classification. The results are shown in Fig. 8. For microcalcification-normal classification PRILF achieved best accuracy of 91.59%, sensitivity of 98.19% and specificity of 82.67% with neighbourhood (8, 1) and channel 2. The results of PRILF are shown in Fig. 9.

The experimental results of the RILF technique gave the highest accuracy with all the three methods. RILF technique is a technique used for noisy images [2].

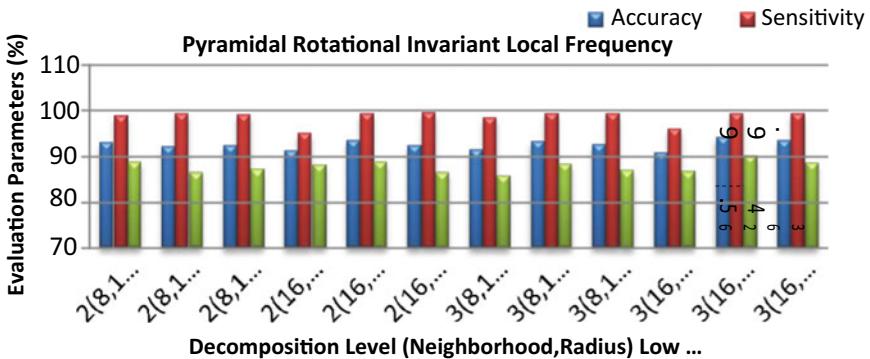


Fig. 8 Results of PRILF (mass-normal)

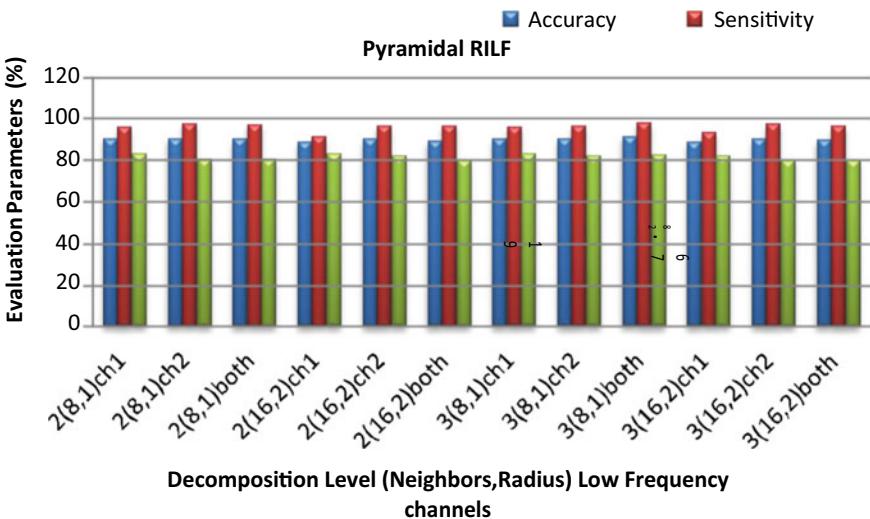


Fig. 9 Results of PRILF (microcalcification-normal)

The robustness to noise is achieved due to the application of 1D Fourier transform on local circular function. In RILF technique the grey level values surrounding the centre pixel are converted into frequency components, this helps in preserving the textural information of the image. Figure 10 gives a comparison of all methods using the RILF technique and we can observe that the evaluation parameters for PRILF are higher compared to BRILF and RILF. The proposed work in this paper evaluates accuracy, sensitivity and specificity with two types of abnormality that are masses and microcalcifications. Table 1 gives a comparison of our method with state of art methods. From that table, we can observe that number of data samples that we have considered is high (2777 samples) and we also calculated values for all evaluation parameters such as accuracy, sensitivity and specificity for breast tissues in comparison with other methods.

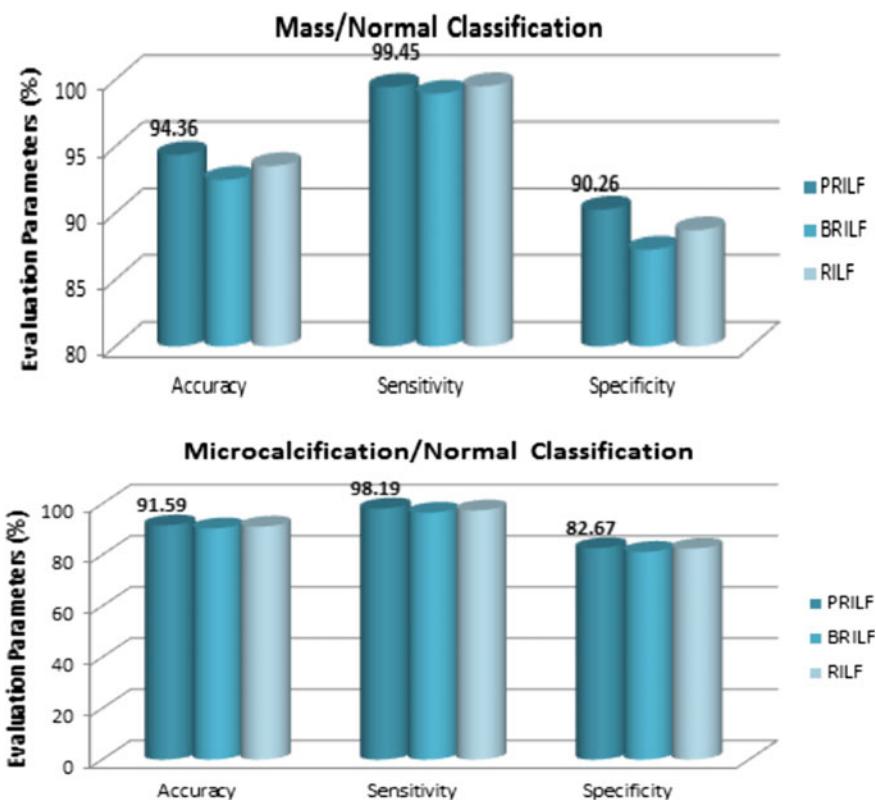


Fig. 10 Comparison of different methods using RILF technique

Table 1 Comparison of our method with state of art methods

Method	Samples	Accuracy (%)	Sensitivity (%)	Specificity (%)
Local seed region growing spherical wavelet transform shape, local seed region growing, spherical wavelet transform, SVM [12] 2013	60	96	—	—
Contourlet-based mammography mass classification using the SVM family [19] 2010	90	96.6	—	—
LBP + SVM [6] 2009	1792	94	—	—
GLCM texture features, Neural network architecture [20] 2011		96	—	—
K-means, GLCM and SVM [21] 2009	1177	92.63	86	94.61
Features [22] 2010	300	95	—	—
Quality thresholding, correlogram and shape features [3] 2014	2033	83.53	92.3	82.2
Proposed method (IRMA)	2777	Mass/normal: 94.36	99.45	90.26
		Microcalcification/Normal: 91.59	98.19	82.67

7 Conclusion

In this paper, the features have been extracted with rotational invariant local frequency (RILF) technique using three methods to delineate normal/abnormal ROIs. Among these methods, pyramidal rotational invariant local frequency (PRILF) achieved highest accuracy, sensitivity and specificity when compared to block wise rotational invariant local frequency (BRILF) and rotational invariant local frequency (RILF). The PRILF outperforms the state of art methods by considering large data samples as input. In future, we plan to develop a new algorithm for the classification of ROIs based on the BIRADS classification and would like to explore the shape properties of masses to classify benign and malignant masses.

References

1. <http://cancerindia.org.in/globocan-2018-india-factsheet/>
2. Maani R, Kalra S, Yang YH (2013) Rotation invariant local frequency descriptors for texture classification. *IEEE Trans Image Process* 22(6):2409–2419
3. de Nazare Silva J, Carvalho Filho AO, Silva AC, De Paiva AC, Gattass M (2015) Auto-matic detection of masses in mammograms using quality threshold clustering, correlogram function and SVM. *J Digit Imaging* 28(3):323–337
4. Sharma S, Khanna P (2015) Computer-aided diagnosis of Malignant mammograms using Zernike moments and SVM. *J Digit Imaging* 28(1):77–90
5. Gedik N, Atasoy A (2013) A computer-aided diagnosis system for breast cancer detection by using a curvelet transform. *Turk J Electr Eng Comput Sci* 21(4):1002–1014
6. Lladó X, Oliver A, Freixenet J, Martí R, Martí J (2009) A textural approach for mass false positive reduction in mammography. *Comput Med Imaging Graph* 33(6):415–422
7. Masmoudi AD, Ayed NGB, Masmoudi DS, Abid R (2015) Robust mass classification-based local binary pattern variance and shape descriptors. *Int J Sign Imaging Syst Eng* 8(1–2):20–27
8. Kendall EJ, Flynn MT (2014) Automated breast image classification using features from its discrete cosine transform. *PloS one.* 9(3):91015
9. Ergin S, Kilinc K (2014) A new feature extraction framework based on wavelets for breast cancer diagnosis. *Comput Biol Med* 51:171–182
10. Ericeira DR, Silva AC, De Paiva AC, Gattass M (2013) Detection of masses based on asymmetric regions of digital bilateral mammograms using spatial description with variogram and cross-variogram functions. *Comput Biol Med* 43(8):987–999
11. Wajid SK, Hussain A (2015) Local energy-based shape histogram feature extraction technique for breast cancer diagnosis. *Expert Syst Appl* 42(20):6990–6999
12. Gorgel P, Sertbas A, Ucan ON (2013) Mammographical mass detection and classification using local seed region growing–spherical wavelet transform (LSRG–SWT) hybrid scheme. *Comput Biol Med* 43(6):765–774
13. de Oliveira FSS, de Carvalho Filho AO, Silva AC, de Paiva AC, Gattass M (2015) Classification of breast regions as mass and non-mass based on digital mammograms using taxonomic indexes and SVM. *Comput Biol Med* 57:42–53
14. Beheshti SMA, Noubari HA, Fatemizadeh E, Khalili M (2016) Classification of abnormalities in mammograms by new asymmetric fractal features. *Biocybern Biomed Eng* 36(1):56–65
15. Mohanty AK, Senapati MR, Beberta S, Lenka SK (2013) Texture-based features for classification of mammograms using decision tree. *Neural Comput Appl* 23(3–4):1011–1017
16. Reyad YA, Barber MA, Hussain M (2014) Comparison of statistical, LBP and multi-resolution analysis features for breast mass classification. *J Med Syst* 38(9):1–15
17. Paramkusham S, Rao KM, Rao BP (2018) Comparison of rotation invariant local frequency, LBP and SFTA methods for breast abnormality classification. *Int J Sign Imaging Syst Eng* 11(3):136–150
18. Wang W, Chen W, Xu D (2011) Pyramid-based multi-scale lbp features for face recognition. In: International conference on multimedia and signal processing, vol 1. IEEE, pp 151–155
19. Moayedi F, Azimifar Z, Boostani R, Katebi S (2010) Contourlet-based mammography mass classification using the SVM family. *Comput Biol Med* 40(4):373–383
20. Nithya R, Santhi B (2011) Classification of normal and abnormal patterns in digital mammograms for diagnosis of breast cancer. *Int J Comput Appl* 28(6):21–25
21. Martins O, Braz A Jr, Correa Silva A, Cardoso de Paiva A, Gattass M (2009) Detection of masses in digital mammograms using K-means and support vector machine. *ELCVIA Electron Lett Comput Vis Image Anal* 8(2):39–50
22. Surendiran B, Vadivel A (2012) Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. *Int J Med Eng Inf* 4(1):36–54

Internet of Things (IOT) Architecture—A Review



Achinta K. Palit

Abstract Internet of Things (IOT) can be considered as the future evolution of the internet, that incorporates Machine to Machine, Business to Machine, Human to Human, etc. It provides connectivity for everyone and everything. It embeds intelligence and automation in internet network services. With the increasing demand for the IOT, its deployment in the various areas has increased. Future requirements of the internet and its network will establish a new service. Its deployment will bring enhancement in the community and social life. Its pervasive presence in the future will have major benefits and support in the research and development environment. Its main objective is to enable things to be connected at any time, any place, with anything and anyone ideally using network services. It is not a single technology, rather it is a mixture of hardware and software technology, based on which integration of information will be processed, managed, operate, etc. the whole system.

Keywords Internet of things (IOT) · Machine to machine (M2M) · Business to machine (B2M) · Human to human (H2H) · Heterogeneity · IOT architecture layer

1 Introduction

IOT is the network of physical devices that are interconnected with each other [1]. It enables each thing to connect, collect and exchange information. Involvements of IOT with internet have extended the network beyond standard devices. Internet network started connecting with computer and laptop have now enabled into communicating with smart devices like vehicles, homes, medical, cities, industries, etc. Patel et al. [2]. It's a new world of the environment that generates smart cities, houses, energy, transport and many other intelligent areas of services. It helps us to understand and control the entire environment in real sense rather than virtually. The terms IOT cannot be only related to Internet, rather it can be referred to as an independent

A. K. Palit (✉)
Faculty in MCA, AMIT, Bhubaneswar, India
e-mail: achintakupalit@yahoo.co.in

communication network which can be managed and controlled digitally without human action. With IOT it is possible to turn small matters of devices into bigger elements of components, by enhancing its existing role [3]. The basic role of IOT is to create autonomous and secure network connectivity between the real world devices and applications. It generates a link between real-world and virtual world by enhancing the role of digital devices.

IOT consists of several devices like sensors, communication devices, processing units, service units, security controls, etc. which are attached with clouds decision making and action-oriented system [4]. These devices have unique features and identification that are accessible through internets which are uniquely designed for autonomous services. Everyday these electronic devices are encountered differently but when they are put together, it becomes a reliable, recognisable, locatable, controlled addressable through smart sensing devices.

2 Characteristics of IOT

Some of the important characteristics of IOT are as follows:

1. Interconnectivity

With IOT anything can be interconnected globally. Information and communication can be shared equally around the globe. It allows us to enhance information without any trouble. It makes the system easy and comfortable.

2. Things Related Services

It is capable of providing things related services within limited capacity. It maintains equals and balanced importance for the associated virtual physical things. Services are basically related to the information gathered and stored. It has the capacity to change the world within seconds.

3. Heterogeneity

IOT devices are heterogeneous. They have the capacity to exchange information and communication through other networks also. They are unable to perform on any platform. They are not fixed or constant in any position. They can coordinate with each other.

4. Dynamic Changes

The state of devices changes dynamically with changes in location and speed of the network. It may be connected or disconnected, its state of condition changes according to it. It changes the environment dynamically with state of condition.

5. Enormous Scale

Management of IOT should be of highest order of magnitude. Magnitude of network communication should be of highest level of scale. So that there should not be any trouble in the communication devices. Major work of the management will be to generate and interpreted data at all levels. It will enhance in building efficient and semantic system. It will make a superior system in handling any application.

6. Security

To gain benefits in using IOT, it must be important to have secured and privacy in all standard. Safety of personal and private information should be taken utmost. It will achieve in securing the endpoints without loss of information. Security and privacy level of data and information should be maintained at all grounds. So that importance of the devices should be maintained. It will enable the development of devices.

7. Connectivity

Connectivity enhances network accessibility and compatibility. Accessibility will allow to track the information from anywhere. Compatibility will increase the device capacity to consume and produce data at no means. It will increase the capacity of the device to manage and handle information at any moments.

3 IOT Architecture Layered

IOT architecture layered design is shown in Fig. 1. It consists of different layered

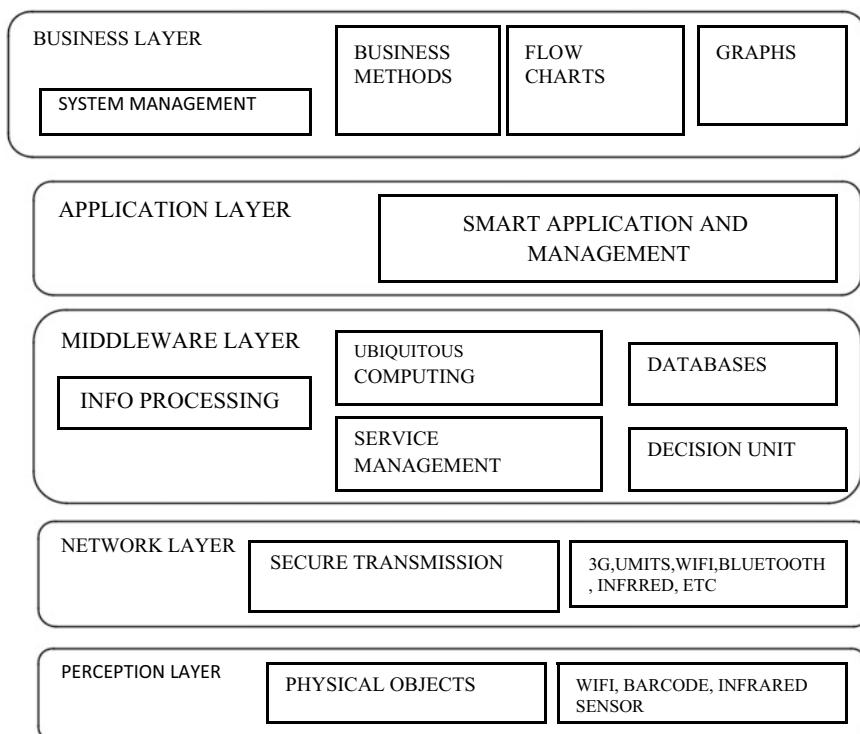


Fig. 1 Architecture of IOT. *Source* Khan et al. [1]

technologies that provide support for developing IOT system. These layers are subdivided by their roles and function. Their importance depends on their function and behaviour in developing the system. This can be described as follows:

1. Perception Layer

This layer is also known as the device layer. This layer is consisting of smart devices and sensors. It is the lowest layer of IOT architecture. There are various types of devices for different purposes. These devices have the capacity to take measurements of temperature, air quality, humidity, pressure, flow, movements, speed, electricity, etc. Depending on the type of data, information can be gathered about location, temperature, orientation, motion, vibration, acceleration, humidity, chemical changes in the air, etc. On some cases, it can be measured with the physical property and converts it into a signal which is understandable to the devices. The devices are grouped according to unique purposes such as environmental sensor, body sensor, home appliances, vehicles telematics sensor, etc. Most of the devices require connecting to the network gateway like a Local Area Network (LAN) such as Ethernet, Wi-Fi, etc. Personal Network (PAN) such as ZigBee, Bluetooth and Ultra Wide Bound (UWB). They are connected to backend server and application through Wide Area Network (WAN) such as GSM, GPRS and LTE. Low power and low data rate are connected through Wireless Sensor Network (WSN). These devices can accommodate more sensor connectivity while retaining adequate battery life and covering large areas of network.

2. Network Layer

This layer is also known as transmission layer. This layer securely maintains the data transmission from sensor devices to the information processing system. Transmission medium can be wired or wireless system. Technology can be 3G, UMTS, Wi-Fi, Bluetooth, etc. depending upon sensor devices. Current network technology is used to support Machine to Machine (M2M), network and application. With an increase in demand for IOT services and applications such as high-speed transactional services, context-aware applications, etc. Multiple networks with various technologies and access protocols are needed to work with each other in a heterogeneous configuration. These networks can be in the form of private, public or hybrid network model which are built to support the communication requirement of IOT in various ways. To have a secure and privacy network various gateways like microcontroller, microprocessor, etc. and network gateway like Wi-Fi, GSM, GPRS, etc. are implemented for transferring information from the Perception layer to Middleware layer.

3. Middleware Layer

This layer is also known as Information Processing layer. This layer renders the processing of information possible through analytics, security controls, process modelling and management of devices. This layer brings interaction of objects and systems together forming into contextual data. This layer is responsible for the service management and link to the database. It receives the information from Network layer and store in the database. Data management is the ability to manage

data information flow. Information can be accessed, integrated and controlled through the management of data. It shields the unnecessary data and reduces the risk of privacy of data sources. It is implemented through data filtering technique which is used to hide the data leak and essential information. Various analytics tools are implemented to extract a large amount of raw data into processed data. Decision about data is relevant or not based on streaming analytics data, which is carried out in real-time speed. This layer is made of ubiquitous computing database, service management and decision unit. This layer supports decision logics and trigger interactive and automated process to enable a more responsible IOT system. This layer is responsible for the acquisition of information and forwarding to remote services for analysis and storage. This layer directly interacts over network layer. This layer manages the privacy and security of information by maintaining highest order of security.

4. Application Layer

The application layer is also known as smart application and management layer. This layer provides global management of applications based on the information processed by the Middleware Layer. This is the domain for Smart Cities, Smart Energy, Smart Transport, etc. Potential increment of the IOT has increased the implementation of applications in various areas. Information is gathered and interacted for global benefits purposes. This layer differentiates the purpose of IOT according to the requirement of the areas where it must be implemented. It helps in accurate measuring and monitoring information for the future and present. It helps in detecting and tracking any location and position. It can even alert the user from future danger. It maintains a complete balance in information exchange. This layer can be expanded in the future according to the requirement of the devices.

5. Business Layer

This layer is also known as System Management. This layer is responsible for the management of the overall IOT system. It represents business models, graphs, flow charts, etc. based on the data received from application layer. The actual development of IOT depends on better business management. Based on the analysis of data, this layer will help to determine the future course of action and business strategies. This layer has various methods to represent data through models, graphs, flow chart, etc. which is easily understandable.

4 Conclusions

The future development of IOT in enhancing the role of IOT. With an increasing demand for cloud computing, data, signal processing, security, privacy, embedded system, etc. will increase the demand for IOT. It provides services to illustrate how various technologies relate to each other and communicate. It is gaining popularity in day to day life. It is an emerging technology in the present environment. IOT has emerged as a major transition between Human to Human, Machine to Machine, etc.

Its automation and intelligence has created a major effect on information exchange around the world. Its deployment will benefit the future research work and increases many other purposes link with it. Integrating the physical world with its virtual world by using a better medium of communication with IOT will benefit in many ways. It is a key factor in the near future which will emerge as a new revolution in the field of communication exchange.

References

1. Khan R, Khan SU, Zaheer R, Khan S (2012) Future internet: the IOTs architecture, possible application and key challenges. IEEE
2. Patel KK, Patel SM (2016) IOT: Definition, characteristics, architecture, enabling technologies, application and future challenges. IJESC
3. Clark J (2016) What is IOT? IBM (Blog)
4. Chen X-Y, Jin Z-G (2012) Research on key technology and application for IOT. Elsevier
5. Calum McChelland (2017) What is IOT? A simple explanation of IOT. J IOT 18394
6. Ning H, Hu S (2012) Technology classification, industry and education for future IOT. IJCS

Robust and Lightweight Control System for IoT Networks: Enabling IoT for the Developing World



Jithu G. Panicker and Mohamed Azman

Abstract The proliferation of smart automated control systems in contemporary environments has led to the realization of various smart and connected products/solutions. For a true “Smart City”, it is of paramount importance to have secure, stable, smart and automated control systems in commercial domains as well as residential domains. Domotics/Home Automation, is the enabling of residential spaces to constitute smartly controlled devices or appliances, say the lighting systems or the HVAC systems, which may have their parameters varied with little to no human intercession. The alteration of their variable parameters may be done based on human action, timers, inputs from sensors or the combination of multiple such triggers. This may be actualized by IoT-enabled devices and appliances that can interact/communicate with the control systems or each other. The solutions available presently provide with excellent utilitarian features that can empower the users to automate much of their daily chores/tasks and, if not completely automatized, at least can reduce human effort by a considerable degree. However, majority of the solutions suffer from certain major drawbacks. They require live Internet to employ the cloud-based APIs that the systems utilize. This paper proposes a secure and robust speech-based automation and control system that functions with or without Internet. The single-board-computer (SBC)-based central automation and control system wirelessly communicates with the connected devices and appliances through local and private low-latency MQTT links. The proposed system does not compromise on its feature set, providing AI-based attributes at reasonable cost factors, and is resistant to total failures, with multiple backup layers.

Keywords Home automation · Control system · Digital assistant · Voice commands · IoT

J. G. Panicker · M. Azman (✉)

Department of Electronics and Communication Engineering, National Institute of Technology,
Warangal, Telangana 506004, India
e-mail: mohamedazmanm1@gmail.com

J. G. Panicker

e-mail: jithugpanicker@outlook.com

1 Introduction

One of the greatest developments [1–3] of the century is the Internet of things (IoT). IoT connects physical items embedded with electronics, software and sensors to the network, qualifying them to accumulate and reciprocate data so as to perform desired tasks with little to no human intervention. Major companies have started directing their R&D towards IoT-enabled devices for interconnected functioning. With the emergence of 5G, a significant hike is expected in the number of devices connected via Internet. According to Statista, the number of connected devices around the world will elevate drastically from 23.14 billion in 2018 to 75.44 billion in 2025, which is around 326.01% increase over a span of 7 years. IoT has gained popularity in the recent years because of the development of critical technologies and its enormous applications in the field of Smart Home Systems (SHS) [4], healthcare, connected cars [5, 6], smart cities, etc. The smart home concept is the one that attracts both academia [4, 7] and industries [8, 9] because they are directly related to people's everyday life.

Developments in SHSs are growing at an exponential rate. It enables the control of electronic and electrical systems (lighting, HVAC, security, entertainment, etc.) remotely, with natural speech or automated partially or completely by employing timers, sensors, counters, etc., or their amalgamations. Among the different types of SHSs, speech recognition-based ones are the most effective and sought-after. They are desirable in general use cases to provide natural and intuitive means of interaction as well as in certain situations such as for the differently abled, who may face difficulties in performing daily activities at home or outside and may require assistance to perform various basic tasks. With speech recognition integrated into these control systems, the need to physically press buttons or use sliders/knobs is eliminated. It is a highly flexible option that may be made compatible with various future technologies, and it may also be quite easily customized for individual requirements. In recent times, SHS has improved drastically with the introduction of various new wireless technologies. Ideally, these systems should be designed such that its implementation is plausible in existing homes, without requiring drastic changes to the existing infrastructure; for this, the physical size and wireless nature are important factors. Recent developments have led to cheaper and smaller devices; single-board computers or SBCs are examples of affordable and compact computing devices that could be used in these systems. One such board is the Raspberry Pi R; it is widely used for prototyping due to its availability and flexibility.

Message Queuing Telemetry Transport, also known as MQTT, a publish-subscribe-based messaging protocol, is used for lightweight communication for sensors and actuators and is enhanced for minimal latency and reliability demanding networks. It is mainly used for M2M communication where low bandwidth usage and dependability are salient. It enables resource-constrained IoT devices to send, or publish, information about a given topic to a server that functions as an MQTT message broker. The use of MQTT ensures a responsive and steadfast network. This type of a setup may also be used with daisy chain networks [10], mesh networks

[11], etc., for short-range configurations such as demonstrated in this work, or even for long-range configurations [12].

The setup posited in this work was designed to counter the issues faced by existing products/prototypes. The noteworthy features are:

- Ability to verbally control the connected devices or interact with the Control and Automation System
- Devices are wirelessly connected to the Central Control System aiding flexibility
- Operable in the absence of internet connectivity
- Ability to control devices remotely with the help of an internet equipped smartphone
- Supports conventional switching and varying of device/appliance parameters as a Backup Mode
- Enhanced security supporting end-to-end encryption and user authentication
- Minimal latency and computationally insubstantial operation due to the usage of MQTT protocol
- Supports control of multi-dimensional parameters [13] of the connected devices
- Artificial Intelligence based personal digital assistant provided allows users to intuitively interact with the system.

2 Related Work

There are products being designed for SHSs; however, in almost all cases, if these systems do support speech recognition, they usually use cloud-based services to control the devices. Products such as Google Home™ or Amazon® Alexa™ are highly potent and comprise of a large feature set; however, they are completely useless when off the Internet. In a smart home where an SHS is primarily used for all switching and controlling purposes, this is a deal-breaker.

Previous research work has greatly contributed to vastly increasing the application base and feature sets of SHS. A formerly designed system [14] is based on adding speech recognition features to a home automation control system that is entirely established on an SBC. This gives path to the potential of embedding controlling and switching functionality in a fully integrated and consolidated system. The work uses a Raspberry Pi® board that is hardwired to a multi-channel relay unit. It uses Wolfram Alpha®'s cloud-based speech recognition API. There are quite a few more previously proposed and designed systems that are similar to this one [15]; however, they fundamentally share a few common disadvantages.

Availability of Internet is mandatory for their functionality, and they rely on a wired connection between the control centre and the device to be controlled. This

means, in case of unavailability of an Internet connection, the entire system may be deemed useless.

Some [16] have used local speech processing, making the system able to function with no access to the Internet; they generally use conventional speech recognition algorithms or sometimes, AI techniques and Natural Language Processing to convert the speech to text. A variety of speech-to-text (STT) engines and frameworks are available; however, having a cloud-based SST API that is strongly trained with an enormous amount of data will almost always provide better results and accuracy, along with smaller computing time and response time delays; this is true only if a stable connection to the Internet is available. In case of inaccessibility to the Internet, a local SST engine is the only option.

Coming to another common drawback, if each device has to be hardwired to the central control unit [14], it would lead to extreme limitations for the smart home and smart appliance designers; further, if it has to be implemented in an existing household, drastic re-wiring may be required. To give the product architects freedom to design solutions that are truly smart, ensuring the wireless nature of the technology is crucial. The rudimentary working of such a concept has been previously demonstrated [17].

It should be noted that having multiple control units would not be an appropriate solution regardless of the cost factor as it would lead to higher network traffic and also would be of concern when it comes to security since they require being connected to the Internet; having multiple entry points to a single network would be undesired, having to guard one door to enter the system is safer than having to guard many.

Most existing systems are based on protocols such as HTTP [18] (most Web applications or applications developed on Android™ or iOS use HTTP or HTTPS), or even SPTM [19]. These may be acceptable unless the latency is of high importance and if numerous devices are connected to the network at any given instance. A fully edged SHS is expected to have a significant number of smart devices connected to it, and to avoid laggy operation, a lightweight alternative has to be used. MQTT, as mentioned in the previous section, makes a great substitute [20], due to its lightweight nature.

Data encryption may be used over MQTT for further enhancing the systems' security aspects [21] additional to basic authorization features such as username and passwords or fingerprints [22].

Another way to annex a simple layer of security is to incorporate voice recognition or speaker recognition [23] such that only the voice of a recognized speaker is authenticated to access the control system. Furthermore, gesture-based control may also be appropriately integrated [24] if desired in addition to speech-based control to further enhance the system's flexibility.

3 Proposed System

To have a strong and robust SHS that is stable in all known possible circumstances, it is a salient necessity to have a system that works both on the Internet, as well as off the Internet, without compromising on the primary features, cost and size factors, flexibility offered to product designers or technical architects and its security.

The proposed system is exhaustive and robust as it focuses on solving many of the issues present in the existing systems. The general system may be designed as an IoT control hub, similar to that of presently available AI Assistant boards that are embedded into speakers with dedicated far-array microphones inbuilt.

However, additional features have been added in order to bolster the holistic functionality, security and reliability of the system.

3.1 The System

Figure 1 shows the constituents of the general nodes/devices. The working has been explained in four separate diagrams, each of which highlights different sets of features that address different issues amongst the existing solutions.

Mode A (Online, with AI Assistant)—Fig. 2 depicts the working of “Mode A”.

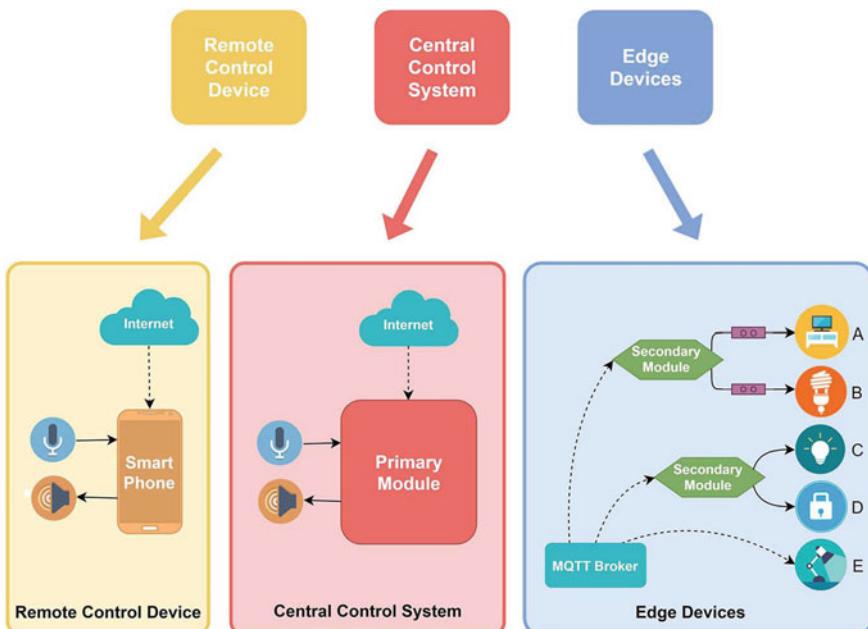
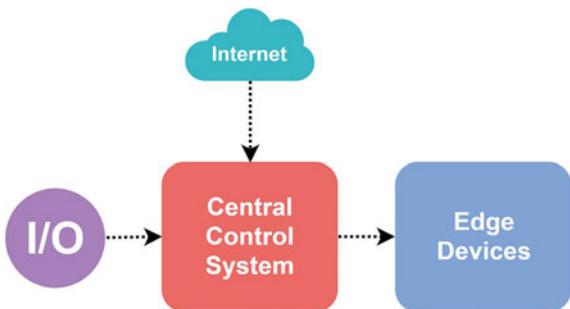


Fig. 1 General prototype block diagram

Fig. 2 Mode A (online, with AI assistant)



This is the most basic version of the system. The functioning is similar to what has been previously designed [13]. It has a central control system which encompasses the primary module, a microphone array and a speaker system for auditory input and for auditory output. This embedded device may be a fully integrated SoC or a programmable microprocessor. It is programmed with a speech recognition system (also known as “speech-to-text” or “Automatic Speech Recognition” (STT or ASR)), an artificial intelligence-based smart assistant and a parameter control system for switching and/or controlling the parameters of the smart devices or appliances. In Mode A, a constant Internet connection (through Wi-fi) is required for the desired functionality.

The online speech recognition system will allow the user to communicate with the primary module in an intuitive manner with speech-based controls and requests. The AI Assistant will provide with features such as search engine results, date and time, weather updates, news updates, timers, reminders, calendars, to-do lists, alarms and other basic tasks. The parameter control system will act as a control hub for the smart and IoT-enabled devices in the network. It will allow for the switching of the device or appliance and also to control its various features and parameters. Since a constant Internet connection is expected in this specific mode (i.e. Mode A), an AI system could be based on a large server elsewhere and may possess a vast variety of features, allowing for a highly sophisticated and a very smart device.

Edge devices would encompass IoT-enabled devices and smart appliances with the ability to connect to the network. The secondary module may be based on a very basic microcontroller with Wi-fi connectivity. It may be interfaced with the device or appliance in multiple ways. From Fig. 1, let device ‘A’ be a typical television with no wireless connectivity options or any smart features; let device ‘B’ be a regular compact fluorescent light bulb (or an incandescent bulb) with no wireless connectivity options or any smart features. In such cases, the devices may be connected to a relay (or a power socket with a built-in relay), and these relay modules may be connected to the microcontroller-based secondary module which will then enable these devices or appliances to be connected to the network. Let devices ‘C’ and ‘D’ be a semi-smart LED bulb and a semi-smart door-lock system, respectively. They possess the ability to have their parameters switched or varied but do not possess the ability to wirelessly connect to the network; i.e., they have microcontrollers built-in, but are not

integrated with a wireless connectivity module (e.g. a Wi-fi chip). These devices may be directly connected to the microcontroller-based Wi-fi-enabled secondary module, hence enabling such devices to connect to the network as well. Let device ‘E’ be an IoT-enabled smart light with wireless connectivity features built-in. This device (or any such IoT-enabled device) may be directly connected to the network and does not need a secondary module.

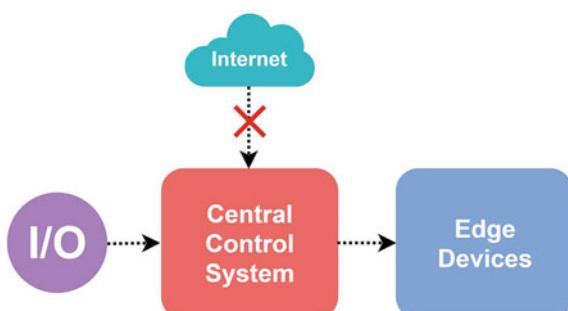
The speech recognition system will translate what the user has verbally said to text; after this, the AI Assistant will understand the user’s request. If the request requires searching the Internet for a certain fact or information, it will do so. If the request requires switching or controlling of whichever of the devices or appliances in the network, it will forward the request to the parameter control system. This system will process the request and, if authorized, will communicate with the edge devices to proceed with the switching or controlling of the device or appliance.

Mode B (Of line, with Local STT)—The second mode, namely, “Mode B” is depicted in Fig. 3. The basic functioning is slightly different from Mode A. The difference is the absence of Internet connectivity. Whenever the availability of an Internet connection has been compromised, the device will toggle from Mode A to Mode B. Due to the absence of a connection to the server, all the desired functionality has to be programmed locally; hence, the features provided by the AI Assistant will be limited. Features such as date and time, reminders, timers and alarms may be locally integrated and programmed; however, weather updates, news updates and search engine results would not be possible in this “Of line Mode”. If the control system is local to the network, most of the actions could be performed without the need for Internet. For example, a sensor in the local network is able to actuate/toggle a mechanism/switch which is also a part of the same network with no Internet access.

The offline speech recognition system will convert the user’s verbal speech into text; after this, the offline mode will scan for keywords (e.g. on, off, bulb one, door, etc.). These keywords when processed by the programmed algorithms will generate meaningful commands that will be passed on to the parameter control system, after which the procedure is the same as described for Mode A.

Mode C (Online, with Remote Control Access)—In the third mode (Fig. 4), i.e. “Mode C”, say the user is absent from the premises and hence is unable to use the microphone(s) built into the primary module; an instance of this may be when the user

Fig. 3 Mode B (offline, with local STT)



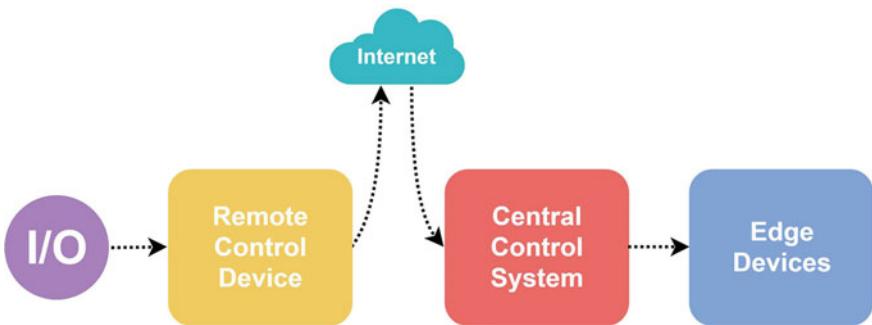


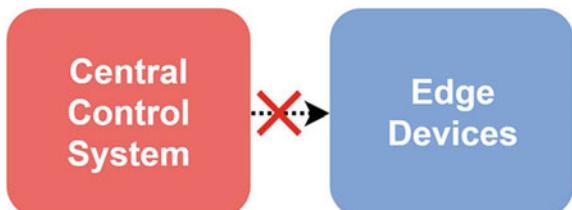
Fig. 4 Mode C (online, with remote control access)

is outside his or her house/office or wherever the system has been based. However, the user is expected to remotely carry his or her Internet-enabled smartphone with an integrated microphone and speaker system. The cellular (or Wi-fi) connection will allow the smartphone to communicate with the primary module.

Since Internet connectivity is a must in this mode, online speech recognition system may be used, after which the text obtained from the speech is used by the AI Assistant, providing the full suite of features, along with the ability to switch/control devices and appliances remotely from a completely different location by communicating with the central control system, in a similar manner as explained previously.

Mode D (Backup Mode)—The fourth mode (Fig. 5), is the “Backup Mode” or the “Conventional Mode” and is completely independent of the Central Control System. A common problem in existing SHSs is the absence of a two-way switching and control system (Staircase Switch/Multiway Switching). Generally, such devices can be controlled via the network only when the main switch is ‘ON’. To counter this in certain desired cases, the main switch may be kept on at all times; additionally, an onboard ‘Push Button’ can be provided (or Knobs, Sliders etc.) that shall be programmed to switch the device’s state directly from the microcontroller (Secondary Module) itself, without needing any sort of wireless communication with the primary module. This mode could be optionally implemented on devices that may benefit from such a system.

Fig. 5 Mode D (backup mode)



3.2 Architecture of Primary Module

The Primary Module has been designed to accommodate all the systems required to provide the functionality discussed above. Figure 6 shows the blocks involved in its architecture.

The online speech recognition system and the AI Assistant System blocks are both dedicated to a use case where constant Internet connection is available, i.e. Mode A. If the AI Assistant understands the user's request as a command to control any of the devices on the network, it will advance the request to the Parameter Control System. In case of Mode B, the local or online speech recognition system will advance the request forward if valid and authorized. In case of Mode C, the Primary Module receives the request from a remote device over the Internet, after which it advances the request to the succeeding system.

The Parameter Control System holds a database consisting of all the smart-devices or appliances connected and registered to the network, and their respective variable parameters. It understands the request and maps it to the address of the device on the network to which the command has been directed to. The data type of the parameter to be altered is known and is accordingly revised. Security features may be implemented within or after these blocks.

For the kinds of applications that are implemented in this project, the Message Queue Telemetry Transport (MQTT) protocol is by far the best option. MQTT works atop the TCP/IP layer, and since it's a very lightweight protocol, it focuses on the extremely efficient use of bandwidth and does not comprise any built-in security features except for username and password authentication. Due to the aggressive

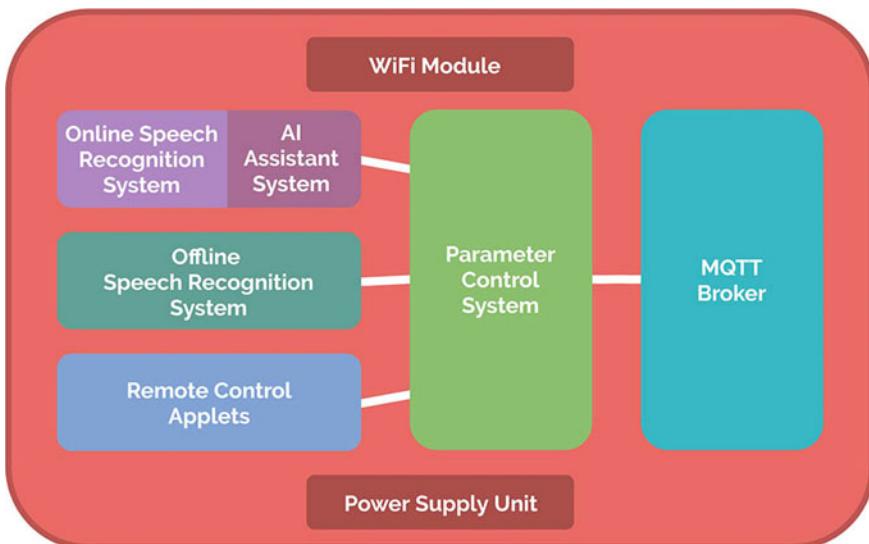


Fig. 6 Architecture of primary module

escalation in IoT development and adoption, securing the networks is critical. The built-in username and password protection offer no encryption whatsoever and leave the network open to exploitation.

One way to solve this is by implementing SSL/TLS connections [25] to securely transfer data over MQTT. Port 1883 is the standard for MQTT while port 8883 is registered and reserved for MQTT over SSL/TLS. Powerful platforms like the Raspberry Pi® or Orange Pi™ may easily support this [26] at the edge of the network. However, this solution comes at a huge cost when being implemented using the limited resources available in compact microcontrollers; the memory consumed by SSL/TLS is relatively sizable, along with relatively high CPU usage due to the communication overhead [27]. This may result in an unstable and frail system, leaving very little memory for the intended application itself. The computing resources would also be high for short-lived connections that IoT applications generally use (Note that SSL should be avoided regardless due to security shortcomings). Often, for machine-to-machine communication applications, end-to-end encryption or payload encryption would be a better choice than link encryption or channel encryption.

So, a workaround to this could be developed by implementing end-to-end AES encryption and hash authorization functions [28, 29] only on the data payload being transferred. To produce the cipher text, initialization vector may be used along with the message and the key to avoid or rather lessen the chances of successful pattern analysis attacks. As for authentication, a session ID and a hash function may be used. This is a lightweight alternative to SSL/TLS secure channels. Since this is done in the application layer, it offers much more flexibility to the possible designs [30]. Furthermore, sleep mode [31] is offered by many of the microcontrollers; this feature may be utilized to reduce standby power usage [32].

3.3 *MQTT Model Configuration*

The Control System is configured according to the MQTT model shown in Fig. 7. The model permits the connected devices to communicate with the Control System in 3 different ways as per requirements.

1. The Control System sends data to the Edge Device via MQTT Broker.

Control System → Edge Device 1.

In Fig. 7, the Control System publishes the control signals within a specific topic, the fan (Edge Device 1) is subscribed to that topic.

2. The Edge Device sends data to the Control System via MQTT Broker.

Control System ← Edge Device 4.

The pollution sensor (Edge Device 4) publishes data under a topic, the Control System is subscribed to that very topic.

3. The Control System and the device exchange data between each other via MQTT Broker.

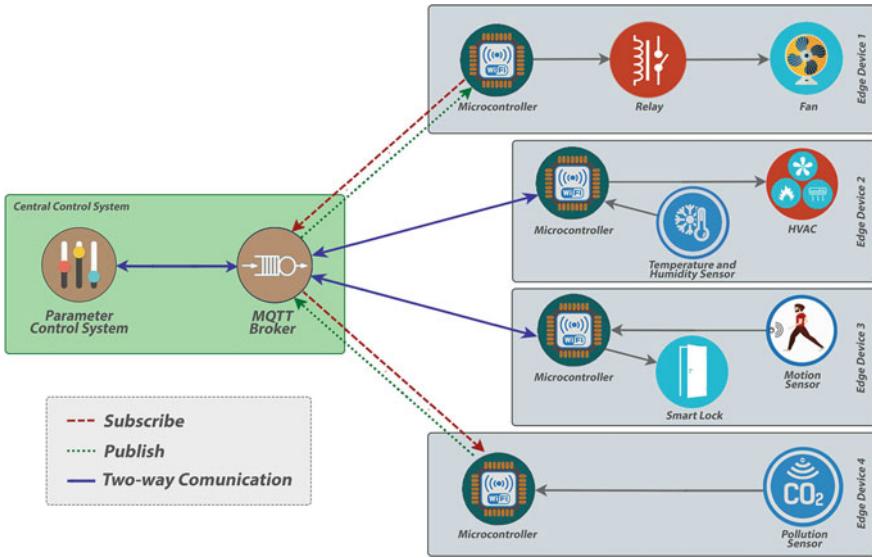


Fig. 7 MQTT model configuration

Control System ↔ Edge Device 2.

The temperature and humidity sensor publishes data under a topic, the Control System is subscribed to the same topic. Based on the input from the sensor, the Control System publishes the control signals under another topic to which the HVAC system is subscribed.

Control System ↔ Edge Device 3.

Similarly, the motion sensor publishes data under a topic, the Control System is subscribed to the same topic. Based on the input from the motion sensor, the Control System publishes the control signals under another topic to which the Smart Lock is subscribed.

4 Prototype Design

For demonstration and testing of the posited system, a prototype was implemented; it is capable of controlling devices in the four different ways explained previously. The prototype is a reasonably affordable and relatively easy mean to construct a voice-controlled home assistant which would help in achieving the smart-home objectives. The implemented system may be used to control home appliances and also as a chatbot for queries amongst other things. Users may ask different types of questions regarding weather, news, facts, current updates on sports, basic calculations and all other general information that could be found using the search engine. The Primary Module was based on a Raspberry Pi® 3 Model B+, while the Secondary Module

was based on a NodeMCU ESP8266. The Primary Module, along with its dedicated microphone array and speaker system were placed inside a 3D printed casing and powered using a powerbank further giving it flexibility in terms of portability. Appropriate wiring cavities were provided in the casing for charging the device.

4.1 Mode A (*Online, with AI Assistant*)

The microphone array is used to receive voice commands from the user; these commands will be used as input to the Primary Module. The speaker is used for auditory acknowledgement to commands and also to deliver answers to various questions asked by the user or to ask follow up questions. The board has a built-in Wi-fi chip that enables the module to connect to the Internet. For speech recognition, in this mode, Google's Cloud Speech API was used. For the smart personal digital assistant features, Google's Assistant API was used. As soon as the system is powered up, it connects to the network and establishes a connection to communicate with the Google servers to run the APIs. Once the connection is established and the system start-up is complete, it listens for a keyword to trigger the voice input and it is processed in the Primary Module. The system remains idle until the keyword is detected. MQTT, the lightweight publish/subscribe messaging protocol is employed to liaise with the edge devices.

An MQTT broker (Eclipse MosquittoTM) was installed in the Primary Module, which acts as a broker to publish/subscribe messages on different topics. The Secondary Module was used for the devices that cannot be controlled wirelessly through the network by itself. These devices were controlled by using a relay (Devices 'A' and 'B' of Fig. 1 and the Bulb, Fan and Heater of Fig. 8). Devices that can be directly connected to the Secondary Module without a relay in between were connected with wires linking the ESP to the devices' inbuilt controller (Devices 'C' and 'D' of Fig. 1). Devices that can be innately connected to the network wirelessly were connected directly to the Primary Module (Device 'E' of Fig. 1) with some software (firmware) tweaks. All the devices were made to be a part of the same local

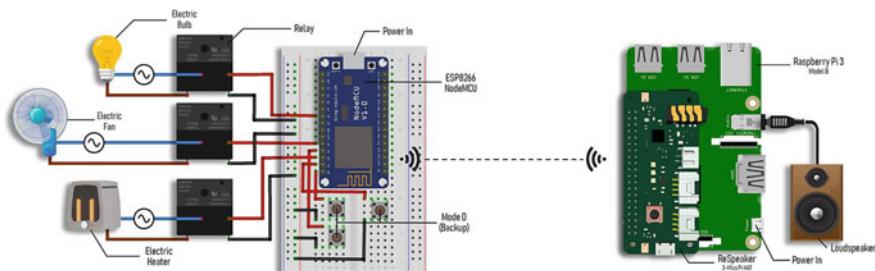
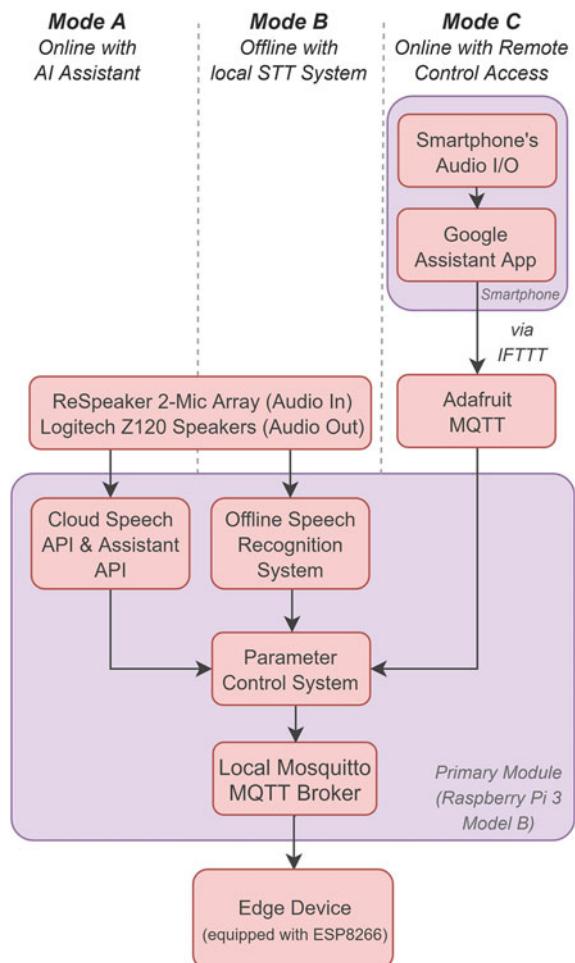


Fig. 8 Prototype circuit diagram

network. A working Internet connection is required in order to ensure uninterrupted service from the Google APIs. The Secondary Module was subscribed to custom topics and could identify the device control operation by identifying the message that was published by the Primary Module on the custom topic (Fig. 9).

The use of MQTT protocol for communication ensures fast and reliable data transfer. The speaker interfaced is programmed to acknowledge the device control response and is also used to provide feedback from the Google server. For client implementation, Eclipse Paho is used on the Raspberry Pi®. The Eclipse Mosquitto™ acts as a link between the Primary Module and the Secondary Module. It is an open-source broker that enables server-implementation of the MQTT protocol and implements a lightweight method of M2M communication using a publish/subscribe model. It can be used in high power to low power machines. It is mainly used in IoT

Fig. 9 Request flow



applications and runs over the TCP/IP layer. The connected devices communicate only when required.

A few sets of custom Python™ phrases were added to the Eclipse™ Paho MQTT open source program's code. If Google's Cloud Speech API detects these phrases, the program publishes a message on the topic which corresponds to the custom phrase. If the custom phrases are not detected, Google's Assistant API is used in order to search for results based on the user's input phrase. If the queries are general queries regarding time, weather etc., then the system responds just like how an Android™ based Google Assistant™ application would. When a custom phrase is encountered, the assistant is paused and the Parameter Control System is activated; it sends a message on a particular topic from the registry dedicated to the registered and connected devices. If the user commands "Turn ON the bedroom light", and this phrase is found to be present in the custom phrase list, then a message '1' (let '1' denote ON and '0' denote OFF) is published on the topic "bedroom light state". Multiple cases could be added on to the custom phrase list in order to perform the same task with a variety of commands. Alongside dedicated custom phrases, the system is also designed to extract keywords from input phrases in order to perform intended tasks; if the phrase "Turn ON the bedroom light" is present in the recognized list, and the phrase "Turn the bedroom light ON" is not, the system extracts and matches the keywords (bedroom light and ON) to proceed with the action. The system is also capable of implementing power commands in order to shut down or reboot the system.

4.2 Mode B (*Offline, with Local STT*)

In case of a weak Internet connection or absence of Internet connectivity, Mode A becomes impotent to perform all tasks and hence the system switches to Mode B. Mode B uses an Offline STT to identify keywords. Sound Pattern Recognition (SOPARE) [33] is an open-source offline speech recognition system and was used to construct Mode B for the prototype. It can be trained in any language; each word/phrase should be trained individually and the dictionary has to be run in order to ensure that the trained words are correctly recognized. SOPARE is just an STT and in order to add features that enable automation, plugin programs/functions are added and called when trained words/phrases are detected. These plugins are configured to publish respective data (using the Parameter Control System) under the required topics through the Mosquitto™ broker to the Edge Devices on the detection of the assigned keywords. A Hotword or wake-word to trigger the STT was added in order to get the system to start listening and taking input.

4.3 Mode C (*Online, with Remote Control Access*)

This mode is applicable if the user is away from the premises but desires to control the devices remotely. In order to implement the mode, an Android™ based smartphone with Google Assistant™ application installed was used. This mode is considered to be an add-on to the Mode A system. However, unlike Mode A, Mode C uses the speakers and microphones built into the smartphone itself for remote access. IFTTT™ (If This Then That) is an easy way to connect different apps/applets together and to establish a chain of events on a conditional basis. IFTTT™ was used to establish the work flow of Mode C by making a recipe with Google Assistant™ and Adafruit™. Another alternate to Mode C is to use Google Assistant™ along with Webhooks to control these devices. By using Webhooks, the security of the system is compromised by the implementation of port forwarding. Hence, Adafruit was considered over Webhooks to build the system. An Adafruit™ account with multiple feeds was created, where each feed denoted a device that is meant to be controlled. The data corresponding to the feed indicates the state of the device. While setting the Adafruit™ applet using IFTTT™, the feed and the data to be saved are mentioned for each phrase/task. If the user commands “Turn ON the kitchen light”, then the Google Assistant™ connects to the Adafruit™ server, then the data corresponding to the feed “kitchen light state” is saved as ‘1’ (‘1’ denotes ON and ‘0’ denotes OFF). All the feeds are stored in the dashboard. The username and key of the dashboard are used to authenticate and establish a connection with the Primary Module. The received data is processed by the Parameter Control System and then the appropriate requests are forwarded to the Edge Device through the Mosquitto™ broker by publishing the data on the respective topics (Ref. Mode A).

4.4 Mode D (*Backup Mode*)

This mode gives the privilege to control devices in a conventional manner using physical buttons or sliders, either when the communication between the Primary Module and the Secondary Module has been compromised or when the user is in close proximity to the device/appliance and desires to use conventional means. Push buttons are connected to some of the unused digital pins of the microcontroller. These buttons are programmed to be used to control the device/appliance directly from the Secondary Module itself, without needing any communication with any other device. The parameters changed are reflected in the registers either immediately (in case of active communication with Primary Module) or when the connections are successfully established.

4.5 Hardware Used for Prototyping

The Central Control System composes of a Raspberry Pi® 3 Model B+ for the Primary Module, a ReSpeaker Dual Microphone Array and a Speaker System, all powered by a 20,000 mAh powerbank and housed in a 3D printed enclosure.

The Edge Devices are composed of the Secondary Module which is based on a NodeMCU ESP8266 Wi-fi-enabled chip, the device/appliance itself, a trans-former setup to power the Secondary Module directly, and push-buttons for Mode D. It's designed with a mounting mechanism that allows the Secondary Module to be embedded into traditional household switchboards or with the device/appliance in an unobtrusive manner.

The Raspberry Pi® 3 Model B+ is where the main processing and switching/controlling takes place as it runs the custom programs. It is powered by a 4-core Broadcom chip running at 1.2 GHz which is plentiful for the expected needs. It has 1 GB of onboard RAM and comes with a built-in Wi-fi chip that allows connection to the local network which is expected to be connected to the Internet.

The ReSpeaker Pi HAT with dual microphones can capture voice within a radius of up to 3 meters. It has been interfaced with the Raspberry Pi® board via the 40 GPIO pins available. Not all the 40 pins are used, the unused pins are redirected as usable I2C or GPIO ports. The Pi HAT does not need to be separately powered, it sips in power from the Raspberry Pi® itself.

As for the Secondary Module, an extremely affordable Wi-fi-enabled chip, the NodeMCU ESP8266 microcontroller was used. It is compact, robust and compact enough to be integrated into any general home appliance or switch-board. However, it should be noted that when it comes to production, all that is required is to feature a programmable Wi-fi chip with the controller that already exists within the appliance. So this, when designed with proprietary chips, will become all the more compact and economical.

4.6 Software Implementation

An instance of the systems holistic (all four modes) execution has been depicted in Fig. 10 and is further discussed.

A lighting device was connected to the network via a relay and a Wi-fi enabled microcontroller. At first, a Trigger Word or a Hotword/Wake-word is used to activate the Control System, then the connected device was turned ON using speech based commands in Mode A. After this, the wireless connectivity was disabled, following which the Control System switches itself from Mode A to Mode B.

To illustrate Mode D, the connected device was switched OFF using the onboard push button. The device turns off, and if and when the device is able to communicate with the Control System, the microcontroller will inform the primary module about the Mode D event, and the primary module will modify the respective registers

```

pi@raspberrypi: ~
main:online:waiting for trigger
main:online:listening...
speech:online:processing...
speech:online:"Turn on lamp 1"
main:online:processing...
main:online:"Lamp 1 has been turned on."
main:warning:"Internet connection failed!"
main:warning:"Switching to Mode B"
main:backup_event:receiving...
main:backup_event:"Lamp 1 turned off."
main:offline:waiting for trigger
main:offline:listening...
speech:offline:processing...
speech:offline:"Turn on lamp 1"
main:offline:hotword detected:on;lamp 1
main:offline:action21
main:offline:"Lamp 1 turned on."
main:warning:"Connected!"
main:warning:"Switching to Mode A"
main:online:waiting for trigger
main:online:remote:receiving...
main:online:remote:"Lamp 1 turned off."

```

Mode A

Mode D

Mode B

Mode C

Switching Mode

Switching Mode

Fig. 10 Screen-shot (control system)

accordingly. In this instance, the connection between the Control System and the edge device is active, and only the Internet connection has been compromised, hence, the Mode D event is immediately registered by the Primary Module.

Since the control system is in Mode B, on activating it with the trigger, a speech based command was given to turn the lighting device ON. The Internet connection was then re-activated causing the Control System to switch back to Mode A from Mode B.

For remotely controlling the lighting device via Mode C, an AndroidTM based Internet-enabled smartphone was used with Google AssistantTM (shown in Fig. 11) which has been linked to the Primary Module via broker. A speech based command to turn the lighting device OFF was given to the Assistant, which forwards the request to

Fig. 11 Screen-shot (remote device)

the Primary Module. Once authorized, the modifications to the respective parameters are registered and the command is executed.

5 Conclusion

This work gives insights on the potential and flexibilities home automation systems could carry regardless of the infrastructural development status of the region. The posited system is economically and practically reasonable yet feasible with off-the-shelf products, with no compromise in terms of functionality or security even when connectivity is absent. It also provides remote control access along with automation and AI capabilities further increasing features and convenience.

Almost all appliances consist of inbuilt microcontrollers innately, adding simple, low-priced wireless connectivity chips to them would vastly improve their utility, efficiency and adaptability with negligible increments in terms of cost. MQTT shows to be an ideal protocol for such M2M use cases, and provides high levels of dexterity and adeptness; end-to-end encryption ensures security and safety of the entire system. This could be applicable to small-scale households or large-scale enterprises. Speech based interaction further gives accessibility to medical ecosystems. Its simplicity gives way to interoperable devices giving consumers a wide variety of options when selecting devices/appliances, ensuring a consolidated and integrated experience, eradicating the need for multiple software applications/installations, without needing engineers or technicians for setup. Appliances from various manufacturers could be controlled from a single platform for a unified UI/UX.

References

1. Kortuem G, Kawsar F, Sundramoorthy V, Fitton D (2010) Smart objects as building blocks for the internet of things. *IEEE Internet Comput* 14(1):44–51
2. Yang R, Newman MW (2013) Learning from a learning thermostat: lessons for intelligent systems for the home. In: Proceedings of the 2013 ACM international joint conference on pervasive and ubiquitous computing, ser. UbiComp ’13. ACM, New York, NY, USA, pp 93–102
3. Lin J, Yu W, Zhang N, Yang X, Zhang H, Zhao W (2017) A survey on internet of things: architecture, enabling technologies, security and privacy, and applications. *IEEE Internet Things J* 4(5):1125–1142
4. Jie Y, Pei JY, Jun L, Yun G, Wei X (2013) Smart home system based on IoT technologies. In: 2013 international conference on computational and information sciences, June 2013, pp 1789–1791
5. He W, Yan G, Xu LD (2014) Developing vehicular data cloud services in the IoT environment. *IEEE Trans Ind Inform* 10(2):1587–1595
6. Song T, Capurso N, Cheng X, Yu J, Chen B, Zhao W (2017) Enhancing GPS with lane-level navigation to facilitate highway driving. *IEEE Trans Veh Technol* 66(6):4579–4591
7. Wang M, Zhang G, Zhang C, Zhang J, Li C (2013) An IoT-based appliance control system for smart homes. In: 2013 fourth international conference on intelligent control and information processing (ICICIP), June, pp 744–747

8. Akerman smart home security (2016) (online). <https://www.ackermansecurity.com/>
9. Nest thermostat (2016) (online). <https://nest.com/thermostat/meet-nest-thermostat/>
10. Azman M, Panicker JG, Kashyap R (2019) Wireless daisy chain and tree topology networks for smart cities. In: 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT), Feb, pp 1–6
11. Kashyap R, Azman M, Panicker JG (2019) Ubiquitous mesh: a wireless mesh network for IoT systems in smart homes and smart cities. In: 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT), Feb, pp 1–5
12. Panicker JG, Azman M, Kashyap R (2019) A LoRa wireless mesh network for wide-area animal tracking. In: 2019 IEEE international conference on electrical, computer and communication technologies (ICECCT), Feb, pp 1–5
13. Kodali RK, Azman M, Panicker JG (2018) Smart control system solution for smart cities. In: CyberC 2018: international conference on cyber-enabled distributed computing knowledge discovery
14. Hidayat S, Firmando SF (2015) Scheduler and voice recognition on home automation control system. In: 2015 3rd international conference on information and communication technology (ICoICT), May, pp 150–155
15. Patchava V, Kandala HB, Babu PR (2015) A smart home automation technique with Raspberry Pi using IoT. In: 2015 international conference on smart sensors and systems (IC-SSS), Dec, pp 1–4
16. Rani PJ, Bakthakumar J, Kumaar BP, Kumaar UP, Kumar S (2017) Voice controlled home automation system using natural language processing (NLP) and internet of things (IoT). In: 2017 third international conference on science technology engineering management (ICONSTEM), Mar, pp 368–373
17. Paul A, Panja M, Bagchi M, Das N, Mazumder RM, Ghosh S (2016) Voice recognition based wireless room automation system. In: 2016 international conference on intelligent control power and instrumentation (ICICPI), Oct, pp 84–88
18. Soumya S, Chavali M, Gupta S, Rao N (2016) Internet of things based home automation system. In: 2016 IEEE international conference on recent trends in electronics, information communication technology (RTEICT), May, pp 848–850
19. Jain S, Vaibhav A, Goyal L (2014) Raspberry pi based interactive home automation system through e-mail. In: 2014 international conference on reliability optimization and information technology (ICROIT), Feb, pp 277–280
20. Nan E, Radosavac U, Matic M, Stefanovic I, Papp I, Antic M (2017) One solution for voice enabled smart home automation system. In: 2017 IEEE 7th international conference on consumer electronics-Berlin (ICCE-Berlin), Sept, pp 132–133
21. Gill K, Yang SH, WI W (2013) Secure remote access to home automation networks. IET Inf Secur 7(2):118–125
22. Jose AC, Malekian R, Ye N (2016) Improving home automation security; integrating device fingerprinting into smart home. IEEE Access 4:5776–5787
23. Vacher M, Lecouteux B, Romero JS, Ajili M, Portet F, Rossato S (2015) Speech and speaker recognition for home automation: preliminary results. In: 2015 international conference on speech technology and human-computer dialogue (SpeD), Oct, pp 1–10
24. Cenedese A, Susto GA, Belgioioso G, Cirillo GI, Fraccaroli F (2015) Home automation oriented gesture classification from inertial measurements. IEEE Trans Autom Sci Eng 12(4):1200–1210
25. Secure communication with TLS/SLS (online). <https://dzone.com/articles/secure-communication-with-tls-and-the-mosquitto-broker>
26. Urien P (2017) Securing the IoT with TLS/DTLS server stacks embedded in secure elements: an eplug usecase. In: 2017 14th IEEE annual consumer communications networking conference (CCNC), Las Vegas, Mar, pp 569–570
27. MQTT security fundamentals: TLS/SSL (online). <https://www.hivemq.com/blog/mqtt-security-fundamentals-tls-ssl>
28. Boyce S (2017) Practical IoT cryptography on the espressif esp8266 (online). <https://hackaday.com/2017/06/20/practical-iot-cryptography-on-the-espressif-esp8266/>

29. Aubin DS (2017) Esp8266 SSL/TLS MQTT connection (online). <https://internetofhomethings.com/homethings/?p=1820>
30. Cope S (2017) Introduction to MQTT security mechanisms (online). <http://www.steves-internet-guide.com/mqtt-security-mechanisms/>
31. Watchdog timer and sleep mode of microcontroller (online). <http://microcontrollerslab.com/watchdog-timer-sleep-mode/>
32. Williams E (2016) Minimal MQTT: power and privacy (online). <https://hackaday.com/2016/06/02/minimal-mqtt-power-and-privacy/>
33. Kauss M. Sopare (online). <https://www.bishoph.org>

Real-Time Detection and Prediction of Heart Diseases from ECG Data Using Neural Networks



K. V. Sai Kiran, Mohamed Azman, Eslavath Nandu,
and S. K. L. V. Sai Prakash

Abstract Portable real-time continuous heart disease detection and/or prediction systems based on AI could be exhaustive, as it addresses the various drawbacks of conventional ECG machines. Real-time ECG monitoring devices generally tend to be large, heavy, and expensive and are placed in hospitals; they require the patient to be bed-bound. Portable systems, on the other hand, offer mobile monitoring, but not in real time and are not continuous but rather require the patient to regularly scan, with each scan taking minutes for analysis of the data. Certain diseases cannot be detected with short-term cardiac monitoring, instead require long-term continuous monitoring. Most of these systems use conventional disease detection algorithms, which tend to be limited to specific diseases and usually have less flexibility. AI systems, however, possess the flexibility of accommodating detection of various diseases using the same algorithm, just by providing additional data, and also have the ability to be vastly improved based on the data provided. In this work, the developed detection system is based on an ANN; the data to train the ANN were obtained from UCI-ML repository. An API was developed to implement the detection system on real-time continuous monitoring portable/wearable devices. The architecture and parameters of the neural network were tuned to get optimal performance. The system's timing analyses and performance metrics like the accuracy of ANN were measured. The neural network achieved an accuracy of 93.01%. Efficiency was given a high priority, to design a feasible solution for implementation on wearable devices to continuously monitor heart activity.

K. V. Sai Kiran · M. Azman (✉) · E. Nandu · S. K. L. V. Sai Prakash
Department of Electronics and Communication Engineering,
National Institute of Technology, Warangal 506004, Telangana, India
e-mail: mohamedazmanm1@gmail.com
URL: <http://www.nitw.ac.in>

K. V. Sai Kiran
e-mail: saikirank025@gmail.com

E. Nandu
e-mail: eslavathnandu@gmail.com

S. K. L. V. Sai Prakash
e-mail: sai@nitw.ac.in

Keywords Artificial neural networks · ECG · Mobile cardiac telemetry · Real-time detection of heart diseases

1 Introduction

A considerable proportion of the deaths in the world is due to heart diseases. In most cases, it is difficult to predict these heart-related issues beforehand. Moreover, a majority of heart diseases, when untreated immediately, might have fatal consequences. Around 17.7 million people die each year [1] due to cardiovascular diseases throughout the world. This accounts for 31% of total deaths globally. These are the major cause of deaths in the world than any other reason. Among them, 7.4 million people are affected by coronary heart disease and 6.7 million people are affected by heart strokes, annually. Hence, there is a need to give utmost importance to detect these heart diseases at their early stages. Almost any kind of heart disease can be detected by analysis of electrocardiogram (ECG) which represents the electrical activity of the heart. There are several heart diseases and it is tedious to detect each type of heart disease from ECG manually. Since a lot of ECG data and corresponding heart activity statistics are available, one may adopt artificial intelligence-based algorithms which leverage these large datasets to give meaningful conclusions and accurate predictions.

1.1 Overview of Heart Functionality

The heart is one of the most important organs in the body. It has four chambers separated by valves. The four chambers are: two atria and two ventricles. The atria collect blood and pump it to the ventricles. The ventricles contract and pump blood to various parts of the body. Right atrium collects the deoxygenated blood from all over the body. This deoxygenated blood is pumped to the right ventricle. From the right ventricle, it is pumped to lungs where the blood gets oxygenated. This oxygenated blood is carried by the pulmonary vein and poured into the left atrium which is then pumped to the left ventricle. This oxygenated blood is pumped from the left ventricle to all the body parts via the aorta. The cycle repeats, keeping the blood flowing throughout the body; this is called the cardiac cycle.

1.2 Heart Diseases

There are several heart-related diseases due to abnormal functioning of the heart. Some of them are myocardial infarction, coronary artery disease, heart arrhythmia, atrial fibrillation, tachycardia, premature ventricular contraction, bradycardia, atrial

flutter, ventricular tachycardia, etc. Myocardial infarction (MI) [2] is the most common type of heart disease. The reason for MI is a blood clot in the ruptured plaque area. Plaque is a solid material consisting of fat, cholesterol, calcium, cellular waste, and fibrin. Plaque is deposited on the inner walls of the arteries over months or years. If blood pressure increases due to some reason, the plaque ruptures and blood starts clotting in that area in little to no time. Oxygenated blood supply is reduced to heart muscles if this clot occurs in coronary arteries (which supply blood to heart muscles). Soon heart tissues start dying. This death (infarction) of heart muscles (myocardial) is called myocardial infarction (depicted in Fig. 1 [3]).

Immediate treatment should be given after the block formation to avoid fatal consequences. General treatment includes giving tissue plasminogen activator (TPA) which can remove the blood clot; TPA should be given before most of the heart muscles die as these damages to heart cells are irreversible. The time before which significant damage to the heart muscle occurs is highly dependent on the percentage of the block, age of the patient, gender, and many other biological parameters. Generally, this time varies between a few minutes to a couple of hours. After this time period, majority of the heart cells die and the heart may stop beating. Even after the heart stops beating for about 10–20 min, defibrillation can be used to give external electric shock and make the heart beat again, and then TPA may be given to remove the clot. The most common symptom of myocardial infarction is chest pain. However, unfortunately, most diabetic patients will not feel the chest pain which may cause fatality without them ever recognizing the symptoms of heart attack. So, there is a need to detect these heart attacks in other ways, quickly without depending on normal biological symptoms. In most diseases, there is a significant amount of time between

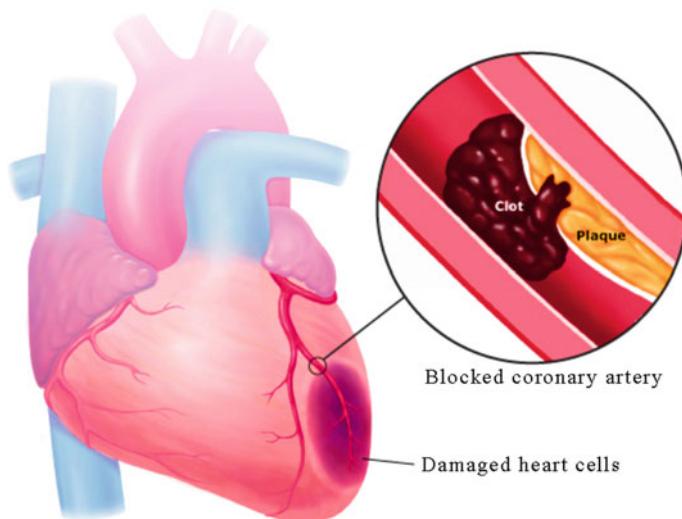


Fig. 1 Myocardial infarction

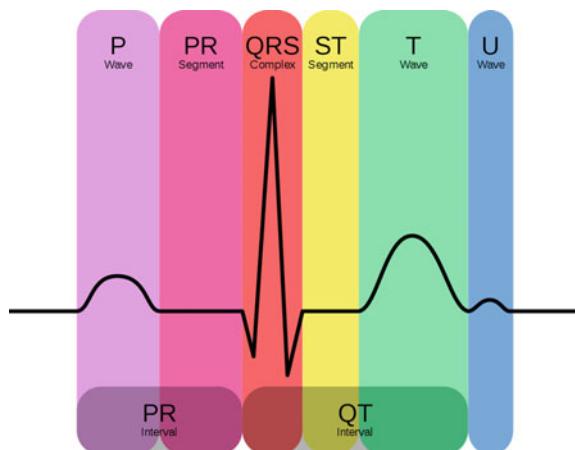
when the imperfection may be detected by ECG data and when the consequences are fatal; if treatment is given within this time, the patient could be recovered completely and may also avoid serious outcomes.

1.3 *Electrocardiogram*

Electrical impulses generated in the heart system are responsible for heartbeats. An electrical impulse is generated by the sinoatrial node (SA node), which is also called the natural pacemaker. The signal generated at the sinoatrial node travels to the atrioventricular node (AV node) through the atria. Due to the passing of electrical pulse through the atria, they will contract. This is called atrial depolarization. Now, the atrioventricular node sends this electrical pulse through the ventricles which results in ventricular contraction. This is called ventricular depolarization. Then, atria and ventricles relax which is called ventricular repolarization. Thus, contraction and relaxation of the heart is a result of these electrical impulses. This electrical activity of the heart when represented as the pulses' amplitude with respect to time on a 2D plane is called electrocardiogram (ECG); Fig. 2 [4] depicts an ideal ECG waveform for one cardiac cycle.

P wave signifies atrial depolarization which is responsible for atrial contraction. QRS complex signifies ventricular depolarization which is responsible for ventricular contraction. T wave signifies ventricular repolarization which is responsible for ventricular relaxation. Ventricles have to pump blood to all parts of the body in contrast to atria which pump blood to ventricles. So, ventricles need high power to pump blood all over the body. Thus, QRS complex has a higher amplitude compared to other waves which give more power to ventricles to do their job. Almost all the

Fig. 2 ECG waveform



known heart diseases can be detected by the analysis of the ECG. Some of the heart diseases and their corresponding ECG changes [5] are shown below:

- Myocardial infarction—Elevated ST segment
- Early stage of myocardial infarction—Peaked T waves
- Ischemia—Depressed or elevated ST segment and inverted T waves
- Sinus bradycardia—Low heart beat rate and low duration of QRS complex
- Sinus tachycardia—High heart beat rate and P wave indistinguishable from T wave
- Right bundle branch block—Widening of QRS complex
- Atrioventricular block—Longer PR interval.

2 Literature Review

Existing real-time ECG monitoring systems are usually substantially large and bulky machines placed in hospitals, requiring the patient to be bed-bound. Existing portable ECG machines do not support real-time monitoring and cannot monitor continuously. They require timely scanning with each scan taking tens of seconds and additional time for analysis of the recorded data. They usually tend to use conventional detection algorithms, not based on machine learning.

The Omron® HCG 801 and the HeartCheck™ MD100 devices require the patient to measure ECG periodically and require 30 seconds for each scan. It then gives some basic information; however for further analysis, the SD card needs to be removed from the device and inserted into a PC after which a proprietary software tool has to be used. Sandor's Spyder ECG is a cloud-based wireless and continuous ECG recorder; however, it uses conventional analysis algorithms to give a basic report with limited detection abilities and limited emergency alert options. Medtronic's SEEQ™ is a wireless adhesive-based monitor that checks for irregularities [6].

A previous work [7] evaluates the behavior and performance of neural networks on the classification or categorization of electrocardiogram sequences according to normality or abnormality of the input data. The results were compared with some other models such as random forests, logistic regression, and SVMs. For testing the models, a computer system with a dedicated graphics card was used, as it accelerates such training processes. Bagging tree [8] is an ensemble method which has certain desirable qualities like highly efficient computational performance and ability to deal with class imbalance problem within the data used for classification. Another method is based on optimization of the genetic algorithm [9], where the system calculates the number of hidden nodes for the neural network which train the network with proper selection of neural network architecture and uses the global optimization of genetic algorithm for initialization of neural network. An algorithm called VF15 [10] for voting feature intervals is an inductive and supervised algorithm for inducing classifications knowledge with examples; it outperforms some standard algorithms like the nearest neighbor and Naive Bayesian classifiers. The feasibility of real-time analysis and classification of electrocardiogram data on a personal digital assistant has

been highlighted previously [11]. It is dated in 2005, since which the efficiency and performance of low-TDP chipsets have exponentially improved. However, the work [11] gives us useful insights regarding implementation of classifiers on low-power devices. An Android™ application [12] which allows real-time ECG monitoring was previously developed; it used data acquired by a Shimmer™ sensor. Real-time ECG monitoring system based on FPGA [13] with VHDL applied to transmit and record data has also been developed. In some projects [14], data extracted were proposed to be transmitted to medical centers via text messages. A similar work [15] proposed a conceptual framework for monitoring of a diver's ECG pulses underwater. It also included an alert system which could warn the diver and/or other persons. The system proposed in the following work is exhaustive, as it addresses all the above-mentioned issues. It supports real-time and continuous detection and/or prediction and uses artificial intelligence to process and analyze the recorded data.

3 Proposed System

The proposed solution is to implement neural networks to classify fed ECG data into one of the different heart diseases or classify as healthy. Any AI-based approach must have a preprocessing step of training the neural network with existing data using some technique or algorithm. There on, any new ECG data can be fed to this trained neural network which will classify this ECG data into one of the classes.

3.1 The System

3.1.1 Block Diagrams

Figure 3 shows all the blocks that are present in the proposed detection system. The block in yellow (Block 1; data extraction) is the data extraction device. This device may be implemented using electronic skin. The sensors or electrodes used are variants of the flexible electronics, they adhere to the human skin due to van der Waals forces, and this was developed by Someya [16] from University of Tokyo, and Professor Rogers [17] from Rogers Research Group. A simple depiction of what Block 1 might look like has been shown in Fig. 4. The raw analog data are filtered and sampled using simple circuits that can be printed onto the e-skin itself. Memory units for these types of e-skin circuits have already been developed.

It could be powered by a small battery which may be recharged using solar cells, inductors, or thermoelectric power harvesting pads (delta T, i.e., potential developed due to the difference in temperature of the two surfaces, one of which is in contact with the skin, and the other is exposed to the atmosphere). Research work on these technologies is ongoing [16–18], as demand seems predictable. The block in blue (Block 2; real-time detection) is the device in users' possession, which may be a

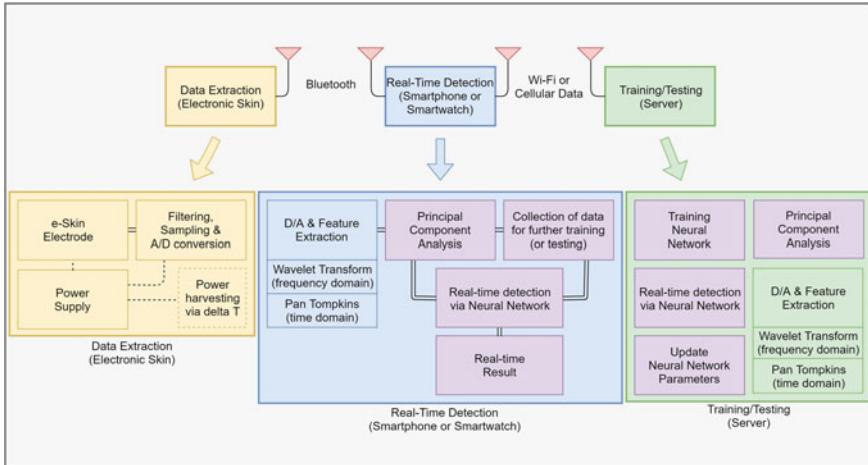
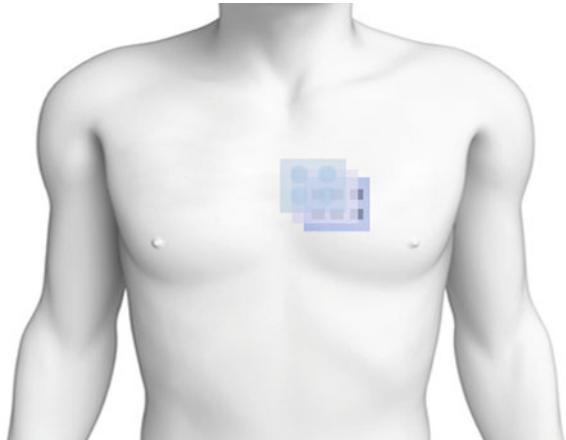


Fig. 3 Full system

Fig. 4 Electronic skin device



smartphone, a smartwatch, or a dedicated fitness or medical device. This unit would be connected to the data extraction unit via Bluetooth™ or any other suitable low-energy and low-range wireless communication protocol. The data extracted by the sensors from the human body by Block 1 will be transmitted to Block 2. The block in green (Block 3; testing/training) is the server where the training of the neural network takes place. Block 2 may be connected to Block 3 via the Internet, using cellular communication or Wi-Fi. The modules highlighted in purple (principal component analysis, collection of data for further training or testing, training neural network, real-time detection via neural network, update neural network parameters, real-time result) in Fig. 3 are the ones that have been implemented in this work.

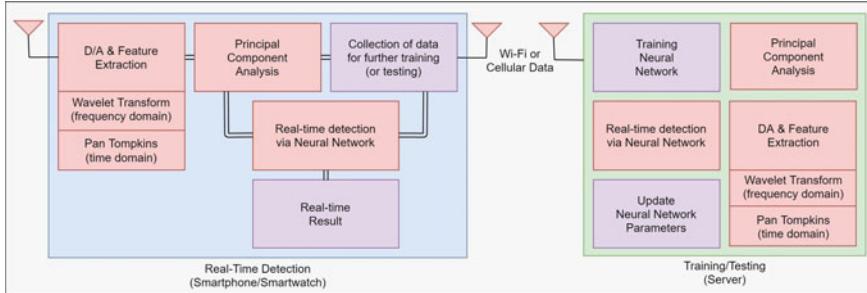


Fig. 5 Flexible modules

The system may be designed in a very flexible manner depending on the requirements of the end product. According to the processing power available and the battery life desired, modules may be moved to and from Block 2 and Block 3. The modules highlighted in pink (D/A & feature extract, wavelet transform, Pan–Tompkins algorithm, principal component analysis, real-time detection via neural network) in Fig. 5 are those which need to be present in at least either one of the two, Block 2 or Block 3. The modules in purple (collection of data for further training or testing, training neural network, update neural network parameters, real-time result) in Fig. 5 are needed to be present in their respective blocks. If the smartphone or smartwatch has enough processing power to compute feature extraction (via wavelet transform for frequency domain features and/or Pan Tompkins algorithm for time domain features), to reduce the number of features being input into the neural network, and to run the scanning/detection through the neural network, those respective modules may be implemented within the users' device. However, if the processing power is not available, or if battery life is of more importance, then some or all of the blocks may be moved to the server-side. The server updates the parameters of the neural network on a timely basis, maybe weekly, monthly, etc. It can be updated as the server trains the neural network with new data as and when new data are available. This way, the accuracy may be continually improved. The architecture of the neural network may also be updated so as to get better performance and/or efficiency. In case all the flexible modules are present in the users' device itself, then the lag or latency introduced by the WAN network will not be of any issue as the entire scanning/detection can be done locally. The total time taken for each test (detection/scanning) loop will be the time consumed for sensors to extract data, filtering of data, transmission via Bluetooth™, extraction of features, reduction of features via PCA, and running the test on the local neural network. Let this model be “Type 1”. Figure 6 represents a model of Type 1; the modules highlighted in purple are actively used, while the modules in white are either not being used or not present at all. In Type 1, when the server updates the parameters of the neural network, the parameters are sent over via timely updates to the users' device. The device will use the neural network for detecting the disease, locally on the users' device. This way, WAN network's performance is

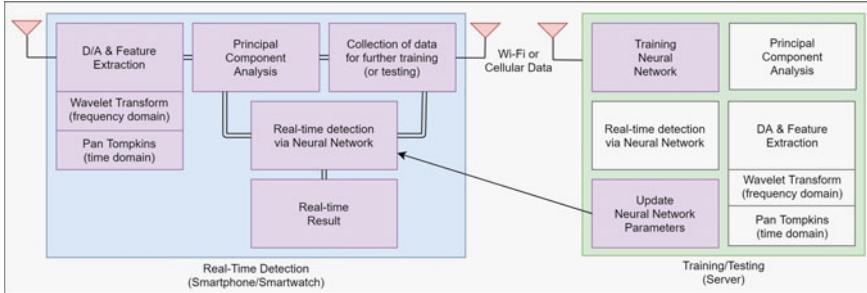


Fig. 6 Type 1 model

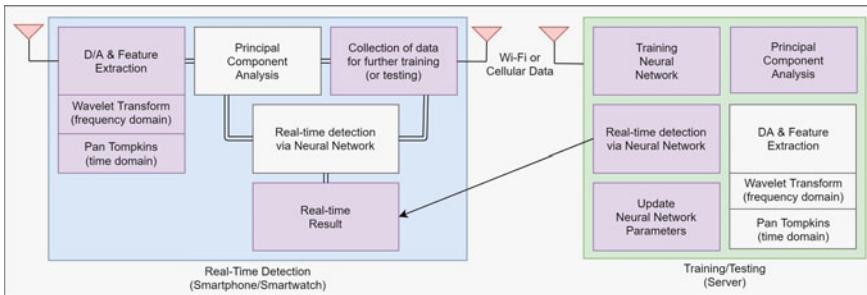


Fig. 7 Type 2 model

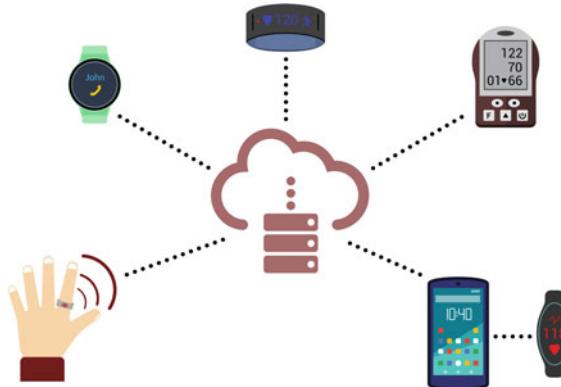
inconsiderate and Internet connectivity is delay-insensitive. This model will be able to alert the patient of any emergency even in the absence of an Internet connection.

Another model defined is of “Type 2”. Figure 7 represents a model of Type 2. In this model, the users’ device is required to function with low processing power and/or the device is required to have a very long battery life, hence the local processing load is minimized. The data received from the Block 1 (data extraction block, i.e., the electronic skin device) will be forwarded to the Block 3 (server) via Block 2 (user’s device). The server will do the training as usual, but in this case, also the scanning/detection. After scanning/detection, when the result is found, it is sent to the Block 2, and there the result is displayed or expressed via other means. In Type 2, the WAN network plays a role in the scanning/detection of the extracted data, hence Internet connection is delay-sensitive. This model, though it will provide greater battery life, will require a constant internet connection to function.

3.1.2 Architecture of the API

Figure 8 depicts a possible and general setup where the element at the center is the server (Block 3), and all the elements connected to it are the users’ devices (Block 2). These devices may be designed in different ways with different features

Fig. 8 General network architecture



and specifications as per different requirements. The API was developed to serve two different kinds of requests and was programmed using a Python™ framework named Django. The first type of request, namely “Request Type 1”, was based on the Type 1 model; it was designed for fast and immediate scanning/detection with minimalistic delay, as all processing required for scanning/detection is done locally, hence not requiring a constant internet connection.

The second type of request, namely “Request Type 2”, was based on the Type 2 model; it was designed for power-efficient products which are expected/required to have very long battery life, or to be very small and compact in physical size. In this model, the processing for scanning/detection is done on the server, after which the results are received at the user’s device (Block 2) and produced in any desired form, maybe a visual display, an audio message, or as vibrations to alert the user in case of possible problems.

In Fig. 9, five sample models have been shown; the top layer elements denote the server (Block 3), the middle layer elements are the users’ devices (Block 2), and the lower layer elements denote the electronic skin device (Block 1). “A” (smartphone) and “B” (smartwatch) may be examples of Type 1 model, where the devices have enough computing power to compute the scanning/detection processes. Request Type 1 could be implemented for such devices. “C” (fitness band) and “D” (activity tracking ring) may be examples of Type 2 model, where the devices’ endurance is of priority or if the device has physical size limitations. Request Type 2 could be implemented for such devices. “E” (smartphone and fitness band) is a hybrid between Type 1 and Type 2 models, where the data extracted are collected by the fitness device and then passed on to the smartphone where some or all of the processing required for scanning/detection may take place. The features and computations may be distributed so as to achieve desired performance and efficiency parameters.

The API was hosted on a DigitalOcean® server, based on Ubuntu 16.04.4 ×64 operating system. The Virtual Machine was allocated 1 GB of RAM and 25 GB of solid state storage.

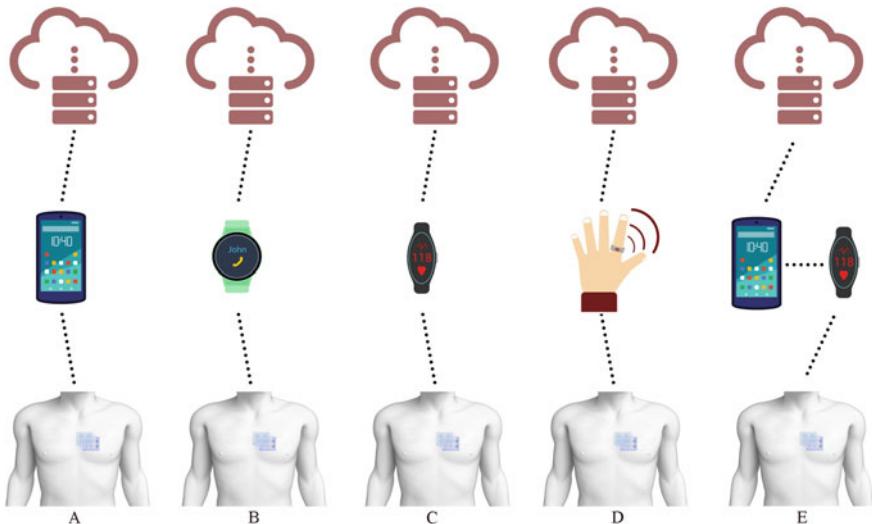


Fig. 9 Sample models

3.2 Data

Data were taken from the UCI machine learning repository [19]. This database consists of 452 ECG records having 16 different classes of heart conditions. Each record has 279 extracted features. These extracted features consist of amplitude-wise features like amplitude of P wave and time-wise features like the average width of P wave. Out of 452 ECG records representing 16 classes of outputs, some of the diseases had a very low number of sample records. If these are to be included in the analysis, there would be a reduction in the accuracy as the neural network cannot be trained well for these specific diseases (due to lack of sample data). A total of 8 diseases spanning over 23 records were eliminated leaving 429 records representing 8 classes of outputs.

3.2.1 Output Classes

Following are the 8 output classes:

- Normal
- Ischemic changes
- Old anterior myocardial infarction
- Old inferior myocardial infarction
- Sinus tachycardia
- Sinus bradycardia
- Right bundle branch block
- Others.

3.2.2 Features

There were 279 features out of which two features, height and weight, were combined and replaced by a single feature, namely body mass index or BMI (BMI equals weight in kilograms divided by the square of height in meters); it has a close relationship with heart diseases. After the combination, the remaining 278 features were used.

3.3 Training

70% of the records were used for training the neural network. The remaining 30% of the records were used to test the neural network and to check whether the system is able to correctly generalize the new data. Backpropagation algorithm [20] was used for training the neural network. First, all weights of the network are randomly initialized. Then, node values are calculated using forward propagation. After that, the error values are calculated by propagating backward from the output layer and hence the name backpropagation. Detailed steps of training neural network using backpropagation algorithm are as follows:

1. Randomly initialize the weights of the neural network.
2. Perform steps 3–5 for each training example.
3. Compute node values by forward propagation using sigmoid activation function.

$$a(l) = 1/(1 + e^{\theta(l-1)^T * a(l-1)}) \quad (1)$$

where $a(l)$ represents the node values of layer l and $\theta(l-1)$ is the weights of the network connecting layer $l-1$ and layer l .

4. Error of the last layer is calculated by subtracting the true value from the obtained value.
5. Errors of nodes in the hidden layers are computed by multiplying the error values of the next layer with the parameters of the current layer followed by multiplication with the derivative of the activation function (a).

$$\delta(l) = ((\theta(l))^T * \delta(l+1)) * a(l) * (1 - a(l)) \quad (2)$$

where $\delta(l)$ represents errors corresponding to nodes in layer l

6. While performing these steps, accumulate the values obtained by multiplying node values by next layer error value.

$$\Delta(l) = \Delta(l) + \delta(l+1) * (a(l))^T \quad (3)$$

where $\Delta(l)$ represents partial derivative of cost function with respect to parameters of layer l of the neural network without considering regularization

7. Add regularization term to the non-bias unit to prevent overfitting.

$$D_{i,j}^{(l)} = (\Delta_{i,j}^{(l)} + \lambda \times j \times \Theta_{i,j}^{(l)}) \div m \quad (4)$$

where $D_{i,j}^{(l)}$ represents partial derivative of cost function with respect to $\Theta_{i,j}^{(l)}$ and considering regularization

Thus, we have obtained partial derivatives of the cost function with respect to all theta values.

8. Optimization techniques like gradient descent will use these partial derivative values and cost function values to find the optimal theta values.

3.4 Architecture of Neural Network

The neural network has three layers, an input layer, a hidden layer, and an output layer; it is depicted in Fig. 10. The number of nodes in input layer must be equal to the number of reduced feature set size in one training example which is equal to 180. The number of nodes in output layer must be equal to the number of output classes, and in this case, it is equal to 8.

The number of nodes in hidden layer can be a variable. This value can be changed and from the results obtained, one may fix this number to an optimum value. In this case, 120 nodes in the hidden layer is a near-optimum choice in terms of accuracy and number of computations required to train the Neural Network.

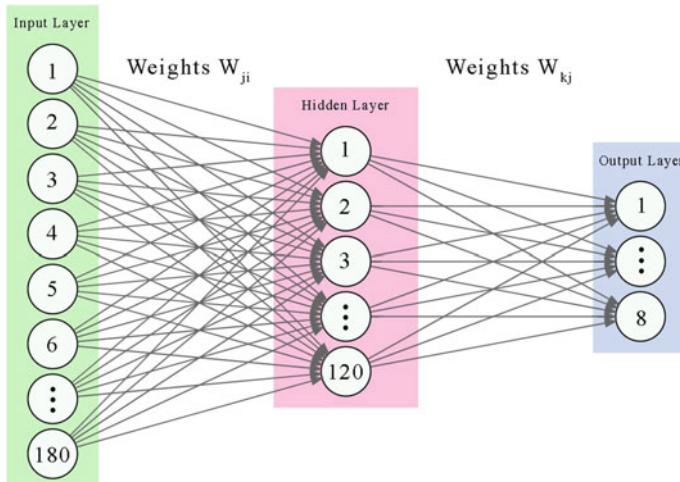


Fig. 10 Architecture of the neural network

4 Results

4.1 Effect of Varying the Number of Hidden Layers

The number of hidden layers is taken as 1 for the first time and 2 for a second time by keeping other variables constant, i.e., hidden layer size is fixed as 120 and the gradient descent will run for 1000 iterations.

As shown in Table 1, neural network with one hidden layer yields better results than the neural network with two hidden layers. This is due to the fact that the neural network with two hidden layers suffers from overfitting which reduces the accuracy. Increasing the number of hidden layers will further increase overfitting and subsequently further decrease the accuracy.

4.2 Effect of Varying the Number of Nodes in the Hidden Layer

By fixing the number of hidden layers as 1 due to better results obtained and also fixing the number of iterations as 1000, the number of nodes in the hidden layer very varied from 20 to 140; the accuracy of the neural network and computational complexity required to train the neural network were observed.

From Fig. 11, it can be observed that the accuracy increases with increase in the number of nodes in the hidden layer, but this will saturate after a certain value.

From Fig. 12, it can be observed that the computational complexity also increases with increase in hidden layer size. So, there is a trade-off between accuracy and computational complexity. Select the optimum value after which accuracy will not increase notably but there still is an increase in the number of computations. This point is 120 in this case. So, consider 120 as optimum hidden layer size.

Table 1 1 hidden layer versus 2 hidden layers

Parameter	1 Hidden layer	2 Hidden layers
Training set accuracy (%)	98.44	75.08
Test set accuracy (%)	78.99	67.40
Number of computations	4.1×10^{10}	6.39×10^{10}

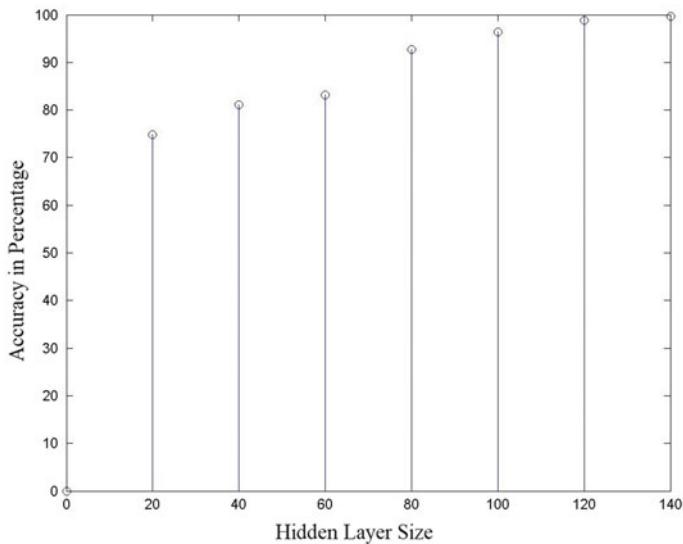


Fig. 11 Result 1 (accuracy vs hidden layer size)

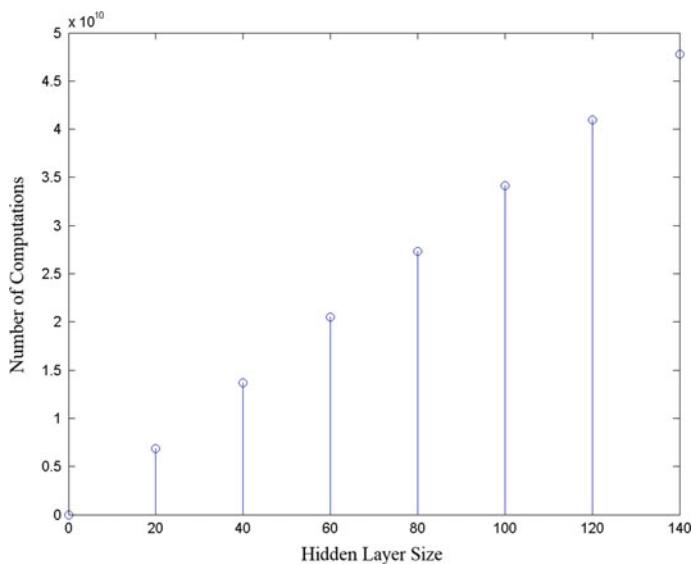


Fig. 12 Result 2 (number of computations vs hidden layer size)

4.3 Early Stopping

Early stopping [21] is used in machine learning techniques to avoid overfitting. If the number of iterations is increasing, training set accuracy will also increase. However, the test set accuracy increases until a certain point and then starts decreasing due to overfitting. So, there is a need to choose this early stopping point. Figure 13 shows a general early stopping scenario; here, Epoch denotes the number of iterations.

Figure 14 shows that the number of iterations should be 1000 to get the best train and test set accuracy without overfitting. This value is dependent on the dataset used.

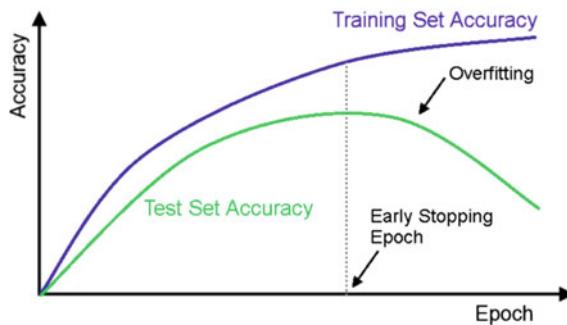


Fig. 13 Early stopping

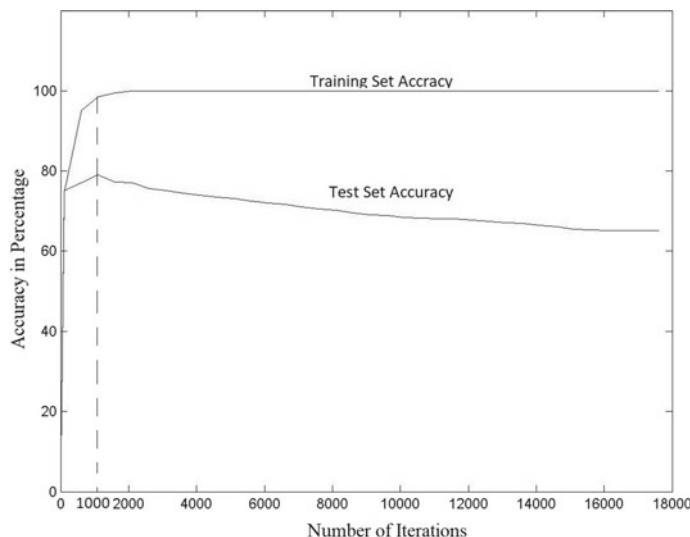


Fig. 14 Result 3 (early stopping)

4.4 Varying both Number of Iterations and Hidden Layer Size

Accuracy increases with increase in both the number of iterations and number of nodes in the hidden layer as shown in Fig. 15. So, accuracy is maximum at that point where both are maximum. Note that both the parameters will saturate upon increasing.

4.5 Dimensionality Reduction Using PCA

To reduce the number of computations, PCA was applied on the data reducing the number of dimensions [22] from 278 to a lesser number without much effect on accuracy. Steps taken:

1. Calculate covariance matrix
2. Calculate eigenvectors and eigenvalues
3. Sort data using eigenvalues
4. Consider first “ k ” dimensions, eliminate the remaining.

The above steps were performed for different values of k (number of dimensions) varying from 10 to 210 with increments of 20 and their corresponding accuracies were found. Plots were drawn between accuracy and number of dimensions. It is observed from Figs. 16 and 17, that there is no considerable variation in accuracy, but computations were varying by a significantly large factor.

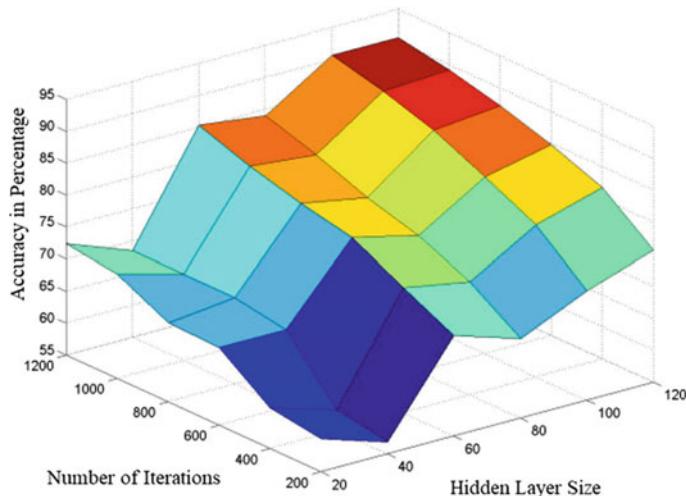


Fig. 15 Result 4 (accuracy vs number of iterations and hidden layer size)

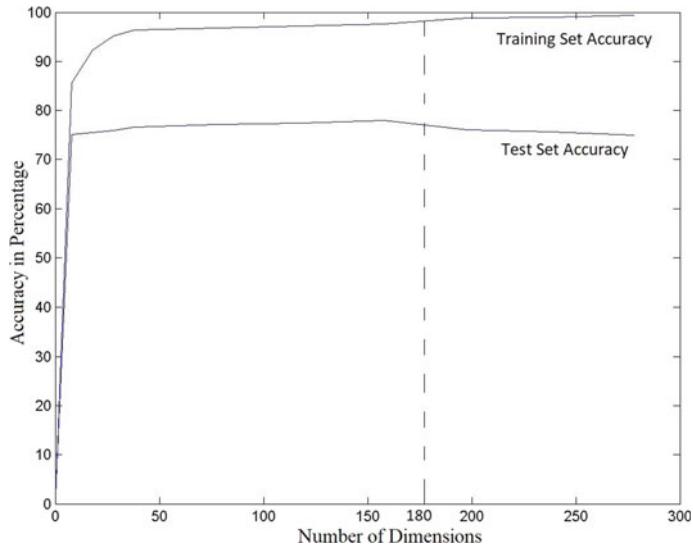


Fig. 16 Result 5 (accuracy vs number of dimensions)

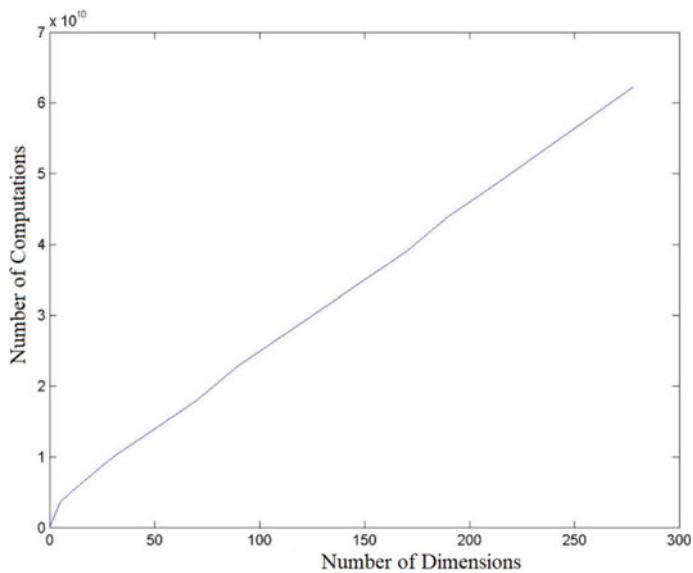


Fig. 17 Result 6 (computations vs number of dimensions)

Table 2 180 Dimensions versus 278 dimensions

Dimensions	180	278
Number of computations	4.1165×10^{10}	6.2186×10^{10}
Test set accuracy (%)	78.99	74.91
Train set accuracy (%)	98.44	99.02

From Table 2, the decrease in the number of computations after reducing the number of features from 278 to 180 can be observed. The application of PCA has resulted in a considerable decrease in the computational complexity.

Computations reduced by 2.1021×10^{10} ; percentage improvement in the number of computations is 33.8%, i.e., for every 100 computations previously, now only 66 computations are occurring.

4.6 Learning Curves

Learning curves [23] depict improvements in performance or error percentage on the vertical axis and training set size or iterations on the horizontal axis. Learning curves were plotted by taking error on the y-axis (vertical) and training dataset size on the x-axis (horizontal). With very little training data, any model can fit that small data. But this model cannot generalize well for the new testing data. This results in small training set error and high test set error.

With an increase in the training data size, the model cannot fit a large amount of data perfectly, however it can predict the test set with better accuracy. So, the training set error increases and the test set error decreases.

If the model has a high bias (as depicted in Fig. 18), the training set and the test set error will be close to each other and both will be values higher than desired. If the model has high variance (as depicted in Fig. 19), there will be a large gap between training and test set errors. To eliminate this, regularization needs to be added. Figure 20 shows the fitting of curves according to the bias and variance.

Fig. 18 High bias



Fig. 19 High variance

4.6.1 High Bias

As shown in Fig. 20, neural network suffering from high bias will not fit the training data well. This occurs due to several reasons such as: having fewer number of features in the model, having fewer number of layers in the model, or using a very high value for regularization parameter.

4.6.2 High Variance

This problem is also called “Overfitting”. As the name suggests, the model will fit extremely accurately to the training data. This occurs due to reasons such as: having more number of layers in the model, having more number of features in the model, or having less amount of data for training.

From Fig. 21, it can be concluded that the implemented model suffers from high variance or overfitting. To avoid this problem, regularization parameter (λ) [24] is introduced in the cost function.

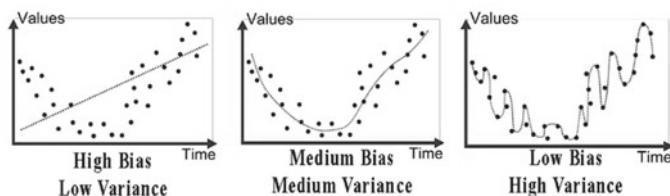


Fig. 20 Underfit, correct fit, and overfit

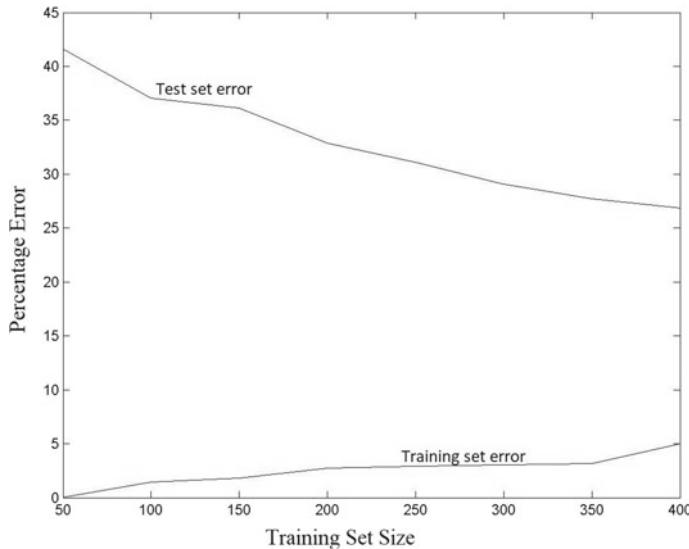


Fig. 21 Result 7 (percentage error vs training set size)

$$\begin{aligned}
 J(\theta) = & -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log(h_\theta(x^{(i)})) \\
 & + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))] + \frac{\lambda}{2m} \sum_{j=1}^n \theta_j^2
 \end{aligned} \quad (5)$$

Even if the model is suffering from overfitting due to some features, the weights of those features are penalized by adding the sum of squares of weights, multiplied by λ to the cost function (Eq. 5 [24]). So, the contribution of these features will be less and the problem of overfitting is relatively decreased.

Several values of λ are tested, among which λ equal to 4 shows the best results in terms of training set accuracy and test set accuracy. The test set accuracy without the regularization parameter was observed to be 76%, while with the regularization parameter was observed to be 78.99%.

4.7 Mean Square Error Versus Number of Iterations

Mean square error (cost function value) is plotted against the number of iterations in Fig. 22. A decrease in error with an increase in the number of iterations is an indication for the correct working of the neural network.

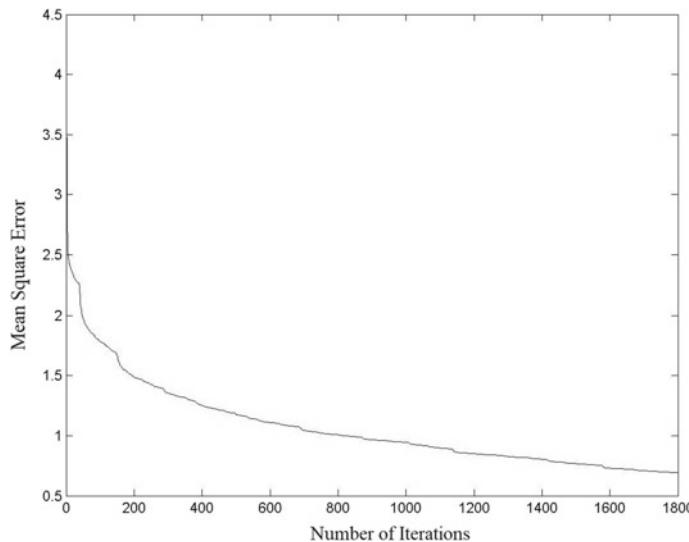


Fig. 22 Result 8 (mean square error vs number of iterations)

4.8 Generation of Confusion Matrix

A confusion matrix [25] is used to understand the neural network's performance. It is a square matrix with the number of rows equal to the number of output classes.

4.8.1 Data

Table 3 consists of the eight output classes, and the respective number of records present in the database.

4.8.2 Confusion Matrix

Table 4 shows the obtained confusion matrix. Here, rows describe the prediction results of the neural network; columns signify the actual output classes.

4.8.3 Class-Wise Accuracy

Table 5 shows the class-wise accuracies obtained for the neural network. Since Class 1 has the maximum available datasets, it shows the highest accuracy of 96.73%, supporting the general trend of obtaining higher accuracy with larger datasets. Accuracy over the entire dataset is found to be 93.01%.

Table 3 Class types

Class no.	Class type	No. of records
1	Normal	245
2	Coronary artery disease	44
3	Old anterior myocardial infarction	15
4	Old inferior myocardial infarction	15
5	Sinus tachycardia	13
6	Sinus bradycardia	25
7	Right bundle branch block	50
8	Others	22

Table 4 Confusion matrix

	1	2	3	4	5	6	7	8
1	237	4	0	0	0	3	1	0
2	6	37	0	0	0	1	0	0
3	0	1	14	0	0	0	0	0
4	1	0	0	14	0	0	0	0
5	2	0	0	0	11	0	0	0
6	2	0	0	0	0	23	0	0
7	4	0	1	0	0	0	45	0
8	2	1	0	0	0	0	1	18

Table 5 Class-wise accuracy

Class	Accuracy (%)
1	96.73
2	84.09
3	93.33
4	93.33
5	84.62
6	92.00
7	90.00
8	81.82

4.9 Timing Analysis of the API

Timing analysis was done for the two types of requests using “cURL” command [26] which is a command prompt tool that is used for data transfer to the server or from the server using commonly used protocols like HTTP, HTTPS, etc.

4.9.1 Timing Analysis of Request Type 1

Command was run 10 times and the averages were taken; the observed results are shown in Table 6.

4.9.2 Timing Analysis of Request Type 2

Similar to the first case, the command was run 10 times and the averages were taken. Several parameters of the API’s performance are shown in Table 7.

To this, the time consumed for computing the test processes must be added. Using a Windows® 10-based PC with an Intel® Core™ i5 4210U processor, 8GiB of RAM and a dedicated Nvidia® GeForce® GT840M graphics card, it took an average of

Table 6 Timing analysis of request type 1

Size downloaded	295.3920 kB
Size of the header	304 bytes
Size of the request	124 bytes
Download speed	183.4219 kB/s
Domain name lookup time	1 us
TCP connection time	7.7060 ms
Total time	1.6719 s

Table 7 Timing analysis of request type 2

Size downloaded	30 bytes
Size of the header	300 bytes
Size of the request	1.0130 kB
Size uploaded	817 bytes
Download speed lookup time	71.7 B/s
Upload speed	1.9671 kB/s
Domain name lookup time	1 us
TCP connection time	1.6009 ms
Total time	0.4266 s

0.44 seconds for running test loops using Python™. The performance metrics and benchmarks for a present-day mobile SoC, are either comparable or better, especially when it comes to AI computations as some newer SoCs have integrated neural processing unit (NPU) alongside integrated graphics processing unit (GPU), specifically for machine learning applications. So for a mobile device, the time consumed for running test loops will be similar. Further, we have to add to this the time consumed for the Bluetooth™ communication, which will vary according to the distance and version of Bluetooth™ used. To further improve the timing performance, MQTT could be used. Overall time consumed will still be significantly low enough to give the user proper treatment in time, hence reducing the chances of serious consequences.

5 Conclusion

Over the last few years, research on mobile cardiac telemetry in ambulatory cardiac monitoring [27] has proven that such a system may be revolutionary as it is vastly better when compared to existing systems in terms of flexibility and features. Future work could explore the possibilities of integrating music recognition techniques like those used by Shazam®, SoundHound Inc., etc., [28–32] where both time and frequency domains [33] are used to create filtered spectrograms and acoustic fingerprints [34, 35] are generated in order to scan for patterns or inconsistencies in patterns in a compute-efficient manner. Considering the severity of heart diseases as observed by the statistics provided by WHO, a highly responsive and low-latency real-time wearable device that monitors the ECG continuously and alerts the patient and/or the family members or concerned doctors instantly in case of emergencies would hugely decrease the possibilities of fatality, as the required drug(s) could be delivered (or actions could be taken) within a small time frame. A highly flexible system like this could enable manufacturers to address the different requirements of the customer. It would also enable researchers to adopt much more accurate algorithms within the same type of system. Furthermore, a system where specific computing blocks may be flexibly moved to different regions in the network would mean the potential to design high-performance, low-latency, and power-efficient systems without compromising any of them. This could be applied to various other fields, such as autonomous vehicle systems and on-premises security systems among others.

References

1. WHO (2017) Cardiovascular diseases—key facts. <http://www.who.int/mediacentre/factsheets/fs317/en/>
2. Macon BL, Yu W, Reed-Guy L, Acute myocardial infarction. <https://www.healthline.com/health/acute-myocardial-infarction>

3. Marker K (2016) Preventing heart failure after acute myocardial infarction. <https://www.labroots.com/trending/cardiology/2335/preventing-heart-failure-after-acute-myocardial-infarction>
4. van Helvete H (2014) Schematic representation of normal ecg. <https://en.wikipedia.org/wiki/Electrocardiography#Theory>
5. Houghton A, Gray D (2012) Making sense of the ECG, 3rd edn. Hodder Education, p 214
6. Zhang EL (2016) Cardiovascular medical device series: introduction and technical trends in the cardiac monitoring industry. <http://ochis.org/CMreport>
7. Haque A (2014) Cardiac dysrhythmia detection with gpu-accelerated neural networks. Stanford University, Computer Science Department
8. Khatun S, Morshed BI (2017) Detection of myocardial infarction and arrhythmia from single-lead ecg data using bagging trees classifier. In: Electro information technology
9. Wale AS, Sonawani PSS, Karande PSC (2017) Ecg signal analysis and prediction of heart attack with the help of optimized neural network using genetic algorithm
10. Guvenir HA, Acar B, Demiroz G, Cekin A (1997) A supervised machine learning algorithm for arrhythmia analysis. Computers in cardiology, pp 433–436
11. Rodríguez J, Goñi A, Illarramendi A (2005) Real-time classification of ECGS on a PDA. In: Information technology in biomedicine, vol 9, pp 23–24
12. Gradi S, Kugler P, Lohmuller C, Eskofier B (2012) Real-time ECG monitoring and arrhythmia detection using android based mobile devices. In: Engineering in medicine and biology society, pp 2452–2455
13. Yang Y, Huang X, Yu X (2007) Real-time ECG monitoring system based on FPGA. In: The 33rd annual conference of the IEEE industrial electronics society (IECON). pp 2136–2140
14. Naeemabadi MR, Chomachar NA, Khalilzadeh MA, Dehnavi AM (2011) Portable device for real-time ECG recording and telemetry. In: Electrical and computer engineering
15. Cibis T, Groh BH, Gatermann H, Leutheuser H, Eskofier BM (2015) Wearable real-time ECG monitoring with emergency alert system for scuba diving. In: IEEE engineering in medicine and biology society (EMBC), pp 6074–6077
16. Yokota T, Zalar P, Kaltenbrunner M, Jinno H, Matsuhisa N, Kitanosako H, Tachibana Y, Yukita W, Koizumi M, Someya T (2016) Ultraflexible organic photonic skin. *Sci Adv* 2(4):e1501856
17. Jeong Y, Kim J, Xie Z, Xue Y, Won S, Lee G, Jin S, Hong S, Feng X, Huang Y, Rogers J, Ha J (2017) A skin-attachable, stretchable integrated system based on liquid GaInSn for wireless human motion monitoring with multi-site sensing capabilities. *NPG Asia Mater* 9:e443–e443
18. Poliks M et al (2016) A wearable flexible hybrid electronics ECG monitor. In: IEEE 66th electronic components and technology conference, pp 1623–1631
19. Dua D, Efi KT (2017) UCI machine learning repository. <http://archive.ics.uci.edu/ml>
20. Ng A, Backpropagation. <https://www.coursera.org/learn/machine-learning/supplement/pjdBA/backpropagation-algorithm>
21. Deeplearning4j, Early stopping. <https://deeplearning4j.org/earlystopping>
22. Chawla M, Verma H, Kumar V (2006) ECG modeling and QRS detection using principal component analysis, pp 1–4
23. Ng A, Learning curves. <https://www.coursera.org/learn/machine-learning/supplement/79woL/learning-curves>
24. Ng A, Cost function. <https://www.coursera.org/learn/machine-learning/supplement/1tJlY/cost-function>
25. Raju NR, Rao VM, Jagadess BN (2017) Identification and classification of cardiac arrhythmia using neural network. *Helix* 7(5):2041–2046
26. Command line tool and library for transferring data with urls. <https://curl.haxx.se/>
27. Real time ECG monitoring. <http://cardiacmonitoring.com/mobile-cardiac-telemetry/real-time-ecg-monitoring/>
28. Mawata C (2015) How does shazam work. <http://coding-geek.com/how-shazam-works/>
29. Wang A (2003) An industrial strength audio search algorithm
30. Harvey F (2003) Name that tune. <https://www.scientificamerican.com/article/name-that-tune/>

31. Cooper T (2018) How shazam works. <https://medium.com/@treycoopermusic/how-shazam-works-d97135fb4582>
32. Wang S (2016) How does shazam work?. <https://www.acrcloud.com/blog/how-does-shazam-work>
33. Tóth L (2014) Combining time- and frequency-domain convolution in convolutional neural network-based phone recognition, pp 190–194
34. Schalkwijk J (2018) A fingerprint for audio. <https://medium.com/intrasonics/a-fingerprint-for-audio-3b337551a671>
35. Cano P, Batlle E, Gómez E, Gomes L, Bonnet M (2005) Audio fingerprinting: concepts and applications, vol 2, pp 233–245

Gender Classification Based on Fingerprint Database Using Association Rule Mining



Ashish Mishra, Shivendu Dubey, and Amit Sahu

Abstract Detection of gender plays an essential part in forensic and medico authorized examinations. Fingerprints are primarily accurate as well as an authentic way for individual and gender classification. Fingerprint identification system design may be used for authentication of right person in real time; however, RFID-based authentication is not reliable because it may be used by anyone. As ladies have a tendency to have a fundamentally higher edge thickness (scaled-down focuses) than men, it may be separated just when the unique mark is in the model frame (great) not in inert shape (not great). Confinement of accessible work is to discover amend individual when inert prints (obscure unique mark) typically accessible technique is has less acknowledgment rate and less edge thickness, for idle fingerprints likewise a time for acknowledgment is additionally high henceforth it is likewise require to decrease the time for acknowledgment and enhance edge thickness.

Keywords Left of center · Right of center · Likelihood ratio · Probability density function · Association rule mining

A. Mishra (✉) · S. Dubey · A. Sahu
Gyan Ganga Institute of Technology and Sciences, Jabalpur, India
e-mail: ashish.mish2009@gmail.com

S. Dubey
e-mail: shivendudubey@gkits.org

A. Sahu
e-mail: amitsahu@gkits.org

1 Introduction

There is no two fingers are establish to include the same feature and it is an overwhelming arithmetical possibility with the aim of no two will ever be establish to match [1]. This is the probability of two people having the same finger impression are regarding individually in sixty-four thousand million of the earth population. The same twins, initiate from one fertilized egg, are possibly as mainly identical of any beings on earth [2]. It is allocate the same DNA profile for the reason that they began their existence as one entity, yet their fingerprints are as unique as any unrelated individual [3]. The increasingly frequency of crime has through fingerprinting an essential tool in the hands of examine officers. If the gender of the person could be established with certainty, the load of the investigating officer would be reduced by half [4]. Fingerprint recognition system design may be apply for authentication of right person in real time situation, but at some time authentication is a not reliable because it may be used by anyone. As ladies have a tendency to have a fundamentally higher edge thickness than men; however, it may be separated just when unique mark is in model frame not in inert shape [5]. Thumb impressions are of distinct importance. They are even used in lieu of signature in India in many important documents including property documents, competitive examinations, etc.

2 Background

The process of extrication of internal knowledge, data relationship among various images and various patterns that don't seem to be expressly keep in images and lots of concepts from computer vision are used, image process and retrieval, mining information, knowledgeable system learning, databases and AI are known as Image mining. Image mining method comprised of following steps as shown in Fig. 1:

- Pre-processing
- Transformations and Feature extraction
- Mining significant patterns and features
- Evaluation
- Interpretation and finally extracting knowledge.

Various techniques that are used and how they are implemented in image mining shows in Fig. 2.

The association rule mining applies an effective implement on behalf of pattern identification in knowledge discovery and data mining. Its most important purpose is to extricate useful information from large datasets. The paper is structured seeing that. Section two presents an explanation of the fingerprint recognition method, process well the foremost necessary steps. In Section three, the related work is presented, showing its many fingerprint recognition approaches. Section four explains the proposed that human fingerprint Acquisition for database and human fingerprint

Fig. 1 The image mining process

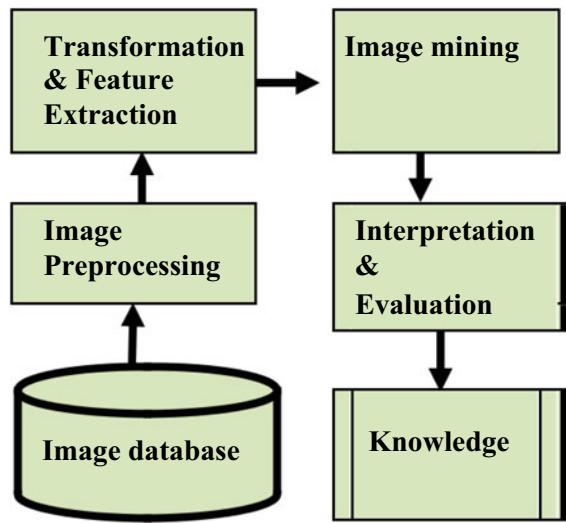
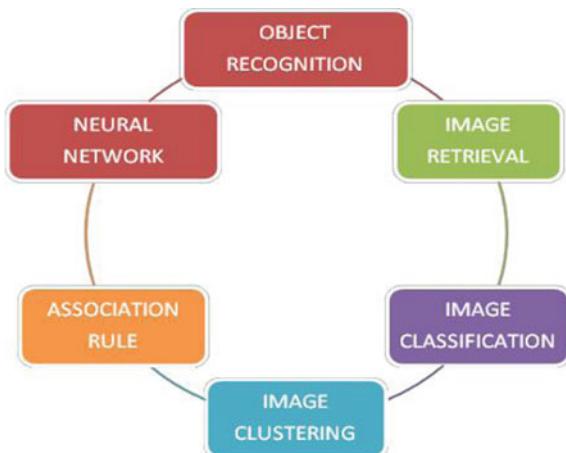


Fig. 2 Image mining techniques



to be recognized. Section five Study the outcome find nearby a discussion of them. Lastly, Section six summarizes the paper.

3 Related Work

Nithin et al. [6] applied 200 subjects (100 men and 100 women) in the age group of 18–30 years. Ridge densities on the right and left-hand thumbprint were determined using a newly designed layout and analyzed statistically. The experiments showed

that women tend to have a higher thumbprint ridge density in both the areas examined, individually and combined. Soanboon et al. [7]: applied the ridge thickness in fingerprints beside with explain that men have coarser finger ridges than women which imply that men will contain less ridges in a specified region than women and thus a lower ridge density. The higher fingerprint ridge density in women is attributed to the fact that women tend to have finer epidermal ridges than men. Men generally have coarser ridges than women and the difference is around 10%.

Mishra et al. [8]: In this paper that the difference between the finger ridge density in men and women in a specified area may be attributed to the fact that on an average body proportions of men are larger than women and thus the same numbers of ridges are accommodated among the men in a larger surface area and thus, a lower density is observed among men.

Tarare et al. [9]: applied, K nearest neighbor classifier is apply as a classifier which uses Euclidean Distance determine for classification as well as classifies testing fingerprint as male or women fingerprint. And describes the overall process of above scheme. DWT transform will give the features of some of the fingerprint images of the dataset (training images) to create a database of features which will be used as a lookup table for classification of unknown fingerprint and other fingerprints (testing fingerprints) will be used for testing. KNN classifier will assign one of two groups to test fingerprint.

Lumini et al. [10] describe several systems and architectures related to the combination of biometric systems, both unimodal and multimodal, classifying them according to a given taxonomy. Moreover, we deal with the problem of biometric system evaluation, discussing both performance indicators and existing benchmarks.

Shinde et al. [11] describe an overall comparison of frequency domain techniques, mainly focusing on DWT and its combinations are presented. And also uses canny edge detector and Haar DWT based fingerprint gender classification technique.

Kapoor et al. [12] determine any significant difference in thumbprint edge thickness of guys and women in a focal Indian populace to empower assurance of sexual orientation, an investigation was led on 200 subjects (100 guys and 100 women) in age gathering of 18–30 years.

Mishra et al. [13] applied Gender arrangement utilizing affiliation administer mining and grouping approach. Fingerprint recognition for gender arrangement approach is done through different systems similar to support vector machines, neural network, fuzzy C means.

4 Thumbprint Ridge Density Calculation

Two straight lines bisecting each other were drawn. This intersect point be located at the center or center of the print. 5 mm above this, another transverse line was drawn. Two squares of 25 mm 2 each were drawn on both sides (left and right). These were our chosen areas for analysis. Ridge counting was performed in these designated areas and the values were tabulated. At the time of counting the number of ridges,

this transparency was superimposed on the print, so that the lower intersection lies on the core of the print, in cases of Whorls and Loops. In Arches, the intersection was kept on the lowest ridge which flows continuously from one side to the other side of the print. The epidermal ridges from one corner of the square to the diagonally contrary corner were counted. Point was not calculated, forks were calculated as two ridges excluding the handle, and a lake was calculated as two ridges.

Figure 3 shown proposed work block diagram of design it may be observed in the diagram that human fingerprint Acquisition for database and human fingerprint to be recognized is shown, after human fingerprint and CASIA database [9].

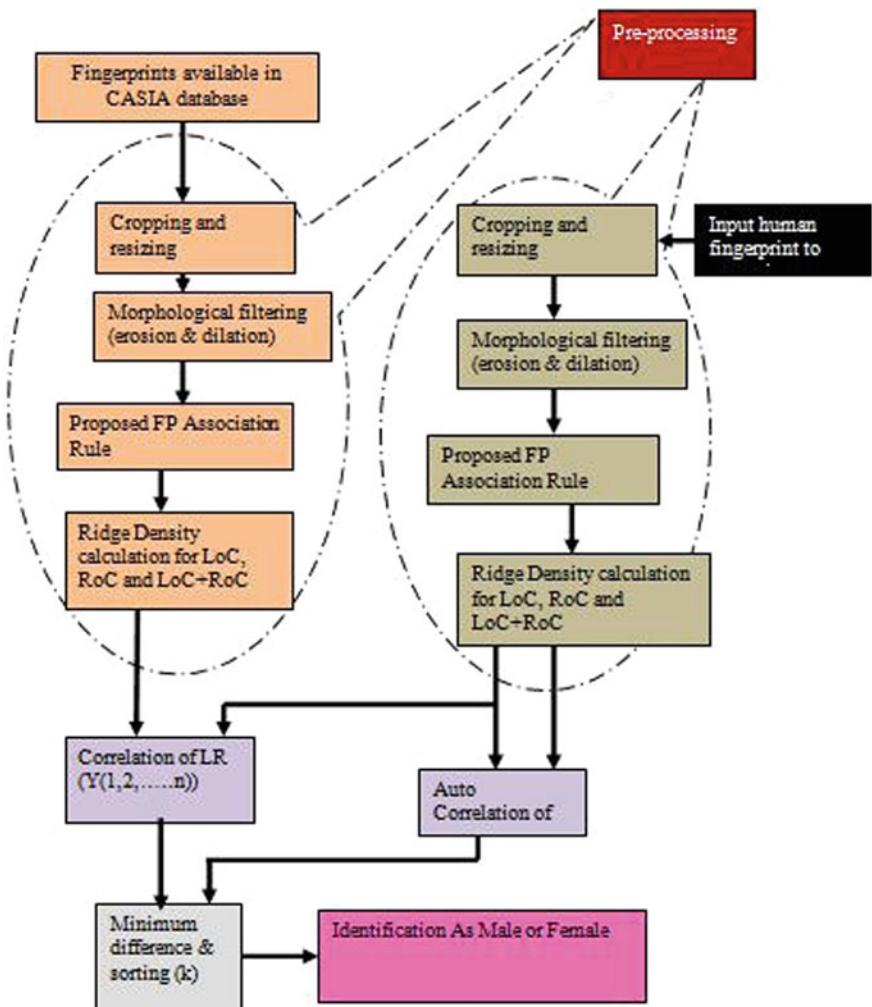


Fig. 3 Block diagram of proposed work

Pre-processing includes Cropping of the central part of human fingerprint then a Morphological filtering which performs dilation than erosion this procedure significantly enhance the quality of human fingerprint especially the quality of a latent fingerprint.

After Pre-processing Correlation based identification procedure gives ‘K’ position in CASIA database.

5 Experimental Work

In the experimental work shown that statistics of ridge densities in men and women is shown in Table 1. In this work consider women, the edge thickness range from 12 to 19 ridges per 25 mm^2 at the LoC with the mean ridge density of 14.6 and 12–18 ridges per 25 mm^2 at the RoC with the mean ridge density of 14.56.

The range of LoC and RoC combined is experiential to be 19–27 ridges with 23.40 as the mean and 24–36 ridges with the mean value of 29.16 in men and women, respectively. Women were originate to have a significantly higher ridge density than men at LoC, RoC and Combined Applying the t-test, the differences in the ridge densities of men and women at LoC, RoC and Combined were found to be statistically significant at $p < 0.01$ levels (Table 2). In men, the ridge density ranged from 9 to 15 ridges per 25 mm^2 at both the Left of Center (LoC) and the Right of Center (RoC) with the mean ridge density of 11.58 and 11.82, respectively. NoS is number of samples.

Table 2 shows that the frequency distribution of ridge densities at the left and right of center per 25 mm^2 in men and women. It is experimental that none of the men have a mean ridge density of more than 15 and there are no women who have mean ridge densities below 12.

Table 3 shows the frequency distribution of mean ridge densities of LoC and RoC combined. It is experiential that none of the men have a mean ridge density of more than 27 and there are no women who have mean ridge densities below 23. Women have a considerably greater combined ridge density than men (Fig. 4).

Table 1 The thumb ridge density in both men and women

Parameter	Male			Women		
	(LoC)	(RoC)	LoC + RoC	(LoC)	(RoC)	LoC + RoC
Mean ridge density	11.58	11.82	23.40	14.6	14.56	29.16
Minimum ridges	9	9	19	12	12	24
Maximum ridges	15	15	27	19	18	36
Standard deviation	1.46	1.37	1.995	1.68	1.54	2.57
Standard error	0.1	0.09	0.14	0.11	0.10	0.18
Range	9–15	9–15	19–27	12–19	12–18	24–36

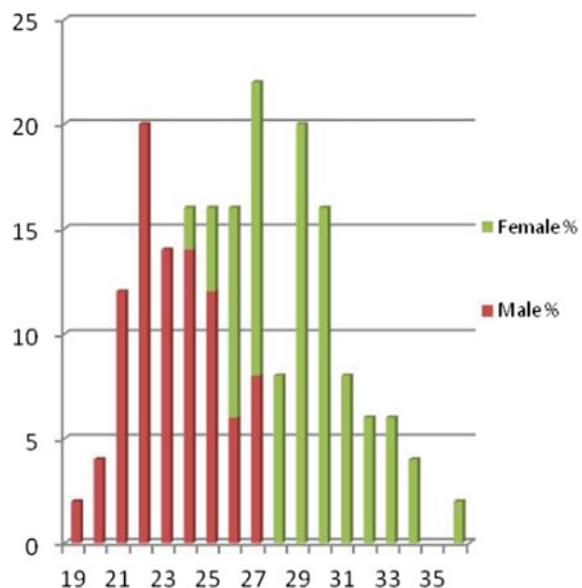
Table 2 Frequency distribution of mean ridge density in male and women thumb prints

Ridge density	Male				Women			
	(LoC)		(LoC)		(LoC)		(LoC)	
	NoS	%	NoS	%	NoS	%	NoS	%
9	12	6	8	4				
10	36	18	32	16				
11	48	24	32	16				
12	64	32	72	36	20	10	16	8
13	20	10	32	16	36	18	40	20
14	8	4	20	10	48	24	52	26
15	12	6	4	2	40	20	28	14
16					28	14	40	20
17					16	8	20	10
18					8	4	4	2
19					4	2		
Total	200	100	200	100	200	100	200	100

Table 3 Frequency distribution in male and women thumbprints

Combined ridge density LoC + RoC	Male		Women	
	NoS	%	NoS	%
19	4	2		
20	8	4		
21	24	12		
22	40	20		
23	28	14		
24	28	14	4	2
25	24	12	8	4
26	12	6	20	10
27	16	8	28	14
28			16	8
29			40	20
30			32	16
31			16	8
32			12	6
33			12	6
34			8	4
35				
36			4	2
Total	200	100	200	100

Fig. 4 Samples based on the combined ridge density



Possibility density for men and women derived from the frequency distribution (at LoC and RoC respectively) were applied to calculate the likelihood ratio and posterior possibility of gender designation for the given ridge count for subjects. At LoC, the statistical analysis of the likelihood ratio and the odds ratio shows that a ridge density of 612 ridges per 25 mm^2 is more likely to be of male origin ($p = 0.90$), whereas a ridge density of P13 ridges per 25 mm^2 is more likely to be of women origin ($p = 0.69$) (Table 4). Posterior probability shows that a fingerprint with a ridge density

Table 4 ratios derived from the observed ridge count at LoC

Ridge density at LoC	Probability density		Likelihood ratio		Favored odd	
	Male (C)	Women (C')	C/C'	C'/C	Male	Women
9	0.06	0.001	60	0.017	0.99	>0.01
10	0.18	0.001	180	0.006	0.99	>0.01
11	0.24	0.001	240	0.004	0.99	>0.01
12	0.32	0.1	3.2	0.313	0.90	>0.10
13	0.1	0.18	0.556	1.8	0.31	<0.69
14	0.04	0.24	0.167	6	0.03	<0.97
15	0.06	0.2	0.3	3.333	0.09	<0.91
16	0.001	0.14	0.007	140	0.01	<0.99
17	0.001	0.08	0.0125	80	0.01	<0.99
18	0.001	0.04	0.025	40	0.01	<0.99
19	0.001	0.02	0.05	20	0.01	<0.99

Table 5 Shown in densities and likelihood ratios derived from the observed ridge count at RoC

Ridge density at RoC	Probability density		Likelihood ratio		Favored odd	
	Male (C)	Women (C')	C/C'	C'/C	Male	Women
9	0.04	0.001	40	0.025	0.99	>0.01
10	0.16	0.001	160	0.006	0.99	>0.01
11	0.16	0.001	160	0.006	0.99	>0.01
12	0.36	0.08	4.5	0.222	0.95	>0.05
13	0.16	0.2	0.8	1.25	0.36	<0.64
14	0.1	0.26	0.385	2.6	0.15	<0.85
15	0.02	0.14	0.143	7	0.02	<0.98
16	0.001	0.2	0.005	200	0.01	<0.99
17	0.001	0.1	0.01	100	0.01	<0.99
18	0.001	0.02	0.05	20	0.01	<0.99

of 610 ridges per 25 mm^2 will have a higher probability of belonging to a male ($p = 0.99$). Also a ridge density of P16 ridges per 25 mm^2 will be more indicative of women ($p = 0.99$).

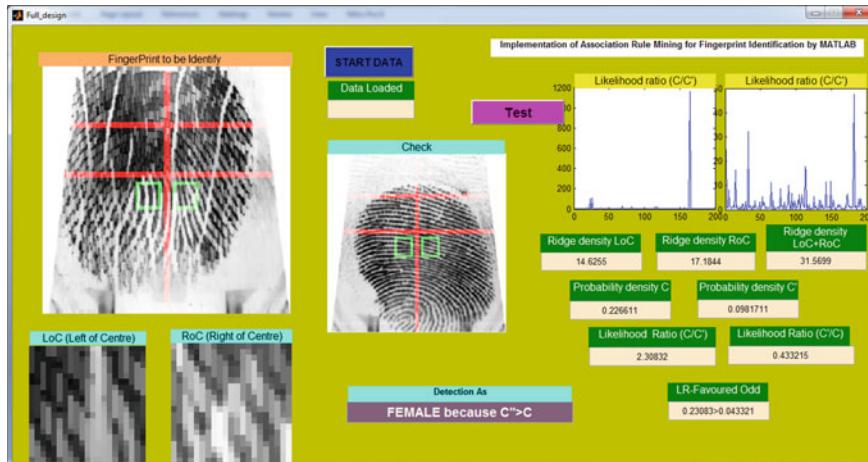
At RoC, the statistical analysis of the likelihood ratio and the odds ratio shows that a ridge density of 612 ridges per 25 mm^2 is more likely to be of male origin ($p = 0.95$), whereas a ridge density of P13 ridges per 25 mm^2 is more likely to be of women origin ($p = 0.64$) (Table 5). Posterior probability shows that a fingerprint with a ridge density of 611 ridges per 25 mm^2 will have a higher probability of belonging to a male ($p = 0.99$). Similarly, a ridge density of P15 ridges per 25 mm^2 will be more indicative of women ($p = 0.98$) (Table 6).

For the Combined ridge density (LoC + RoC), the statistical investigation of the likelihood ratio and the odds ratio shows that a ridge density of 625 ridges per mm^2 is more likely to be of male origin ($p = 0.96$). Statistically significant gender difference are observed in the thumbprint ridge density in the LoC and RoC areas analyzed in this paper. The women have a higher thumbprint ridge density than men in both these areas. Our results are in agreement with the recent studies conducted on fingerprint ridge density (Figs. 5, 6 and 7).

From the Table 7 and Fig. 8 it can be observed that proposed work Time for identification is better and sp overall throughput is also better. Proposed work identification rate is also better than earlier work, and proposed work found better Density difference than earlier work as proposed work average LR is 0.635 and earlier work average LR is 0.62 only.

Table 6 Probability densities and likelihood ratios derived from the observed combined ridge count

Combined ridge density [Left + Right]	Probability density		Likelihood ratio		Favored odd	
	Male (C)	Women (C')	C/C'	C'/C	Male	Women
19	0.02	0.001	20	0.05	0.99	>0.01
20	0.04	0.001	40	0.025	0.99	>0.01
21	0.12	0.001	120	0.008	0.99	>0.01
22	0.2	0.001	200	0.005	0.99	>0.01
23	0.14	0.001	140	0.007	0.99	>0.01
24	0.14	0.02	7	0.143	0.98	<0.02
25	0.2	0.04	5	0.2	0.96	<0.04
26	0.06	0.1	0.6	1.667	0.36	<0.64
27	0.08	0.14	0.5714	1.75	0.33	<0.67
28	0.001	0.08	0.0125	80	0.01	<0.99
29	0.001	0.2	0.005	200	0.01	<0.99
30	0.001	0.16	0.0063	160	0.01	<0.99
31	0.001	0.08	0.0125	80	0.01	<0.99
32	0.001	0.06	0.0167	60	0.01	<0.99
33	0.001	0.06	0.0167	60	0.01	<0.99
34	0.001	0.04	0.025	40	0.01	<0.99
36	0.001	0.02	0.05	20	0.01	<0.99

**Fig. 5** Data base upload for ridge LoC and RoC density extraction and calculation

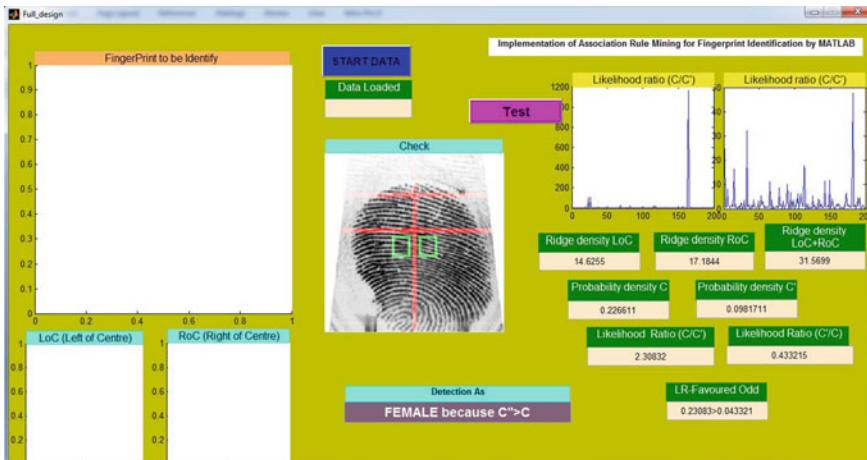


Fig. 6 Fingerprint recognized as women

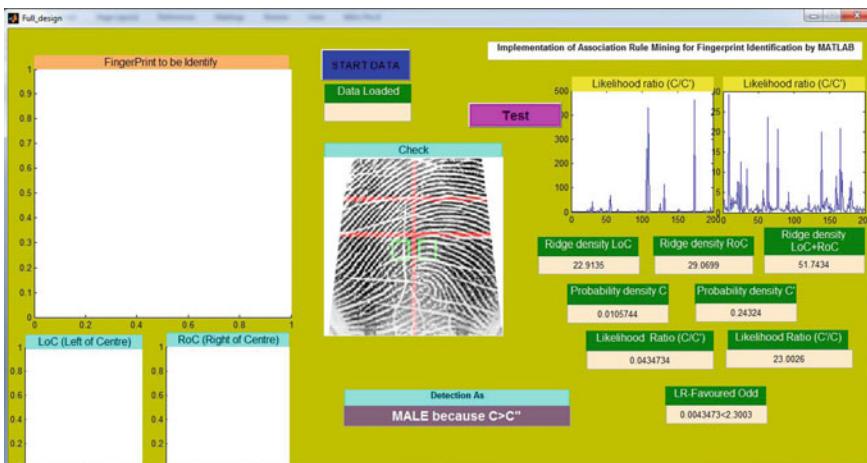


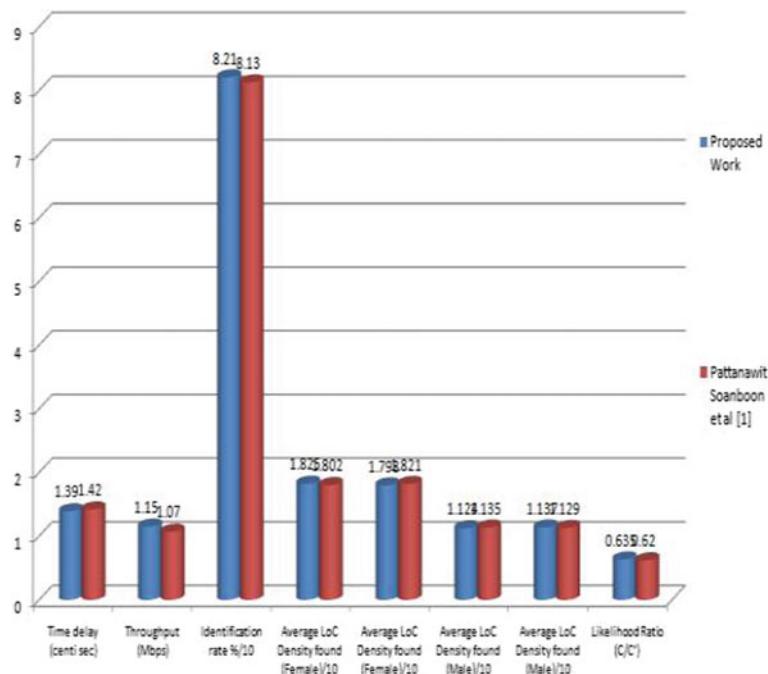
Fig. 7 Fingerprint recognized as Male

6 Conclusions

This paper shows that girls of the MP population of Central India have a considerably higher thumb ridge density than men. The variations among men and women fingerprint ridge density are statistically important. The outcome of this paper are support and would promptly act as a supportive tool for forensic consultants and in law enforcement, 14, 23 as they will be apply as presumptive indicators of the gender of an unknown print left at against the law scene 21 this could be attain simply

Table 7 Comparative results

Parameter	Proposed work	Soanboon et al. [1]
Time delay (s)	139.123	142.58
Throughput (Mbps)	1.15	1.07
Identification rate %	82.1	81.36
Average LoC density found (Women)	18.25	18.02
Average LoC density found (Women)	17.98	18.21
Average LoC density found (Male)	11.24	11.35
Average LoC density found (Male)	11.37	11.29
Likelihood ratio (C/C')	0.635	0.62

**Fig. 8** Identification rate comparisons with proposed method

by qualitatively examining if prints appear to be coarse or fine then quickly quantifying ridge density during a manner analogous to ways correspond to during this paper. The findings may also be helpful in identification of mutilated remains once a dismembered hand is brought for medico-legal examination. Finally this paper overcomes the concentrated limitation 14 wherever all 10 fingerprints were needed for the determination of the gender.

References

1. Cunliffe F, Piazza PB (1980) Criminalistics and scientific investigation. Prentice-Hall, Inc., New Jersey, p 266
2. Modi JP (2002) Modi's medical jurisprudence and toxicology, 22nd edn. Lexis Nexis Butterworths, Noida, pp 37, 39, 40, 72
3. Agnihotri AK, Jowaheer V, Allock A (2012) An analysis of fingerprint ridge density in the Indo-Mauritian population and its application to gender determination. *Med Sci Law* 52(3):143–147
4. Mishra A et al (2017) A survey: gender classification based on fingerprint. *Int J Pure Appl Math (Scopes)* 117(20):985–992. ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
5. Nayak VC et al (2010) Sex differences from fingerprint ridge density in the Indian population. *J Forensic Leg Med* 17:84–86
6. Nithin MD et al (2011) Gender differentiation by finger ridge count among South Indian population. *J Forensic Leg Med* 18:79–81
7. Soanboon P, Nanakorn S, Kutanan W (2016) Determination of sex distinction from unique mark edge thickness in northeastern Thai young people. *Egypt J Forensic Sci* 6:185–193 (ScienceDirect, Elsevier)
8. Mishra A et al (2017) A novel technique for fingerprint classification based on Naive Bayes classifier and support vector machine. 169(7): 0975 – 8887
9. Tarare S et al (2015) Fingerprint based gender classification using DWT transform. In: IEEE international conference on computing communication control and automation, 978-1-4799-6892-3/15
10. Paulino Alessandra A, Feng Jianjiang, Jain Anil K (2015) Latent unique mark matching using descriptor-based correlation. *IEEE Trans Inf Forensics Secur* 8(1):31–45
11. Shinde SR et al (2015) Gender classification with KNN by extraction of Haar Wavelet features from canny shape fingerprints. In: IEEE international conference on information processing Vishwakarma Institute of Technology. 978-1-4673-7758-4/15
12. Kapoor N et al (2016) Sex differences in thumbprint ridge density in a central Indian population. *Egypt J Forensic Sci* 5:23–29
13. Mishra A et al (2015) A review on gender classification using association rule mining & classification based on fingerprints. In: 2015 fifth international conference on communication systems & network technologies. IEEE Computer Society, 978-1-4799-1797-6/2015
14. National Center for Biotechnology Information. <http://www.ncbi.nlm.nih.gov>

Cyber Terrorism-Related Multimedia Detection Using Deep Learning—A Survey



Himani Mandaviya and Snehal Sathwara

Abstract In this era of technological know-how crime has changed from bodily assault to virtual assault and due to increase of Internet applications, i.e., social media has made conversation simpler as nicely as given an open supply for terrorist to layout their undertaking with the assist of social media structures it is easy to terrify and create chaos in society which leads in excessive cyber terrorism ratio from last few years. Due to this extend in cyber terrorism ratio we would like to discover such exercise the usage of multimedia dataset which could become aware of such undertaking the usage of live API of social media in combination with deep learning to obtain accurate solution to this growing trouble of cyber terrorism due to the fact keyboards used through terrorists are extra dangerous than a bomb.

Keywords Cyber terrorism · Social media crime · Deep learning · Cyber crime · Multimedia crime · Artificial intelligence · Convolutional neural network · Cyber attacks

1 Introduction

Cyber terrorism is multiplied in past few years and due to extensive boom in technologies as we recognize but it have properly effect as properly as awful effect and at present awful use is increasing greater than exact ones therefore for that identification, detection, analysis of this is crucial which can warn us so we can take preventive steps or process to stop that activities earlier than it occurs. There are many options given the usage of desktop learning, deep learning, information mining strategies however all the options are given on a pre-defined dataset of cyber terrorist.

H. Mandaviya (✉) · S. Sathwara
Marwadi University, Rajkot, Gujarat, India
e-mail: himanimandaviya.hm@gmail.com

S. Sathwara
e-mail: snehal.sathwara@marwadieducation.edu.in

1.1 Cyber Terrorism

It is an act of Internet terrorism the place a motion is carried out using computers, networks, mobiles or using equipment like virus, worms, malicious software/hardware, which motive violence, suffering, injuries or death, damage to public property, instability in political and social life.

2 Factors Affecting Cyber Terrorism

Cyber Terrorism has many factors because of which it takes place such affective cyber terrorist attack factors are explained below.

Social Factors

Social elements broadly speaking of upbringing of a specific person, political view, cultural background, persona traits, and faith of person on any precise theme of society, religion, politics, training and other.

Components

Component consists of Domain of the victim, technique of motion to be used to target the victim, actors to be participated in the attack, motivation in the back of planning of this attack, equipment and methods observe to target the victim and ultimately the impact of the assault on the victim.

Support Function

The guide characteristic is the way with the aid of which terrorist crew gets assist for their mission for that they require recruitment of the attackers, education the attackers, intelligence, reconnaissance, planning of attack, logistics, finance of the mission, propaganda or agenda of the mission, and social services.

Objectives

It specifies the intention at the back of the attack whether the attack is planned for protest, to disrupt, kill, terrify, intimidate, demands, achieve sensitive information or for soliciting money. To unfold these objectives, they use exclusive verbal exchange channels like news, social media, advertisements, speeches, encoded messages, speeches, pics and videos.

Effects

This activities of cyber terrorists create big effect by which society gets affected this results or we can say consequence of this terrorists activities can be injury to property, concord of nation, fitness and safety, violence, serious injuries, monetary loss and political stability (Fig. 1).

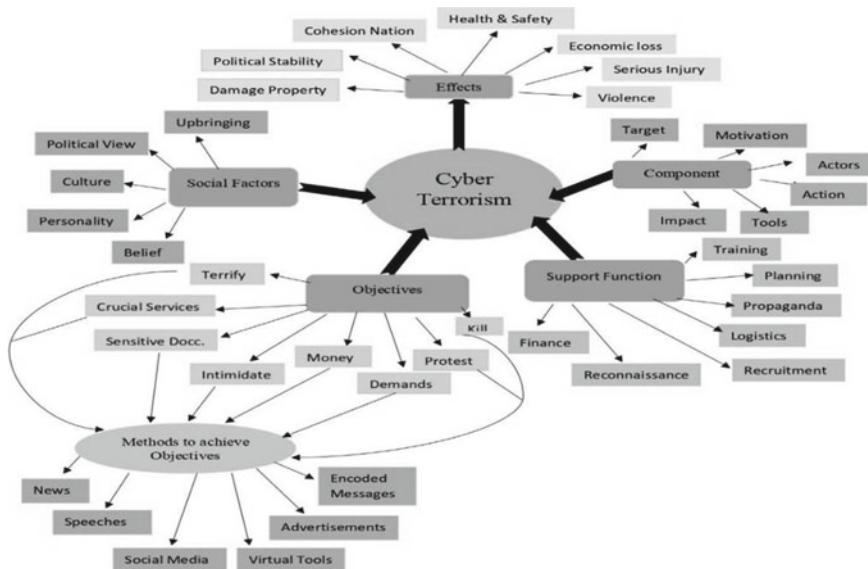


Fig. 1 Factors of cyber terrorism

3 Phases of Cyber Terrorism

It essentially consists of three phases as described below:

Practices

It describes the methods by means of which the attain their target to assault like the practice to corrupt data, spread worms and virus using DOS, disrupt crucial systems, disinformation, unfold propaganda, steal deposit playing cards for finance, deface Web sites these are few frequent practices terrorist observe before accomplishing their goal.

Modes of Operation

It describes the operational modes for how they execute their diagram like they set propaganda, perception, administration of the mission, destructive and disruptive attack.

Attack Levels

It includes of the have an effect on of an assault or stage of attack like unstructured attack, structured attack, advanced, coordinated, complex or simple assault which has least impact (Fig. 2).

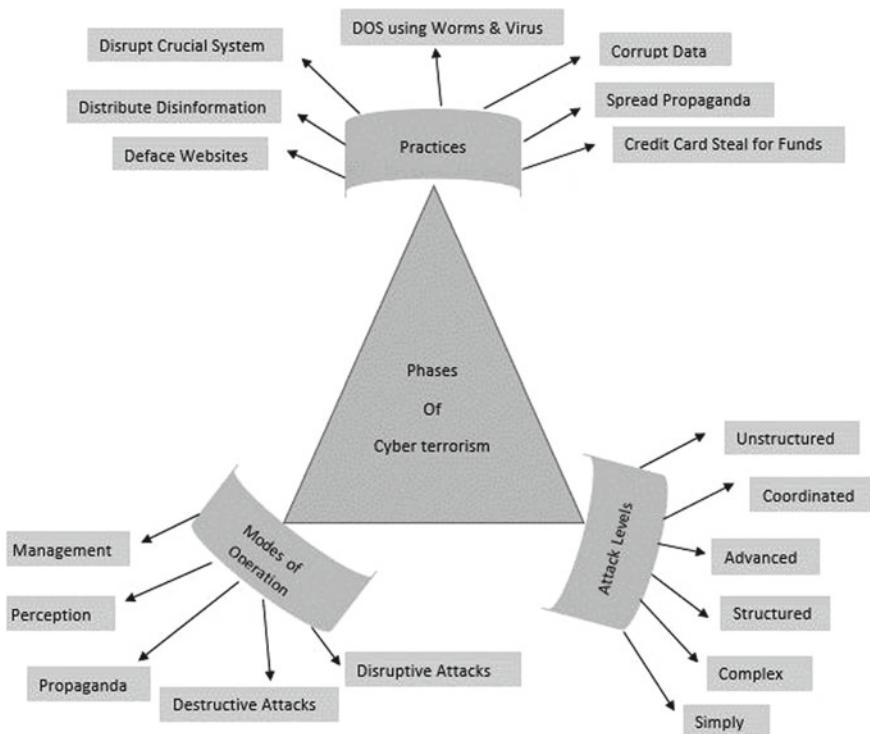


Fig. 2 Phases of cyber terrorism

4 Deep Learning

It is described as the characteristic of artificial intelligence which mimics like working of human thought in pattern advent and information processing for decision making. It is the subfield of Machine Learning (M.L.) it is additionally regarded as deep neural mastering or deep neural network as it has capability of studying networks which are unsupervised from statistics, i.e., unlabeled or unstructured.

Deep Learning makes use of a hierarchical stage of artificial neural network which can process of laptop learning. The neural networks are created in such a way that it works like human thought where neuron nodes are connected collectively like Web. It helps in detecting frauds, thefts & exceptional crook activities. Deep mastering has exclusive architectures which have been applied on speech and audio recognition, computer vision, natural language process, community filter of social media (Fig. 3).

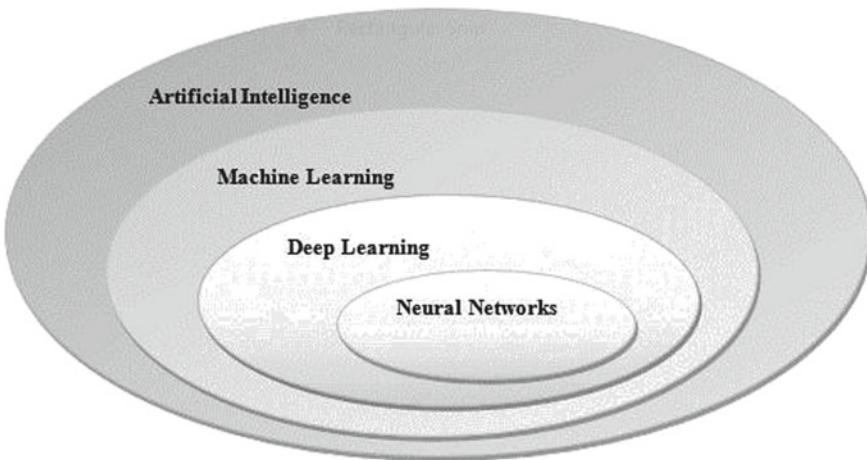


Fig. 3 Artificial intelligence phases

4.1 CNN (*Convolutional Neural Network*)

CNN consist of one input and output layers with a couple of hidden layers interior it. These hidden layers consist of collection of convolutional layers which convolve using multiplication or dot product. It has activation characteristic which is observed by way of ReLU layer followed by using convolutions such as pooling layer, full connected and normalization layers this layers are hidden because they are masked by activation characteristic and closing convolution.

In CNN backpropagation is used for the accuracy of quit product or remaining product. Technically it uses sliding dot product and move correlation. It's magnitude of indices in matrix and additionally affects how weight is determined through precise index point. Convolutional layer has specific input tensor with shape (No. of images) * (image width) * (image height) * (image depth). Input passes to its subsequent layer which technique data solely for receptive field (Fig. 4).

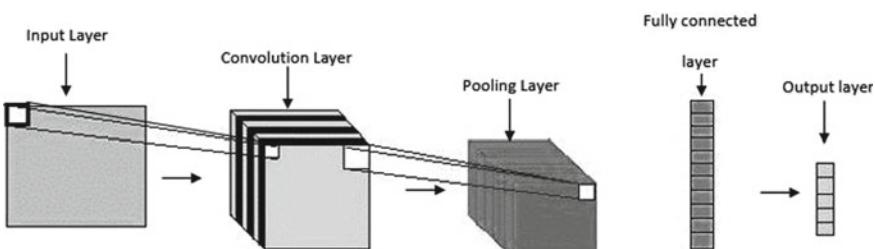


Fig. 4 CNN (convolutional neural network) process

Pooling layer has 2 computation neighborhood an global computation this layer reduces the dimensions of the data by means of combining output of neuron clusters at one layer into a single neuron in subsequent layer. Local pooling the use of small clusters commonly $2 * 2$ whereas in global pooling all neurons are related to the convolutional layer. In addition it has max and common computation the place max uses the most cost from every cluster of neuron and average computation makes use of average value from each cluster of neuron at prior.

In fully linked CNN it connects each one layer to each another layer where each neuron receives input from each and every element from previous layer and the enter location of neuron is regarded as receptive fields.

5 Literature Survey

There are many methods for evaluation of cyber terrorism like online videos containing terrorist agenda to analyze, extract conclusions two using combined Web scrapping and video assessment method [1]. CNN is used on social media for photo retrieval using probability two score the use of soft-max classifier having decreased dimension CNN for retrieving photograph facts sets having ISIS logos, Guy Fawkes Masks for picture classification and object detection [2]. CNN method used to predict risky items like blood, knife, gun in picture to discover whether or not crime has befall or now not the use of Rectified Linear Unit (ReLU) utterly related CNN [3].

CNN used for detection localization and for bizarre events in surveillance videos [4]. Based on transferring gaining knowledge of and mix CNN feature low-level picture characteristic to higher describe CSI photographs using CNN model on giant -scale image net and CSI picture database for satisfactory tuning [5]. Crime Intrusion Detection System detects in real time videos, images and additionally indicators human supervisor to take essential movement, it is experimented on videos and photos dataset amassed from you tube and Google [6].

Terrorism is spreading online the use of text, speeches, videos, pictures and to recognize these Internet properties and ban naturally device uses 2 modes teaching mode and detection mode to cease suspicious words and net sites are mined [7]. Image representation feature learned by CNN and fed to ELM (Extreme Learning machine) for classification which has been benchmarked on MNIST and has effectively improved accuracy in comparison of single hybrid CNN-ELM classification up to 99.33% Accuracy [8]. Human machine collaborative, semi-supervised learning system that can effectively identify malicious social media posts a dynamic classifier resulting in reasonably accuracy nearly 80% [9]. Multilanguage approach deep semantic technology helps to collect and analyze huge amount of heterogeneous and complex multimedia which automatic translates to English [10].

These are few approaches by means of which online social media records are analyzed to become aware of terrorism from them.

6 Proposed Architecture

As per literature survey methods and tools on multimedia, i.e., images, videos, texts, it is clear that all work on dataset of terrorist or we can say after the terrorist assault took place. There is no device but which ought to pre-identify the attacks. Here we have designed a proposed architecture which will help us in collecting stay information multimedia from which we are going to extract photographs and movies the usage of a filtering mechanism, which will then classify into 2 components pictures and 2D videos. That pictures and videos will in addition work on visual and technical classification and then shape a dataset containing all suspicious facts involving terrorism, which will be then given as enter ton CNN and then in the end offers the resultant output (Fig. 5).

7 Conclusion

A survey related to cyber terrorism detection on social media is described the place one-of-a-kind technique and dataset the place used for identification of cyber terrorism the use of images, movies stay surveillance, text. All the identification is executed after crime has occurred right here there is one proposed structure layout which will use live dataset of terrorist activities from social media which can assist in safeguarding our society, country and any major crises so that we can be warned and even can take preventive measures.

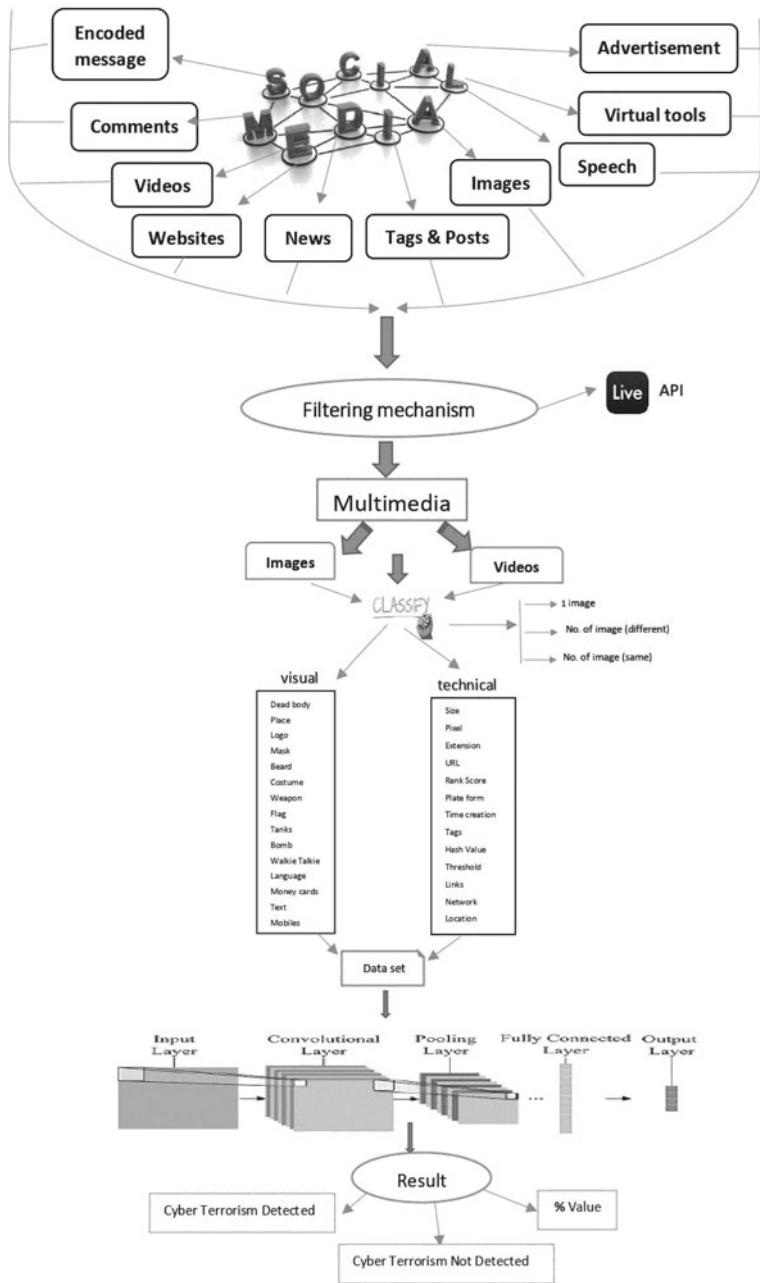


Fig. 5 Proposed architecture

References

1. García-Retuerta D, Bartolomé Á, Chamoso P, Corchado JM (2019) Counter-terrorism video analysis using Hash-based algorithms
2. Chitrakar P, Zhang C, Warner G, Liao XL (2016) Social media image retrieval using distilled convolutional neural network for suspicious e-crime and terrorist involvement detection. In: 2016 IEEE international symposium on multimedia
3. Nakib M, Hasan MS, Khan RT. Crime scene prediction by detecting threatening objects using convolutional neural network
4. Boundour S, Hittawe MM, Mahfouz S, Snoussi H (2017) Abnormal event detection using convolutional neural networks and 1-Class SVM classifier. In: 8th international conference on imaging for crime detection and prevention, ICDP-2017, Madrid 13–15 Dec 2017. IET Digital Library
5. Liu Y, Peng Y, Hu D, Daxiang L, Lim K-P, Ling N (2018) Image retrieval using CNN and low-level feature fusion for crime scene investigation image database. In: Proceedings, APSIPA annual summit and conference, Hawaii, 12–15 Nov 2018
6. Navalgund UV, Priyadarshini K (2018) Crime intention detection system using deep learning. IEEE
7. Ashwariya S, Janani A, Alekhya M. Online terrorist detection systems. Int J Adv Res Ideas Innov Technology
8. Kannoja SP, Jaiswal G (2018) Ensemble of hybrid CNN-ELM model for image classification. In: 2018 5th international conference on signal processing and integrated networks (SPIN)
9. Bhattacharjee SD, Balantrapu BV, Tolone W, Talukder A (2017) Identifying extremism in social media with multi-view context-aware subset optimization. In: 2017 IEEE international conference on big data (BIGDATA)
10. Mencarini M, Sensidoni G. Multilanguage semantic behavioral algorithms to discover terrorist related online contents. In: First Italian conference on cybersecurity (ITASEC17), Venice, Italy. Copyright 2017 for this paper by its authors. Copying permitted for private and academic purposes

The Role of Technologies on Banking and Insurance Sectors in the Digitalization and Globalization Era—A Select Study



Venkamaraju Chakravaram, Sunitha Ratnakaram, Nitin Simha Vihari, and Neelakantam Tatikonda

Abstract Soon the world is entering into the third decade of the twenty-first century (3D/21C) with advanced technologies (such as FinTech, InsurTech, and Blockchain Technologies) in all the fields. Globally 3D/21C Technologies are going to affect the financial services in a multitude of ways in the Globalization and Digitalization era. The present research study is an attempt to list out and study the role of the advanced technologies which are going to drive and affect greatly the financial services especially in the Banking and Insurance sectors in the period of 3D/21C globally. We tried to describe various aspects of the phenomenon in this context; hence, we followed the descriptive type of research methodology in the present study. Although there are many existing studies in nature, Banking and Insurance fields are the main focused sectors in the current research study. Researchers examined the role of FinTech and InsurTech on these sectors and concluded.

Keywords Banking · Financial engineering · Financial services · FinTech · Insurance · InsurTech · Technologies

V. Chakravaram (✉) · S. Ratnakaram
Jindal Global Business School, O. P. Jindal Global University, Sonipat, Haryana, India
e-mail: vchakravaram@jgu.edu.in

S. Ratnakaram
e-mail: sratnakaram@jgu.edu.in

N. S. Vihari
BITS Pilani, Dubai Campus, Dubai, United Arab Emirates
e-mail: nitinvihari@dubai.bits-pilani.ac.in

N. Tatikonda
Wolkite University, Wabe Bridge, Ethiopia
e-mail: neelakantmtati@gmail.com

1 Introduction

The advanced functionality and utility of the newly evolving and innovative technologies have changed the perception of traditional methods. The innovative technologies are bringing changes and helping in revolutionizing the industries of financial services sector particularly dominating the sub-sectors like Banking and Insurance. Advanced technologies like Financial Technologies and Insurance Technologies which are in short known as “FinTech” and “InsurTech” are already transforming the financial sector globally such as in the business fields of Banking and Insurance sectors. Fintech was also known as back-end technology which is generally used to run traditional financial services of organizations. But has morphed into a term primarily that can be used to describe disruptive financial technologies.

The traditional practices and procedures in both the sectors in the landscape are set to rapidly change in the next decade of the twenty-first century (3D/21C). Particularly, the safety and security code based technological features, like advanced cryptography and biometrics etc. will help the customers and other key stakeholders of sectors to protect against financial scams, claim scams in both the sectors respectively. These advanced technologies and remote applications are making easier the life in financial transactions when compare with banking and insurance transactions in the past two decades. This sort-out the issues of visiting the bank physically and meeting the Insurance agent/adviser. If we start practicing these technologies once, the experience which we are going to get is likely to be much more customer-friendly and easy to users.

1.1 Literature Review

In 2017, Vantage et al. [1] presented an article, where the author beautifully explained about the FinTech, InsurTech and tried to address how these technologies are changing the scene, caused by the tremendous growth of both banking and insurance sectors worldwide. In the same year, Goldstein et al. [2] proposed his research paper titled “5 trends to watch in Banking Technology in 2018”, which was well examined and presented about the impact of technologies like APIs, IoTs, New formats of security authenticable technologies, artificial intelligence and etc.

In 2018, Szakil et al. [3] explained briefly how to guide the beginners on FinTech, definition, Technologies involved in this and various benefits of FinTech to the financial services in the banking sector. Furthermore, Carey et al. [4] tried to introduce the latest trends and technologies in the Banking sector following the UK banking sector. They also briefed and explained the industrial experts’ predictions nicely in his research work.

New Jen App et al. [5], where the author listed seven advanced technology trends which are going to change the banking, insurance and finance sectors. Examined the technologies like Cloud Services, Artificial Intelligence, Mobile Banking,

Blockchain technology, Updated ATMs, Security technologies and etc., Smith et al. [6], listed the impact of technologies on the insurance sector, particularly on changing the face of liability, claims settlement process, price discrimination.

In 2019, Csiszar et al. [7], tried to explain how eight new technologies are going to change the banking sector in coming five years, examined Blockchain technologies, upgraded ATMs, Proliferation of Non-Banks, Apple Store Style experience, Automated Financial Services employee, Mobile and Digital Banking and Partnerships and etc.

1.2 Our Contributions

- The researchers studied the advanced technologies of 21C/3D which are playing a major role in the smooth function of business operations in both the banking and insurance sectors.
- We attempted to bring all the technologies which come under FinTech and InsurTech in the said two sectors.
- Tried to analyze the level of significance of these technologies (FinTech and InsurTech) in both sectors.

1.3 Significance of the Study

Globally, majority of the financial services are providing by the Banking and Insurance Sectors (IBEF 2019). These two sectors are growing rapidly head to head in terms of digitalization and globalization. Also, these two are high growth sectors in terms of business volume. Both sectors are growing rapidly using advanced technologies available in the market. The name of these advanced technologies referred by the experts of the Banking sector is “FinTech”. It stands for Financial Technology (FinTech) and refers to the technologies which are transforming the financial services with digitalization colors over the past numbers of years globally. Similarly, “InsurTech” is the name which is using in the Insurance sector for the usage of Insurance Technologies by the Insurance Sector. Usage of these two advanced technologies at all stages of their financial and insurance services and at operations is one of the main reasons for customer satisfaction (Automation Edge 2018). The companies serving in the financial sectors are enabled them to offer the greatest array of innovative technology-enabled services globally. These are like digitalized online instruments, products, schemes with financially engineered models to allow customers to manage their risks easily, to create wealth and to meet their various financial needs.

Based on the above review of literature, by 3D/21C time period (After 2020), FinTech experts and IT professionals of both the Banking and Insurance sectors

need to address the following themes and issues to go for proper strategic planning and to formulate and to execute various strategic decisions in both the sectors.

As per the observations, we could draw out the following points:

- The ever-expanding rapidly growing, sharing natured economy will be embedded in every part of the financial system. Hence IT experts have to give main priority to the issue while launching or introducing new technologies in the sectors.
- Blockchain technology is looking like a giant leader and highly dynamic nature among all 3D/21C technologies. It is going to shake things up. Hence IT experts need to be ready to adapt and design the process as per features of this giant technology.
- Digital Technologies also may become mainstream technologies in core operations of these sectors. The technologies must have the capacity of assessing and analyzing skills of the Customer intelligence, which will be an important point in upcoming years to helps a company in maximization of their revenues, growth ultimately profitability.
- Technology advancements in robotics and Artificial Intelligence also not negligible technologies, these technologies definitely will start a wave of re-shoring and localization. Hence experts should be ready to address waves of these technologies.
- One more technology going to rock in 3D/21 period i.e., Public Cloud Technologies, it may become the dominant infrastructure model among all the leading alternatives.
- Safety and Security is the first concern in the financial services sector particularly from the hackers and online robbers, attackers. Hence cybersecurity mechanism will be one of the top priority points in 3D/21C to overcome the risky situations of all financial institutions.
- Be ready to enjoy and the traffic of these advanced technologies in the 3D/21st period as Asia will be in dominating mode by emerging as a key center of the advanced technology-driven innovative region in the globe (Tech Trends, Deloitte 2018)
- Experts also should be ready to connect the regulating bodies, those who are also changing their regulating bodies and filling with complete advanced technologies.

Based on the above review of literature, discussions and significance of the studies of the topics, we set the following objective and hypothesis.

2 Objective

To study the role of Financial Technologies (FinTech and InsurTech) on Banking and Insurance sectors.

3 Hypotheses

H1: There is a significant role of FinTech, InsurTech on the Banking and Insurance Sectors.

4 Methodology

Researchers used the descriptive research methodology to discuss the facts on stated research objectives, to test and prove the listed hypotheses. In continuation of this exercise, researchers reviewed many articles, research papers, published annual reports of the two sectors, presentations, discussions, lectures of experts of the fields. Finally tried to conclude the results.

4.1 Impact of Technologies on the Banking and Insurance Sectors

The technologies which are using in the banking sector worldwide, which are caused by the tremendous growth in the sector globally known as FinTech. The sudden and rapid explosion of the Internet, Intranet and the Mobile Internet has influenced and the main cause for the speed and rapid growth and development of financial technologies. Even the industries which are highly regulated and highly cautious like Insurance have started to migrate to incorporate and to embrace the innovative technological advances and opportunities for the enhanced effectiveness offered by FinTech solutions. This profound acceptance and promotion of FinTech innovations in the financial sectors attracted a large number of investments and focus by the investors and IT professionals in the FinTech. This sudden change caused the rapid development, growth, and implementation of FinTech throughout the financial services sector.

4.2 Key Trends in the FinTech

Some of the professional experts of the sectors viewed, the FinTech broadly the combinations of the following listed technologies. Insurance Technology (InsurTech), Regulatory Technology (RegTech), Financial Data APIs, Payments Technology, Online Banking Applications, Mobile Banking Applications through various Apps and etc. The above each categorized technologies represents a distinct category of finance-specific technology.

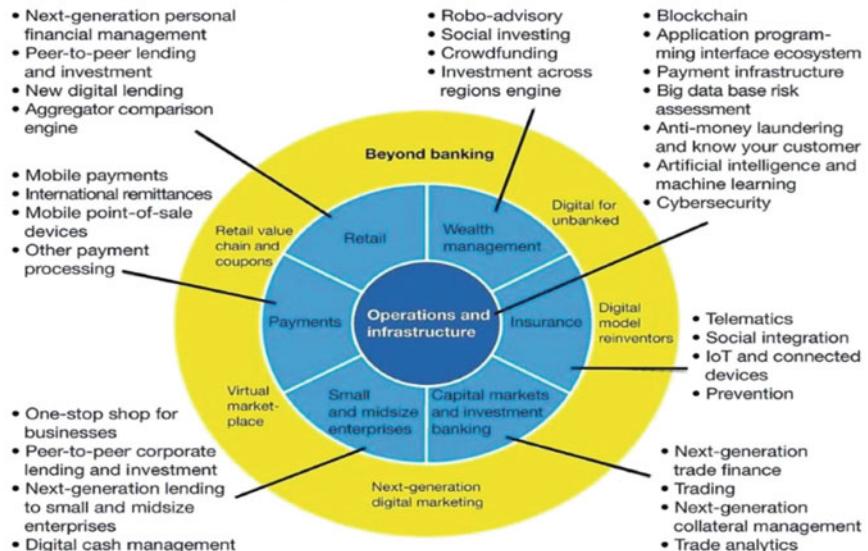
4.3 The Focus Areas of FinTech

As indicated in Fig. 1 Key FinTech Trends, we can segregate technologies used within the banking operations, technologies used beyond the banking operations. Technologies used within the banking sector for Wealth Management, Insurance operations, Capital market, investment banking, special application windows created for small and midsize enterprises, technologies to make a payment, fund transfers and the transactions using for retailing transactions in the banks. Beyond the banking sector are technologies for next-generation digital marketing (3D/21C), virtual market place, retail value chain and coupons, digitalization for unbanked and digital model inventors.

5 Technologies in FinTech

This section deals with the technologies in FinTech that will dominate every technology in the near future. Each technology in the FinTech is described as follows:

Key fintech trends (Areas emerging as new norms in banking)



Source: Panorama by McKinsey

Fig. 1 Key FinTech trends [8]

5.1 Blockchain Technology

Blockchain technology is expected to have an enormous impact on the services provided by FinTech companies. Use of Blockchain Technology in Banking Transactions: Globally 42% of Block Chain technology users are from Banking and Finance Sector (Universa, 2018) alone and stands these two sectors as number one among all the Blockchain users across the globe. As shown in Fig. 2, a banker can easily transfer the money or funds by introducing Blockchain technologies in the banking transactions and operations. If a branch raised the transaction at its' "A" point, this will be visualized to everyone those who are accessing the network of the same bank, irrespective of their accessing distance, accessing place and accessing time. The same concept was well explained in Fig. 2.

Regardless of whether it is the system of directing exchanges or upgrading resources, Blockchain innovation is going to make every one of these strategies just as an assortment of different business forms snappy and simple.

Worldwide FinTech overview results demonstrate that constantly 2022, 77% of the worldwide organizations in the money related segment intend to take Blockchain into live creation.

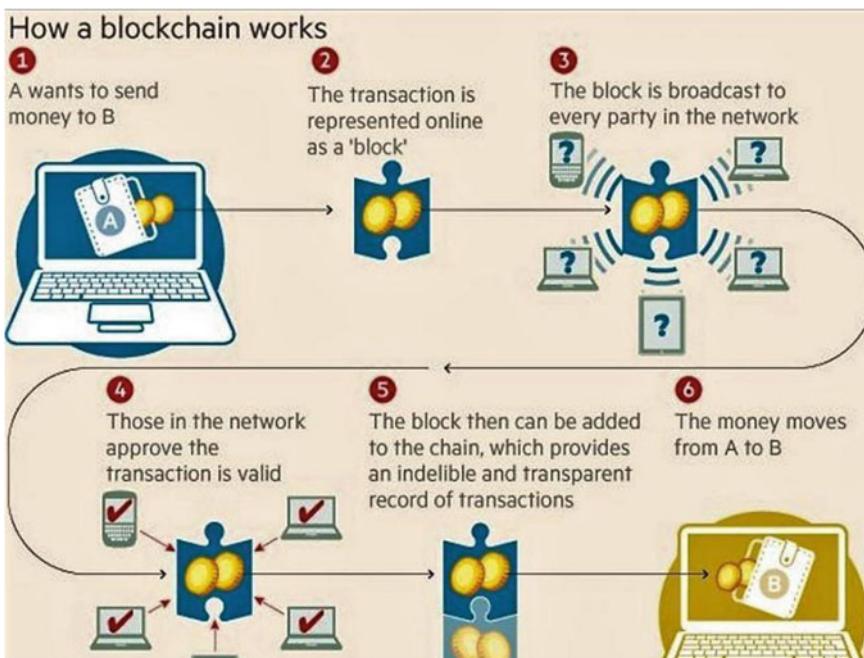


Fig. 2 Working of blockchain technologies [9]

Sooner rather than later, Blockchain is probably going to build speed, productivity, security and lower the expenses of a greater part of activities of the FinTech organizations.

5.2 Cryptocurrency and Digital Lending

In the ongoing years, paper cash has lost its esteem and individuals are getting settled with the plastic or computerized digital cash. Computerized monetary standards like Bitcoin, Ripple, Litecoin and so forth have picked up fame in a brief span and are generally acknowledged by the general population everywhere throughout the globe.

The methods for loaning have additionally changed and all the loaning procedures have turned out to be advanced. This trend of digital lending is expected to grow in the year 2019 and computerized monetary forms are required to turn out to be progressively prominent [8].

5.3 Artificial Intelligence

In the coming years, Artificial Intelligence is aimed to be the top trend as the rise in popularity of the advanced cognitive solutions of AI is expected to reduce the overheads and costs of the financial services.

Furthermore, to stay in the competition, all FinTech companies will be actively using AI to deal with their competitors. FinTech organizations will make the most out of AI and will be utilized for recognizing cheats, doing the examination and for providing various anti-money laundering solutions to the companies [8].

5.4 More Financial Startups and Solutions

Many FinTech organizations are emerging time and now by following the reaction of clients toward the computerized administration of the cash. Top Big-shot organizations also are putting their investments and cash into new and rising FinTech organizations. Furthermore, the top organizations are dismissing the traditional method for subsidizing and offering inclination to Initial Coin Offerings (ICO's) as their financing channel.

The year 2018 has seen numerous new FinTech organizations in India as well as everywhere throughout the globe and the achievement of new organizations has unquestionably pulled in more interests in this division.

5.5 Rise of Mobile Technology

With cell phones being the focal point of everyone's life at the present, FinTech segment also is exploiting the equivalent. Due to an expansion in the utilization of mobile devices by individuals over the globe, monetary administrations are likewise going versatile. Mobile banking gives clients a chance to deal with their funds with only a couple of snaps as opposed to setting off to the banks actually. The people find it very much convenient and have shown a great response toward the mobile banking services and the comfortability in accessing the account transactions are managed well with just a few clicks. Mobile FinTech exchanging applications like Matador and so forth have empowered the clients to make investments in the stock market without paying the middleman with the broker fees. The clients' life is so happy as each and every activity is done using smartphones with great comfortability [8].

It is trusted that the FinTech organizations will begin building up their very own portable technique and take the game to a whole new level.

5.6 Increased Investments in Improving Cybersecurity

Cybersecurity has certainly turned into a point of convergence in the gathering for the organizations and governments everywhere throughout the world. Recent breaches of information and the effect of WannaCry Ransomware has made cybersecurity a critical worry for the organizations. These breaches have deeply affected around 143 million individuals and with the technological advancements made each day, information breach of such sorts are relied upon to increment and productive in their endeavors.

FinTech organizations are giving unified consideration in improving the cybersecurity as cybersecurity speculations came to around \$5 billion in recent years. The financial specialists and undertakings are relied upon to put resources into cutting edge innovation to check the digital burglaries.

As indicated by measurements, the financial related industry has received \$17.4 billion interest in the year 2017 and is flourishing internationally. The FinTech trends are all innovation upheld and all the FinTech organizations must conquer the final barriers and barricades to spread the selection and get a change the financial services.

5.7 Upgraded ATMs

ATMs first introduced in the world in the year 1967. Now the drastic transformation is taking place in these ATMs operations. At present ATMs introduced in the world with the latest technologies, which we can operate and avail services using Biometric, iris recognition. Also using mobile phone devices messages. These ATMs are upgraded

with the latest technologies to deposit money and checks without going to the bank branch. Soon we can avail all kind of banking services from ATMs itself (Ref: Website of IndusInd Bank) These Advanced technologies are going to change the environment in ATMs.

5.8 User-Friendly Banking Apps

No need to go banks, all the transaction we can do using Apps. If cash and money are required, bankers are ready to send them to specially appointed courier services to their customers. Soon customers will get experience at banks like apple showroom experience to do their transactions with a pleasant atmosphere to do banking.

5.9 Automated Financial Services

These technologies will minimize the employees cost on organizations. Automated machines will deliver all kind of financial services to their customers effectively.

5.10 Mobile and Digital Banking

Bankers are investing heavily on the digitalization of their banking operations. Already some of the banks are making very good returns on investments by utilizing the technologies with high customer satisfaction on their services.

5.11 Partnerships and Collaborations

Bankers are going to have collaborations and partnership with FinTech companies for them continues implementation of new and advanced technologies. Same time they are going to have tie-ups with media houses to digitalize their advertisements social site platforms.

5.12 Wearable Technologies

By providing some digitalized wrist watches or cooling glasses, bankers can easily grab the information about the customer while entering into the bank premises with these device “Bluetooth signal” technologies. Due to this technological innovation, bankers can save valuable time of customer by providing quick service from their end.

5.13 The Core Benefits of FinTech

Minimization of costing: Companies operational costs drastically changed and reduced in their day to day operations due to the introduction of FinTechs at each and every stage of business processes.

5.14 Qualities in the Decision-Making Process

Computerization or digitalization of processes always gives accurate and fast results. Also, these technologies are helping in various ways to analyze the business situations in multiple business operations and occasions in companies in a short period. These qualitative analytical reports help strategists, key decision makers in the decision making the process and to maintain qualities.

5.15 Focus on More Transparency

Due to the digitalization of all the financial transactions and services in the banking sector, banks and financial institutions can easily maintain transparency in their each and every transaction, operations of various services with less time at an optimum cost both the parties in the service transaction.

6 Test of Hypotheses

From the above all discussions, we could able to infer that, the new advanced technologies (FinTech, InsurTech) has good amount of impact on both the Banking and Insurance sectors. Therefore, our study shows that there is a significant impact of the technologies on the said two sectors. Hence proved.

7 Conclusion

Based on the above discussions, after check with various technologies and innovations used and practicing in both sectors, the researchers tested the hypothesis and proved that there is a significant impact of FinTech technologies on Banking and Insurance Sectors. These technologies will be broadly in the form of FinTech and InsurTech. Therefore, there is a significant impact of FinTech, InsurTech on the Banking and Insurance sectors.

References

1. FinTech, InsurTech. What does this mean and What Can it do for Global Insurers? <https://www.lexology.com/library/detail.aspx?g=38d3e0e8-eba9-43f0-9747-4cf72449be98>
2. Top 5 Trends in the Insurance Industry. <https://www.wns.com/insights/articles/articledetail/590/top-5-trends-in-the-insurance-industry>
3. Szakil P et al. <https://learn.g2.com/fintech>
4. Trends to watch in banking technology in 2018. <https://biztechmagazine.com/article/2017/12/5trends-watch-banking-technology-2018>
5. New Jen set. <https://www.newgenapps.com/blog/techtrends-banking-insurance-finance-tech-solutions>
6. Smith C. <https://knowtechie.com/the-impact-of-technology-on-the-insurance-industry/>
7. Csiszar J. <https://www.gobankingrates.com/banking/technology/new-banking-technology/>
8. Key FinTech Trends in 2018. <https://www.sigmainfo.net/6-key-fintech-trends-2018/>
9. Lokesh A. How does blockchain technology works. <https://mindmajix.com/how-does-blockchain-technology-work>

Venkamaraju Chakravaram is a triple post graduate and is currently pursuing his doctoral studies from Osmania University, Hyderabad. His research interests include Financial Engineering in Insurance Business. At present, he was working as a Senior Manager—Study Abroad Programs in OP Jindal Global University, NCR, New Delhi.

Sunitha Ratnakaram is a triple post graduate and is currently pursuing her doctoral studies from Indian Institute of Management, Lucknow, in the area of Marketing and working as teaching faculty in Jindal Global Business School. Her research interests include social media and qualitative research. She is a Assistant Dean.

Nitin Simha Vihari is currently working as an Assistant Professor at BITS Pilani, Dubai, UAE. His research interests are Human Resource Management, Corporate Sustainability and Responsible Business Practices.

Neelakantam Tatikonda is currently working as Professor at Wolkite University, Ethiopia. His research interests are Marketing Management, Consumer Behavior.

Review of Recent Plagiarism Detection Techniques and Their Performance Comparison



Manpreet Kaur, Vishal Gupta, and Ravreet Kaur

Abstract With the explosive growth of technology and the easy availability of content on the web, it creates new challenges to discriminate against the original work from plagiarized material. Content is said to be plagiarized when it is taken from other original sources without giving its reference. To address this issue Plagiarism detection tools are required. Over the years, extensive work has been done in the development of anti-plagiarism tools. This paper presents the types of plagiarism with an aim to review Extrinsic Plagiarism detection techniques using Linguistic-based features, Syntactic-based features, and Semantic-based features. Further, an overview of some current state of art methodologies and their results has been discussed on the dataset of PAN-PC 2009, PAN-PC 2010, and PAN-PC 2011. This paper also analyzes the pros and cons of some existing systems and by comparing results it also identifies that some of the systems have less potency to detect the manual and highly shuffled complex types of plagiarism such as translation obfuscation.

Keywords Plagiarism detection · Extrinsic plagiarism detection · Intrinsic plagiarism detection · PAN-PC datasets

1 Introduction

World Wide Web provides access to data present in the documents, databases, and other sources of information using internet service. The availability of knowledge and information in the digital form leads to “Plagiarism” by “Plagiarist”. Plagiarism

M. Kaur · V. Gupta (✉) · R. Kaur
UIET, Panjab University, Chandigarh, India
e-mail: vishal@pu.ac.in

M. Kaur
e-mail: baidwanreet1510@gmail.com

R. Kaur
e-mail: ravreetkaur@pu.ac.in

[1] is defined as the act of stealing and copying the intellectual work, ideas, results, or language of another person without giving credit to the original author and presents it as one's own original work. Plagiarism is not only a major concern in the field of academics but other domains as well such as politics, journalism, music industry, art, medical and scientific research are few to mention here. For this reason, in the academic and research world, various institutions like Elsevier, Springer, and many incorporate anti-plagiarism tools. The function of the plagiarism detection system is to capture the plagiarized content.

Plagiarism detection can be applied to two types of documents, namely Natural Language and Programming Language. Plagiarism detection for Natural language known as Text plagiarism detection and for Programming Language is known as Software or source code plagiarism detection [2].

Plagiarism detection on the basis of usage of original resources and reference collection can be further categorized as Extrinsic and Intrinsic Plagiarism detection.

Extrinsic plagiarism detection technique uses reference collection for pair-wise comparison between suspicious document and source document based on features such as semantic features, syntactic features, and so on [3]. Intrinsic plagiarism detection systems do not take into account the reference of sources for plagiarism detection. Intrinsic plagiarism detection [4] techniques catch plagiarism cases when no reference collection is available for comparison between suspicious documents and source documents. It uses features such as writing style of the author, vocabulary richness including other stylometric features such as deviation in writing style [5], most common words [6] used by the author, and their frequency and vocabulary richness [7].

On the basis of similarity of a text document in respect of language, plagiarism detection can be either monolingual or multilingual. Monolingual plagiarism detection deals with the detection of plagiarism cases where the source and suspicious documents are in the same language such as English–English, whereas Multilingual or cross-language plagiarism detection is used when the original and corresponding plagiarized document uses different languages such as English-Chinese.

The rest of the paper is organized as follows: Sect. 2 presents the Types of Plagiarism. Section 3 describes the steps involved in Plagiarism detection. Section 4 classified the Plagiarism detection systems as Extrinsic and Intrinsic Plagiarism detection systems and also demonstrates the features for Extrinsic Plagiarism detection systems. It also throws light on the current state of art methods. Section 5 discusses the results and performance of systems on PAN-PC 2009, PAN-PC 2010, and PAN-PC 2011 dataset. Section 6 presents future work and Sect. 7 draws a conclusion.

2 Types of Plagiarism

Various researchers define and categorize the plagiarism types according to their studies and analysis. Plagiarism types can be Copy and paste, Disguised, Shake and

paste, Structural, Plagiarism by translation, Metaphor, Patchwork paraphrasing, and Idea plagiarism [8].

Besides these, Plagiarism types are given below:

2.1 Intentional Plagiarism

Intentional or deliberate plagiarism takes place when plagiarists copy the content, steal the idea or work done by others deliberately, and present it as their own ideas. The reason for practicing this plagiarism could be laziness among plagiarists, lack of confidence, stress, or anxiety due to competition and a lack of knowledge about the subject.

2.2 Unintentional Plagiarism

Unintentional or accidental plagiarism takes place when proper citations and references are not given. The reason for practicing this kind of plagiarism could be a lack of knowledge for citing the original sources or unintentionally represent the idea with similar words.

2.3 Self Plagiarism

Self-plagiarism [9] is an act of reusing the own previously published material without giving citations that it has used earlier and presented it as new. The reason behind practicing self-plagiarism could be to save time and efforts for publishing more work.

2.4 Mosaic Plagiarism

Mosaic plagiarism [10] is an act of making use of phrases and use of synonyms in place of original words from source but the idea remains the same without giving credit to the original author.

3 Steps in Plagiarism Detection

Extrinsic plagiarism detection mechanism is shown in Fig. 1. Suspicious document and Source documents are given as input to the system; these documents are

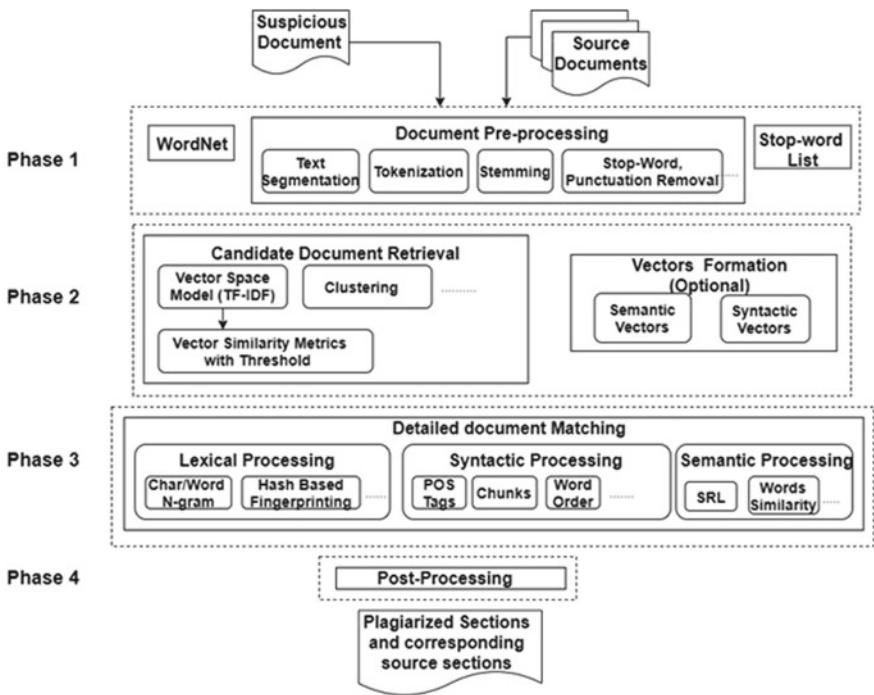


Fig. 1 Steps in Extrinsic plagiarism detection

processed through four phases before production of output as plagiarized section and corresponding source sections.

3.1 First Phase

In the first phase, pre-processing is performed by subjecting input documents to Natural Language Processing techniques using Text Segmentation, Tokenization, etc.

3.2 Second Phase

It consists of candidate document retrieval (a subset of input source documents) followed by optional vectors formation. Candidate document retrieval retrieves the source documents for every query document that is globally similar to a suspicious or query document. The purpose of the candidate document retrieval section is ignoring the irrelevant documents and minimizes the processing time in further

stages. Approaches used for candidate document retrieval are the vector space model (VSM) using term frequency-inverse document frequency (tf-idf) with n -gram representation, Clustering, etc. After VSM (tf-idf), vector similarity metrics such as the Jaccard coefficient [11], Cosine coefficient, etc. can be applied for calculating the similarity score between suspicious document and source documents. Further, a similarity score is compared with the pre-defined threshold value. Document having a similarity score above the threshold will be considered a set of candidate source documents. Another processing section in this phase is Vectors formation which is used for constructing syntactic and semantic vectors of suspicious and source sentences.

3.3 *Third Phase*

In this phase, suspicious and source documents are compared in a detailed manner at sentence or word level using some similarity measures. For detailed document matching input can be either candidate source documents selected in the previous phase and query document or vectors obtained from the vectors formation section. Plagiarism detection system utilizes the Lexical processing for lexical features, Syntactic and Semantic processing for evaluating syntactic and semantic similarity, respectively. In recent research, some systems blend syntactic and semantic features for achieving efficient results. The output of this phase is plagiarized content in fragmented form.

3.4 *Fourth Phase*

Once the detailed matching is done then output is pieces of plagiarized content then post-processing takes place for boundary detections in which obtained matched fragments are converted into passages by splitting and merging the fragments based on some conditions and the threshold value. Passages that overlap are also removed in post-processing. As a result, the output is plagiarized passages and their corresponding sources.

4 Plagiarism Detection Systems

With an increase in activities of plagiarism and dishonesty practiced by plagiarists, need for more advanced systems arises for handling the complex type of plagiarism in which words are highly shuffled, synonyms of words are used and translated text. Many state-of-art methods fail to detect the complex type of plagiarism. To quantify

the cases of plagiarism, detection systems are formally classified as Extrinsic Plagiarism detection systems and Intrinsic Plagiarism detection systems. The efficiency and performance of systems depend on the type of features used for detection.

4.1 Extrinsic Plagiarism Detection Systems

Extrinsic plagiarism detection systems match the suspicious documents with source documents available in reference collection using various features. Some of the features are described below:

Linguistic-based features Linguistic-based features [6] are used for comparison at the character level, word level, and sentence level. These features produce efficient results in the detection of simple plagiarism cases like no obfuscation and those cases where complexity level is low. But it fails to detect the plagiarism as the toughness of plagiarism increased where words are highly shuffled and summary obfuscation. These features can produce good results when merged with other methods. It includes approaches such as Character-based n -gram, Word-based n -gram, and Hash-based fingerprinting.

In *Character-based n -gram*, consecutive n -characters sequence of text in suspicious and source documents are considered whereas in *Word based n -gram* approach consecutive n -words sequence of text in suspicious and source documents are considered for comparison purposes. Number of n -grams [12] that can be generated by using N number of characters/words in text and n size of n -gram is given by Eq. 1.

$$\text{No. of } n\text{-grams} = (N - n + 1) \quad (1)$$

Hash-based fingerprint [13] of the document is a set of hashed values of text segments (n -gram) generated by hash function such as MD5 to represent the document in a more compact form. Fingerprint uniquely distinguishes the one document from other documents.

Syntactic-based features Syntactic-based features analyze the document at the syntax level by segmenting it into paragraphs, lines, and sentences. It includes the approaches Part of Speech (POS) tagging, Chunking, and word order.

Part of speech tagging [14] method assigns a tag to each word in the source and suspicious sentence according to its class such as noun, verb, preposition, etc. *Chunking* scheme is used to represent phrases such as noun phrases, verb phrases, etc. in the form of a parse tree.

Word order [15] is used to measure the word similarity between suspicious sentence and source sentence based on positional index value in the syntactic vector.

Semantic-based features Semantic-based features [16] deals with the meaning of words, synonyms, and hyponyms using dictionaries or semantic databases to find semantic relatedness among documents.

Semantic Role Labeling technique analyzes each part of the sentence and catches the semantic relationship among arguments present in the sentence. Most of the systems used SRL to identify the arguments present in the text of source and suspicious document and labeling them according to the role of each argument.

Word Similarity utilizes the dictionaries, semantic webs, and lexical databases such as WordNet [17] to compute the similarity between pair of words.

For identifying plagiarism [12] author Wielgosz et al. (2017) used the winnowing [18] algorithm for the generation of compressed form fingerprints of a document in which n -grams were subjected to some hash function which generated the hashed values for all n -grams. This method obtained the best results with window size = 40 and n -gram size = 32. A System [19] took into account the stop-word n -gram with a stop-word list. To represent the documents full fingerprinting was used in which each n -gram participate for representation. System [20] in addition to stop-word n -grams [19] utilized two other types of n -grams namely named entity and n -grams of all words for detection of plagiarized segments from the text document. Similar to the n -gram based approach, work has been done to address the problem of downloading a large number of original documents from the web for comparison purposes [21]. The author Velásquez (2017) used shingles [22] of size 1 word and 3 words defined in the text. Shingles of size 1 word convey less information as compared to shingles of size 3 word. The author Osman et al. (2012) has adopted SRL with weighted arguments for matching the suspicious and original documents at the sentence level [16]. In contrast to approach [20] where each n -gram of a suspicious document is compared with each n -gram of a source document, Osman et al. (2012) reduced the unnecessary comparisons and matched similar kinds of arguments. Research work [14] incorporates the use of POS, Chunks, SRL, and combined approaches Chunk with POS and SRL with POS. The comparison was made between similar kinds of POS tags and SRL roles of words in suspicious and source sentences. Another system [23] was able to detect the plagiarism in those cases also where a bag of words of the source sentence and suspicious sentence are similar but convey different meanings. Similar work was carried forward by the author Abdi et al. (2017) where SRL was also used in combination with semantic and syntactic similarity evaluation [15].

Plagiarism Detection System [24] in which a joint vector of sentences was formed for finding the syntactic and semantic relatedness between sentences then Tanimoto Coefficient [25] was deployed for measuring semantic relatedness score and syntactic vectors similarity. For this, relatedness was calculated at word level first by measuring relatedness as a function of features with assigned weights. Combination of inverse path, local density, and depth estimation features gave outstanding results. For overall plagiarism detection Eq. 2 was used [24].

$$\text{Rel}_{\text{overall}}(S_q, S_x) = \Phi \cdot \text{Syn_rel}(S_q, S_x) + (1 - \Phi) \cdot \text{Sem_rel}(S_q, S_x) \quad (2)$$

Best performance was given by the system with the optimal value $\Phi = 0.8$.

4.2 Intrinsic Plagiarism Detection Systems

The intrinsic plagiarism detection system aims to capture plagiarized sections in those cases where no reference collection of the source documents is given. As it is not possible always that real sources are accessible in digital form. Thus, it is advantageous in plagiarism detection when sources are not available in digital format.

5 Results and Discussions

This paper analyzed the performance of various plagiarism detection approaches evaluated by exploiting PAN-PC 2009, PAN-PC 2010, and PAN-PC 2011 dataset [26] as shown in Table 1. Results are expressed in terms of measures viz., Precision, Recall, F1-Score, Granularity and Plagdet for these systems.

$$\text{Granularity} = \sum_{i=1}^n \frac{\text{Number of True Positive cases}}{\text{Precision} + \text{Recall}}$$

(n is Number of True Positives)

$$\text{Plagdet} = \frac{F_1\text{-Score}}{\log_2(1 + \text{Granularity})}$$

Based on the literature survey, this study also analyzes the pros and cons of some of the plagiarism detection systems shown in Table 2. Figure 2 depicts the comparison of systems evaluated on PAN-PC 2009 dataset. CDPDS [27] boost the performance of plagiarism detection system in terms of Precision (+0.019), Recall (+0.028) and F1_Score (+0.024) better than SRLPDS. CDPDS aimed to improve the time efficiency by selecting important keywords.

Figure 3 indicates the performance comparison among various systems evaluated on PAN-PC 2010 dataset. SWNG (Stamatatos 2011) showed outstanding results with reference to simulated plagiarism cases when compared with other state-of-art methods [19] but has lower (below 0.5) recall, F1-score, and Plagdet score than for artificial and verbatim plagiarism. It indicates that the majority of simulated plagiarism cases remain undetected. SWNG showed best results for the verbatim type of plagiarism with the highest Plagdet score (0.94) in comparison to other systems in Fig. 3.

TSPDS [28] improved the performance of SWNG by eliminating the stop words from documents and exploit synonyms of words in sentences. TSPDS reported results

Table 1 Performance of plagiarism detection techniques

PAN-2009					
PDS system	Features used		Results	Authors	Reference
SRLPDS	SRL and Weighted arguments scheme		Set of 100 documents Pre = 0.895, Rec = 0.830, F ₁ = 0.861	Osman et al. (2012)	[16]
CDPDS	Weighted scheme based on positions of keywords (Keyword matching)		Set of 100 documents Pre = 0.914, Rec = 0.858, F ₁ = 0.885	Sharma and Sharma (2014)	[27]
PAN-2010					
PDS system	Features Used		Results	Authors	Reference
SWNG	Stop word 11-gram (Candidate Retrieval), Stop word 8-gram (Passage Boundary), Character 3-gram (Passage Similarity), Full- Fingerprinting		Simulated Pre = 0.89, Rec = 0.27, Plag = 0.41, F ₁ = 0.41 Artificial: High Pre = 0.97, Rec = 0.79, F ₁ = 0.87, Plag = 0.85 Artificial: Low Pre = 0.95, Rec = 0.84, F ₁ = 0.89, Plag = 0.89 Verbatim Pre = 0.96, Rec = 0.93, F ₁ = 0.94, Plag = 0.94	Stamatatos (2011)	[19]
TSPDS	Word level matching, Using Synonyms without Stop words		Pre = 0.953, Rec = 0.726, F ₁ = 0.901, Plag = 0.901	Grman and Ravas (2011)	[28]
UTPDS	SRL + POS tagging, Syntactic-Semantic similarly metric		Simulated Pre = 0.952, Rec = 0.836, F = 0.901, Plag = 0.901	Vani and Gupta (201E)	[14]

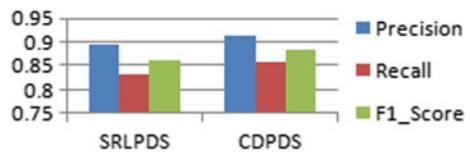
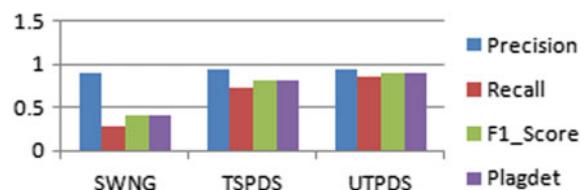
(continued)

Table 1 (continued)

PAN-2011	PDS system	Features used	Results	Authors	Reference
PDLK	Semantic similarity (Words, Sentences) metric, Word Order similarity metric	Pre = 0.902, Rec = 0.702, $F_1 = 0.790$, Plag = 0.789	Abdi et al. (2015)	[23]	
IEPDM	SRL, Semantic similarity metric, Word Orders similarity metric	Pre = 0.921, Rec = 0.622, $F_1 = 0.743$, Plag = 0.737	Abdi et al. (2017)	[15]	
NTPDS	Inverse Path length, Local density, Depth, Estimation	Pre = 0.956, Rec = 0743, $F_1 = 0.836$, Plag = 0.836	Sahi and Gupta (2017)	[24]	
UTPDS	POS tagging, Syntactic-Semantic similarity metric	Pre = 0.925, Rec = 0.793 $F_1 = 0.854$, Plag = 0.846	Vani and Gupta (2018)	[14]	

Table 2 Pros and Cons of Existing plagiarism detection systems

Title	Pros	Cons
Plagiarism Detection Using Stop word n -grams [19]	Able to detect synonyms replacement, exact boundary detections of plagiarized passages	Not able to detect rephrased a large amount of content
Plagiarism Detection in Text using Vector Space Model [29]	Employ corpus with documents written in English, German and Spanish	Lower recall value, System cannot detect complex cases with high obfuscation
PDLK: Plagiarism detection using linguistic knowledge [23]	It detects the meaning of sentences with a different bag of words, paraphrasing, restructuring of sentences and exact copy of a text	Not able to distinguish between active and passive sentences and cover limited words for evaluating semantic similarity between words using WordNet
A Novel Technique for Detecting Plagiarism in Documents Exploiting Information Sources [24]	Discard false-positive cases using the depth estimation feature	Only noun [n,n] and verb [v,v] pairs are accepted as a part-of-speech
Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: Comparison, analysis and challenges [14]	Reduced irrelevant comparisons by comparing similar kind of roles, chunks, and tags	Performance of system reduced for PAN 2014 (all obfuscation) in terms of Plagdet

Fig. 2 Comparison on PAN-PC 2009**Fig. 3** Comparison on PAN-PC 2010

as Precision (+0.063), Recall (+0.456), F1-Score (+0.414), and Plagdet (+0.414) better than SWNG. UTPDS [14] achieved the promising results for the simulated type of plagiarism and reported Precision (+0.062), Recall (+0.586), F1-score (+0.491) and Plagdet (+0.491) more than SWNG.

Figure 4 presents the performance comparison among various recent plagiarism detection approaches assessed on PAN-PC 2011 dataset. PDLK [23] used semantic

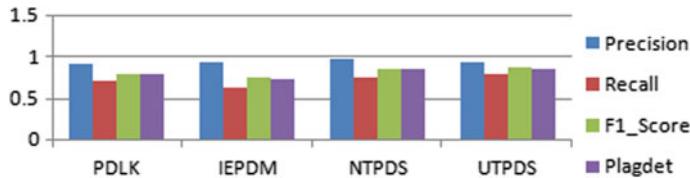


Fig. 4 Comparison on PAN-PC 2011 dataset

similarity metrics and word order similarity metrics for plagiarism detection. Thereafter, in IEPDM [15] SRL was also involved additionally with two techniques resulted into the enhanced performance of PDLK in terms of Precision only by (+0.019) but performance dropped in terms of some measures as Recall (-0.08), F1_score (-0.052) and overall plagiarism detection score Plagdet (-0.052) accordingly. NTPDS [24] uplift the performance measures as Precision (+0.054), Recall (+0.041), F1_score (+0.046) and Plagdet (+0.047) better than PDLK. UTPDS [14] indicated the encouraging with Recall (+0.05) more than NTPDS to avoid False-Negative cases and improved Plagdet score by (+0.01).

6 Future Work

Future work will therefore, include time reduction for extraction of keywords [27], Exploitation of location or position of stop-words, and their frequency for further research work [19]. The candidate document retrieval unit has a great impact on the overall effectiveness of the system so, the selection of optimal threshold is important to avoid the absence of relevant documents. Techniques used by systems to detect plagiarism influence the effectiveness of the system. In addition to this PAN-PC Dataset contains a wide collection of documents, existing plagiarism detection systems showed the best results for a limited number of suspicious documents with respect to available documents for detecting plagiarism. This survey encourages researchers to build a plagiarism detection system for processing the utmost suspicious documents by bearing in mind the time efficiency.

Besides the above aforementioned future directions, machine learning can be used for finding optimal parameter values and the use of stop-word n -grams for intrinsic plagiarism detection systems.

7 Conclusion

The vast amount of data available on the internet in digital form encourages plagiarists to copy the content for the sake of taking benefits without acknowledging original sources. To quantify the cases of plagiarism, detection systems are formally classified

as Extrinsic Plagiarism detection systems and Intrinsic Plagiarism detection systems. This paper explored types of plagiarism and it is observed that most of the plagiarism activities take place in the academic and research world. Further, we describe four phases in plagiarism detection in which actual plagiarism is detected in the detailed document matching step. Traditional and recent techniques in this domain have also been discussed.

In recent research, most of the systems are the fusion of two or more techniques and boost system performance by producing encouraging results. Next, we compare the performance of systems evaluated on PAN-PC 2009, PAN-PC 2010, and PAN-PC 2011 dataset. It can be concluded for PAN-PC 2009 dataset that CDPDS showed good results as shown in Fig. 2 because it attempted to capture the copy detection whereas SRLPDS made efforts to detect the active to passive voice transformed plagiarized and semantically similar sentences as well. It is observed that SRLPDS has the potency to capture the copy and paste, synonym replacement, structural changes, active to passive voice modification, and CDPDS capable of copy detection.

For PAN-PC 2010 dataset our survey shows that SWNG has attained lower performance for simulated cases of plagiarism but achieved good results in case of artificial and verbatim plagiarism. TSPDS took into account combined documents of different types of plagiarism, unlike SWNG. So, it has an impact on performance of the system. It also conveyed that UTPDS performed better than SWNG for simulated plagiarism. For PAN-PC 2011 dataset, this paper observed that UTPDS avoid the false-negative cases to a greater extent when compared with other approaches in Fig. 4 and detect most of the plagiarized cases with the highest performance in terms of plagdet score. UTPDS outperformed the rank 1 system on PAN-PC 2011 dataset and showed effective results.

Although a large number of systems have been introduced, this survey feels that plagiarism detection techniques are still in its infancy for translation, summary obfuscation, and simulated cases.

References

1. Halak B, El-Hajjar M (2016) Plagiarism detection and prevention techniques in engineering education. In: 2016 11th European workshop on microelectronics education (EWME). IEEE, pp 1–3
2. Alzahrani SM, Salim N, Abraham A (2011) Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans Syst Man Cybern Part C (Appl Rev)* 42(2):133–149
3. Gupta D (2016) Study on extrinsic text plagiarism detection techniques and tools. *J Eng Sci Technol Rev* 9(5):8–22
4. Zu Eissen SM, Stein B (2006) Intrinsic plagiarism detection. In: European conference on information retrieval. Springer, Berlin, pp 565–569
5. Oberreuter G, Velásquez JD (2013) Text mining applied to plagiarism detection: the use of words for detecting deviations in the writing style. *Expert Syst Appl* 40(9):3756–3763
6. AlSallal M, Iqbal R, Palade V, Amin S, Chang V (2017) An integrated approach for intrinsic plagiarism detection. *Future Gen Comput Syst* 700–712

7. Zu Eissen SM, Stein B, Kulig M (2007) Plagiarism detection without reference collections. In: Advances in data analysis. Springer, Berlin, pp 359–366
8. Naik RR, Landge MB, Mahender CN (2015) A review on plagiarism detection tools. *Int J Comput Appl* 125(11):16–22
9. Samuelson P (1994) Self-plagiarism or fair use. *Commun ACM* 37(8):21–25
10. Das N, Panjabhi M (2011) Plagiarism: why is it such a big issue for medical writers? *Perspect Clin Res* 2(2):67–71. <https://doi.org/10.4103/2229-3485.80370>
11. Jaccard P (1912) The distribution of the flora in the alpine zone. 1. *New Phytol* 11(2):37–50
12. Wielgosz M, Szczepka P, Russek P, Jamro E, Wiatr K, Pietroń M, Źurek D (2017) Evaluation and Implementation of n-Gram-based algorithm for fast text comparison. *Comput Inform* 36(4):887–907
13. Stein B, zu Eissen SM (2007) Fingerprint-based similarity search and its applications. Universität Weimar, pp 85–99
14. Vani K, Gupta D (2018) Unmasking text plagiarism using syntactic-semantic based natural language processing techniques: comparisons, analysis and challenges. *Inf Process Manage* 54(3):408–432
15. Abdi A, Shamsuddin SM, Idris N, Alguliyev RM, Aliguliyev RM (2017) A linguistic treatment for automatic external plagiarism detection. *Knowl Based Syst* 135:135–146
16. Osman AH, Salim N, Binwahlan MS, Alteeb R, Abuobieda A (2012) An improved plagiarism detection scheme based on semantic role labeling. *Appl Soft Comput* 12(5):1493–1502
17. Miller GA (1995) WordNet: a lexical database for English. *Commun ACM* 38(11):39–41
18. Schleimer S, Wilkerson DS, Aiken A (2003) Winnowing: local algorithms for document finger-printing. In: Proceedings of the 2003 ACM SIGMOD international conference on Management of data. ACM, pp 76–85
19. Stamatatos E (2011) Plagiarism detection using stopword n-grams. *J Am Soc Inform Sci Technol* 62(12):2512–2527
20. Shrestha P, Soloria T (2013) Using a variety of N-grams for the detection of different kinds of plagiarism-lab report for PAN at CLEF 2013. In: Proceedings of 5th International Workshop PAN-13, Valencia, Spain, pp 1–8
21. Velásquez JD (2017) Docode 5: building a real-world plagiarism detection system. *Eng Appl Artif Intell* 64:261–271
22. Broder AZ (1997) On the resemblance and containment of documents. In: Proceedings. Compression and complexity of SEQUENCES 1997 (Cat. No. 97TB100171). IEEE, pp 21–29
23. Abdi A, Idris N, Alguliyev RM, Aliguliyev RM (2015) PDLK: plagiarism detection using linguistic knowledge. *Expert Syst Appl* 42(22):8936–8946
24. Sahi M, Gupta V (2017) A novel technique for detecting plagiarism in documents exploiting information sources. *Cogn Comput* 9(6):852–867
25. Huang A (2008) Similarity measures for text document clustering. In: Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand, vol 4, pp 9–56
26. PAN Data. <https://pan.webis.de/data.html>
27. Sharma R, Sharma D (2014) Copy detection mechanism for documents using position based weighted scheme. In: 2014 5th international conference-confluence the next generation information technology summit (confluence). IEEE, pp 521–526
28. Grman J, Ravas R (2011) Improved implementation for finding text similarities in large collections of data. In: Notebook for PAN at CLEF 2011. Notebook Papers of CLEF, pp 1–6
29. Ekbal A, Saha S, Choudhary G (2012) Plagiarism detection in text using vector space model. In: 2012 12th international conference on hybrid intelligent systems (HIS). IEEE, pp 366–371

A Combination of 2DLDA and LDA Approach for Fruit-Grade Classification with KSVM



Yogeswararao Gurubelli, Malmathanraj Ramanathan,
and Palanisamy Ponnusamy

Abstract This paper uses a non-destructive methodology to explore new simulation techniques to address the problem of classification of pomegranate. The recognition of healthy/damaged fruit is accomplished in this article. The approach uses two new feature extraction based adaptive mathematical principles of investigation on two-dimensional linear discriminant analysis (2DLDA) and a combination of 2DLDA and linear-discriminant analysis (LDA), which is 2DLDA-LDA. In order to classify the extracted features in both the methods, the support vector machine (SVM) with the training of polynomial, radial basis function, multilayer perceptron classifiers, quadratic programming and linear classifier kernel functions are used. The two implemented proposed techniques were tested using cofilab pomegranates database comprising of healthy/damaged fruit images. The performance metric such as the accuracy, reduced dimension, runtime and mean square error are evaluated to demonstrate the effectiveness of the method. The simulation results show that 2DLDA-LDA is manifold superior among the two new techniques.

Keywords Non-destructive technique · Feature extraction · Linear discriminant analysis · Kernel support vector machine

1 Introduction

The techniques for feature extraction and dimensionality reduction play an important role in machine learning applications [1], datamining [2], computer vision [3] and Bioinformatics [4]. In the last two decades, there have been many feature extraction algorithms developed. The LDA was started to be the most commonly used technique for data reduction and guaranteed for class discrimination [5]. The aim of this technique is to map the data features from higher dimensional to lower-dimensional sample space. Also, it maximizes the ratio of between-class to within-class variances.

Y. Gurubelli (✉) · M. Ramanathan · P. Ponnusamy

National Institute of Technology Tiruchirappalli, Tiruchirappalli, Tamil Nadu, India

e-mail: yogi.gurubelli@gmail.com

Primarily the LDA faces two challenges. (1) Small sample size (SSS) problem, (2) linearity problem, Ref. [6] proposed that the kernel functions are one of the solutions to overcome the above-listed problems. Another solution, representation of an image sample in matrix form. Based on this approach, many algorithms have been developed. Reference [7] proposed a two-dimensional principal component analysis (2DPCA) algorithm, and subsequently, Ref. [8] developed a bilateral-projection-based 2DPCA (B2DPCA). Two-dimensional linear discriminant analysis (2DLDA) has also been developed in [9]. Some other modifications extended on 2DLDA [10], Similar to this, Fisher's LDA [11] have been reported. There is always a demand for good quality fruits and vegetables to be processed into juice, syrup and wine in today's highly competitive market. Therefore, the market for healthy and fresh pomegranate fruit has been increased. Since ancient times, the use of pomegranate and accounts of its medical qualities have echoed throughout the ages [12]. According to a recent study [13] drinking pomegranate juice a day can improve learning and memory.

The present paper describes the identification of healthy fruits, using the extracted features from pomegranate digital database. The approach uses two feature extraction techniques called 2DLDA and a combination of 2DLDA and LDA methods. Kernel SVM (KSVM) algorithm has been used to identify the discriminate features obtained in both techniques. KSVM is a supervised machine learning algorithm, it can be used in a discriminative classifier known as hyperplane separation and its statistical learning theory has been discussed in [14, 15]. SVM uses the training kernels like Linear, Polynomial, Quadratic programming, Radial Basis Function and Multilayer perception in all experiments. This analysis is performed as a basket of cultivated pomegranates, which could be represented by a few sample fruits. The accuracy depends on the computational methods involved.

The organization of the remaining paper is as follows: Feature extraction techniques and kernel SVM classifiers were described in Sect. 2. The results and discussion are reported in Sect. 3. Finally, the conclusions are presented in Sect. 4.

2 Feature Extraction Techniques and Kernel SVM Classifiers

2.1 Two-Dimensional Linear Discriminant Analysis (2DLDA)

Let $(x_1^1, X_1^1, C_1), (x_2^1, X_2^1, C_1), \dots, (x_{N1}^1, X_{N1}^1, C_1), (x_1^2, X_1^2, C_2), (x_2^2, X_2^2, C_2), \dots, (x_{N2}^2, X_{N2}^2, C_2), \dots, (x_1^L, X_1^L, C_L), (x_2^L, X_2^L, C_L), \dots, (x_{NL}^L, X_{NL}^L, C_L)$ be the image samples from L classes. $x_i^k \in \Re^n$ is the n-dimensional vector from k^{th} class i^{th} sample and $X_i^k \in \Re^{\text{row} \times \text{col}}$ is its corresponding image matrix. $N = \sum_{i=1}^L N_i$ be the total sample size, where N_k is the number of training samples of class C_k . $M_k = \frac{1}{N_k} \sum_{i=1}^{N_k} X_i^k$ is the mean matrix of k^{th} class and $M = \sum_{k=1}^L \frac{N_k}{N} M_k$ is the

centroid of all class means. The optimal vector w^{2d} for 2DLDA is in the form of

$$w^{2d} = \arg \max_{w^{2d}} \frac{w^{2dT} S_B^{2d} w^{2d}}{w^{2dT} S_W^{2d} w^{2d}} \quad (1)$$

where $S_B^{2d} = \sum_{k=1}^L \frac{N_k}{N} (M_k - M)(M_k - M)^T$ and $S_W^{2d} = \frac{1}{N} \sum_{k=1}^L \sum_{i=1}^{N_k} (X_i^k - M_k)(X_i^k - M_k)^T$ are two dimensional between-class scatter matrix (SB) and within-class scatter matrix (SW), respectively. Equation (1) can be equivalent in the form of $S_B^{2d} X = \lambda S_W^{2d} X$ for $\lambda \neq 0$ a generalized eigenvalue problem. In order to get the solution, apply an eigendecomposition to the matrix $S_W^{2d-1} S_B^{2d}$, and select K eigenvectors that have nonzero largest eigenvalues $w^{2d} = \{v_1^{2d}, v_2^{2d}, \dots, v_K^{2d}\}$. The selected eigenvectors represent the projection space of 2DLDA. Therefore, the reduced dimension of 2DLDA is $\text{col} \times K$ i.e. $Y_i^{k^{2d}} = w^{2dT} X_i^k \in \Re^{\text{col} \times K}$.

2.2 Combination of 2DLDA and LDA (2DLDA-LDA)

After extracting features of the image sample using 2DLDA, i.e. $Y_i^{k^{2d}}$ is to be shaped as N images in the form of a vector $((\text{col} \times K) \times N)$ and apply LDA to the entire matrix which is a combination of 2DLDA and LDA. $N^{\text{pop}} = \sum_{i=1}^L N_i^{\text{pop}}$ be the total sample size, $m_k^{\text{pop}} = \frac{1}{N^{\text{pop}}} \sum_{i=1}^{N_k^{\text{pop}}} y_i^{k^{2d}}$ as the mean vector of samples of class C_k , N_k^{pop} is the number of training samples of class C_k and $m^{\text{pop}} = \sum_{k=1}^L \frac{N_k^{\text{pop}}}{N^{\text{pop}}} m_k^{\text{pop}}$ be the mean vector of all samples, then the objective is to maximize the ratio of SB and SW. The optimal vector w_{pop} for proposed algorithm is in the form of

$$w_{\text{pop}} = \arg \max_{w_{\text{pop}}} \frac{w_{\text{pop}}^T S_B^{\text{pop}} w_{\text{pop}}}{w_{\text{pop}}^T S_W^{\text{pop}} w_{\text{pop}}} \quad (2)$$

where $S_B^{\text{pop}} = \sum_{k=1}^L \frac{N_k^{\text{pop}}}{N^{\text{pop}}} (m_k^{\text{pop}} - m^{\text{pop}})(m_k^{\text{pop}} - m^{\text{pop}})^T$

$$\begin{aligned} S_W^{\text{pop}} &= \frac{1}{N^{\text{pop}}} \sum_{k=1}^L \sum_{i=1}^{N_k^{\text{pop}}} (y_i^{k^{2d}} - m_k^{\text{pop}})(y_i^{k^{2d}} - m_k^{\text{pop}})^T \\ &= \frac{1}{N_k^{\text{pop}}} \sum_{i=1}^{N_k^{\text{pop}}} (y_i^{k^{2d}} - m_k^{\text{pop}})(y_i^{k^{2d}} - m_k^{\text{pop}})^T \end{aligned}$$

are between-class scatter matrix and within-class scatter matrix, respectively. Equation (2) can be equivalent in the form of $S_B^{\text{pop}} Y^{2d} = \lambda^{\text{pop}} S_W^{\text{pop}} Y^{2d}$ for $\lambda^{\text{pop}} \neq 0$ as

a generalized eigenvalue problem. In order to calculate the eigenvalues and eigenvectors, apply an eigen decomposition to the matrix $S_W^{\text{pop}^{-1}} S_B^{\text{pop}}$ and select $K1(\neq K)$ eigenvectors that have the nonzero largest eigenvalues $w^{\text{pop}} = \{v_1^{\text{pop}}, v_2^{\text{pop}}, \dots, v_{k1}^{\text{pop}}\}$. Therefore, the reduced dimension of 2DLDA-LDA is $(\text{col} \times K) \times K_1$ i.e. $Y_i^{k^{\text{pop}}} = w^{\text{pop}T} Y_i^{k^{2d}} \in \text{IR}^{(\text{col} \times K) \times K_1}$.

2.3 Kernel SVM Based Classifiers

Let Ψ be some feature space and Φ is the non-linear mapping to Ψ . In order to find the linear discriminant in feature space Ψ , we need to maximize the ratio between the SB and SW, now $w^{2d}, w_{\text{pop}} \in \Psi$ and S_B^Φ, S_W^Φ are the corresponding matrices in Ψ . If Ψ is very high dimensional, it is impossible to solve directly. Hence to overcome this constraint or limitation, the data is mapped explicitly to look for a formulation of the algorithm called Kernel SVM which uses the dot-product $k(x, y) = \Phi(x) \cdot \Phi(y)$. The possible choices for ‘ k ’ in SVMs are linear kernel $k(x, y) = (x^T y)$, a polynomial kernel with degree ‘ d ’ $k(x, y) = (x^T y)^d$ or $(1 + x^T y)^d$, quadratic kernel $k(x, y) = (x^T y)^2$ or $(1 + x^T y)^2$, Gaussian RBF $K(x, y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$ and for multilayer SVM kernel in the two-layer architecture, the hidden-layer representation is $f(x/\theta)$ and hidden layer representation to output is $g(f(x/\theta))$ where ‘ x ’ is an input pattern. Here ‘ θ ’ denotes the trainable parameters in the hidden-layer SVMs are derived in [11] which are shown in Eqs. (3) and (4)

$$f(x/\theta) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K_h(x_i - x) + b_h \quad (3)$$

$$g(f(x/\theta)) = \sum_{i=1}^l (\alpha_i^* - \alpha_i) K_o(f(x_i/\theta), f(x/\theta)) + b_o \quad (4)$$

where α_i^*, α_i are support vector coefficients, b_h and b_o are bias and kernel function for the hidden-layer and output SVMs are

$$K_h(x_i, x) = \exp\left(-\frac{\sum_{h=1}^D (x_i^h - x^h)^2}{\sigma_h}\right)$$

and

$$K_o(f(x_i/\theta), f(x/\theta)) = \exp\left(-\frac{\sum_{o=1}^d (f(x_i/\theta) - f(x/\theta))^2}{\sigma_o}\right)$$

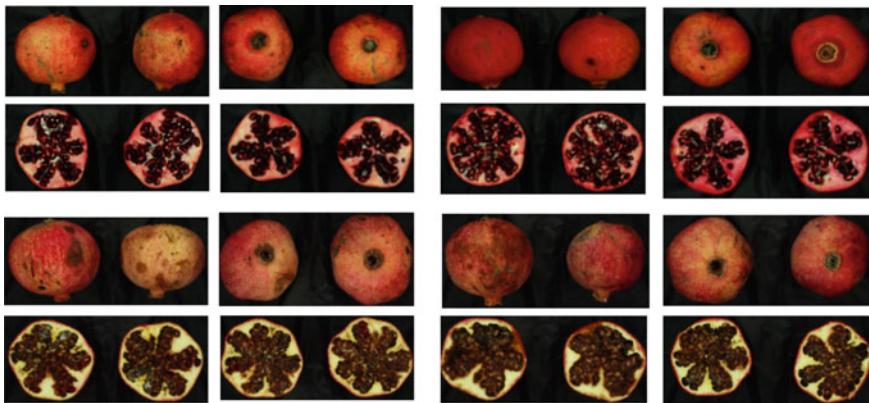


Fig. 1 Side, top and opened pomegranate sample images for healthy and damaged fruit classes

3 Results and Discussions

In this section, the effectiveness of our proposed approach has been tested, with cofilab digital images of pomegranates public datasets. The dataset consists of 328 colour pomegranate images of pixel size (768×1024) and sample images are shown in Fig. 1. In order to appraise the performances of 2DLDA technique and 2DLDA-LDA technique, the accuracy of both methods is compared with pomegranates images.

The accuracy of the two methods (2DLDA and 2DLDA-LDA) was explored through testing different training images of each pomegranate type, i.e. class label. The training images were randomly selected from the database and the remaining images were used as testing images. The accuracy, CPU time and mean square error of these approaches shown in Figs. 2, 3 and 4.

In this paper, we investigated the success of proposed algorithms in terms of classification accuracy, CPU time and mean square error. Also, the results are compared with 2DLDA. It can be observed that the LDA stage in the proposed algorithm reduces the dimension and mean square error and also increase the accuracy. Another observation is that the CPU running time is reduced in 2DLDA-LDA as compared with 2DLDA because of reduced dimension in transformed space.

4 Conclusions

This article discusses two methods of recognizing the healthy pomegranate fruit images using investigation on 2DLDA features and SVM. Classes are categorised using five types of SVM classifiers. The simulation results demonstrate 2DLDA-LDA-SVM is affective and best discriminant feature extraction method among

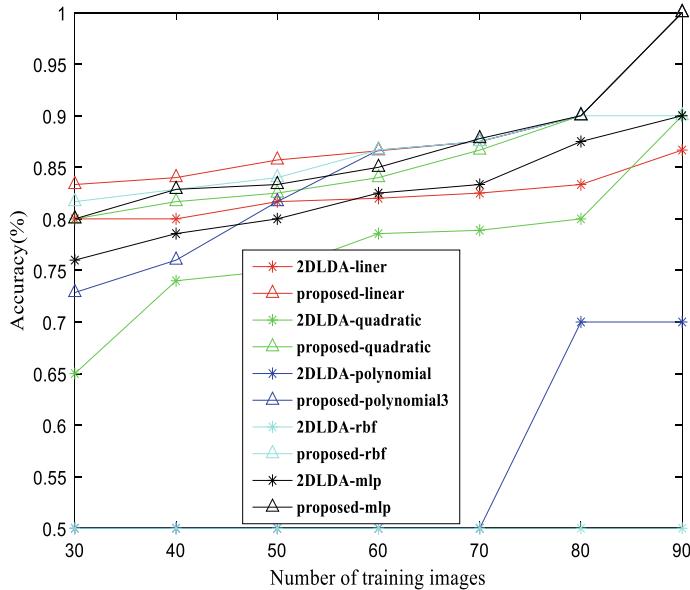


Fig. 2 Accuracies of 2DLDA and 2DLDA-LDA with different training pomegranate images with five different kernels

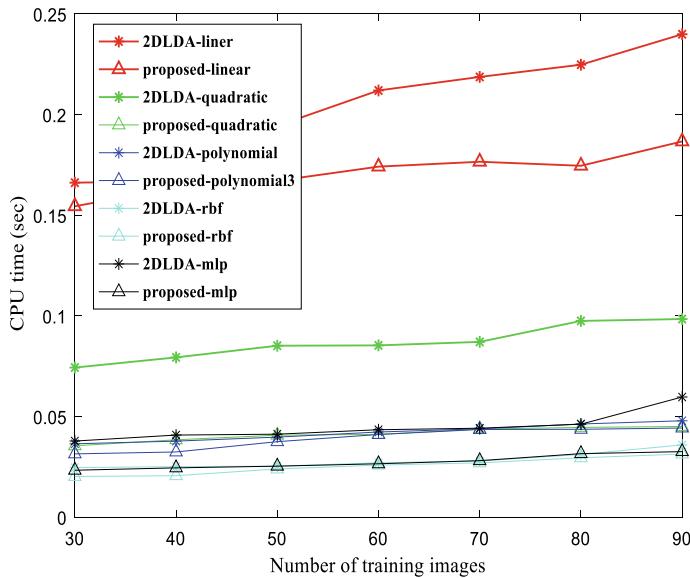


Fig. 3 CPU times of 2DLDA and 2DLDA-LDA with different training pomegranate images with five different kernels

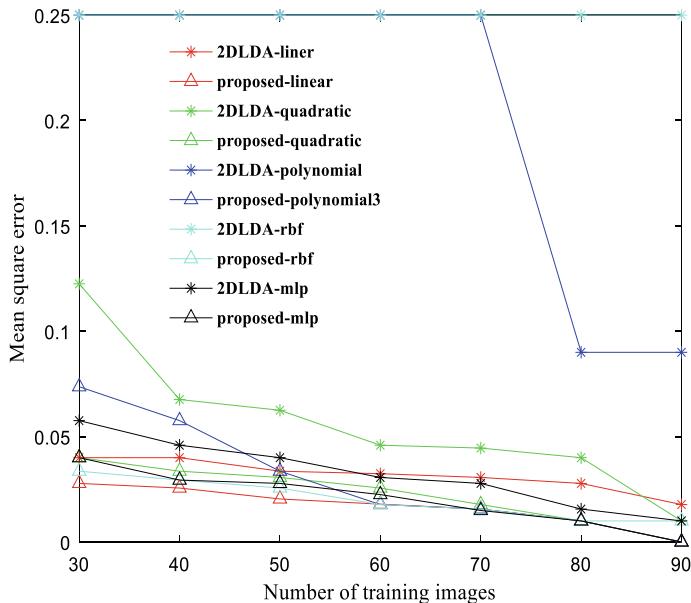


Fig. 4 Mean square errors of 2DLDA and 2DLDA-LDA with different training pomegranate images with five different kernels

2DLDA-LDA-SVM and 2DLDA-SVM. Furthermore, SVM with training kernels like Liner, Polynomial, Quadratic programming, Radial Basis Function, and Multi-layer perception classifiers along with 2DLDA-LDA combination provides superior results.

References

1. Duda RO, Hart PE, Stork DG (2012) Pattern classification, 2nd edn. Wiley
2. Bramer M (2013) Principles of data mining, 2nd edn. Springer
3. Nixon MS, Aguado AS (2012) Feature extraction & image processing for computer vision, 3rd edn. Academic Press (Elsevier)
4. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinformatics 23(19):2507–2517. <https://doi.org/10.1093/bioinformatics/btm344>
5. Pan F, Song G, Gan X, Gu Q (2014) Consistent feature selection and its application to face recognition. J Intell Inf Syst 43(2):307–321. <https://doi.org/10.1007/s10844-014-0324-5>
6. Schölkopf B, Mullert K-R (1999) Fisher discriminant analysis with kernels. In: Proceedings of the 1999 IEEE signal processing society workshop neural networks for signal processing IX, Madison, WI, USA, pp 41–48
7. Yang J, Zhang D, Frangi AF, Yang J (2004) Two-dimensional PCA: a new approach to appearance-based face representation and recognition. IEEE Trans Pattern Anal Mach Intell 26(1):131–137
8. Kong H, Wang L, Teoh EK, Wang JG, Ronda V (2005) Generalized 2D principal component analysis. In: IEEE conference on IJCNN, Canada

9. Li M, Yuan B (2005) 2D-LDA: a novel statistical linear discriminant analysis for image matrix. *Pattern Recogn Lett* 26(5):527–532
10. Noushatha S, Hemantha Kumar G, Shivakumara P (2006) (2D)2 LDA: an efficient approach for face recognition. *Pattern Recogn* 39:1396–1400
11. Zheng W-S, Lai JH, Lid SZ (2008) 1D-LDA vs. 2D-LDA: when is vector-based linear discriminant analysis better than matrix-based? *Pattern Recogn* 41:2156–2172
12. Viuda-Martos M, Fernández-López J, Pérez-Álvarez JA (2010) Pomegranate and its many functional components as related to human health: a review. In: 2010 Institute of Food Technologists, comprehensive reviews in food science and food safety, vol 9, pp 635–654
13. Product profile of pomegranate Agri exchange, APEDA and National Horticulture Board of India. Accessed 11 Sept 2017
14. Vapnik VN (1995) The nature of statistical learning theory. Springer, New York; Nicole R. Title of paper with only first word capitalized. J Name Stand Abbrev (in press)
15. Vapnik VN, Chervonenkis AY (1971) On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab Appl* 17:264–280

Facial Expression Extraction and Human Emotion Classification Using Convolutional Neural Network



Amruta Khot and Anmol Magdum

Abstract Human facial expression extraction and analysis are called as facial imaging which is to be popular research area. It gives rise to face recognition and feature extraction, face shape detection, defining facial expression with emotions to identify user emotional state from images or video feed. Facial expression extraction extends analysis of key components which can be implemented by learning mechanism. The proposed system attempts to look at task of human emotion recognition using convolutional neural network and active appearance model (AAM) which is analytical model that works on geometric point of face. The system will detect face in real-time video streaming along with number of person present and classifying emotions of persons, i.e. neutral, happy, sad, angry, and surprise. The network is improved with effect of variation in hidden layer numbers.

Keywords Facial expression extraction · Convolutional neural network · Active appearance model · Emotion classifications

1 Introduction

1.1 Overview and Motivation

The past years have seen computers come into every aspect of our lives. An active area of research is in improving the interactions between humans and computers. Our project aims to improve the human–computer interaction by providing techniques for a computer to identify human emotion, and to tailor its behaviour accordingly. Detection of human emotion can improve interactions with machines in everyday

A. Khot (✉) · A. Magdum
Walchand College of Engineering, Sangli, India
e-mail: amruta.khot@walchandsangli.ac.in

A. Magdum
e-mail: anmol.magdum@walchandsangli.ac.in

life. For instance, a personal robot can detect the emotions of its user and respond accordingly. Smart houses can detect the mood of the residents and adjust parameters like lighting, air conditioning, and power usage of personal equipment accordingly [1].

Using a face detection module to detect a face and to build a convolution neural network-based classifier to classify the emotions of a person based on facial features extracted using digital image processing [1, 3].

Emotions analysis and detection using facial expression and classification from video streaming helps in real-time responses by the computers. Consideration of facial analysis by machine is emerging and useful in lots of applications which helps to detect face and extract facial feature for analysis of facial appearance and detection of emotions with classifications. This facial expression analysis is divided into three steps:

- (i) Face recognition
- (ii) Facial feature extraction
- (iii) Classifications of emotions.

1.2 Convolutional Neural Network

CNN is based on convolution function. In mathematics, convolution is operation of two functions F and G.

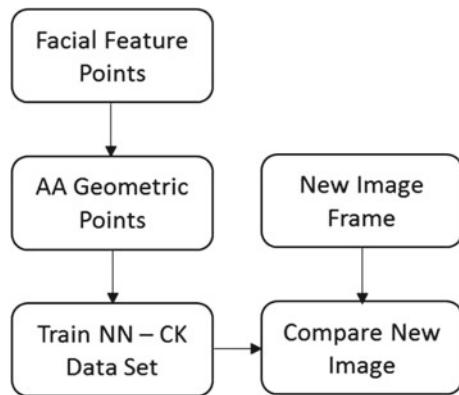
$$F(y) * G(y) = \int F(y - x)dG(y) \quad (1)$$

These two functions defines one shape is reformed by other. CNN is neural network with certain level of complexity having multiples of hidden layers. Also there is activation function at the end of each hidden layer. Each CNN layer learns filters of increasing complexity. The first layer learns basic feature detection filters, edges, corners, etc. The middle layer detects parts of objects. The last layer having higher representations recognizes full objects indifferent shapes and positions [3].

1.3 Active Appearance Model

Active appearance model (AAM) is geometric model of geometry which considers geometric points of face like nose, lips, and eyes. It is generative model which retrieves parametric description of object with optimization that matches statistical model of object dimensions with appearance of newly generated images [2, 4] (Fig. 1).

The algorithm aims to map pre-trained images to image to be generated or predicted by minimizing difference between two images by least square techniques.

Fig. 1 AA Model

Images are generated during training phase. The biggest advantages of using AAM is face get detected even if face is not straight or little bit tilted.

AAM forms an image outline which is considered to be pattern and customs a analytical model to compare the geometry points and texture of a newly captured image to original defined template. AAM extracts a number of geometric points on a face which gives facial features and is to be converted as a vector.

AAM can be defined as:

$$x = \bar{x} + Pb \quad (2)$$

x is the geometric dimension vector, e.g. a face as collections of points [2].

2 Proposed System

1. Using a face detection module to detect faces with different variations.

The first objective is to detect and crop an image from the video frame which could be from different people from different regions.

OpenCV programming library of real-time computer vision can be used.

2. To extract facial expressions using digital image processing and active appearance model algorithm.

First, the cropped image is converted to greyscale instead of a 3D array. The 3D array consists of RGB values which is quite difficult to process and takes a lot of time.

3. To build a convolution neural network to classify and validate the emotions of a person.

Then, this cropped grey-scaled image with only the edges in place of eyebrows, eyes, and mouth is passed through the classifier. The classifier is a pre-trained convolutional

neural network model which is trained on the Cohn-Kanade dataset [5]. The classifier predicts the accurate emotion based on the pre-trained model. Then, we display the label of emotion on the face in real time.

Cohn-Kanade Dataset

CK dataset consists of 100 faces whose age is in range of 18–30 years. 65% of face dataset are feminine, 15% of African-American and 3% of Asian-Latino [5].

3 System Architecture

I. Figure 2, indicates the architecture of CNN model used in our system. Our system takes live video stream from wireless camera and this feed to system. The laptop processes the video into frames and converts the whole stream into greyscale. 3D image is array of R, G, and B, and grey-scaled image is obtained using mean of R, G, and B values.

II. The model first detects faces using haarscde file and OpenCV library from these greyscale images. These detected faces are cropped and resized. And resized image is sent to network model for prediction which is series of hidden layers.

III. We start with sequential model with 16 filters, and kernel size is 7. The padding is set to ‘same’, and image array with input shape $48 * 48$ is passed to first hidden layer. There are five layers. In each layer, we have used batch normalization. It applies transformation such that it maintains the mean activation close to 0 and activation deviation to 1.

IV. The activation function for each hidden layer is RELU (Rectified Linear Unit). RELU being activation function in fully connected neural networks ranges values between 0 and 1 classifying emotions. RELU does not face gradient vanishing problem.

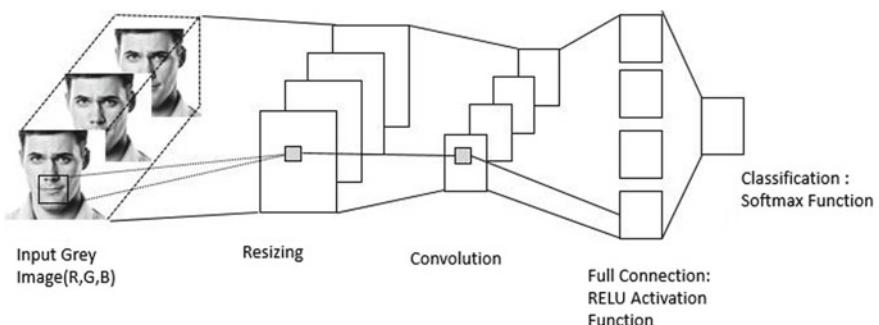
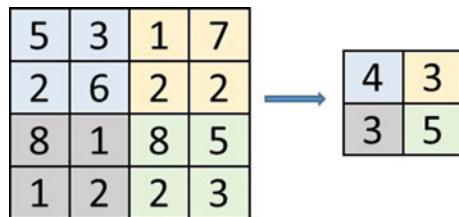


Fig. 2 System architecture using CNN

Fully connected network uses Softmax function as = $\max(0, \cdot)$.

V. Then, we have average pooling with pool size $2 * 2$ as:



The kernel size is gradually decreased from 7 to 5 then to 3 for remaining layers, while filters are increased gradually from 16 to 32, 64, 128, and 256, respectively. After each function, we add a drop out function 0.5 after each hidden layer so that the model is not over fitted to training dataset.

VI. As the final layer has to predict between five classes {Neutral, Happy, Sad, Surprized, Angry}, we use softmax activation function at the end of each layer.

The CNN model is combination of 96 epochs to generalize the learning process and enable better back propagation for weight adjustment.

After classifying the accurate emotions, a circle is plotted around face in video and then text of predicted emotions is displayed on screen.

4 Implementation

1. **Capturing the video:** The system can be coupled with the computer's webcam or any wireless webcam for this purpose.
2. **Image standardization:** It includes making uniform size image conversion of RGB to greyscale images for analysis purpose.
3. **Face detection:** Face detection from input image removes all the surplus things as surrounding of image and retains only relevant information from the face. This leads to different image processing techniques such as face segmentation and curvature features [1].
4. **Facial elements detection:** Regions of interests from nose to mouse are detected. Due to variations in faces and its alignment, the system uses active appearance model.
5. **Active appearance model (AAM):** The AAM algorithm interprets images unseen before by searching for the best match between the current example and that state of the model that minimizes the differences to it [4].
6. **Deep learning:** The system is built on a CNN model assisted with active appearance model for face detection and emotion classification. To reduce the complexity of the input data (in the form of feature map) to the CNN model, global average pooling is used [1, 3].



Fig. 3 Happy emotion

7. **Classify:** The classifier trained using above neural network and with the help of AAM will predict the emotion and draw a circle around it [3].

5 Result Analysis

5.1 Happy

The result in Fig. 3 shows, person with emotions happy. Where the image with yellow circle shows that emotion is completely happy, while green circle represents the change of a emotion of a person from one emotion to neutral.

5.2 Neutral

The result in Fig. 4, shows that the women in the is in the emotions neutral. Where the image with green circle shows that emotion is completely neutral, while the emotion of particular women is totally depend on the position of the eyebrows, nose, and a mouth. Green circle truly indicates that the women in the frame is completely neutral.

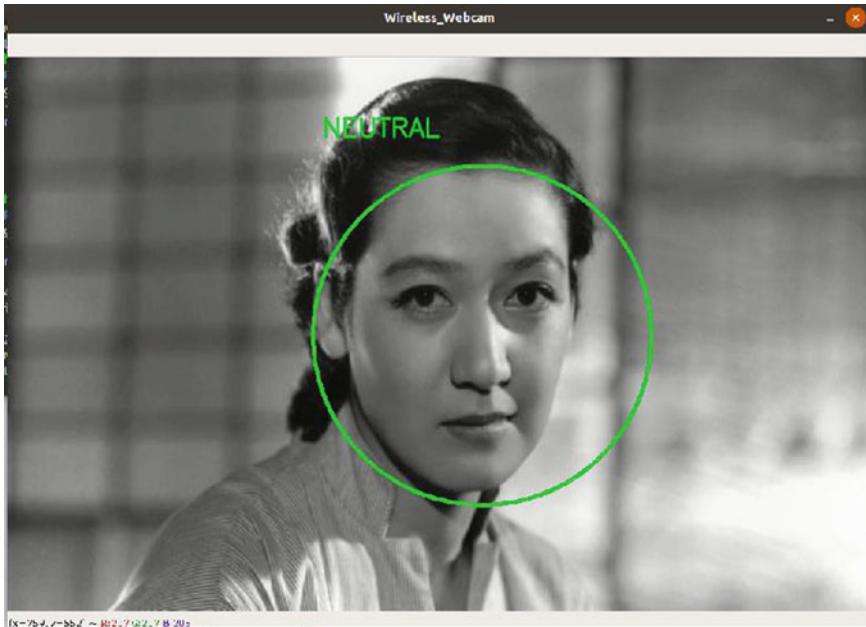


Fig. 4 Neutral emotion

5.3 Surprised

The result in Fig. 5 shows that the persons in the frame is in the emotions surprise. Where the image with cyan circle shows that emotion is completely surprise. But the emotion of particular person present in the frame is totally depended on the position of the eyebrows, nose, and a mouth.

6 Summary

For best results in case of emotion detection, the light conditions must be optimal (good) and delay between the systems must be less. If the webcam of a laptop is used, as a wired connected hardware, the real-time results displayed are excellent. Using the wireless camera of a smart phone might result in some delay due to the network latency. Although results could be pretty satisfying as today's smart phone has 8–13 mega-pixels camera. It also depends on the resolution of the camera used, for a person sitting or standing very far from the camera can give satisfactory results if the resolution of video is increased but then compromising on the time taken for the prediction.

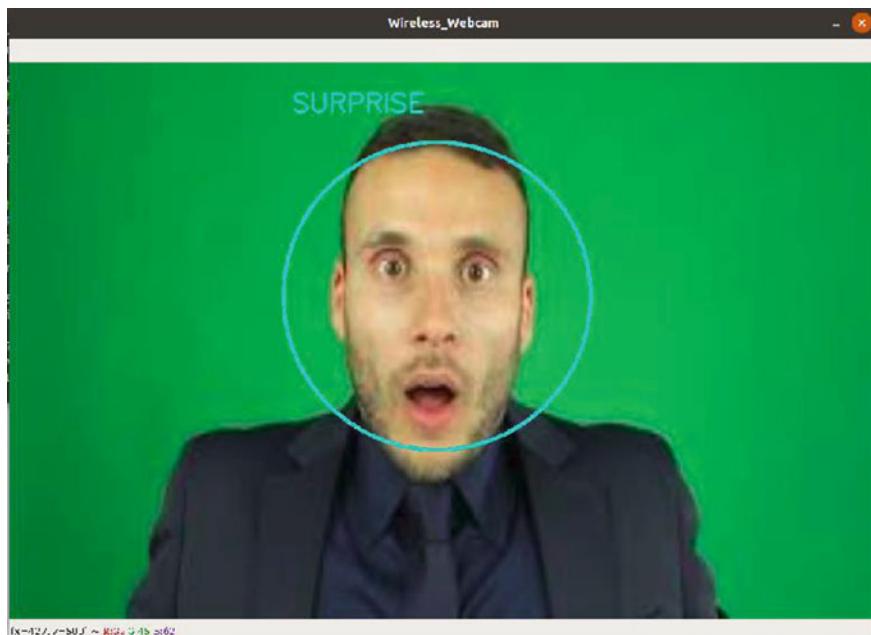


Fig. 5 Surprised emotion

The system also detects various faces in a single frame and gives highly accurate results for different emotions given by different people. To improve upon the further accuracy, we suggest using a high-end hardware machine such that it can use its GPU core as well for tensor flow for best results. In case of a person wearing spectacles, emotions like anger or sadness could be hampered as it depends on the eyes and eyebrows specifically.

References

1. Porzi L, Sangineto E, Zen G, Ricci E (2016) Learning personalized models for facial expression analysis. *IEEE Trans Multimedia*
2. Edwards GJ, Taylor CJ, Cootes TF (1998) Interpreting face images using active appearance models. In: Proceedings 3rd IEEE international conference on automatic face and gesture recognition
3. Matsugu M, Mori K, Kaneda Y (2003) Subject independent facial expression recognition with robust face detection using convolutional neural. *Neural Netw.* [https://doi.org/10.1016/S0893-6080\(03\)00115-1](https://doi.org/10.1016/S0893-6080(03)00115-1)

4. Li T, Zhou J, Tuya N (2017) Recognize facial expression using active appearance model and neural network. In: International conference on cyber-enabled distributed computing and knowledge discovery
5. The Extended Cohn-Kanade Dataset (CK+): a complete dataset for action unit and emotion-specified expression. [https://www.semanticscholar.org/paper/The-Extnded-Cohn-Kanade-Dataset-\(CK%2B\)3A-A-Complete-Lucey-Cohn/4d9a02d08636e9666c4d1cc438b9893391ec6c7](https://www.semanticscholar.org/paper/The-Extnded-Cohn-Kanade-Dataset-(CK%2B)3A-A-Complete-Lucey-Cohn/4d9a02d08636e9666c4d1cc438b9893391ec6c7)

Exploring Opportunities in Hydro Electric Power Plant with Heron's Fountain



Khude Anupam Tanaji and Patil Manoj Dhondiram

Abstract The impact of conventional hydroelectric power generation plants has lead to diminishing agricultural regions. Also, the construction of dams and reservoirs which essentially requires lots of lands has forced the transition of people and consumed forest areas as well. Hydroelectric power plants can affect the quality and the flow of water in the dam or river. Because of the hydropower plant, the oxygen level in the water decreases so this is very harmful to the biodiversity presented under the river water, stream, and dam or the area in which the power plant is been established. Another characteristic of hydropower plants is that the water used for hydropower generation is hardly recycled, i.e., fed back to the water reservoir. This paper explores the opportunity for use of Heron's Fountain to generate electrical power. MATLAB modeling of Heron's Fountain and simulation results to determine requisites for hydropower generation are discussed.

Keywords Hydroelectric power plant · Heron's fountain · Hydropower plant · Biodiversity effect · Kinetic energy · Electrical energy · Generation system · Natural affects · Clean energy · Dam affected people

1 Introduction

Heron's Fountain is nothing but the perpetual type motion machine or it is also a hydraulic machine. Normally hydropower generation plant is constructed in the region of dam, river, streams which stores the bulk amount of water. But this leads to some disadvantages as the water once is used for the generation of electrical energy cannot be used again or recycled. This makes the hydropower generation is dependent

K. A. Tanaji (✉) · P. M. Dhondiram

Department of Electrical Engineering, Annasaheb Dange College of Engineering and Technology, Ashta, Sangli, India

e-mail: anupamkhude@gmail.com

P. M. Dhondiram

e-mail: ndpatileps@gmail.com

on the season (rain & monsoon). The hydroelectric power plant is also affected by draughts, i.e., when water is not available in the reservoir then the plant cannot produce electricity. For water reservoir, land acquisition is the key that sweeps the agricultural land. Because of the hydropower plant more land goes under the water so the community reestablishment problem arises. Due to the insufficient water supply in the summer season, the hydropower plant is unable to produce electricity. To get rid of these drawbacks a Heron's Fountain is proposed to ensure continuous water flow. Some modifications are proposed to make this Heron Fountain a near-perpetual motion machine. Flow rate, water head, water pressure, kinetic energy of water are some of the parameters which play an important role in hydropower generation are modeled with MATLAB, and simulation results are discussed [1].

1.1 Hydropower Generation System

This is a renewable energy source that depending on the water flow cycle. Hydropower generation is cost-effective more reliable and mature energy generation system. Hydropower is been so largest generation system which contributes total of 16% of electrical energy in the world. More than 25 countries are 90% of demand fulfills by this hydropower generation. The hydropower plant is a more flexible source of generation it will fast responding to the fluctuations in load demand within a minute when the reservoir is available then the generated power electricity can be stored for the week, month and year also. In the conventional mainly the thermal power plant is not as much as flexible of hydropower generation plant. In the thermal power generation, more amount of coal is been used due to this the more carbon is been released in the environment in more amount so the thermal power generation is not eco-friendly or it cannot be constructed in the rural area on the other side in hydropower generation power plant the power is been generated by using the kinetic energy presented in the flow of water. Due to this kinetic energy the turbine blade shafts are started to rotate means there is been kinetic energy is converted into mechanical energy. This shaft is directly connected to the generator so there is been the conversion of mechanical energy into the electrical energy takes place. In this way, there is been a generation of clean energy so this generation plant can be constructed in the rural areas also because of its simple and harmless operation [2].

Figure 1 shows the layout of a conventional hydroelectric power plant. Following are the different types of hydropower generation based on various components:

- Based on Generation Capacity:
 - (1) Large hydropower plant- more than 100 MW
 - (2) Medium Hydropower plant—15–100 MW
 - (3) Small Hydropower plant—1–15 MW
 - (4) Mini Hydropower plant—100 KW to 1 MW
 - (5) Micro Hydropower plant—up to 100 KW
 - (6) Pico Hydropower plant—up to 5 KW

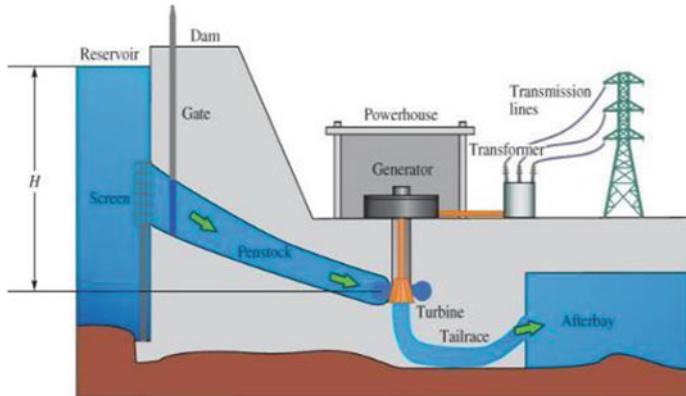


Fig. 1 Layout of hydroelectric power plant

- Based on Water Head:
 - (1) High Head—100 m & above
 - (2) Medium Head—30–100 m
 - (3) Low Head—2–30 m
- Based on Water Supply:
 - (1) Runoff river without poundage
 - (2) Runoff river with poundage
 - (3) Storage type plant
 - (4) Pumped storage plant
- Based on load nature:
 - (1) Base load power plant
 - (2) Peak load power plant

1.2 Heron's Fountain

The physicist, mathematician and inventor named Heron of Alexandria invents the Heron's Fountain which worked as the hydraulic machine that works without any kind of external force or any supply of energy. Firstly the Heron's Fountain has been constructed using hills to ensure the pressure but heron used vertical shaped table pot. It contains a tube and the airtight chambers to construct Heron's Fountain. It contains three chambers each contains equally air and the water at the same level. As the water pores in the first upper chamber or container, it will come in the next second chamber from the pipe using the gravitational force. As the water level is been changed from certain level the air pressure also increased from the chamber with the using of this unequal air pressure the water will flow from the tube to another chamber and so on

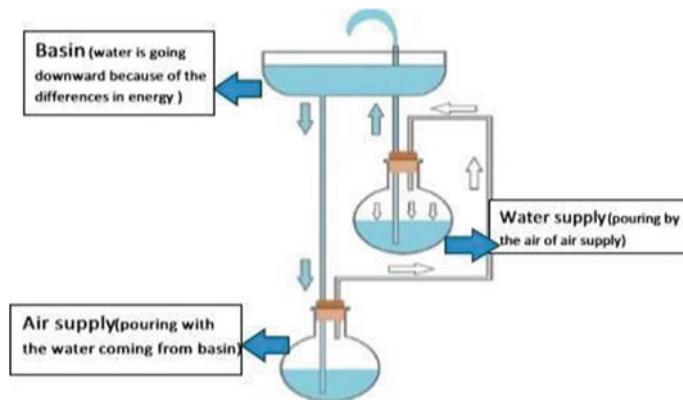


Fig. 2 Heron's fountain

the fountain is been worked with the help of using only air pressure and the earth's gravitational force. So this fountain is called as first perpetual motion machine [3]. Figure 2 has shown below displays conventional Heron's Fountain.

As we discussed in the above point hydropower generation plant requires the continuous supply of water flow with the pressure so we can use this modified heron's fountain to ensure the water supply to the turbine and the shaft blade to generate the electricity.

2 System Development

In conventional hydropower plant, all require continuous water supply, i.e., water dams, river, stream, large water reservoir to operate turbine to power production. In this, re-designed closed process of water supply to hydraulic turbines for energy power production. The main components selections will depend on or based on the capacity of the turbine it will be verified on the basis of the analysis of variance There are three main components of the design.

2.1 Hydraulic Ram Pump

This is the main component used to initiating re-circulation process of water in the system which will drive the hydraulic turbine to produce mechanical force. The component of hydraulic ram pump are waste valve and other moving parts is delivery check valve and the stationary parts are drive, delivery pipe, which will supply water from the source and deliver water in more height, respectively.

First delivery valve is been closed and waste valve is opened the water supply is supplied from an elevated source the water came from drive pipe make kinetic energy when waste valve closed the water open delivery check valve due to compressed air the water gets high elevation after this the water flow slows down because of the lower air pressure so water gets reversed in this manner this ram pump is worked.

$$qh = QH \quad (1)$$

where

Q —Output flow of system (gpm)

h —Output height (feet)

Q —Drive flow (gpm)

H —Drive head (feet)

After considering efficiency then the Eq. 1 becomes,

$$\dot{\eta} = qh/QH$$

Therefore,

$$q = QH\dot{\eta}/h \quad (2)$$

To calculate hydraulic ram pump power,

$$P = 9.81(q/60) h$$

where,

P —Power (watt)

q —Water flow (lt/min)

h —Head delivery (m)

9.81—coefficient of gravity.

2.2 Hydraulic Turbine

The water flow towards the blades by using these force blades spins the turbine shaft due to this water floating power transferred in the mechanical power and the turbine shaft connected directly to the generator. Mechanical energy is transferred in the electrical energy in this manner.

2.3 Modified Heron's Fountain

The original Heron's Fountain invented in the first century by AD inventor, physicist and Mathematician Heron of Alexandria. But it doesn't have a perpetual capability for circulating the water so modify the design of fountain to make more efficient perpetual water flow motion [3].

Now container A contains the head of water which will come into container B by making the use of the earth's force of gravity. Container B and container C both are having the same air pressure and the same level of the water which are connected to each other in an air compact format. In starting container B and container C are fixed with the same level but the water came from container A to the container B the water level in the container B increased so the air pressure and the water pressure in container B and C separated due to this the water will flow from pipe to the container C and due to surface tension of water and the increased air pressure the water started to flow from the pipe to the container A with high pressure and this cycle continuously works. Figure 3 shows the Modified Heron's Fountain below.

To find the pressure in the water we use Bernoulli's principle which states that in the liquid dynamics as the pressure decreases then the speed of water is been increased, which means the speed of water is inversely proportional to the pressure of air presented in the container. The equations are as given below [4].

$$P_1 + \frac{\rho V_1^2}{2} + Pgh_1 = P_2 + \frac{\rho V_2^2}{2} + Pgh_2$$

To find the pressure difference in two finds use the Pascal's law

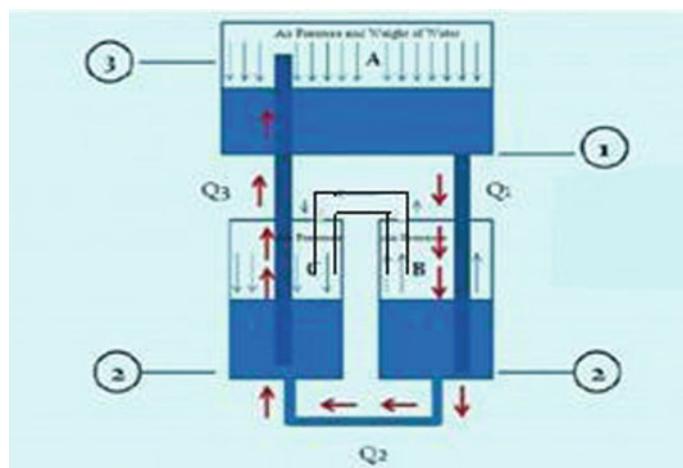


Fig. 3 Modified Heron's fountain

$$\Delta P = P_2 \cdot P_1 = \rho g h_2 \cdot \rho g h_1 = \rho g (h_2 - h_1)$$

Pressure of the water and air while water came downward from container A to container C at point Q in the above-given figure is

$$P_{\text{air}} = P_0 + \rho g h_1$$

The pressure of water at point Q₂ is,

$$P_{\text{water}} = P_0 + \rho g h_2$$

The pressure of water at the jets out or at the tip of the fountain pipe is,

$$\Delta P = P_{\text{air}} - P_{\text{water}} = \rho g (h_1 - h_2)$$

This way found the pressure and the flow rate of water at the tip of fountain theoretically [5].

3 Matlab Modeling

So the principle of Heron's Fountain worked or simulated in the MATLAB software so that purpose proper blocks are selected from the directory of the MATLAB and the block diagram is as shown in Fig. 4.

The selected blocks are from the Simscape library discussed deeply below,

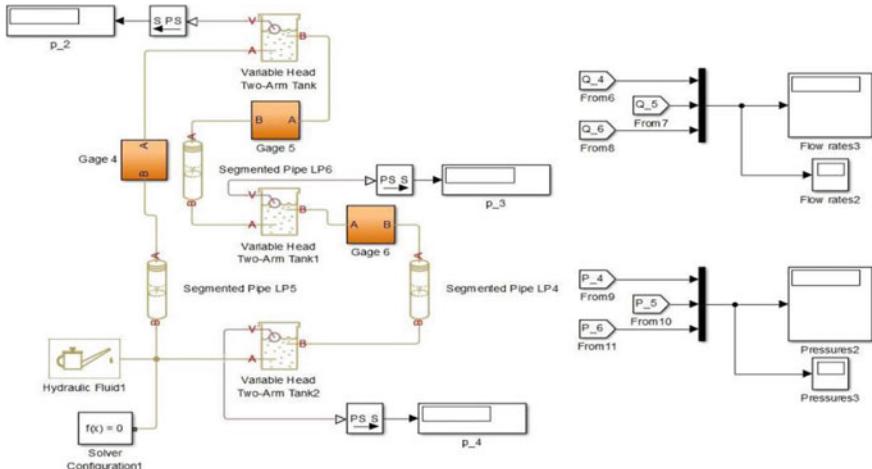


Fig. 4 Block diagram of program in MATLAB

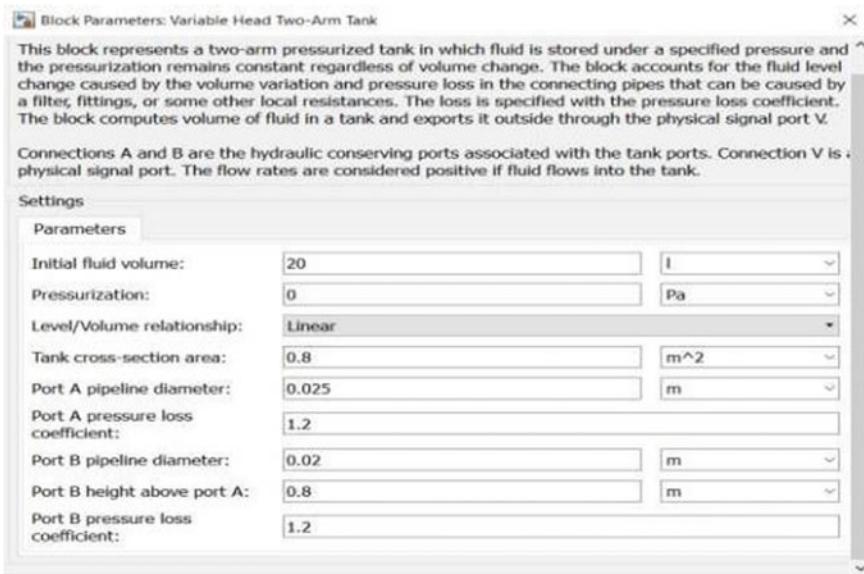


Fig. 5 Parameters of reservoir tank

3.1 *Variable Head Two Arm Tank 1*

There are three tanks used and all connected in the manner of the Heron's Fountain design. The parameter of tank 2 and tank 3 are the same and the parameter of tank 1 is different and they are as shown in Figs. 5 and 6.

3.2 *Hydraulic Fluid*

The hydraulic fluid tank is used to choose the type of fluid which will flow from the pipe and we choose the water as a flowing fluid an all parameters are inbuilt for the water as shown in Fig. 7.

3.3 *Segmented Pipes*

This pipe is connected for the way to flow of the water and the characteristics of the pipes the roughness of the pipe inner surface, internal diameter, pipe length, shape, dimensions of pipes, etc. Figure 8 shows the parameters of the segmented pipe.

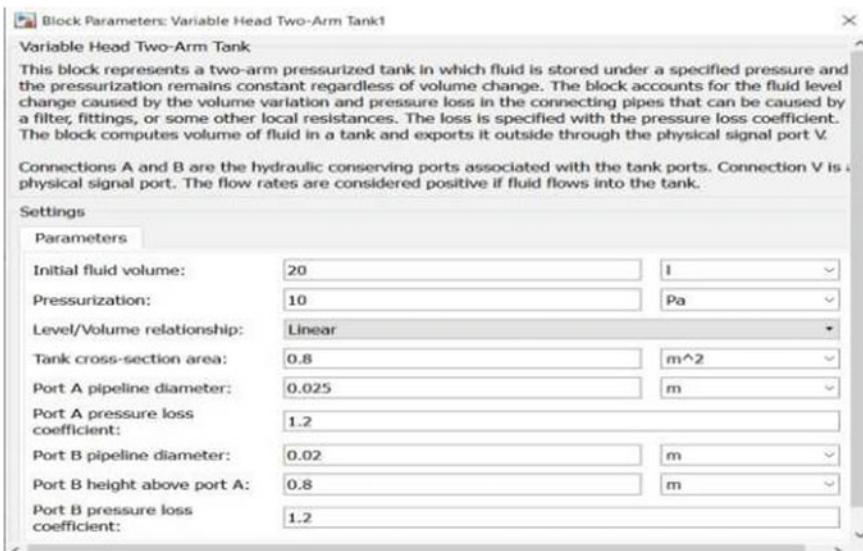


Fig. 6 Parameters of other two tank

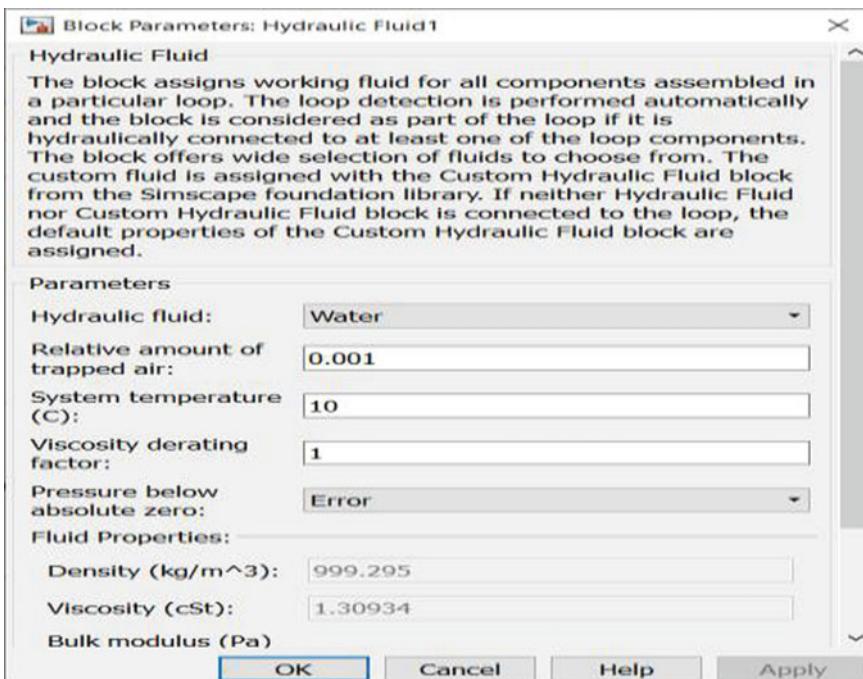


Fig. 7 Parameters of the hydraulic fluid

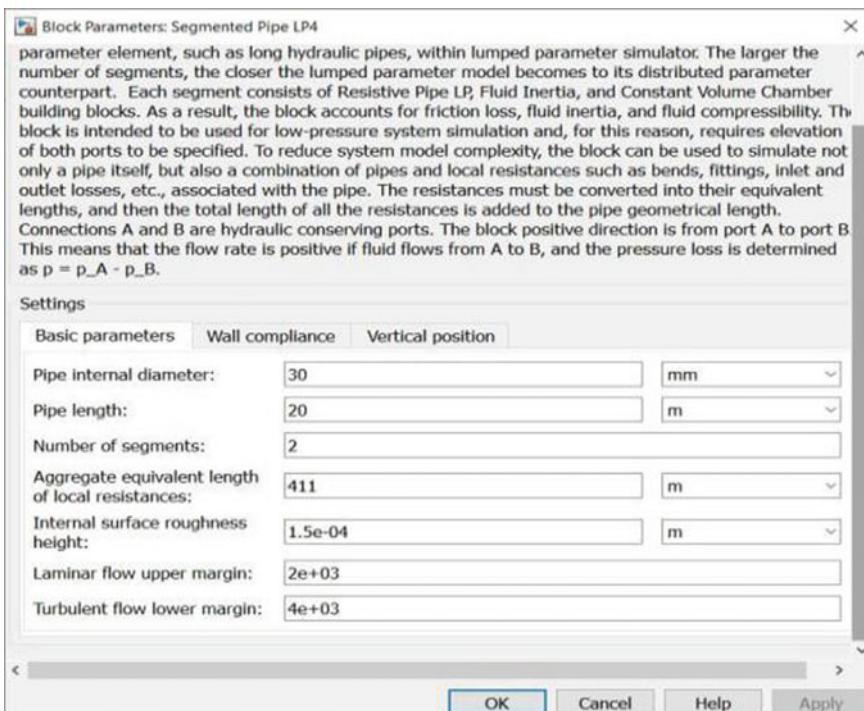
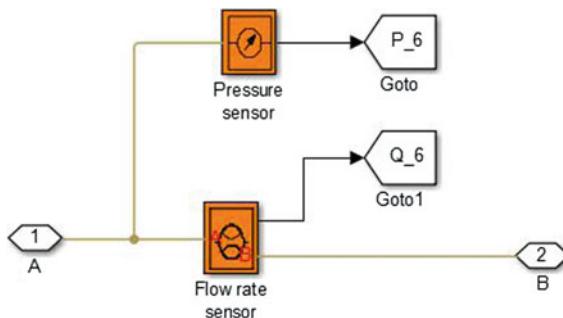


Fig. 8 Parameters of the segmented pipe

3.4 Gages

These blocks are used to measure the pressure and the flow of water in the pipes for this the subsystems are created in that the flow rate measurement meter and the pressure meter is connected and it is as shown in Fig. 9.

Fig. 9 Subsystem in the gages



3.5 PS Simulink Converter

Using this block the converter converts all readings or outputs to unit less only the magnitude value is been displayed.

3.6 Go to and from Block

The go to blocks are connected in the subsystem to give the call to from bus to display the values in the display.

3.7 Displays

These blocks are been displayed in used to display the output the flow rate and the pressure value is this blocks in the form of numerical value

3.8 Scopes

These blocks are used to display the results and the graphical representation to display the results

4 Results and Discussion

The values of flow rate and the pressure of water are displayed in the blocks of display and they are as given Fig. 10.

After running successfully MATLAB program got the pressure of water and the flow rate of water at the end of every tank and display it separately. The blocks Q4, Q5 and Q6 have displayed the flow rate of water presented at the end of tank 1, tank 3 and tank 2, respectively and the blocks P4, P5, and P6 displays the pressure of water presented at the end of tank 1, tank 3, and tank 2, respectively. Flow rate and the pressure of water presented at the end of tank 3 are the same as the flow rate and pressure of water which is emerging from the tip of fountain and used to power generation.

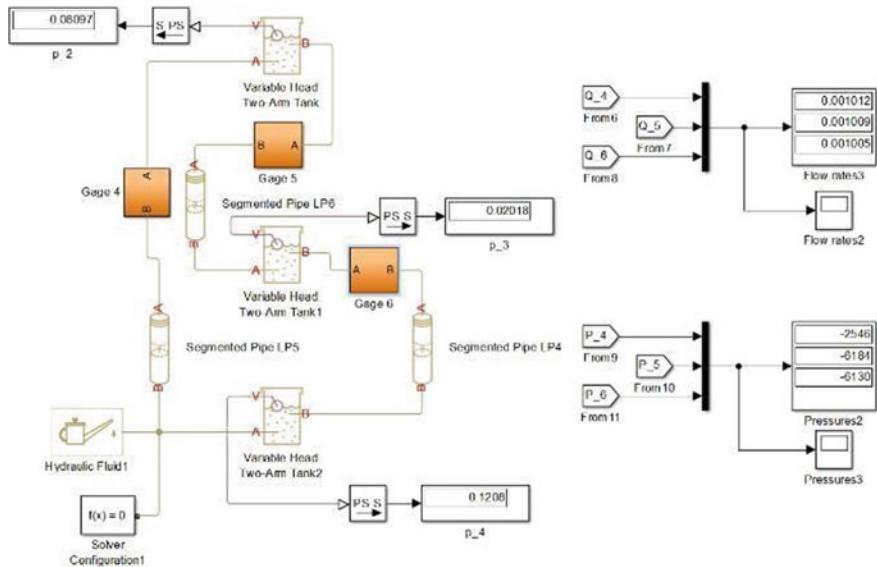


Fig. 10 Output of the MATLAB simulation

5 Conclusion

The flow rate of the water and the pressure of the water presented in pipe at the tip of the fountain is more efficient for generating the power with the proper selection of turbines we can generate the efficient power with compact design and the water is also reused in this way we can design the hydropower plant in the region where the dam, river, and stream are absent and also called as a renewable power source as the water is reused to generate power and it is eco-friendly and generates clean energy and also it is very economical in cost. So this is totally compacted construction so it does not affect nature and also the required reservoir is also fitted in the small region so there is no wastage of agricultural lands for the water reservoir and this model can be constructed in rural areas also.

References

- AK Yahya, Munim WNWA, Othman Z, Pico-Hydro power generation using dual pelton turbines and single generator. In: IEEE conference paper, Mar
- International Renewable Energy Agency, Renewable energy technologies: cost analysis series, Irena Working Paper, vol. 1, Power Sector Issue 3/5, 4–13, June
- Farzanegan RS (2018) Heron's fountain as a hydraulic machine, Tehran/Iran, IYSIE, pp 14–16
- Agbanlog RC, Chen G (2014) Mini hydro-electric power plant with recirculated water power source. In: Guan Y, Liao H (eds) Proceedings of the 2014 industrial and systems engineering research conference. United States of America

5. Rahman MS, Nabil IM, Alam MM (2017) Global analysis of a renewable micro hydro power generation plant. In: AIP conference proceedings 1919, Dec



Mr. Anupam Tanaji Khude was born in Satara, Maharashtra, India, in 1997. He has received his B.E. Degree in Electrical Engineering from Mumbai University Mumbai, Maharashtra, India in 2018. He is currently working M-tech Degree in Electrical Power Systems from Annasaheb Dange College of Engineering and Technology, Ashta, Sangli, Maharashtra.



Mr. Manoj D. Patil was born in Sangli, Maharashtra, India, in 1987. He has received his B.E. Degree in Electrical Engineering from Shivaji University Kolhapur, Maharashtra, India in 2009, and the M.E. degree in Electrical Power Systems from Government College of Engineering Aurangabad (which is affiliated to Dr. Babasaheb Ambedkar Marathwada University Aurangabad), Maharashtra, India in 2011. He is working as Assistant Professor at Annasaheb Dange College of Engineering and Technology, Ashta, Sangli, Maharashtra since July 2011. He is currently working toward the Ph.D. degree with the Division of Electrical Engineering at Dr. Babasaheb Ambedkar Technological University, Lonere, Raigad, Maharashtra, India

A Comprehensive Survey on Application Layer Protocols in the Internet of Things



Hemant Sharma, Ankur Gupta, and Madhavi Latha Challa

Abstract The number of connected IoT devices deployed globally is increasing at a tremendous rate. The IoT has a wide range of application domains providing extensive IoT-based services. The basic idea is to deliver a new set of applications where smart devices collaborate without any human interference. The variability and visibility of IoT based services lead to the development of the full spectrum of protocols. With the exponential growth of applications, it is essential to analyze the existing application layer protocols being used to exchange information among devices. In this paper, a detailed analysis of existing popular application layer protocols like Constrained Application Protocol (CoAP), Message Queue Telemetry Transport (MQTT), Advance Message Queuing Protocol (AMQP), Extensible Messaging and Presence Protocol (XMPP), etc. has been done to categorize them based on well-known properties such as architecture, energy consumption, reliability, QoS, and security aspects.

Keywords Internet of things · CoAP · MQTT · XMPP · AMQP

1 Introduction

The Internet of Things is becoming very popular nowadays both from the technical and commercial point of view due to its simplicity, low cost, and easy deployment [1]. IoT is emerging at this rate because of its various real-time applications deployed in different domains like transportation, energy, healthcare, agriculture, education,

H. Sharma · M. L. Challa
University of Gondar, Gondar, Ethiopia
e-mail: hem.s1209@gmail.com

M. L. Challa
e-mail: saidatta2009@gmail.com

A. Gupta (✉)
MNIT Jaipur, Jaipur, Rajasthan, India
e-mail: gupta1990.cs@gmail.com

smart vehicles, environment, industrial automation, etc. The IoT can be defined as “Enabling the physical objects to sense, think, communicate, control, and have opportunities to interact and collaborate with other physical objects over the network.” To enable these abilities, the physical objects are embedded with electronic computers, sensors, software, and network to monitor collecting and exchanging data among them, and hence called smart objects. These smart objects must be low-cost, energy efficient, secure, and interoperable with software applications. Figure 1 shows the domain-specific applications with smart objects performing their supposed task in the vast field of applications of IoT [2].

According to Cisco IBSG [3], IoT starts from a point when the number of devices connected to the internet exceeds the number of people connected. In 2003, 500 million devices were connected to the internet, and the world population was about 6.3 billion giving the ratio of 0.08 devices to people. Therefore, IoT didn't exist at that time based on CISCO's definition. IoT came into existence between the years 2008 and 2009, where the number of devices exceeds the world population. CISCO predicted that 50 billion devices would be connected by the year 2020 in a world with a projected entire population of 7.6 billion. It gives a ratio of nearly seven devices to people having approximately 80 times growth overpopulation increase, as shown in Fig. 2. It may be noted that the internet never remains static, these estimates are based on the information known to be true today, and also the ratio prediction is based on

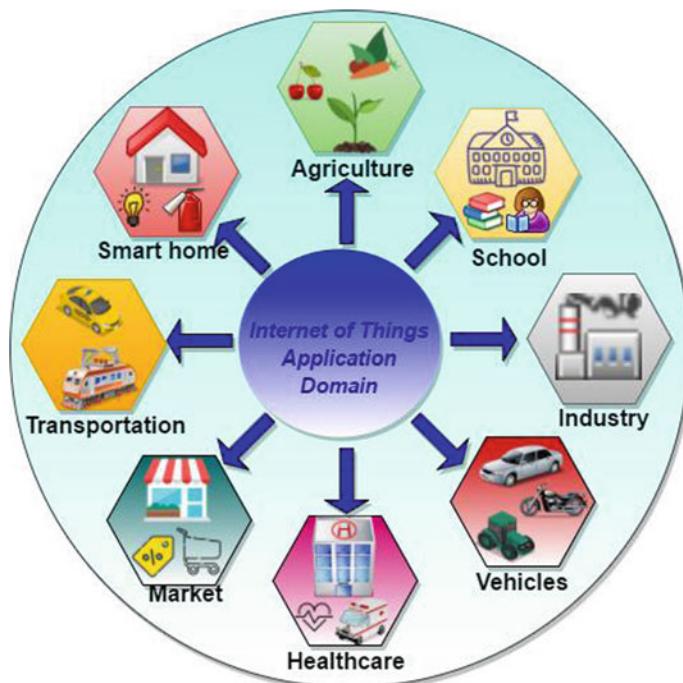


Fig. 1 IoT application domain with the potential market

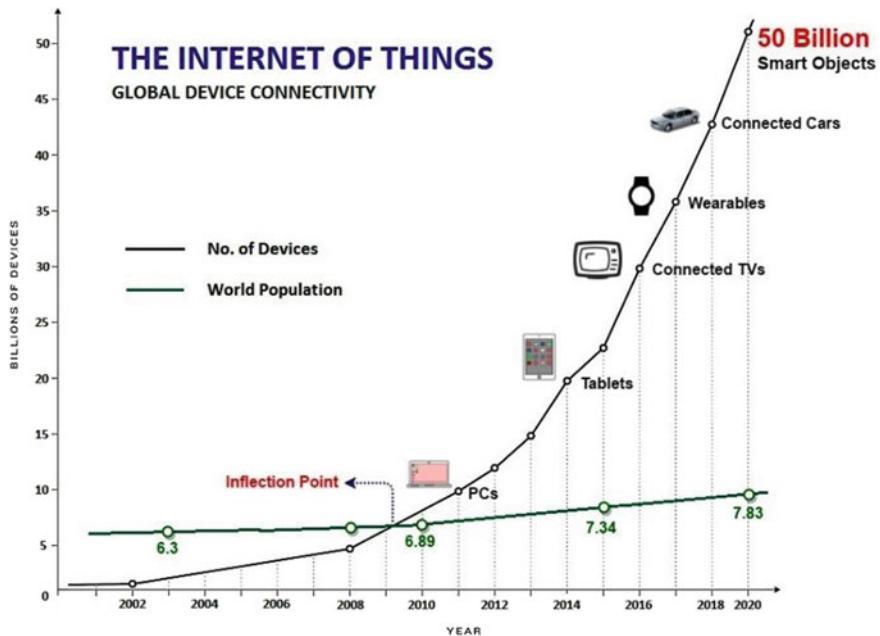


Fig. 2 Cisco IBSG IoT prediction

the population of the whole world although much of which is not connected to the network yet.

The remaining part of this paper is structured as follows. Section 2 describes the industrial opportunities for IoT device developers. Section 3 gives an in-depth survey of specific application layer protocols and their uses in the real world. Section 4 provides the comparative evaluation and analysis of these different application layer protocols. In the end, Sect. 5 concludes the paper with pointers to future work.

2 Industrial Opportunity

Billions of physical devices or objects connecting to the internet are attracting many big organizations to invest their trillions of money in IoT projects. According to the McKinsey Global Institute [4], in the last 5 years, the number of increase in connected devices reaches 3 times. This also gives a great market and industrial opportunities for device manufacturers, software developers, and service providers. IoT has the biggest economic impact on healthcare [5] and manufacturing applications [6]. Many medical applications including remote health monitoring services, diagnosis, treatment, fitness programs, etc. are predicted to create about 2.5 trillion dollars annual growth to the global economy by 2025, as shown in Fig. 3.

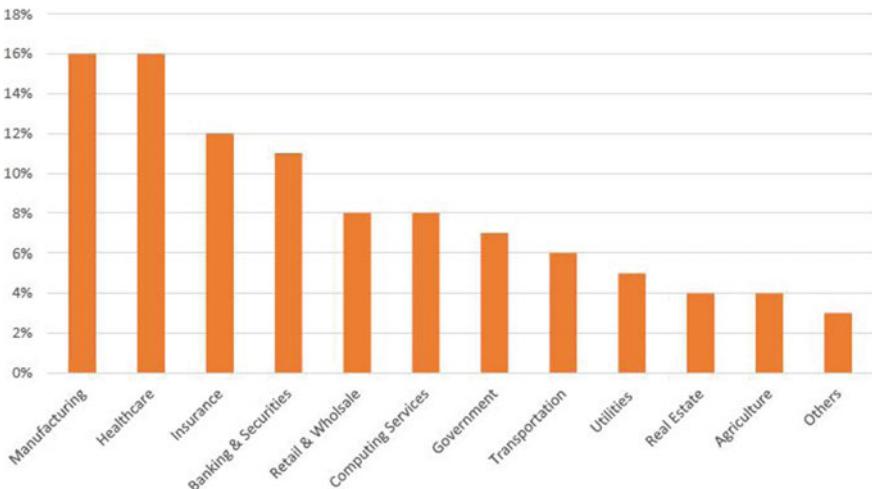


Fig. 3 IoT market

All these facts point to the fast growth and potentially powerful impact of the IoT in the coming years. By considering the above estimates, the challenge in IoT is not a single technology challenge but the entire area of computing. IoT redefines the thinking of the fundamentals of computation and communication paradigms, software, security, and privacy issues. The main central issue in IoT is how to make the full interoperable interconnected devices possible. To provide them with a higher level of smartness by enabling them to adapt to dynamic and autonomous behavior, guaranteeing trust, privacy, and security.

3 IoT Application Protocols

The application layer is the core layer of every communication system model. It is responsible for the effective communication services within the networks and their interconnection. There are many different application protocols that exist with different functionalities in different perspectives. IoT application layer protocols can be categorized into two categories based on the communication-based architecture model. These models are Publish/Subscribe model and Request/Response model. Some of the protocols existing today are IETF's CoAP [7], IBM's MQTT [8], Jabber's XMPP [9], OASIS's Standard AMQP [10], MQTT-SN [11], etc. Next, we will discuss these protocols in detail with their advantages and disadvantages.

3.1 CoAP (*Constrained Application Protocol*)

The Internet Engineering Task Force (IETF) Constrained RESTful Environments (CoRE) Working group has created a specialized web transfer protocol for constrained devices, known as CoAP [12], for M2M and IoT applications. CoAP is designed to implement a subset of REST architecture [13] with common HTTP for simplified integration with the web while offering specialized features for M2M applications such as built-in discovery, multicast support, low overhead, asynchronous message exchanges and simplicity for constrained environments. Unlike HTTP, CoAP uses UDP as a transport layer protocol making it more suitable for IoT applications.

CoAP is a request/response interaction model similar to client/server architecture and uses HTTP methods such as GET, POST, PUT and DELETE to request an action on a resource (using Uniform Resource Identifiers) on a server. CoAP uses REST-CoAP proxies to interconnect with HTTP which acts both as a server and a client for internetworking communication. Figure 4 shows the overall functionality and the mapping between HTTP and CoAP enabling traditional web clients to access CoAP servers transparently.

CoAP logically uses a two-layer approach, the request/response sub-layer, and the messaging sub-layer. The request/response sub-layer handles the REST communications using Method and Response Codes, and the messaging sub-layer deals with UDP transport layer and the asynchronous message interactions. Figure 5 shows the CoAP protocol stack generally use in the constrained environment.

CoAP defines four types of messages—Confirmable message, Non-Confirmable message, reset message, and Acknowledgment message. Some of these messages include Method and Response codes which allows them to carry requests or responses. Since the exchange of messages is asynchronous between two endpoints over UDP for communication, it marks a message as Confirmable (CON) to provide

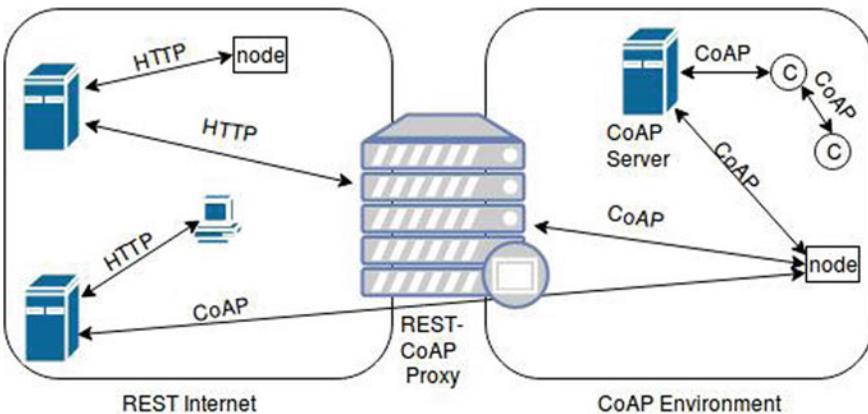
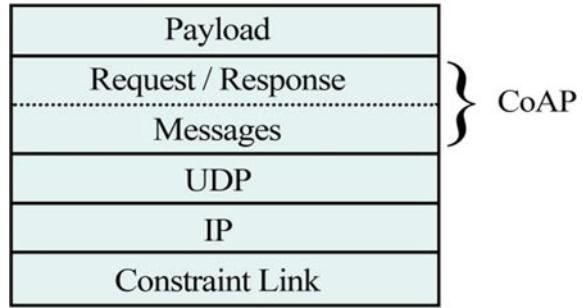
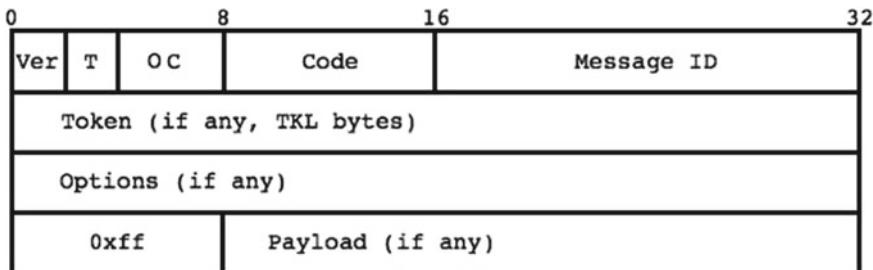


Fig. 4 CoAP architecture

Fig. 5 CoAP protocol stack

reliability. Sender waits for an Acknowledge (ACK) message and retransmits the confirmable message with exponential back-off after the default timeout. A recipient may confirm the CON message by sending a corresponding ACK message or it may reject by sending Reset (RST) message if it is unable to process the confirmable message because of a lack of context information. A non-confirmable message (NON) is sent if the CoAP message does not require reliability. Duplicate Non-confirmable messages are detected by Message-ID which is always present in all header formats. CoAP messages are simple and use binary format to encode. Figure 6 [7] shows the header format of CoAP message. Each message contains a 4-byte fixed-size header followed by a Token value of variable length between 0 and 8 bytes long, again followed by an optional payload field in datagram packet.

CoAP is suitable for low constrained devices because of features like resource observation, Block-wise resource transport, resource discovery, interacting with HTTP and DTLS security [7, 12, 14]. CoAP client and server are implemented in many languages to be directly used in any programs. Some of the implementations are libcoap in C, aiocoap in python 3, californium in java, cantcoap in C/C++, copper in javascript, icoap in objective C, coaprs in rust [15, 16], etc.

**Fig. 6** CoAP header format

3.2 MQTT (*Message Queue Telemetry Transport*)

In 1999, IBM's Andy Stanford-Clark and Eurotech's Arlen Nipper collaboratively developed an asynchronous publish/subscribe “lightweight” messaging transport protocol for Machine-to-Machine communications in wireless networks. MQTT [8] became an OASIS standard in 2014 and recently approved as an ISO standard (ISO/IEC 20922) in January 2016. MQTT protocol is well suited for the IoT applications which require less bandwidth and low power consumption in unreliable networks.

The Publish/Subscribe messaging is an event-driven and requires a broker to push messages to clients. MQTT broker acts as a central point of communication and is responsible for dispatching all the messages between the senders and the interested recipients based on the topic contained within the message. A sender publishes a message with a certain topic to the broker and all the clients that are subscribed to that topic with the broker receive the message. Clients do not require to know each other rather all the communication takes place over the topic. Figure 7 clearly shows the architecture of MQTT using a publish/subscribe pattern providing simple and highly scalable solutions for the IoT applications without having dependencies between the Publisher and the Subscriber.

In contrast to HTTP where the client pulls the information from the server, MQTT broker pushes all the information to the interested clients. Therefore, each MQTT subscriber has to maintain a permanent open TCP connection to the broker. If the subscriber's connection is broken due to any reason, the MQTT broker buffers all the messages and deliver it later when the client is back online. To ensure reliability, MQTT uses three levels of QoS messages. QoS 0 (At most once delivery) means that client sends the packet only for one time and does not wait for the ACK packet from the broker. If the data gets lost, there will be no retransmission of that packet. QoS 1 (At least once delivery) means that the user must ensure the delivery of packet for at least one time. After transmitting the PUBLISH message, client waits for the predefined amount of time to receive PUBACK packet or it retransmits the data. QoS 2 (At most

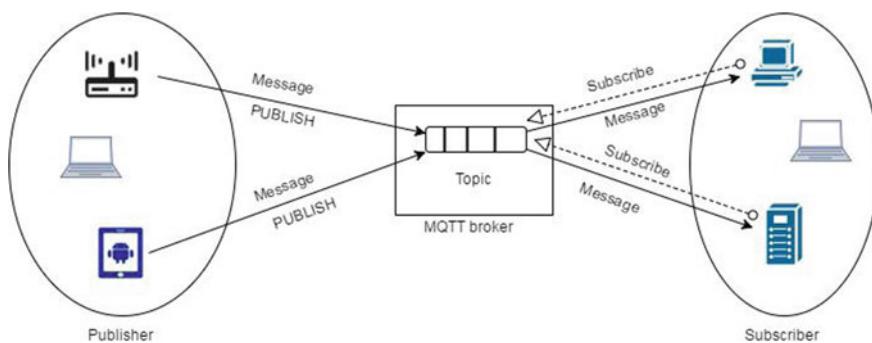


Fig. 7 MQTT architecture

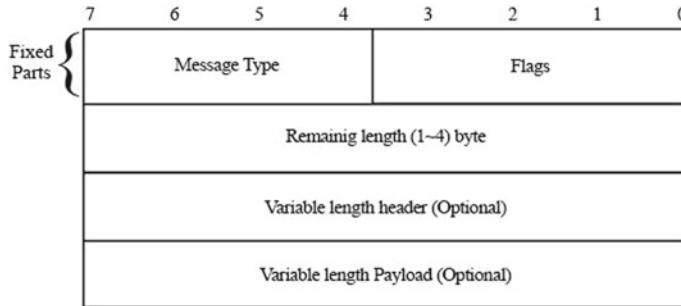


Fig. 8 MQTT header format

once delivery) means that packet must be delivered to the subscriber-only one time. The client sends the PUBLISH message and waits for the PUBREC message. After receiving PUBREC message, the Client sends the PUBREL message to the broker and discards the references to the published data.

MQTT protocol works efficiently by exchanging series of MQTT control packets in a systematic way. All the MQTT Control packet format contains three parts in the order illustrated in Fig. 8 [8].

The first part is the fixed header of 2 bytes which is present in all packets. The first byte of a fixed header tells the MQTT Control packet type and their specific flags. The remaining length field tells the total number of bytes present within the current packet including data in the variable header and the payload. The next field is the variable header which is present in some of the MQTT control packet types. The last field is the payload which is also optional.

MQTT protocol is an ideal messaging protocol for the various applications in IoT and M2M communications such as Health care, Monitoring, smart electricity metering, etc. and used by many companies for their communications. Some of the real-world applications are IBM mobile messaging [17], Facebook messenger, EVRY-THNG IoT platform, Amazon Web Services [18], AdafruitIO cloud service [19], etc. and many more.

3.3 MQTT-SN (MQTT for Sensor Networks)

MQTT-SN is a publish/subscribe protocol [11] specially designed for the wireless sensor networks having the main focus on the constrained devices. It is a variant of the MQTT protocol adapted to the nature of the wireless communication environment. MQTT-SN does not require TCP/IP stack for its operations therefore it is supported by more protocols like Zigbee, Z-Wave, and so on.

MQTT-SN uses 2-byte Topic ID instead of Topic name in the header which makes it lighter than MQTT and ideal for implementation on low-cost, battery-operated devices. It uses MQTT-SN gateway and forwarder to translate packets into

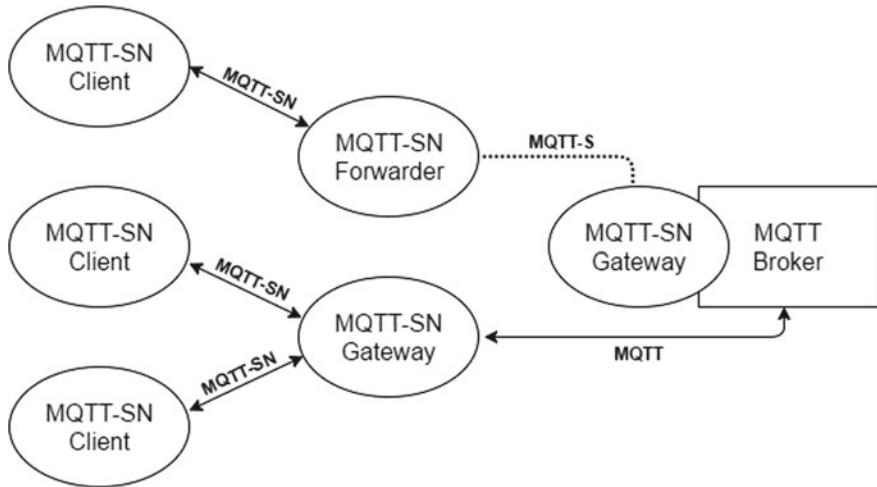


Fig. 9 MQTT-SN architecture

the standard MQTT format [20]. The Gateway then connects to the broker using TCP connection (Fig. 9).

3.4 AMQP (*Advance Message Queuing Protocol*)

AMQP is an open standard protocol [10] built by John O'Hara at JPMorgan Chase in the UK. The protocol efficiently supports a wide range of messaging applications and communication patterns. AMQP is a message-oriented middleware protocol including features like message orientation, routing, queuing, security, and reliability.

Like MQTT, AMQP is an asynchronous Publish/Subscribe messaging [21] which requires a third-party broker to deliver messages to the interested clients. AMQP broker uses two key components for communication: Exchanges and Queues, as shown in Fig. 10.

Exchange routes the messages to their appropriate queue. Message queues are used to store the messages and then send them to the subscribers. AMQP exchanges ensure the interoperability between different clients.

AMQP is a binary, wire-level protocol providing reliable communication with message-delivery guarantees using QoS such as at least once, at-most once, and exactly once delivery and uses TCP as reliable transport layer protocol. AMQP defines the messaging scheme as bare message and annotations may be added by the intermediaries before or after the bare message during transmit to ensure end-to-end encryption and integrity check as shown in Fig. 11.

AMQP is widely used in business applications and financial trading, RabbitMQ is an open-source broker for AMQP. Some of the real-world applications of AMQP

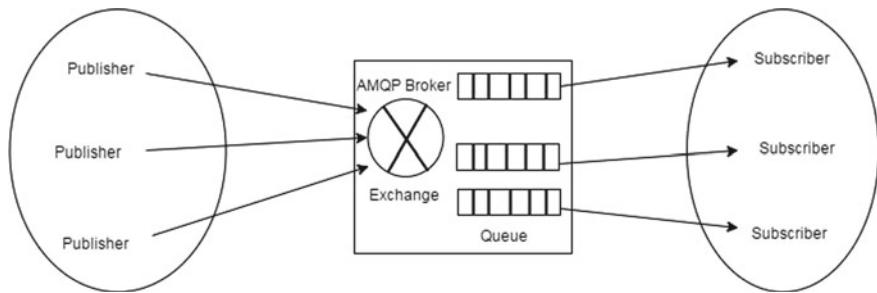


Fig. 10 AMQP architecture

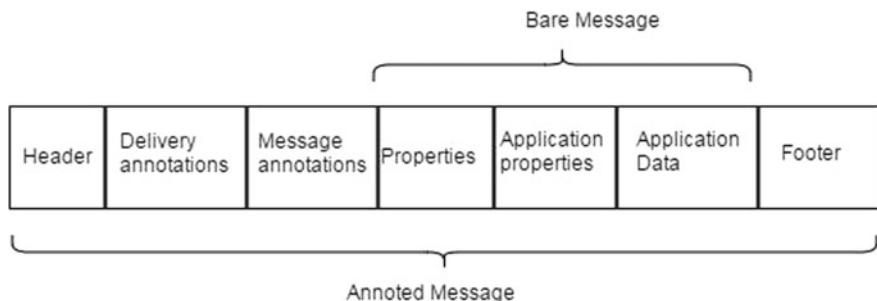


Fig. 11 AMQP message format

are Apache Qpid, IIT Software SwiftMQ messaging product, INETCO's AMQP protocol analyzer, JORAM, Kaazing's AMQP web Client, Microsoft's Windows Azure Service Bus, StormMQ, MQlight, etc. [22].

3.5 XMPP (*Extensible Messaging and Presence Protocol*)

XMPP is an application layer communication protocol based on XML for message-oriented Middleware and is mainly used for Instant Messaging, group chat, gaming, VoIP, video calling, file transfer, presence information, and social networking services. In 1999, Jeremie Miller of Jabber Technology developed an open-source software “jabberd server” which led to the formation of XMPP protocol. IETF standardized XMPP protocol in 2004 and used as an IETF instant messaging and presence technology.

XMPP uses decentralized Client-Server architecture, i.e., anyone can create his or her own XMPP server and can communicate with the rest of the network using DNS. A unique XMPP ID, also called JID (Jabber ID), is given to every user of XMPP in a network from their server. This user ID looks similar to an email address containing username and Domain name such as user@example.com. XMPP is

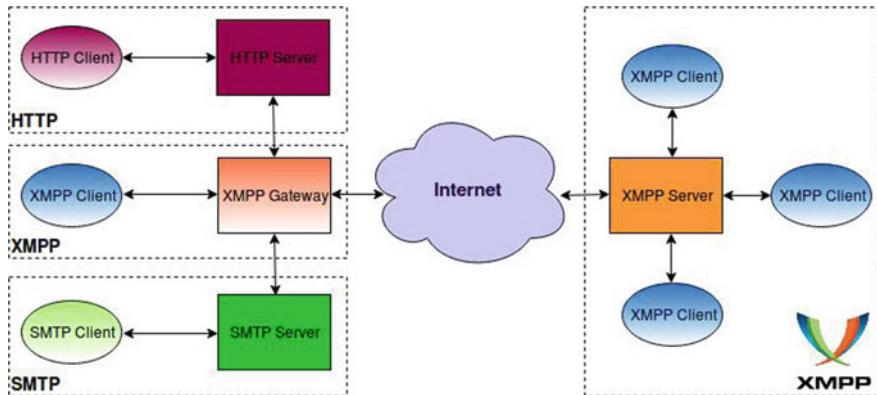


Fig. 12 XMPP communication

secure, scalable, extensible, flexible, standard which provides in-built support for authentication and encryption (Fig. 12).

The core technology of XMPP is streaming XML stanzas over the network. As the XMPP runs TCP on a transport layer, it maintains an always open TCP connection between client and server which exchanges XML snippets over the stream. XML snippets or stanzas are of three types: message, presence, and iq.

4 Evaluation

There are many different application layer protocols exist which can be used in different IoT scenarios. Choosing the best protocol among the pool of protocols is a hard task and depends on the requirement of the applications in which protocol has to be used. No real-time evaluation has been done till now which can compare all the existing protocols but pairwise comparisons and evaluation have been done in many surveys. Since each protocol can perform widely different in different scenarios and platforms. In a constraint environment with low power and computation, REST-based CoAP can perform efficiently while in the case where devices run on battery and require only exchange of messages MQTT is considered a good choice, and for the business applications where devices are not low constraint AMQP works more efficient. So it is not at all justifiable to give a single prescription of one protocol for all the IoT applications but a comparison based on common parameters can be given. Figure 13 shows the most common parameters for all IoT application layer protocols.

Application Protocol	CoAP	MQTT	MQTT-SN	XMPP	AMQP
REST Architecture	Yes	No	No	No	No
Publish/Subscribe	No	Yes	Yes	Yes	Yes
Request/Response	Yes	No	No	Yes	No
Transport	UDP	TCP	TCP	TCP	TCP
Security	DTLS	TLS/SSL	TLS/SSL	TLS/SSL	TLS/SSL
Header Size (Byte)	4	2	2	-	8
QoS	Yes	Yes	Yes	No	Yes

Fig. 13 Comparison between different IoT application protocols

5 Conclusion and Future Work

In this paper, we have given a brief introduction about the IoT and its applications and also the great impact of IoT in the global economy. Then we have focused on the different application layer protocols which are used most commonly nowadays. We have addressed the strengths and weaknesses of each protocol with its practical implementation in industries. In the future, we are aiming at implementing these protocols in a simulation testbed which can give an accurate comparison.

References

1. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. *Comput Netw* 54(15):2787–2805
2. Al-Fuqaha A, Guizani M, Mohammadi M, Aledhari M, Ayyash M (2015) Internet of things: a survey on enabling technologies, protocols, and applications. *IEEE Commun Surv Tutorials* 17(4):2347–2376
3. Evans D (2011) The internet of things: how the next evolution of the internet is changing everything. CISCO White Paper 1(2011):1–11
4. Manyika J, Chui M, Bughin J (2013) Disruptive technologies: advances that will transform life, business, and the global economy. McKinsey Global Institute, www.mckinsey.com/mgi

5. Islam SMR, Kwak D, Kabir MH, Hossain M, Kwak KS (2015) The Internet of things for health care: a comprehensive survey. *IEEE Access* 3:678–708
6. Bi Z, Xu LD, Wang C (2014) Internet of things for enterprise systems of modern manufacturing. *IEEE Trans Industr Inf* 10(2):1537–1546
7. Bormann C, Hartke K, Shelby Z (2014) The constrained application protocol (CoAP). RFC 7252
8. Banks A, Gupta R (2014) MQTT Version 3.1.1, OASIS Standard. <http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html>
9. Saint-Andre P (2011) Extensible messaging and presence protocol (XMPP): core. RFC 6120
10. OASIS Advanced Message Queuing Protocol (AMQP) (2012) Version 1.0, OASIS Standard. <http://docs.oasis-open.org/amqp/core/v1.0/os/amqp-core-complete-v1.0-os.pdf>
11. Stanford-Clark A, Truong HL (2013) MQTT for sensor networks (MQTT-SN) protocol specification version 1.2. http://mqtt.org/new/wp-content/uploads/2009/06/MQTT-SN_B_742_E_1_2_73_5
12. Bormann C, Castellani AP, Shelby Z (2012) Coap: an application protocol for billions of tiny internet nodes. *IEEE Internet Comput* 16(2):62–67
13. Fielding RT (2000) Architectural styles and the design of network-based software architectures. AAI9980887
14. Lerche C, Hartke K, Kovatsch M (2012) Industry adoption of the internet of things: a constrained application protocol survey. In: Proceedings of 2012 IEEE 17th international conference on emerging technologies factory automation (ETFA 2012), pp 1–6, Sept 2012
15. Constrained Application Protocol—Wikipedia, The Free Encyclopedia, 2017. (Online; accessed 1 May 2017)
16. Bormann C (2016) Constrained application protocol implementations. <http://coap.technology/impls.html>
17. Lampkin V, Leong WT, Olivera L, Rawat S, Subrahmanyam N, Xiang R (2012) Building smarter planet solutions with MQTT and IBM WebSphere MQ telemetry. IBM.com/redbooks, Sept 2012. ibm.com/redbooks
18. Amazon Web Services (2017) Message broker for AWS IoT. <http://docs.aws.amazon.com/iot/latest/developerguide/protocols.html>
19. Cooper J (2016) Adafruit products and adafruit IO. <https://learn.adafruit.com/adafruit-io/mqtt-api>
20. Govindan K, Azad AP (2015) End-to-end service assurance in iot mqtt-sn. In: 2015 12th Annual IEEE consumer communications and networking conference (CCNC), pp 290–296
21. Eugster PT, Guerraoui T, Sventek J (2000) Distributed asynchronous collections: abstractions for publish/subscribe interaction, pp 252–276. Springer Berlin Heidelberg, Berlin, Heidelberg
22. OASIS (2017) Advanced message queuing protocol: products and success stories. <http://www.amqp.org/product/realworld>. (Online; accessed 1 May 2017)

Empirical Laws of Natural Language Processing for Hindi Language



Arun Babhulgaoonkar, Mahesh Shirsath, Atharv Kurdukar,
Hrishikesh Khandare, Adwait Tekale, and Manali Musale

Abstract Empirical laws are the statistical laws that describe the relation between entities in a large dataset. They are readily found in nature, and findings have been proven by observations [1]. The primary objective of this study is to verify some of the empirical laws such as Zipf's law, Mandelbrot's approximation, and Heap's law for Hindi language corpus. This involves collecting a corpus, performing text normalization, tokenizing it to get a list of words, identifying word types and their frequency, sorting and ranking the data based on frequency, and representing the relation between the frequency and rank of the word types to validate Zipf's law and Mandelbrot's approximation. For Heap's law, the relation between the number of word types and tokens for different subsets of the corpus is considered. Based on our observations, the Hindi language satisfies the laws mentioned above.

A. Babhulgaoonkar · M. Shirsath · A. Kurdukar · H. Khandare (✉) · A. Tekale · M. Musale

Department of Information Technology,

Dr. Babasaheb Ambedkar Technological University, Lonere, India

e-mail: hrishikesh0408@gmail.com

A. Babhulgaoonkar

e-mail: arbabbulgaoonkar@dbatu.ac.in

M. Shirsath

e-mail: maheshshirsath106@gmail.com

A. Kurdukar

e-mail: 3atharvkurdukar@gmail.com

A. Tekale

e-mail: adwaitpt@gmail.com

M. Musale

e-mail: manalimusale1998@gmail.com

1 Introduction

Natural language processing (NLP) is a research and application field which is used to make a machine understand and manipulate natural language text or speech in order to perform useful things [2]. The goal of NLP researchers is to gather information on how people understand and use language so that developers can build appropriate tools and techniques to make computer systems understand and manipulate natural languages so that they can perform desired tasks.

A natural language text, commonly referred as “corpus,” is a set of paragraphs, whereas a paragraph is a series of sentences. A sentence is a structured set of words which conveys a meaning. a word represents a single distinct meaningful element of speech or writing. Thus, in NLP, words play a significant role. These words exhibit a statistical relation with the length of the corpus [3].

The term “token” refers to the each individual word in a corpus. In a corpus, the distinct words are called “word types.”

2 Empirical Laws

An empirical statistical law represents a type of behavior that has been found across a number of datasets and across a range of types of datasets [4]. These are the observations derived from natural phenomena and have been proved as statistical and probabilistic theorems. An empirical statistical law differs from a formal statistical theorem in a way that these patterns simply appear in natural distributions, without any theoretical reasoning about the data.

In the field of natural language processing, the empirical laws are basic statistical properties which are exhibited by different languages.

The common empirical laws in NLP are:

2.1 Zipf's Law

Zipf's law is an empirical law, based on mathematical statistics, named after the linguist George Kingsley Zipf, who first proposed it [5]. Zipf's law claims certain implementation about human nature. Zipf's law describes the word behavior in entire corpus and roughly describes the characteristics of certain empirical facts.

Zipf's law states that for a given corpus, the frequency of any word is inversely proportional to its rank in the term frequency table,

$$f \propto \frac{1}{r^\alpha} \quad (1)$$

where

$\alpha \approx 1$,

f : the frequency of a word type in a large corpus (number of occurrences),

r : the rank of the word according to frequency.

Equivalently, there exists some constant k such that,

$$f \times r^\alpha \approx k \quad (2)$$

2.2 Mandelbrot's Approximation

The Mandelbrot's approximation is an extension of Zipf's law as Mandelbrot derived a more generalized law to fit the frequency distribution in language by adding offset to the rank.

$$f \propto \frac{1}{(r + \beta)^\alpha} \quad (3)$$

where

$\alpha \approx 1$,

β is offset added by Mandelbrot as $\beta \approx 2.7$.

2.3 Heap's Law

Given the total number of tokens(T), Heap's law is used to estimate the number of distinct word forms or terms(M) in corpus. Heap's law can be stated as, "The dictionary or the size of vocabulary increases linearly with the total number of tokens/terms in the corpus."

$$M \propto T^b \quad (4)$$

where,

$b \approx 0.49$ [6].

3 Proposed Approach

3.1 Corpus

The corpus for Hindi language we used was a collection of various Wikipedia pages covering various topics such as culture, history, politics, agriculture, sports, and tourism. The length of the corpus had been around 3.97 million characters. Using

regular expressions, this data was standardized. Upon normalization, the data size has been reduced to 3.75 million characters. We have chosen the Shakespeare's corpus [7] as the English corpus, for comparison. The original Shakespeare's corpus is 5.45 million characters which was reduced to 4.69 million characters when condensed.

3.2 Tokenization

Tokenization is the process of dividing a string, a text into the token list. A token is an individual word in the text. In this, we split the text by regular expressions, using separators, that produces a tokens list.

3.3 Counting Type Frequency

Next, we used the list of tokens to find unique words, also known as types, from the list of tokens. In addition, we calculated the frequency of every type. For different sizes of the corpus, indicated by the number of tokens, a number of types and tokens were logged. An interval of 200 tokens was maintained between the logs.

3.4 Sorting and Ranking the Types

We sorted the types in decreasing order of their frequencies and then ranked them accordingly such that the word with highest frequency will be ranked as the first.

3.5 Applying Zipf's Law

To validate Zipf's law, we plotted a graph of frequency over rank adjusted using $\alpha = 0.85$ as shown in Fig. 1a. The graph is hyperbolic in nature.

The graph once adjusted with logarithmic axes generates an approximately straight line as shown in Fig. 1b, thus proving Zipf's law.

3.6 Applying Mandelbrot's Approximation

To validate Mandelbrot's approximation, we plotted a graph of frequency over rank adjusted using $\alpha = 0.85$ and $\beta = 2.7$ as shown in Fig. 2a. The graph is hyperbolic in nature.

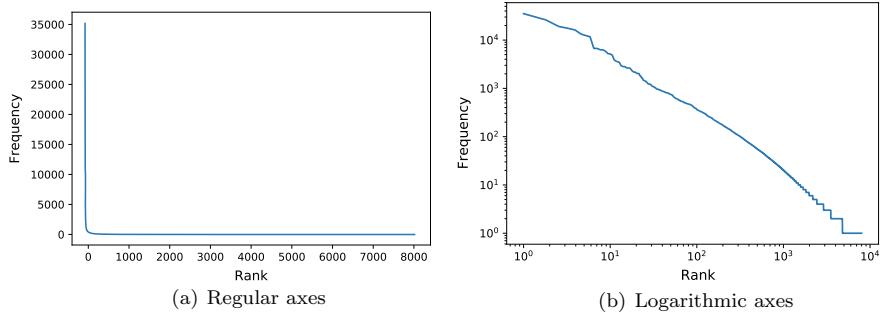


Fig. 1 Zipf's law implementation

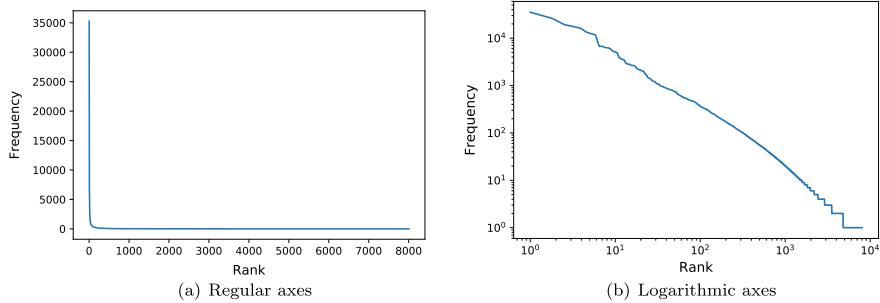


Fig. 2 Mandelbrot's approximation implementation

The graph once adjusted with logarithmic axes generates an approximately straight line as shown in Fig. 2b, thus proving Mandelbrot's approximation.

3.7 Applying Heap's Law

To validate Heap's law, we plotted a graph of M over T^B where M is number of types, T is number of tokens, and $B = 0.49$. As shown in Fig. 3, the graph represents a relatively straight line, thus proving Heap's law.

4 Comparison with English Corpus

As shown in Fig. 4, the result produced for Zipf's law by Hindi corpus is quite similar to that of English corpus. The same is true for Fig. 5, which represents Mandelbrot's comparison, and Fig. 6, which represents Heap's law.

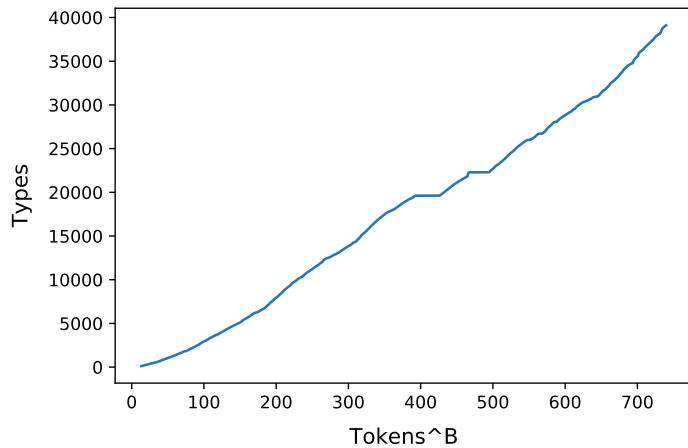


Fig. 3 Heap's law

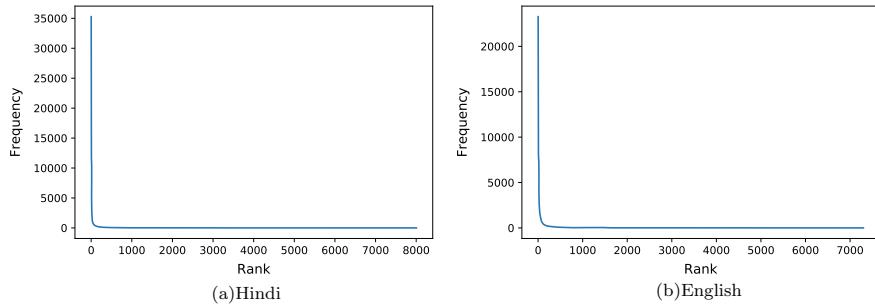


Fig. 4 Showing the comparison between Zipf's law for languages

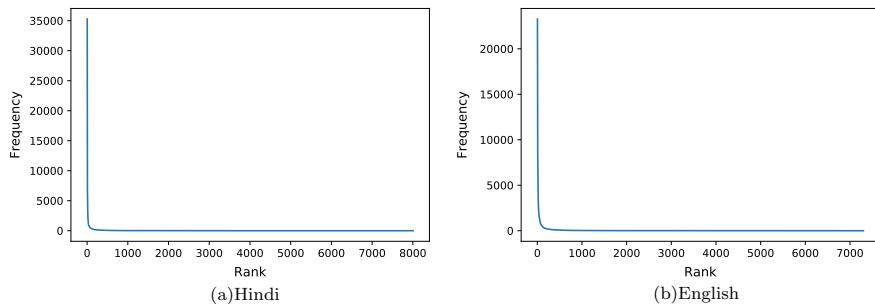


Fig. 5 Showing the comparison between Mandelbrot's approximation for languages

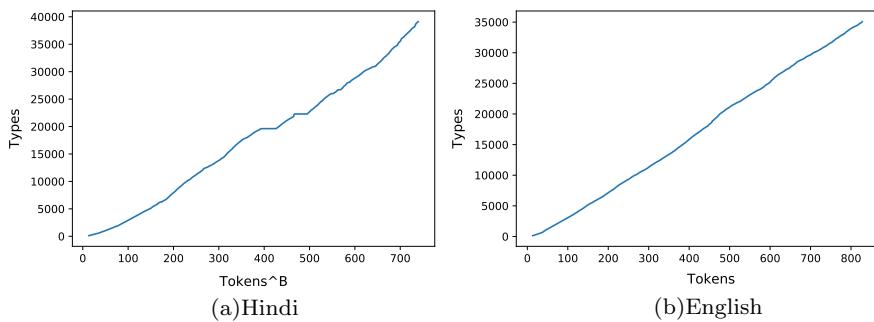


Fig. 6 Showing the comparison between Heap's law for languages

5 Conclusion

Therefore, based on the results found from our research, we may infer that the empirical laws for natural language processing, namely Zipf's law, Mandelbrot's approximation, and Heap's law, are also applicable for Hindi language as they exhibit similar pattern to that of empirical laws for English language.

References

1. Manning C, Schütze H (1999) Foundations of statistical natural language processing. MIT Press, Cambridge
2. Chowdhury GG, Natural language processing. Dept. of Computer and Information Sciences, University of Strathclyde, Glasgow
3. Charniak E (1993) Statistical language learning. MIT Press, Cambridge
4. James A (1996) Natural language understanding, 2nd edn. Benjamin/Cummings, San Francisco
5. Piantadosi ST (2014) Zipf's word frequency law in natural language: a critical review and future directions. Psychon Bull Rev 21(5):1112–1130. <https://doi.org/10.3758/s13423-014-0585-6>
6. Gelbukh A, Sidorov G (2008) Zipf and Heaps laws coefficients depend on language. In: Computational linguistics and intelligent text processing. Springer, pp 332–335. ISBN 978-3-540-41687-6
7. Shakespeare's Corpus. <https://ocw.mit.edu/ans7870/6/6.006/s08/lecturenotes/files/t8.shakespeare.txt>

Classification Method to Predict Chances of Students' Admission in a Particular College



Danny Joel Devarapalli

Abstract There are many Predictor/Prediction systems or applications on the internet, but most of these prediction systems do not use any Machine Learning algorithms in their systems, it's just mere if and else conditions that run these predictions. We see predictor systems are gaining more popularity in predicting colleges a student can get placed into, most of these have if-else methods, but a good number of them use Machine Learning algorithms but they'd show the set of colleges a student can be placed in. This research involved creating a model that would accurately classify if a student can get admission in a particular college give a set of attribute values that would act as the feature set for the model. The two objectives are to determine the chances of a student getting accepted by a college using past data of college-wise allotment based on certain attributes or parameters which highly influence the class attribute or have high-value dependency. The second objective to give an insight of the number of students willing or showing interest to join the college, their contact number, gender, the department or branch they are interested in, to the respective college management/Admission panel which would help them understand and draw certain conclusions on the students' interest, the college's popularity, etc. This will help the management work on these students who are interested in getting accepted/enrolled in the college. This would benefit both the student and the college management, and all of this achieved by MACHINE LEARNING!

Keywords Machine learning · Data science machine learning · Classification · RandomForest · Prediction system · College admission prediction · Data analytics

D. J. Devarapalli (✉)

Department of Computer Science, Vignan Institute of Technology and Science, Hyderabad, India
e-mail: dannyjoeldevarapalli@gmail.com

1 Introduction

There are many admission fairs that happen during the beginning of the new academic year and many students who are unaware of the admission process in many colleges attend these fairs to either know the admission process or to check their eligibility with the college's admission accepting criteria. The educational/admission fairs charge close to ₹50–₹100 for the previous year's college-wise allotment details for students who are interested to check their possibility to get admission in the desired college with the desired course. But with a predictor system, a student can check these details for him, free of charge, and also stay directly in touch with the college's Admission Team, and clear their doubts directly without any involvement of a third-party.

However, Machine Learning comes to the rescue by providing the best classification algorithms, using python and its famous Scikit Learn module—sklearn, which made using machine learning much easier.

1.1 *Motivation*

To explore the field of Data Science and learn what exactly Machine Learning was all about has led me to collect data, preprocess it and use various classification models and evaluate/research on each classification model. Train a model to predict the chances of students' admission confirmation. This has further motivated me to collect legitimate data, preprocess it and use various classification models to train this data and evaluate each classification model and to choose the best one.

1.2 *Objectives*

- a. **Primary Objective:** To determine the chances of a student getting accepted by a college using past data of college-wise allotment based on certain attributes or parameters which highly influence the class attribute or have high-value dependency.
- b. **Second Objective:** Our second objective to give an insight of the students, their number, their gender, the department or branch they are interested in, to the respective college management/Admission panel which would help them understand and draw certain conclusions on the student's interest, the college's popularity, etc. This will help the management work on these students who are interested in getting accepted/enrolled in college.

1.3 Significance and Advantages

To my knowledge, there has been no study or project that has yielded or focused on making predictions on one particular college. Focusing on one specific college gives us an advantage of better accuracy as we rely more on the college's previous admission data. The primary advantage is both the college can initiate their connection with the freshmen and the freshmen are given the accurate results as the prediction is done by using a Machine Learning algorithm.

2 Literature Review

There hasn't been any previous literature that I have found which deals with a similar problem statement, whereas I would want to provide a critical appraisal of previous studies with a similar approach of predictions using Machine learning classification algorithms.

Existing System—There are some existing systems that list colleges for students according to their rank with predefined data. Only few existing systems use machine learning algorithms to predict the chances of students getting seats according to rank but not for specific colleges. The disadvantage is that it only displays all the possible college names but doesn't predict the chances of getting a seat in the college. As there is a chance for them to miss specific college's criteria or restrict to show only the first five records of the results. Many existing systems require the users to login in order to use their system, this would be laboring for the user to enter the details, and to confirm his details and perform all the necessary authentications.

After an extensive search for a similar topic and objective, I have come across this paper that addresses the approach to the same idea of predicting admission chances for a student in a college, College Admission Predictor [1], also predicts are gives a list of colleges a student is eligible for, this is an advantage to a student who has no interest in particular college but is just ready to go to any college which matches with his qualifications and desired branch, but one disadvantage is that the predictor system lacked the usage of any machine learning algorithm or any data mining tool in order to do this. The data that is required to be filled takes time as the system requests the user to fill out an application form, where students might be slack in doing it. Since we want everything to be done quicker.

3 Proposed System

This project is used to determine the chances of a student getting accepted by a college based on certain attributes or parameters. This will help the college management to understand the count of the students who are interested in getting admission in the

college. Based on their inputs for the parameters we can draw certain conclusions on the set of students showing interest, etc. Since the Machine Learning algorithms are used to make the predictions it is important to understand which model best fits the data. The metrics of each model define the accuracy of the system and much better accuracy can be achieved by updating the new data of the official allotted students joining the college every year and adding that data to train the model again by avoiding overfitting error.

Scope and Perspective: Since we are developing this and have only gathered data with respect to our college, so only our college can use this. But we are very much interested to do it for other colleges based on their request to having it on their college's official website. It's always better when the focus is on one thing, so having this for each college exclusively for the college is better. Having it for all colleges under one single platform would require much data and we cannot be certain of the results being accurate. We can achieve it using Big Data Analytics yes, but the results of the students' change each year with respect to the change in the college's infrastructure and facilities offered by the college.

Advantages of the Proposed System: The student can get the most accurate result from the details entered. As we make use of the best classifier after training and evaluation of each classification algorithm. Management can stay in contact with the student as the details the student enters during the time of predictions and his personal details are stored in an Excel sheet file and the values that are required for prediction are fed to the classifier for prediction. By contacting these students who are interested in joining the college helps the college's admission panel to improve its admission process. Also, this gives an insight to the college's management to understand and draw certain conclusions regarding the college's face value in the local area or beyond and can help them improve their advertising, select the proper audience for their advertising with respect to their geographical location/locality and also helps in understanding the statistics of (for instance) the number of students who are interested in joining Computer Science and Engineering department, etc.

4 Methodology

The whole world-flow of this research is by five steps, a clear explanation and description are given with respect to every step.

4.1 *Gathering Data*

The data collection is an important step in machine learning as we need to gather real-time data in order to get accurate results. I collected the college-wise allotment details from the official website of Telangana State EAMCET—<https://tseamcet.nic.in>.

in/Default.aspx and selected “College-wise allotment details” option, which consists of the data of all students who have been allotted a college and department based on their Rank, Sex, Caste-category, Region, Seat Category and desired department. The provisional allotment list contains two menu lists where we need to select the college and the branch and it will display the list of students who have been allotted seats in the college specified and brand.

As I have intended to have two class values <’Yes’, ’No’>, for the class value ‘Yes’ the data of students who have been allotted to our college is collected and for the other value of ‘No’ the data is collected from colleges which have the same departments with respect to our college and which have a lower ranking than our college have been collected. The data displayed is copy-pasted into a Text Document and is later imported to an Excel Sheet by selecting “Data” from the quick access toolbar and “From Text” is selected which will import a text file. The importing is done by using tab space as a new excel cell. Now the data which is loaded into the Excel sheet must undergo the data preprocessing step.

4.2 Data Preprocessing

In this step, data preprocessing is one of the most important steps in machine learning. It is the most important step that helps in building machine learning models more accurately. In machine learning, there is an 80/20 rule. Every data scientist should spend 80% time for data preprocessing and 20% time to actually perform the analysis. We delete the unwanted columns or attributes of data and replace the values which are more appropriate and understandable. The data that is collected has attribute values <’S.no’, ‘Hall Ticket number’, ‘Rank’, ‘Name of the Candidate’, ‘Sex’, ‘Category-Caste’, ‘Region’, ‘Seat Category’>. There are eight attribute values or head values, we only require a few among these. A careful observation of attributes that are needed can be chosen, the “S.No.” which is the serial number is not needed, and it is excluded, next is the “Hall ticket number” this is not going to help us in anyway, and does not contribute to marks or rank, so it is excluded, “Rank” this attribute is important as we are making predictions based on rank, next comes “Name of the Candidate” this can be removed as allotment is not done by names, “Sex” as we have reservations based on the sex of the candidate such as there are special seats allotted for women, hence we keep this attribute, next comes “Category-Caste”, major allotments are done based on the caste of the candidate by reservation, this is not disturbed, “Region” this plays a main role as well, as there will allotment done by region where the student belongs in, the local students are given more priority as such, and the last attribute we see is the “Seat category” has many values which are constructed based on the attribute information that is not provided, such as “female category in sports quota”, etc., hence this attribute is also removed. Two years of data have been collected and I have collected data from our college’s admission office of the previous batch.

Table 1 Attributes and their respective values in the dataset

Attributes	Values
Rank	Range <0 to 100,000>
Sex	‘Male’, ‘Female’ <1, 0>
Caste	OC, BC_(A to E), SC, ST <1 to 8>
Region	AU, NL, OU, SVU <1 to 4>
Department	CSE, ECE, EEE, ME, CE, EIE, IT <1 to 7>
Admitted <class-label>	‘Yes’, ‘No’ <1, 0>

Values

Each attribute has its own set of discrete-values, which are changed to number values to understand them better. This table below shows the set of possible values of each finalized attribute (Table 1).

The data is randomized using the “=rand()” function in excel and the five attributes <‘Rank’, ‘Sex’, ‘Category-Caste’, ‘Region’, ‘Department’> are saved in one file “attributes.csv” and the class label which is <‘Admitted’> is saved into another file called the “class.csv”. The data after this step is loaded into a python script using modules and is fed to the machine learning algorithms, which is clearly explained in the next step in the process.

Now the raw data which was collected from the internet is meaningfully defined as we know that Machine Learning models can only handle numeric features, hence categorical and ordinal data must be somehow converted into numeric features.. We further preprocess data by using methods such as Standardization [2] is a useful technique to transform attributes with a Gaussian distribution and differing means and standard deviations to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1.

Note

1. **Missing data:** There was no missing data in the final dataset as it is complete even before it was placed on the website.
2. **Inconsistent data:** There was no chance of human-error in this dataset as this is computer generated even before it was placed in the website.

4.3 Analyzing Approach

The machine learning approach must be carefully researched and examined based on the data that is ready to use and the final result that we desire. Since our data is labeled and has a class label, we go with a supervised learning approach in machine learning.

Supervised Learning—In Supervised Learning, we train a classifier or a machine learning model that is fed data with labels, which means to say data with feature

names and a class label that is a discrete-valued or categorical values and every data instance has one class label value. We use classification methods/algorithms in supervised learning such as Decision Trees, Support Vector Machines, k-Nearest Neighbors, Random Forest classifier, Naive Bayes classifier, etc. We use these models and they are trained with labeled data then the algorithms predict the output (class value) from input(attribute values) data.

We use these set of supervised learning algorithms, and train our data and evaluate them based on their accuracy score and other validation methods.

4.4 Classification

Before we feed our dataset into the python file and into the classifier we must first divide the complete dataset into two sets, the training set, and the testing set. The splitting is usually done in the ratio 80:20, which means 80% of the data will be used for training the model which is giving the model set of examples to understand the nature of the data and the models construct their very own structures and process on how the data is learned, which will be clearly explained in the coming section. The remaining 20% of the data is used for testing, which is fed into the model to predict these test sets which has the five attribute values. We evaluate the model's performance based on the predicted output.

Loading the data—For data manipulation the library pandas is used which is imported into the python script using the “read_csv (file_address)” method.

Train and test split—The randomized dataset is divided into attributes.csv and class.csv. This dataset is divided or split in the proportion of percentages 80 and 20. The training set is 80% and the testing set is 20%.

In order to split the dataset we can use the module “train_test_split” from sklearn.model_selection.

Here the X_tr contains the instances of the attribute which are used for the model to train on and the Y_tr is the class instances that are used for the training as well. The X_te and Y_te are used for testing, where the model is fed these datasets and the model predicts the class of the X_te set, the algorithm or the model which correctly classifies them has the highest performance.

Training and Testing of the classification models—Before we can use these machine learning methods, we must import the necessary modules into our python script. We have used the following modules to perform the classification training on our dataset.

Decision Tree [3]

A Decision tree is a flowchart-like tree structure, where each internal node represents a test on an attribute, each branch represents an outcome of the test, the leaf nodes are the final nodes where the class label is represented by each leaf node. The attribute values of the tuple are tested against the decision tree. A path is traced

from the root to a leaf node that holds the class prediction for a particular tuple or data record these traced ways are the rules by which classification is done. It is easy to convert decision trees into classification rules. Decision trees can be constructed relatively fast compared to other methods of classification. There are three methods to select the root node and the next child node in the decision tree, but since there are many types of decision trees, the most common and popular ones are:

CART uses Gini Index as metric.

ID3 uses Entropy function and Information gain as metrics.

- Step 1 To use the Decision Tree Classifier we must use the `sklearn` module to import the Decision Tree Classifier from `tree`.
- Step 2 A classifier object is created for the Decision Tree Classifier
- Step 3 The next step would be fitting the training datasets of attributes and class to the classifier to train on it. The training of the decision tree involves building a tree by making the best split of data, choosing the root node. To fit the training data into the model classifier we use the `fit()` method which takes the training datasets of attributes and class values.
- Step 4 Once the training is done the tree is constructed and the predicting the testing set is done, this is done by using the `predict()` method, which takes the `X_te`, the attribute set, which was set apart during the training of the model.

k-Nearest Neighbors (KNN) [4]

k-NN is a simple and basic classification technique. k-NN is also referred to as lazy learning and it is an instance-based learning technique. k-NN is employed within the classification and regression on the applications of the method in many areas. In the classification method, k-NN algorithm is a method for classifying the objects based on the closest training data. It uses the Euclidean distance method to find the distance and the final class is decided based on the voting of maximum positive or negative votes.

- Step 1 Similar to how we have created a class object to the Decision Tree Classifier, we must also do the same with the kNN classifier.
To use the k-NN classifier we must use the `sklearn` module to import the `kNeighborsClassifier` from `neighbors`.
- Step 2 A classifier object is created for the `KNeighborsClassifier`
- Step 3 To fit the training data into the model classifier we use the `fit()` method which takes the training datasets of attributes and class values.
- Step 4 Once the training is the distances are calculated for each instance and predicting the testing set is done by using the `predict()` method, which takes the `X_te`, the attribute set, which was set apart during the training of the model.

Support Vector Machines [5]

SVM is a supervised machine learning algorithm that can be used for either classification or regression challenges. However, it is mostly used in classification

problems. We plot each data item as a point in n-dimensional space, where n is considered as the number of attributes, with the value of each feature being the value of a particular coordinate. After this, we perform classification by finding the hyper-plane that differentiates the two classes very well.

- Step 1 To use the SVM classifier we must use the sklearn module to import the Support Vector Classifier(SVC) from svm.
- Step 2 A classifier object is created for the Support Vector Classifier
- Step 3 To fit the training data into the model classifier we use the fit() method which takes the training datasets of attributes and class values.
- Step 4 After the completion of the model training on the example set which is the training set, it is then fed the testing set to predict the values correctly from the past experience gained from the training. The predict() method is used to predict the class values of the testing set.

Random Forest Classifier [6]

Random forest is like a bootstrapping algorithm with Decision tree (CART) model. Say, we have 1000 observations in the complete population with 10 variables. Random forest tries to build multiple CART models with different samples and different initial variables. For instance, it will take a random sample of 100 observations and 5 randomly chosen initial variables to build a CART model. It will repeat the process (say) 10 times and then make a final prediction on each observation. Final prediction is a function of each prediction. This final prediction can simply be the mean of each prediction.

The whole data set consists of 5302 records, which is pretty huge. Random Forest algorithm works very well with huge datasets, and this is the primary reason for me to choose this algorithm. Since there are over five thousand records in the dataset, and chances of overfitting we choose RandomForest as there won't be a chance for overfitting in this algorithm. Random forest gives much more accurate predictions when compared to simple Decision tree or regression models in many scenarios. These cases generally have high number of predictive variables and huge sample size. This is because it captures the variance of several input variables at the same time and enables high number of observations to participate in the prediction.

The two important features of RandomForest algorithm are: [7]

1. Random sampling of training data points when building trees
2. Random subsets of features considered when splitting nodes

In order to use this RandomForest Classifier, we must follow the same steps that we have followed for the other algorithms.

- Step 1 To use the RandomForest Classifier we must use the sklearn module to import the RandomForest Classifier from the ensemble.
- Step 2 A classifier object is created for the RandomForestClassifier
- Step 3 To fit the training data into the model classifier we use the fit() method which takes the training datasets of attributes and class values

Step 4 After the completion of the model training on the example set which is the training set, it is then fed the testing set to predict the values correctly from the past experience gained from the training. The predict() method is used to predict the class values of the testing set.

Program in Python 1:

```
#1 from sklearn.ensemble import
RandomForestClassifier #2 clf_RDF =
RandomForestClassifier() #3 clf_RDF.fit(X_tr, Y_tr)
#4 prediction_RDF = clf_RDF.predict(X_te)
```

5 Evaluation

Evaluation of classifiers is an important step, as it gives us the performance of each model. Evaluating classifiers help us choose the best working model, by measuring the accuracy of each model, the model that gives us the best accuracy among all other models can be chosen as the best model and be used in further process.

There are many methods for evaluation, out of which confusion matrix is widely used.

Confusion Matrix [8]

Confusion Matrix is a matrix which is of order 2×2 , to define positive examples and negative examples, i.e. number_of_possible_class_values \times number_of_possible_class_values. However, based on the number of possible class values, the order of the matrix is defined. For example, for iris dataset, has three class values Iris Setosa, Iris Versicolour, Iris Virginica so the confusion matrix for the iris dataset would be of the order 3×3 .

Since the dataset that is being used has only two classes, Yes and No, our confusion matrix would be of order 2×2 . To understand how to measure the accuracy and error rate of a model, we need to understand the terms TP, FN, FP, TN.

The accuracy score is calculated as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Classification report [2]

The classification report gives us a clear picture of certain evaluating measures of a classifier such as precision, recall, f_score, and support.

Precision is defined as the relevance of positive examples among all of the examples which were predicted.

```
The Accuracy score of RandomForest is : 93.903206
Confusion matrix for :
RandomForest
[[1304  26]
 [ 71 190]]
Classification Report for : RandomForest
precision    recall   f1-score   support
No          0.95     0.98     0.96     1330
Yes         0.88     0.73     0.80      261

accuracy          0.94
macro avg       0.91     0.85     0.88     1591
weighted avg    0.94     0.94     0.94     1591

RANDOMFOREST CLASSIFIER has highest accuracy: 93.9032055311125
```

Fig. 1 Showing the classification report and the highest accuracy score for random forest classifier

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is defined as the ratio of the number of correctly classified positive examples(TP) divide to the total number of positive examples(TP + FN). A class is rightly defined if there is high recall.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F-measure or F score represents the measures of both recall and precision.

$$\text{F-measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

To check the confusion matrix and accuracy for the models, we must import accuracy_score, classification_report, confusion_matrix modules from sklearn.metrics

The confusion_matrix, accuracy_score and the classification_report for the four following models were,

RandomForest Classifier Fig. 1.

6 Experimental Results

The accuracy measures of the used classifiers have shown that Decision tree and RandomForest classifier has given us the maximum accuracy. Decision tree gave us an accuracy of 91.7661 and RandomForest Classifier has given us an accuracy of 93.90320, which is close to 94% and the highest of all the classifiers used.

To plot these accuracy measures on a graph we import a matplotlib module into our python script.

Program in Python 2:

```
index_n = accuracy_results.index(max(accuracy_results))
print(list_clfs[index_n].upper(), 'has highest
accuracy:', max(accuracy_results)) plt.title('Accuracy
Comparison')

colors = ['r', 'y', 'b', 'g']
plt.bar(clf_names, accuracy_results, color = colors)
plt.show()

plt.plot(clf_names, accuracy_results)
```

The list accuracy_results contains the accuracy scores of all the classifiers, the list_clfs contains the list of all classifiers. The color list plots the graph of classifiers' accuracy based on their position in the list with respect to the position in the list_clfs.

The plot function plots the bar graph with respect to axes of classifier names in the list_clfs and the list of accuracy_results which contains all the accuracy scores of each classifier.

The output bar graphs of the accuracy of all the models are shown in Fig. 2.

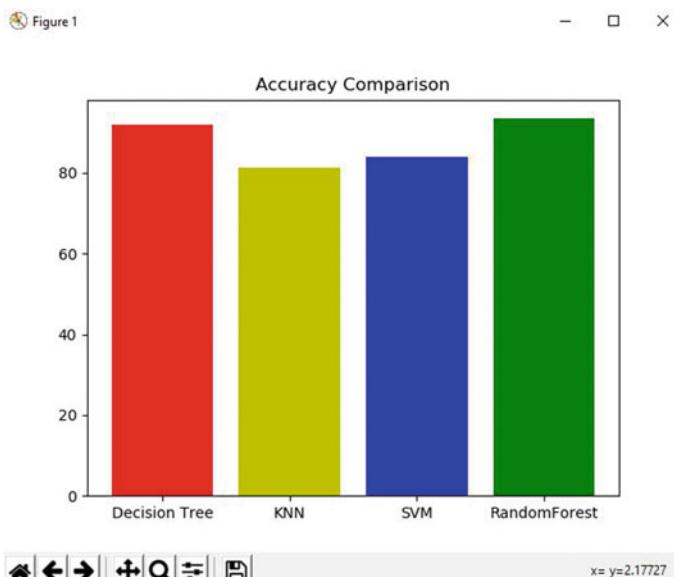


Fig. 2 Showing the accuracy graph of different classification algorithms used

7 Conclusion and Future Work

Since the dataset contained a huge number of records, Randomforestclassifier has worked well with the data. The project's key goal is the deployment of a predictive model has achieved, the project continues. The next step would be to make sure to track the performance of the deployed model and keep updating the existing dataset every year with the new academic allotment details to maintain the model. Performance metrics used for model evaluation can also become a valuable source of feedback by measuring and training the models again with the new set of allotment details. The major task would be to make sure the data would not become outdated. This can further be developed to each and every college, mainly focusing on their college's performance and admission. Students from other colleges can also develop their very own Student Admission Predictors for their very own colleges and deploy it as a web application and put that in their college main website. This can help the college management draw more insights about their college's popularity among students and in public and also promote the college in this way.

Appendix

Acronyms and Abbreviations

EAMCET	Engineering Agricultural and Medical Common Entrance Test
AU	Andhra University, (Students who are from Andhra Pradesh come under this region)
NL	Non–Local
OU	Osmania University
SVU	Sri Venkateswara University
CSE	Computer Science and Engineering
ECE	Electronics and Communications Engineering
EEE	Electrical and Electronics Engineering
ME	Mechanical Engineering
CE	Civil Engineering
EIE	Electronics and Instrumentation Engineering
IT	Information Technology
TP	True Positives
FN	False Negatives
FP	False Positives
TN	True Negatives

References

1. Roa AM, Dharani N, Raghava AS, Buvanambigai J (2018) College admission predictor. *J Netw Commun Emerg Technol (JNCET)* 8(4) www.jncet.org. Referenced online: <http://www.jncet.org/Manuscripts/Volume-8/Issue-4/Vol-8-issue-4-M-32.pdf>
2. Data Flair Team (2018) Data preprocessing, analysis and visualization—python machine learning. <https://data-flair.training/blogs/python-ml-data-preprocessing/>
3. Sharma H, Kumar S (2016) A survey on decision tree algorithms of classification in data mining. *Int J Sci Res (IJSR)* 5(4). Referenced online: https://www.researchgate.net/publication/324941161_A_Survey_on_Decision_Tree_Algorithms_of_Classification_in_Data_Mining
4. Mitchell TM (1997) Machine learning, McGraw-Hill Education, 1st edn (Mar 1, 1997), pp 231–233
5. Ray S (2017) Understanding support vector machine algorithm from examples. <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>
6. Srivastava T (2014) Introduction to random forest—simplified. <https://www.analyticsvidhya.com/blog/2014/06/introduction-random-forest-simplified/>
7. Koehrsen W (2018) An implementation and explanation of the random forest in python. <https://towardsdatascience.com/an-implementation-and-explanation-of-the-random-forest-in-python-77bf308a9b76>
8. Visa S, Ralescu A, Ramsay B, van der Knaap E (2011) Confusion matrix-based feature selection. In: Proceedings of the 22nd midwest artificial intelligence and cognitive science conference 2011, Cincinnati, Ohio, USA, Apr 16–17, 2011. Referenced online:https://www.researchgate.net/publication/220833270_Confusion_Matrix-based_Feature_Selection

Doppler Shift Based Sampling Rate Conversion for GFDM Underwater Acoustic Communication



V. S. Kumar, A. Chakradhar, M. Shiva Prasad, and Y. Shekar

Abstract In Generalized Frequency Division Multiplexing (GFDM) underwater acoustic communication systems, the conventional Doppler shift factor estimation will reduce the transmission rate of the system. The sampling rate conversion of the received signal usually adopts the resampling method, thereby increasing the computational complexity. Aiming at the above problems, a Doppler shift factor (DFS) estimation algorithm based on oversampling technique and a sampling rate conversion algorithm is proposed. The Doppler shift factor is estimated by comparing the transmitted signal with the received signal samples. On this basis, the above-sampled data and a linear interpolation algorithm are used to convert the sampling rate of the received signal. The theoretical analysis and simulation results show that the improved algorithm can greatly reduce the computational complexity of the receiver while ensuring the performance of the system and is suitable for high-speed real-time underwater acoustic communication systems.

Keywords Underwater acoustic (UWA) communication · Generalized frequency division multiplexing (GFDM) · Doppler shift (DS) · Oversampling · Linear interpolation · Sample rate conversion

V. S. Kumar (✉) · A. Chakradhar · Y. Shekar
Department of ECE, SR University, Warangal, India
e-mail: kumar.s.vngl@gmail.com

A. Chakradhar
e-mail: adupa.chakradhar@gmail.com

Y. Shekar
e-mail: shekar4b@gmail.com

M. S. Prasad
Department of ECE, Vaadevi College of Engineering, Warangal, India
e-mail: shiva.martha@gmail.com

1 Introduction

In Under Water Acoustic (UWA) communication system, Doppler Shift (DS) is mainly caused by relative motion between transmitter and receiver, i.e. complex movement of water body itself, which results in frequency deviation of carrier frequency between transmitter and receiver [1]. Besides, the narrow bandwidth of underwater acoustic channel also causes the Doppler shift to have a significant impact on underwater acoustic communication system. Multicarrier modulation has been widely used in the field of UWA communication; this modulation can solve the problem of frequency selective fading due to multipath effect. However, due to signal delay and carrier frequency offset caused by severe Doppler shift, the receiver of GFDM UWA communication system generates a large inter-carrier interference (ICI) [2, 3]. Thus, an accurate estimation of the Doppler frequency shift is a key issue in reducing ICI.

The Doppler shift estimation first needs to estimate Doppler shift factor (DSF) at the receiver, and then use DFS to perform sampling rate conversion on the received signal, thereby eliminating the Doppler shift. Therefore, the DSF estimation and sampling rate conversion algorithm is key to DFS estimation. The traditional DSF estimation algorithm uses the method of inserting auxiliary data. In [4], a linear frequency modulated signal is mainly used for DSF estimation. In [5], the DSF is estimated by using a cyclic prefix in OFDM symbol as auxiliary data. The literature [6, 7] uses the power of the receiver signal to estimate the Doppler shift with channel variation. In [8], the Doppler shift compensation of underwater acoustic channel is divided into two steps, namely wideband compensation and narrowband compensation. For the sampling rate conversion algorithm, the usual method is to use the resampling technique, i.e. estimated DSF is used to compensate for the Doppler shift generated by received signal resampling, thereby recovering original data. This algorithm not only affects the effectiveness of the system but also requires a large amount of computations. In addition, the above Doppler factor estimation algorithm will also cause system estimation performance to drop sharply in the case of multipath propagation [9].

In this paper, the Doppler shift factor is estimated by oversampling, no auxiliary data is needed. The Doppler shift at the receiver is compensated directly by combining the linear interpolation and sampling rate conversion process. Doppler shift compensation process reduces a large number of computational complexities by avoiding the process of resampling. Thereby, oversampling and linear interpolation methods don't add much computational complexity to the system. In, this paper an improved algorithm for DSF estimation and Doppler shift sampling rate conversion is proposed to improve the transmission efficiency of GFDM underwater acoustic system.

2 Theoretical Model of Doppler Frequency Shift

Carrier detection, symbol synchronization of UWA communication are severely affected by Doppler interference [10]. For broadband UWA signals, Doppler Effect will cause received signal to generate Doppler shift in the frequency domain and compress or expand in the time domain. The effect of Doppler on the signal is usually modelled as:

$$y(t) = x[(1 + a)t] \quad (1)$$

where $x(t)$ represents transmitted signal, $y(t)$ represents received signal with Doppler frequency shift, $a = v/c$ represents the Doppler frequency shift factor, c represents the velocity of sound in water. v indicates a relative radial velocity of transmitter and receiver, which is considered only in a symbol period where v remains unchanged. The transmitter and receiver are close to each other with $a > 0$, and the transmitters and receivers are far away from each other with $a < 0$. Due to the serious multipath effect in the UWA channel, it is assumed that each transmission path is the same.

The discrete-time signal after being sampled at the receiving end:

$$y[nT_s] = x[n(1 + a)T_s] \quad (2)$$

where T_s is the sampling period; n is an integer. If the Doppler shift factor is known, the received signal is resampled with factor ‘ a ’, and the original signal is recovered at the receiving end:

$$x[nT_s] = y\left[\frac{nT_s}{1 + a}\right] \quad (3)$$

The following is a theoretical analysis influence of Doppler Effect on frequency and time domain. Assuming that transmitted signal frequency is f_t and length is T_t , the frequency of received signal affected by Doppler and length T_r is expressed as:

$$f_r = (1 + a)f_t \quad (4)$$

$$T_r = \frac{T_t}{1 + a} \quad (5)$$

The Doppler Effect is given by:

$$a = \frac{f_r}{f_t} - 1 = \frac{T_t}{T_r} - 1 \quad (6)$$

3 Doppler Frequency Shift Factor Estimation Algorithm

In a GFDM underwater acoustic communication system, the Doppler shift factor is usually obtained from auxiliary data. The Doppler shift factor estimation based on oversampling Doppler shift theoretical model is

$$\begin{cases} f_{ns} = \frac{N_0}{T} = N_0 \Delta f \\ T_{ns} = \frac{T}{N_0} = \frac{1}{N_0 \Delta f} \end{cases} \quad (7)$$

where T is GFDM symbol interval, $\Delta f = 1/T$ carrier frequency interval, N_0 is number of subcarriers. f_{ns} is Nyquist sampling frequency, and T_{ns} is GFDM symbol sampling period, respectively. At the transmitting end, number of Nyquist samples for GFDM symbol with a Cyclic Prefix (CP) is $(N_{CP} + N_0)$ where N_{CP} represents the number of Nyquist samples of CP. At the receiver, GFDM symbol is still sampled at Nyquist frequency, and an actual number of Nyquist samples is N . Since the signal is stretched during transmission, length of GFDM symbol with CP at the transmitter and the receiver end is:

$$\begin{cases} T_t = \frac{(N_0 + N_{CP})T}{N_0} \\ T_r = NT_{ns} = \frac{NT}{N_0} \end{cases} \quad (8)$$

where $N \neq N_{CP} + N_0$ due Doppler effect, assuming that the received signal is ideally synchronized, then according to (6), the following equation is given by:

$$a = \frac{N_0 + N_{CP}}{N} - 1 \quad (9)$$

According to theoretical analysis, the value ‘ a ’ calculated by (9) will decrease with the increase of sampling frequency. Therefore, if a more accurate ‘ a ’ value is obtained according to (9), the sampling frequency should be increased, and sampling rate can reduce the errors effectively.

Based on above assumptions and analysis, the received GFDM symbol with CP is oversampled at r times the Nyquist sampling frequency. The oversampling frequency f_{rs} and the sampling period T_{rs} is given by:

$$\begin{cases} f_{rs} = r \cdot \frac{N_0}{T} = rN_0 \Delta f \\ T_{rs} = \frac{T}{r \cdot N_0} = 1/(r \cdot N_0 \Delta f) \end{cases} \quad (10)$$

The actual oversampling point on the receiver is N_r , the GFDM symbol with CP is represented as:

$$\begin{cases} T_t = \frac{(N_0 + N_{CP})T}{N_0} \\ T_r = N_r T_{rs} = \frac{N_r T}{r N_0} \end{cases} \quad (11)$$

According to (6)

$$a = \frac{r(N_0 + N_{CP})}{N_r} - 1 \quad (12)$$

Equation (12) is an estimate of DSF obtained by the oversampling method at the receiver. It is known from comparison with the result obtained by (9), that the error of DSF estimation will be correspondingly reduced due to the decrease of the sampling period, but the effect of this error reduction is not obvious. Besides, the oversampling method is used at the receiving end, and the main purpose is to design an improved algorithm for sampling rate conversion based on oversampling technique.

4 Improved Sampling Rate Conversion Algorithm Based on Oversampling Technique

The Doppler shift factor ‘ a ’ is obtained by oversampling method (12), it is possible to resample received data according to (3), and converting sampling rate of received data to eliminate the Doppler Effect. Since resampling requires a large number of operations, the oversampling combined with improved algorithm without any resampling is used directly through linear interpolation, to complete sampling rate conversion of received data, which not only save a lot of receiver computations, but it also improves the system performance. The following is N_r oversampled data represented by vector Y :

$$Y = [y(1), y(2), \dots, y(n), \dots, y(N_r)] \quad (13)$$

The sampling time corresponding to each discrete data in (13) is:

$$T = \left[0, \frac{T}{rN_0}, \frac{2T}{rN_0}, \dots, \frac{(n-1)T}{rN_0}, \dots, \frac{(N_r-1)T}{rN_0} \right] \quad (14)$$

The discrete data obtained after the sampling rate conversion is further represented by the vector Y' :

$$Y' = [y'(1), y'(2), \dots, y'(k), \dots, y'(N_{CP} + N_0)] \quad (15)$$

The sampling moment corresponding to each data in (15) is expressed as:

$$T' = \left[0, \frac{T}{N_0(1+a)}, \frac{2T}{N_0(1+a)}, \dots, \frac{(k-1)T}{N_0(1+a)}, \dots, \frac{(N_{CP} + N_0 - 1)T}{N_0(1+a)} \right] \quad (16)$$

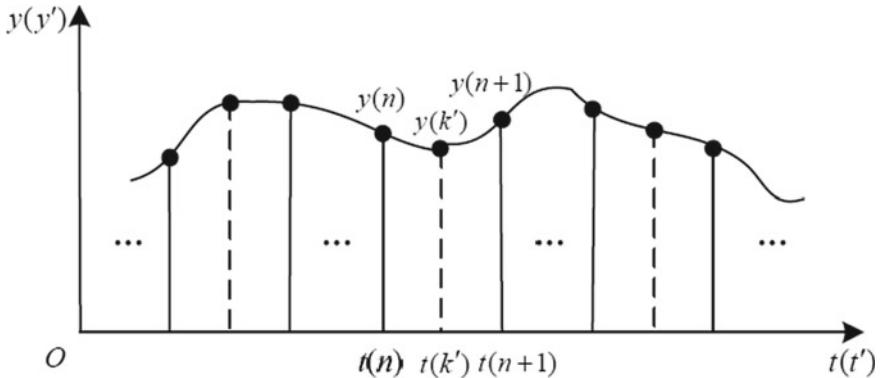


Fig. 1 Sample rate conversion based on linear interpolation

From (13) to (16), it is observed that, if data in (15) is obtained by using oversampled data in (13), the sampling rate conversion is completed. To obtain data in (15) from (13), data of (13) is linearly interpolated to obtain the data of each sampling in (16).

Figure 1 illustrates sampling rate conversion using linear interpolation process. The interpolation process is given below. The sampling time corresponding $Y'(k)$ in (15) is:

$$t'(k) = \frac{(k-1)T}{N_0(1+a)} \quad (17)$$

Assuming that $Y'(k)$ in (13) is between $y(n)$ and $y(n+1)$, the moment $t(n)$, $t(n+1)$ of $t'(k)$ in (14), $y(n)$ and $y(n+1)$ correspond to $t(n)$ and $t(n+1)$, respectively. According to (14), $t(n)$ and $t(n+1)$ is represented as:

$$\begin{cases} t(n) = \frac{(n-1)T}{rN_0} \\ t(n+1) = \frac{nT}{rN_0} \end{cases} \quad (18)$$

$y'(k)$ is estimated by linear interpolation as follows:

$$y'(k) = \left(1 - \frac{t'(k) - t(n)}{t(n+1) - t(n)}\right)y(n) + \left(\frac{t'(k) - t(n)}{t(n+1) - t(n)}\right)y(n+1) \quad (19)$$

Considering (14) to (16), (19) is obtained as:

$$y'(k) = \left[1 - r\left(\frac{(k-1)}{(1+a)} - \frac{(n-1)}{r}\right)\right]y(n) + r\left(\frac{(k-1)}{(1+a)} - \frac{(n-1)}{r}\right)y(n+1) \quad (20)$$

In (20), a positive integer n satisfies the following equation:

$$\frac{r(k-1)}{1+a} < n < \left(\frac{r(k-1)}{1+a} + 1 \right) \quad (21)$$

Through the above interpolation process, sampled data y' is obtained by removing N_{CP} samples.

The traditional sampling rate conversion algorithm has a large amount of computation and requires a large number of hardware devices. For example, a large number of polyphase filters will be used, so when the number of subcarriers is very large, the traditional algorithm is large in terms of computational complexity and hardware overhead. The linear interpolation sampling rate conversion algorithm can obtain the new sample value of the signal by linear interpolation according to existing data, thereby completing the conversion of the sampling rate. Since the linear operation does not require a large amount of calculations, the sampling rate conversion by this method can greatly improve the system efficiency. For narrowband systems, the use of this algorithm in baseband can save considerable computational complexity, while for underwater acoustic signals; the difference between the two is small. It is sometimes more convenient to apply this linear algorithm to signals. In addition, it should be noted that the selection of the sampling factor r should be large enough to take into account, system errors, complexity, signal to noise ratio (SNR) and the signal-to-distortion ratio (SDR) of the linear interpolator. So, to reduce the error caused by linear interpolation, the signal-to-distortion ratio can refer to the following formula [11]:

$$\text{SDR}(dB) \cong 40 \ln\left(\frac{f_{rs}}{2f}\right) = 40 \ln(r) \quad (22)$$

where f_{rs} is the oversampling frequency; f is the frequency of the signal.

5 Simulation Results

The simulation results of the proposed algorithm are carried out for Doppler shift factor estimation and sampling rate conversion. UWA communication system adopts the ray theory [12, 13], the maximum delay spread of UWA channel is set to 8 ms. Corresponding to actual shallow water communication channel in the range of 2–7 km, the receiver path is considered with a delay >8 ms as noise processing. The carrier frequency is set to 24 kHz, the maximum Doppler shift that is processed is 82 Hz, which is equivalent to Doppler frequency generated when relative speed between transmitter and receiver is 8 knots shift. The system uses comb pilots and all

Table 1 GFDM simulation parameters

Parameters	Value
FFT size	256
Data subcarriers	186
Pilot carriers	64
Empty carriers	24
Carrier frequency	18 KHz
Cyclic prefix length	8.7 ms
Symbol period	42.3 ms
Oversampling factor	$r = 2, 4, 8$
Modulator	4QAM; 16QAM
Data rate	16.3 (4QAM)/KHz; 21.2 (16QAM)/KHz
Maximum doppler shift	62 Hz
Carrier frequency interval	23.44 Hz
Bandwidth	11 KHz

pilot subcarriers have associated power and same frequency spacing. The oversampling factor r used is $2 \times$, $4 \times$, and $8 \times$ Nyquist sampling rates, respectively. Table 1 illustrates a detailed description of system simulation parameters. In the simulation process, conventional algorithm is compared with an improved algorithm in terms of a received signal constellation and the bit error rate (BER) curve. To minimize the complexity, the simulation process in this paper adopts the non-coding method.

Figure 2 illustrates the 16QAM constellation diagrams with oversampling factors $r = 2 \times, 4 \times, 8 \times$. It is observed that with different r values, the receiver performance is different. The estimation accuracy of DSF and linear interpolated SDR are greatly improved with $r = 4$. Figure 3 illustrates the performance comparison of proposed, conventional methods with 4QAM and 16QAM modulation, respectively. It is observed that under 4-QAM modulation, the proposed and conventional methods perform similarly without much improvement. The performance of the proposed method is significantly improved with 16-QAM modulation. From Figs. 2 and 3 it is observed that the proposed method performs better when compared with conventional methods under a higher modulation scheme. The performance improvement of the system is more prominent under large oversampling rates and high modulation, which reflects the advantages of the improved algorithms in processing high-speed data.

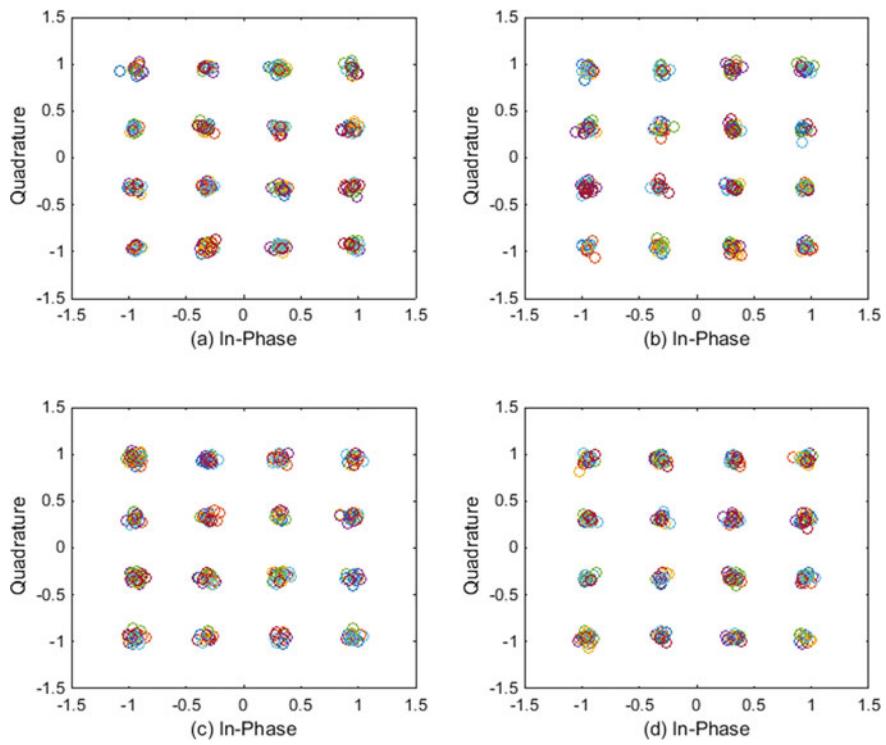


Fig. 2 Comparison of 16-QAM constellations with different oversampling rates, **a** original signal constellation, **b** proposed algorithm with $r = 2$, **c** proposed algorithm with $r = 4$, **d** proposed algorithm with $r = 8$

6 Conclusion

In this paper, an improved algorithm is addressed to solve the Doppler shift problem. The main idea is to use the oversampling technique as the basis for solving the problem. Using oversampling the data is combined with a linear interpolation algorithm, to estimate Doppler shift factor and sampling rate conversion. From simulation results and analysis, the performance of the proposed method is improved at higher data rates with less computational complexity than conventional methods.

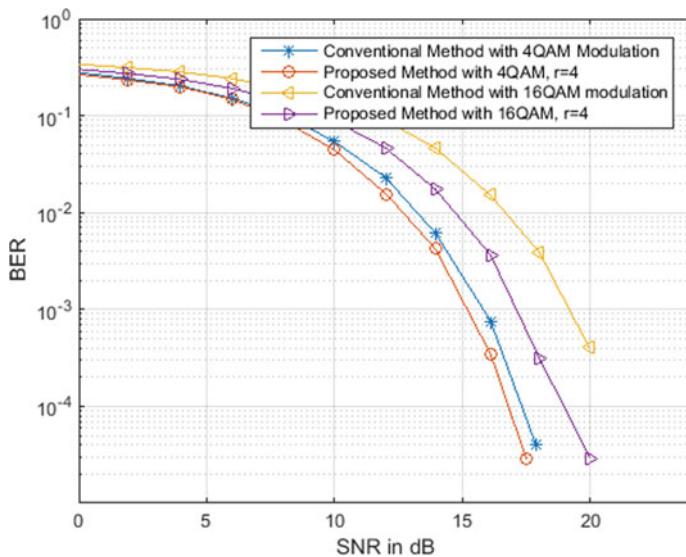


Fig. 3 BER performance comparison

References

- Huang J et al (2010) Comparison of basis pursuit algorithms for sparse channel estimation in underwater acoustic OFDM. In: OCEANS'10 IEEE Sydney, pp 1–6. IEEE
- Lim B, Ko YC (2017) SIR analysis of OFDM and GFDM waveforms with timing offset, CFO, and phase noise. *IEEE Trans Wireless Commun* 16(10):6979–6990
- Ara M (2013) Reliable and secure wireless communications systems: a physical-layer approach
- Kumar JT, Kumar VS (2020) Novel distance-based subcarrier number estimation method for OFDM system. In: International conference on Modelling, Simulation and Intelligent Computing. Springer, Singapore pp 328–335
- Kollam S, Reddy KRL, Rao DS (2019) Denoising and segmentation of MR images using fourth-order non-linear adaptive PDE and new convergent clustering. *Int J Imaging Syst Technol* 29(3):195–209
- Ren G et al (2007) An efficient frequency offset estimation method with a large range for wireless OFDM systems. *IEEE Trans Veh Technol* 56(4):1892–1895
- Wu J et al (2017) Influence of pulse shaping filters on PAPR performance of underwater 5G communication system technique: GFDM. *Wireless Commun Mobile Comput*
- Li B, Zhou S, Stojanovic M et al (2008) Multicarrier communications over underwater acoustic channels with nonuniform doppler shifts. *IEEE J Oceanic Eng* 33(2):198–209
- Kumar VS, Tarun Kumar J (2019) NC-OFDM/OQAM based cognitive radio network. LAP Lambert Academic Publishing (2019) 13
- Rao AS, Garige, SV (2019) IoT based smart energy meter billing monitoring and controlling the loads. *Int J Innov Technol Exploring Eng* 8(4):340–344
- Sharif BS, Neasham J, Hinton OR et al (2000) A computationally efficient doppler compensation system for underwater acoustic communications. *IEEE J Oceanic Eng* 25:52–61

12. Kumar VS (2020) Joint iterative filtering and companding parameter optimization for PAPR reduction of OFDM/OQAM signal. *AEU-Int J Electr Commun* 153365
13. Hebbar RP, Poddar PG (2017) Generalized frequency division multiplexing for acoustic communication in underwater systems. In: 2017 International conference on circuits, controls, and communications (CCUBE). IEEE

A Novel Recommendation System for Housing Search: An MCDM Approach



Shreyas Das, Swastik Ghosh, Bhabani Shankar Prasad Mishra, and Manoj Kumar Mishra

Abstract Recommendation system is the part and parcel of modern days' information retrieval system. It takes input from the users and comes up with a personalized list of choices or alternatives to the user. This process makes the selection easy for the users. Recommendation system is helpful where the number of available choices is too large. In this paper, we have proposed a rank-based flat recommendation system. We took importance factors against the housing attributes from the users along with some preference values. These were processed through PROMETHEE II method to generate a rank-based recommendation system.

Keywords Recommendation system · PROMETHEE II · Housing search · E-commerce

1 Introduction

In recent days, people move a lot from one place to another due to their jobs or other affairs. The very first problem they face in a new city is the problem of accommodation. Here comes the need of housing recommendation system. Recommendation system is one of the key components of modern days' information retrieval systems. It provides the solutions of information overload problems. Recommendation systems are used where both the search space and number of available alternatives are typically vast [1]. Currently, recommendation systems are being used in a large scale because it comes up with a list of personalized solutions, which is time efficient for the users. Recommendation systems are used in different domains like e-learning, tourism services, e-commerce, social media, housing etc.

In our study, we have proposed a housing recommendation system. By help of this recommendation system, both the home-seekers and real-estates will be helpful. Housing recommendation system will give an idea to the real-estate where to make

S. Das (✉) · S. Ghosh · B. S. P. Mishra · M. K. Mishra
KIIT Deemed to be University, Bhubneswar, India
e-mail: shreydas29@yahoo.com

housing units and which types of people are mostly going to stay there. Such business patterns will be helpful for them.

Our proposed model is an MCDM approach, as it involves multiple criterions. There are conflicting criteria too. Everyone wants to avail maximum facilities at a minimal cost. In order to solve this problem, we applied PROMETHEE II method in our work.

2 Literature Survey

Recommendation system is one of the key components of modern days' information retrieval system. Recommendation system can be of three types based on different mechanisms it uses. These are collaborative, content-based and hybrid. There are different soft-computing techniques which can improve the overall performance of recommendation systems [2].

Conventional housing recommendation systems can not ease the pain of a home-buyer while the person searches for a house. It produces an in-depth search results based on the person's needs. Finalizing a house from this vast result set takes a good amount of time. So, the researchers investigated a general search pattern of the home-buyer [3]. Using this search pattern, they implemented a user-oriented housing recommendation system. They considered both ontological structure and case-based reasoning to employ the recommendation system.

In recommendation systems, there are different techniques to filter the data it has. Based on different filtering techniques used, recommendation systems can be broadly classified into three types. These are collaborative, content-based, hybrid. In case of collaborative filtering, the recommendation system uses the ratings provided by the users against the items they bought. This can further be classified into two types. The first one is item-based filtering. In this case, if a user rates similar or almost similar to two similar items, then those items are grouped together. Another type of collaborative technique is user-based filtering. In this method, if an item is rated similar by two different users, then these users are grouped together. In content-based filtering, the semantics associated with the items are used to generate the recommendation. These two types of recommendation systems have its own pros and cons, which can be avoided by using hybrid filtering. Hybrid filtering takes the advantages of both collaborative and content-based filtering techniques [1].

Badriyah et al. [4] proposed a content-based recommendation system for housing search. It groups the housing properties like region, country, ownership status, size, amenities offered etc. In this study, users' profiles were created using tf-idf method. Each time a user visited the semantics of the advertisements, the data was stored in the database. This data was further processed using Apriori algorithm to generate frequent itemsets, from which the final recommendation was shown.

Alrawhani et al. [5] proposed a case-based housing recommendation system which uses collaborative filtering. In this method user-based collaborative filtering method was used. In case of case-based reasoning, it is seen that whether a similar type of

problem was solved beforehand or not. If yes, then the recommendation system comes up with the same solution. But if the problem is not known to the recommendation system, then it tries to solve it from scratch and generate recommendation and this recommendation goes to the system's database for future reference.

In [6], the researchers implemented PROMETHEE II method in an e-commerce recommendation system. They considered four attributes like price, brand, customer's interest and reviews given by the users.

3 Proposed Approach

3.1 PROMETHEE

PROMETHEE is the abbreviated form of Performance Ranking Organization Method for Enrichment of Evaluations. It was first introduced by Prof. Jean-Pierre Brans in 1982. There are different forms of it such as PROMETHEE I, II, III, IV, V, VI, TRI, Cluster and Fuzzy. In this study, we have used PROMETHEE II method. This method provides ranking based list of the alternatives.

The PROMETHEE II method consists of seven steps [7, 8].

Step 1: Normalization of the decision matrix using the following equation:
for beneficial criteria,

$$\frac{X_{i,j} - \min(X_{i,j})}{\max(X_{i,j}) - \min(X_{i,j})} \quad (1)$$

for non-beneficial criteria,

$$\frac{\max(X_{i,j}) - X_{i,j}}{\max(X_{i,j}) - \min(X_{i,j})} \quad (2)$$

Step 2: Calculate the pair-wise differences of the i th alternative with respect to the other alternatives.

Step 3: Calculate preference function, $P_j(i, i')$. The preference function is calculated using the formula enlisted below.

$$\begin{aligned} P_j(i, i') &= 1, && \text{if } R_{i,j} > R_{i'j} \\ &= 0, && \text{otherwise} \end{aligned} \quad (3)$$

Step 4: Evaluate the aggregated preference functions.

$$\pi(i, i') = \left[\sum_{j=1}^m P_j(i, i') w_j / \sum_{j=1}^m w_j \right] \quad (4)$$

w_j is the importance factor of the j th attribute.

Step 5: Calculate the positive and negative outranking flows as follows:

Positive or leaving outranking flow for i th alternative,

$$\varphi^+ = \frac{1}{n-1} \sum_{j'=1}^n \pi(i, i') \quad (5)$$

Negative or entering outranking flow for i th alternative,

$$\varphi^- = \frac{1}{n-1} \sum_{j'=1}^n \pi(i', i) \quad (6)$$

where $i \neq i'$ and n is the number of alternatives

Step 6: Calculate net outranking flow accordingly.

$$\varphi(i) = \varphi^+(i) - \varphi^-(i) \quad (7)$$

Step 7: Generate the rank-based list of the alternatives. The higher value of $\varphi(i)$ means the higher rank of i th alternative.

3.2 Proposed Model

The work-flow of the proposed architecture follows several steps.

- Step 1: User provides his preferred choice and importance factors for the housing attributes.
- Step 2: The system filters its database based on the preferred choices of the highly important attributes (user-specific).
- Step 3: Applied PROMETHEE II method over filtered dataset to generate the rank-based recommendation.

The four housing attributes that we dealt with were location, price, flat furnishing and flat size. There are two types of criterion in PROMETHEE II method. One is beneficial criteria and another one is non-beneficial criteria. Here price is the only non-beneficial criteria, as we prefer to avail maximum facilities at a minimal cost. The preferred price refers the upper budget of the user while he searches for housing properties. In order to get the decision matrix we assigned all the variables numeric

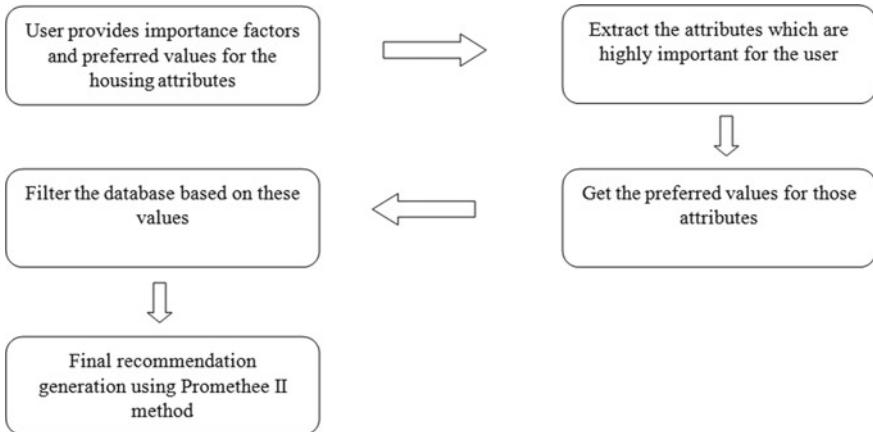


Fig. 1 Work-flow model of the proposed recommendation system

values. For example, we assigned ‘unfurnished’ flats as 1, whereas we assigned 2 and 3 to the ‘semi-furnished’ and ‘fully furnished’ flats, respectively. The same thing we did for the 1 BHK, 2 BHK and 3 BHK flats. We assigned 1 for 1 BHK, 2 for 2 BHK and 3 for 3 BHK flats. But for the locations, we adapted a different mechanism. We found number of different distinct locations from the dataset we had. In our dataset, we encountered different locations like Magarpatta, Hadapsar, Mundhwa and Amanora Park Town. We found the distances between each location. We have assigned the maximum value to the preferred location of the user. The second highest value was assigned to the location, which is closest to the user’s preferred location. For example, consider the preferred location for the user is ‘Magarpatta’. So, we assigned 4 to it, considering 4 is the highest value. We found that the distance of ‘Amanora Park Town’ from ‘Magarpatta’ is minimal. So, we assigned 3 for ‘Amanora Park Town’ when the preferred location is ‘Magarpatta’ (Fig. 1).

4 Experiment and Result Analysis

We extracted 30 housing units from www.magicbricks.com using webscraper.io. In our study, we considered 4 housing attributes, i.e., locations, flat size, flat type and price. There are 4 different locations. Flat type signifies how furnished those flats are. Based on furnishing there are fully furnished, semi-furnished and unfurnished flats are available. Flat size denotes number of bedrooms is there along with hall and kitchen (BHK).

In the Fig. 2, we have shown how number of houses gets changed with the change of importance factors for the housing attributes. In this case, we have kept the preferred values intact for all the cases. For first column, price and flat size are the two most important attributes for the user. As, 1,50,000 is the maximum price

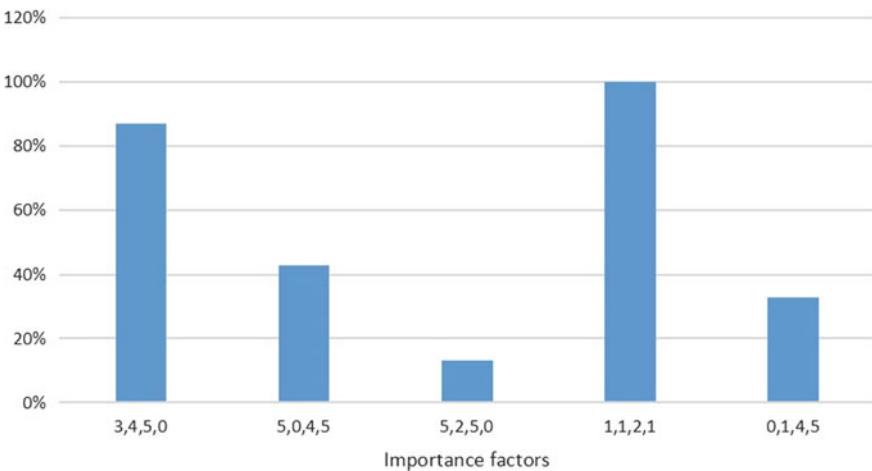


Fig. 2 Percentage of number of houses shown to the user for different importance factors

value in the entire dataset, all the houses fall in this category. For this instance, more than 80% houses are shown to the user. Here the dataset gets filtered where flats are of 2 BHK. But, for the 4th instance all the houses are shown to the user. The reason behind the same is all the housing attributes are less important to the user.

In Table 1, we have shown how the ranking of recommended flats get changed with the change of importance factors by using PROMETHEE II method. In the above mentioned table, the first column contains the importance factors of ‘location’, ‘price’, ‘flat size’ and ‘flat type,’ respectively. Here, we kept the preferred values same for all the cases. In the third row, we can see only 4 flats are recommended. As the importance factors of ‘location’ and ‘flat size’ are very high, the database got filtered accordingly and came up with a comparatively less number of flats.

In Fig. 3, we can see the importance factors for the attributes. The user has rated 5, 2, 4 and 3, respectively, for the housing attributes like location, price, size and flat type, respectively. Higher importance factor denotes that the attribute is highly important to the user. The preferred choices for the same attributes are ‘Mundhwa’, 1,50,000, 2 BHK and unfurnished, respectively. As the most important attributes for the user are location and size. So, we filtered our database based on the housing properties which are of 2 BHK and located at ‘Mundhwa’.

Table 1 Top 5 rank-based flat ids with different importance factors

Importance factors	Top 5 rank-based flat ids
3, 4, 5, 0	16, 10, 6, 12, 5
5, 0, 4, 5	12, 10, 16, 1, 5
5, 2, 5, 0	16, 10, 12, 25
1, 1, 2, 1	12, 16, 10, 13, 9
0, 1, 4, 5	1, 6, 5, 30, 17

```

Important factors for the corresponding attributes are:
[5, 2, 4, 3]
User's preferences are:
[('Mundhwa'), [150000], ['2'], ['Unfurnished']]
Most important pref values for the user is/are:
[['2'], ['Mundhwa']]
Recommended list of flats are:
[['Mundhwa', 21000, 3, 'Semi-Furnished'], ['Mundhwa', 20000, 2, 'Furnished'], ['Mundhwa', 22000, 3, 'Semi-Furnished'], ['Mundhwa', 16000, 1, 'Unfurnished']]

```

Fig. 3 Sample output

Post that, we applied PROMETHEE II method to get the final rank-based recommended list. As discussed in the previous chapter, there are two types of criterion considered in case of multi-objective decision making approach. Here we considered, ‘price’ as non-beneficial criteria and all other attributes were considered as beneficial criteria, i.e., users want maximum facilities in a minimal cost. In the example, we have shown, we can see the user wants a flat at ‘Mundhwa’ only. So, locations of all the recommended flats are at ‘Mundhwa’ only. Moreover, the user has a strict preference over 2 BHK flats. But he is ready to negotiate with it a bit as the attribute is rated as 4. That’s why our recommendation system has recommended some flats of size 3 BHK and 1 BHK along with 2 BHK.

5 Conclusion

In this study, we have proposed a rank-based recommendation system for housing search. We have used PROMETHEE II method to get the rank-based recommended list. As no real user is involved in this study, we are planning to implement it in real market, so that we can measure the accuracy of our proposed model.

References

1. Isinkaye FO, Folajimi YO, Ojokoh BA (2015) Recommendation systems: principles, methods and evaluation. Egypt Inf J 16:261–273. <https://doi.org/10.1016/j.eij.2015.06.005>
2. Das S, Mishra BSP, Mishra MK, Mishra S, Moharana SC (2019) Soft-Computing based recommendation system: a comparative study. Int J Innov Technol Exploring Eng (IJITEE) 8(8):131–139
3. Lee J-H, Yuan X, Kim S-J, Kim Y-H (2013) Toward a user-oriented recommendation system for real estate websites. Inf Syst 38:231–243. <https://doi.org/10.1016/j.is.2012.08.004>
4. Badriyah T, Azvy S, Yuwono W, Syarif I (2018) Recommendation system for property search using content based filtering method. In: 2018 International conference on information and communications technology (ICOIACT), pp 25–29. <https://doi.org/10.1109/icoiaact.2018.8350801>
5. Alrawhani EM, Basirona H, Saáyaa Z (2016) Real estate recommender system using case-based reasoning approach. J Telecommun Electron Comput Eng 8(6):177–182

6. Niknafs A, Charkari NM, Niknafs AA (2008) A PROMETHEE-based recommender system for multi-sort recommendations in on-line stores. In: 3rd International conference on digital information management, ICDIM 2008, pp 399–404. <https://doi.org/10.1109/icdim.2008.4746743>
7. Athawale VM, Chakraborty S (2010) Facility location selection using promethee ii method. In: Proceedings of the 2010 international conference on industrial engineering and operations management Dhaka
8. Amaral TM, Costa AP (2014) Improving decision-making and management of hospital resources: an application of the promethee ii method in an emergency department. Oper Res Health Care. <https://doi.org/10.1016/j.orhc.2013.10.002>

Health-Related Tweets Classification: A Survey



Kothuru Srinivasulu

Abstract With the rise of social media platforms, a lot of data is available online in the form of tweets, reviews and posts. The data shared includes text related to health domain also. Due to the richness of information available in the shared texts, research community started to utilize this shared text in various applications like Pharmacovigilance. Even though research community started to develop systems to automatically classify health-related tweets, there is no paper which provides a review of various systems developed for automatic health-related tweets classification. In this survey paper, we provide a review of systems developed for health-related tweets classification.

Keywords Health tweets · Survey · Text classification · Natural language processing

1 Introduction

Evolution of Internet and social media platforms like Twitter, Facebook, Reddit offered general public a medium to share information. Information is shared in the form of tweets in Twitter, posts in Facebook and Reddit, question and answers in online discussion forums. This information shared by general public includes text related to medical domain also. For example, (a) patients share their health experience, reviews on the medications they use in the form of tweets; (b) health-related questions and answers in online medical discussion forums like Askapatient etc. As the text shared is freely accessible and contains rich medical information, text corpus can be mined to get valuable insights [1].

Handling user generated texts is challenging as the texts are noisy with (a) irregular abbreviations and irregular grammar, (b) lot of misspelled words. Moreover general public express their reviews or opinions mostly using colloquial words in a casual

K. Srinivasulu (✉)
National Institute of Technology, Tiruchirappalli, India
e-mail: sinu.kothuru@gmail.com

language. Further, length of text in platforms like Twitter is very short. A tweet is allowed to have a maximum of 140 characters. So, the context expressed in tweets is sometimes ambiguous because of lack of enough contextual information.

Research community started to utilize health-related tweets because of their rich medical information context, to provide better health services. Health Language processing laboratory of university of Pennsylvania organized a series of shared tasks [2–4] to develop systems to leverage rich medical information in health-related tweets. The laboratory organized shared tasks to identify tweets reporting (a) Adverse Drug Reactions, (b) Personal Health Mentions, (c) Vaccination behavior and (d) Drug names. Further, they organized shared tasks to extract and normalize ADR mentions i.e., to identify the spans of adverse drug reactions and then map them to standard medical concepts. Even though there is a rising interest in research community to develop systems to automatically classify health-related tweets, there is no paper which provides a review of various systems developed for automatic health-related tweets classification. In this survey paper, we provide a review of systems developed for health-related tweets classification.

1.1 Literature Selection

Due to the recent popularity of automatic tweets classification task, we collected papers related to SMM4H 2017 [2], 2018 [3] and 2019 [4] shared tasks. SMM4H 2017 organized shared tasks related to (a) ADR tweets classification, (b) Medication intake tweets classification and (c) ADR normalization. SMM4H 2018 organized shared tasks related to (a) ADR tweets classification, (b) Medication intake tweets classification, (c) Drug tweets classification and (d) Vaccination behavior tweets classification. SMM4H 2019 organized shared tasks related to (a) ADR tweets classification, (b) Extraction of ADR mentions, (c) Normalization of ADR mentions, (d) Personal Health mentions tweets classification. In this survey paper, we provide a review of systems submitted for the shared tasks (a) ADR tweets classification, (b) Medication intake tweets classification, (c) Drug tweets classification and (d) Personal Health mentions tweets classification. In these tasks, Personal Health mention tweets classification is a three way text classification problem while others are binary text classification.

2 Embeddings

Machine learning systems require numerical vectors as input. So, text has to be converted into numerical vectors before machine learning algorithms. Various localized representations based on word frequency, N-grams etc. are used. In these representations, there is a one-to-one mapping between word and its representation. The main disadvantages in these are (a) large size of vectors, (b) completely different

vectors even for similar words, (c) require more computational resources and time to process these large vectors. Distributed vector representations or embedding inspired from distributional hypothesis represent words in low dimensional vector space in a way that similar words are closer. In embedding, the meaning of a word is distributed across many dimensions. Some of the popular embedding models are Word2vec, Glove [5], FastText [6], ELMo [7] and BERT [8].

Word2vec proposed by Mikolov et al. attracted research community with its simple and effective architecture to learn embeddings. Continuous Bag of Words (CBOW) and Skipgram are the two variants in Word2vec. Both these models are based on three layered neural network. CBOW learns vectors by predicting focal word using context words. CBOW model is trained using

$$J = \frac{1}{M} \sum_{i=1}^M \log p(w_i | w_{i-n}, \dots, w_{i-2}, w_{i-1}, w_{i+1}, w_{i+2}, \dots, w_{i+n}) \quad (1)$$

where M is the number of unique words in training corpus. Here w_i is the focal word and $w_{i-n}, \dots, w_{i-1}, w_{i+1}, w_{i+2}, \dots, w_{i+n}$ are context words. Skipgram learns vectors exactly opposite to CBOW models. It learns word vectors by predicting context words based on focal words. Skipgram model is trained using

$$J = \frac{1}{M} \sum_{i=1}^M \sum_{-n \leq j \leq n, j \neq 0} \log p(w_{i+j} | w_j) \quad (2)$$

To reduce the overhead in learning vectors in both these models, methods like Hierarchical SoftMax and negative sampling were used.

The main drawback in Word2vec is that it is unable to leverage global counts in training corpus. Glove model proposed by Pennington et al [5]. leverages both local context information like Word2vec as well as global count values like LSA. It is basically log bilinear model trained using weighted least square objective. The weight function in objective function regulates weight values for rare as well as frequent co-occurrences. Here objective function is

$$J = \sum_{i,j=1}^M f(C_{ij})(u_i^T v_j + b_i + b_j - \log C_{ij})^2 \quad (3)$$

Here C_{ij} represents the number of times word ‘ i ’ co-occurs with word ‘ j ’ and it is calculated from the training corpus, while u_i and v_j are focal and context word vectors which are learned by the model.

Both Word2vec and Glove models learn vector representations for words by treating them as atomic units. As a result (a) no embedding for OOV words i.e., no vector representation for words that are missing in the training corpus. These words have to be assigned with zero or random vectors, (b) lack of sub word information in learned vector representations. To overcome these two drawbacks, Bojanin

et al. introduced FastText model [6] by improving Skipgram model with character n-grams. In this model, embeddings are learned for character n-grams instead of words. A word representation is obtained from the sum of vector representations of its character n-grams. Here objective function is

$$\sum_{i=1}^M \left(\sum_{c \in C_i} l(s(w_i, w_c)) + \sum_{n \in N_{i,c}} l(-s(w_i, n)) \right) \quad (4)$$

where M is vocabulary size, C_i and $N_{i,c}$ represents set of all the context for w_i and set of all negative examples for word ‘ i ’, ‘ l ’ and ‘ s ’ represents logistic loss and scoring functions respectively. Here the scoring function finds similarity between the given words.

Word2vec, Glove and FastText models assign context insensitive representation to a word i.e., these models assign same representation to a word irrespective of its context. To encode context information into word vector, models like ELMo and BERT were proposed. ELMo consists of two layer BiLSTM on the top of CNN+ Highway and it is trained using a language modeling objective. CNN generates context independent word representations from character embeddings by applying convolution and max pooling operations. BiLSTM predicts the words with word representing as input. ELMo model generates word vector as weighted average of three vectors namely context independent word vector and vectors from each of the BiLSTM layers.

$$\text{ELMo}_k^{\text{task}} = r^{\text{task}} (s_0^{\text{task}} h_{k0}^{LM} + s_1^{\text{task}} h_{k1}^{LM} + s_2^{\text{task}} h_{k2}^{LM}) \quad (5)$$

where s_i^{task} , $i = 0, 1$ and 2 are task specific weights and r^{task} is scaling factor.

The two main drawbacks in ELMo model are (a) the vectors learned are shallow bidirectional i.e., the vectors are obtained by concatenating vectors from forward and backward LSTM, (b) it is sequential in nature as it is based on LSTM. BERT overcomes these two drawbacks with Masked Language Modeling and Transformer Encoder. Masked Language modeling objective allows the model to encode bidirectional information while Transformer encoder which is based on self-attention can be run in parallel. BERT model initiated a new era in deep learning based Natural Language Processing by eliminating the need to train a downstream model from scratch.

3 Health-Related Tweets Classification

3.1 *ADR Tweets Classification*

Adverse Drug Reaction is the unwanted consequence of consuming a drug or combination of drugs as per the prescribed dosage. In many cases, deaths occurring after the treatment are because of adverse drug reactions. So, identifying adverse drug reactions in clinical text is of great importance. Automatic classification of ADR tweets is designed as a two way classification where the developed system should be able to identify whether a tweet contains ADR or not. The main challenge in this task is that the developed system should be able to differentiate ADR and the reason to use the drug. Table 1 summarizes the details of all systems developed for this task. As reported in Table 1, majority of the systems [9–24] are based on deep learning models.

3.2 *Health Mention Tweets Classification*

In general, Internet users share information related to their health as well as discuss information related to various general health issues. The objective of this task to develop a system which will identify the tweets reporting personal health mentions. Here, the challenge is to distinguish personal health mentions from general health issues. Table 2 summarizes the details of various systems developed for automation health mention tweets classification.

3.3 *Drug Tweets Classification*

The objective of this task is to develop systems which can differentiate tweets with and without drug names. Formally, it is a binary text classification where the system is to assign the label ‘0’ for tweets without drug names and ‘1’ for tweets with drug names. It is challenging as the system has to effectively make use of the context information in the tweets to identify the drug mentions. Table 3 summarizes the details of various systems developed for automation drug mention tweets classification.

Table 1 Summary of systems developed for adverse drug reaction tweets classification

System	Model	Embedding	Syntactic features	Semantic features
[9]	BERT	WordPiece	–	SIDER 4.1 features
[10]	BERT+Logistic regression	WordPiece	–	–
[11]	Ensemble of (a) CNN + BiLSTM (b) BERT	Word2vec Twitter embeddings and WordPiece	–	MedDRA features
[12]	ULMFiT	Pretrained word embeddings	–	–
[13]	BiLSTM	Twitter glove	–	–
[14]	BiLSTM+Multi Head Self-attention	Character and Word2Vec Twitter word embeddings	POS	SentiWordNet and SIDER 4.1 medical lexicon features
[15]	CNN	Twitter Glove	–	–
[25]	SVM	–	–	cTAKES features
[27]	SVM	Twitter Glove	Negation	ADR lexicon features
[16]	BiLSTM	Word2vec and BERT	–	–
[17]	ULMFiT	–	–	–
[18]	LSTM	Character, Twitter Glove and BERT	–	–
[19]	BiLSTM+MHSA	Character, Word2vec Twitter	POS	SentiWord Netscore and SIDER 4.1 lexicon features
[21]	BiLSTM	FastText	–	MetaMap features
[28]	BiLSTM+SVM	FastText	–	–
[22]	CNN+ attention	Word2vec twitter	–	–
[26]	K2 Bayesian network	Word2vec embeddings	–	cTAKES, Sentiment score and ADR Tags
[23]	CNN	GoogleNews Word2vec embeddings	–	ADR Lexicon features and SentiWord score
[24]	CNN	Health Twitter embeddings	–	–
[30]	Naïve bayes	–	POS	ADR Lexicon features

Table 2 Summary of systems developed for automation health mention tweets classification

System	Model	Embeddings	Syntactic features	Semantic features
[11]	BERT	WordPiece	–	–
[12]	ULMFiT	Pretrained word embeddings	–	–
[16]	BiLSTM+SVM	Word2vec and BERT embeddings	POS and modality features	–

Table 3 Summary of systems developed for drug tweets classification

System	Model	Embeddings	Syntactic features	Semantic features
[18]	BiLSTM+MHSA	Character and Word2vec Twitter	POS	SentiWordNet score and SIDER 4.1 lexicon features
[21]	BiLSTM	FastText	–	MetaMap features
[31]	SVM	–	–	–
[28]	NBSVM	–	–	–
[29]	CNN-LSTM	Character embeddings	–	–

3.4 Medication Intake Tweets Classification

It is a three way text classification problem where the system has to classify given tweets into one of the three classes namely ‘Intake’, ‘Possible Intake’ and ‘Non-intake’. Intake means, tweets contains mentions of personal medication consumption, ‘Possible Intake’ means it is not sure but there is possibility of medication consumption by the user and ‘Non-intake’, means tweets contains mention of medication names only but not the consumption. Table 4 summarizes the details of various systems developed for automatic medication intake tweets classification.

4 Discussion

Shift from CNN or RNN based models to BERT: There is a shift from CNN or RNN based models to BERT in developed systems for automatic health-related tweets classification. CNN or RNN based models are to be trained from scratch using task specific labeled dataset. Most of the datasets in clinical domain are small in size as it is time consuming and expensive to label large number of instances. By training CNN or RNN based downstream models from scratch using small datasets, the parameters of the models are not fully learned. So, the performance of model is limited. BERT is pretrained using large unlabeled text corpus and then the pretrained model is fine-tuned using task specific labeled dataset. As the weights of BERT models are

Table 4 Summary of systems developed for medication intake tweets classification

System	Model	Embeddings	Syntactic features	Semantic features
[20]	Stacked BiLSTM with context aware attention	Word2vec Twitter	–	–
[21]	BiLSTM	FastText	–	MetaMap features
[29]	CNN-LSTM	Character embeddings	–	–
[22]	CNN+ attention	Word2vec Twitter	–	–
[24]	CNN	Health Twitter embeddings	–	–
[30]	Naïve bayes	–	POS	ADR Lexicon features

pretrained, the model can be fine-tuned even with small datasets. So, BERT based models perform better than CNN or RNN based models, in general.

Use of semantic features: Some of the systems utilized domain specific semantic features (a) from tools like cTAKES, MetaMap [21, 25, 26], (b) from knowledge sources like MedDRA [11] and SIDER 4.1 [9, 14, 18, 19]. Use of semantic features from domain knowledge sources helps the model to learn more and as a result, the performance of model improves.

Out Of Vocabulary (OOV) words: The main issue when dealing with noisy text like tweets is misspelled words. Use of embeddings inferred from less noisy text like News corpus or Web Crawl results in more number of vocabulary words. As embeddings are missing for these words, these words have to be assigned with zero or random vectors. By assigning zero or random vectors, the meaning of these words is ignored. With large number of OOV words and random or zero vectors assigned vectors to these words, the model is unable to learn enough patterns from the training instances and hence wrongly classify many of the instances. So to handle this problem, systems used (a) embeddings from tweets [11, 13–15, 18–20, 27], (b) embeddings models like FastText which represent a word vector a sum of vectors of its character n-grams [21, 28], (c) embeddings from health-related tweets [24] and (d) generated word vectors from character embeddings [29]. By inferring embeddings from tweets, the number of OOV words is reduced.

5 Conclusion

In this survey paper, we presented a review of various systems developed for health-related tweets classification. We discussed systems developed for (a) ADR tweets

classification, (b) Health mention tweets classification, (c) Drug tweets classification and (d) Medicine intake tweets classification. In these systems, we observed three trends namely (a) Shift from CNN or RNN based models to BERT, (b) use of semantic features and (c) use of embeddings inferred from tweets or FastText embeddings or ELMo or BERT embeddings to handle out of vocabulary words.

Acknowledgements I would like to thank K. V. Iyer for his valuable suggestions in writing the paper.

References

1. Subramanyam KK, Sivanesan S (2019) SECNLP: a survey of embeddings in clinical natural language processing. *J Biomed Inf* 103323
2. Sarker A, Gonzalez-Hernandez G (2017) Overview of the second social media mining for health (SMM4H) shared tasks at AMIA 2017. *Training* 1(10,822):1239
3. Weissenbacher D, Sarker A, Paul M, Gonzalez G (2018) Overview of the third social media mining for health (SMM4H) shared tasks at EMNLP 2018. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 13–16
4. Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, Gonzalez G (2019) Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 21–30
5. Pennington J, Socher R, Manning C (2014) Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp 1532–1543
6. Bojanowski P, Grave E, Joulin A, Mikolov T (2017) Enriching word vectors with subword information. *Trans Assoc Comput Linguist* 5:135–146
7. Peters ME, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. arXiv preprint [arXiv:1802.05365](https://arxiv.org/abs/1802.05365)
8. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805)
9. Chen S, Huang Y, Huang X, Qin H, Yan J, Tang B (2019) HITSZ-ICRC: a report for SMM4H shared task 2019-automatic classification and extraction of adverse effect mentions in tweets. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 47–51
10. Miftahutdinov Z, Alimova I, Tutubalina E (2019) KFU NLP team at SMM4H 2019 tasks: want to extract adverse drugs reactions from tweets? bert to the rescue. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 52–57
11. Ellendorff T, Furrer L, Colic N, Aepli N, Rinaldi F (2019) Approaching SMM4H with merged models and multi-task learning. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, University of Zurich, pp 58–61
12. Dirkson A, Verberne S (2019) Transfer learning for health-related Twitter data. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 89–92
13. Cortes-Tejada J, Martinez-Romo J, Araujo L (2019) NLP@ UNED at SMM4H 2019: neural networks applied to automatic classifications of adverse effects mentions in tweets. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 93–95

14. Ge S, Qi T, Wu C, Huang Y (2019) Detecting and extracting of adverse drug reaction mentioning tweets with multi-head self attention. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 96–98
15. Úbeda PL, Galiano MCD, Martín-Valdivia MT, Lopez LAU (2019) Using machine learning and deep learning methods to find mentions of adverse drug reactions in social media. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 102–106
16. Bagherzadeh P, Sheikh N, Bergler S (2019) Adverse drug effect and personalized health mentions, CLaC at SMM4H 2019, tasks 1 and 4. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 123–126
17. Mahata D, Anand S, Zhang H, Shahid S, Mehnaz L, Kumar Y, Shah R (2019) MIDAS@ SMM4H-2019: identifying adverse drug reactions and personal health experience mentions from twitter. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 127–132
18. Aroyehun ST, Gelbukh A (2019) Detection of adverse drug reaction in tweets using a combination of heterogeneous word embeddings. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 133–135
19. Wu C, Wu F, Liu J, Wu S, Huang Y, Xie X (2018) Detecting tweets mentioning drug name and adverse drug reaction with hierarchical tweet representation and multi-head self-attention. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 34–37
20. Xherija O (2018) Classification of medication-related tweets using stacked bidirectional LSTMs with context-aware attention. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 38–42
21. Minard AL, Raymond C, Claveau V (2018) IRISA at SMM4H 2018: neural network and bagging for tweet classification
22. Han S, Tran T, Rios A, Kavuluru R (2017) Team UKNLP: detecting ADRs, classifying medication intake messages, and normalizing adr mentions on twitter. In: SMM4H@ AMIA, pp 49–53
23. Jain S, Peng X, Wallace BC (2017) Detecting twitter posts with adverse drug reactions using convolutional neural networks. In: SMM4H@ AMIA, pp 72–75
24. Magge A, Scotch M, Gonzalez G (2017) CSaRUS-CNN at AMIA-2017 tasks 1, 2: under sampled CNN for text classification. CEUR Workshop Proc 1996:76–78
25. Vydiswaran VV, Ganzel G, Romas B, Yu D, Austin A, Bhomia N, Chan S, Hall S, Le V, Miller A, Oduyebo O (2019) Towards text processing pipelines to identify adverse drug events-related tweets: university of michigan@ SMM4H 2019 task 1. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 107–109
26. Tsui FC, Shi L, Ruiz V, Barda AD, Ye Y, Xue D, Mi F, Jain U (2017) Detection of adverse drug reaction from twitter data. In: SMM4H@ AMIA, pp 64–67
27. Wang CK, Dai HJ, Wang BH (2019) BIGODM system in the social media mining for health applications shared task 2019. In: Proceedings of the fourth social media mining for health applications (# SMM4H) workshop & shared task, pp 117–119
28. Aroyehun ST, Gelbukh A (2018) Automatic identification of drugs and adverse drug reaction related tweets. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 54–55
29. Tokala S, Gambhir V, Mukherjee A (2018) Deep learning for social media health text classification. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 61–64
30. Wang CK, Chang NW, Su ECY, Dai HJ (2017) NTTMU system in the 2nd social media mining for health applications shared task. CEUR Workshop Proc 1996:83–86
31. Çöltekin Ç, Rama T (2018) Drug-use identification from tweets with word and character N-grams. In: Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task, pp 52–53

Smart Farming



Nihar Sardal, Ankit Patel, and Vinaya Sawant

Abstract Agriculture plays a vital role in the development of an agricultural country like India. As the population soars from today's 1.3 billion to an estimated 2 billion by 2050, the demand for food is expected to more than double. In an agricultural process, the farmer aims to achieve increased yield at the least cost. The number of factors affecting the farm is high which complicates the decision-making process. The proposed system aims to assist farmers in selecting the crop for cultivation using sensor data collected from the field (Shirsath et al. in 2017 International conference on intelligent computing and control (I2C2), Coimbatore, pp 1–5 2017 [1]). Sensors are connected to the Cloud using IoT. Analytics is performed on the real-time data stored in the cloud to analyze sensor data and identify any outliers. In the cloud, Machine Learning based real-time analytics is performed to analyze sensor data and identify any outliers. This system uses Machine Learning and IoT to develop an intelligent and affordable farming product. The techniques incorporated within this system improve the precision of the result and automate crop monitoring thus reducing human involvement. Real-time data collected can be utilized to predict disease using machine learning algorithms. The real-time update will alert the farmer by indicating which crop is in trouble, so the expenses on insecticides, pesticides will reduce.

Keywords Farming · Precision agriculture · Cloud application

N. Sardal (✉) · A. Patel · V. Sawant

Department of Information Technology, Dwarkadas J. Sanghvi College of Engineering, Mumbai, India

e-mail: nihar.sardal@djsce.edu.in

A. Patel

e-mail: ankit.patel@djsce.edu.in

V. Sawant

e-mail: vinaya.sawant@djsce.ac.in

1 Introduction

Agriculture contribution to the Indian economy is substantial. Indian agriculture sector accounts for 18% of India's gross domestic product (GDP) and provides employment to 50% of the country's workforce [2]. India has the capacity to grow enough food to meet the needs of its entire population, yet is unable to feed millions of them.

Farmers in agricultural nation like India use traditional techniques or depend on their intuition to decide the crop to be cultivated. These methods require constant human intervention.

Precision agriculture is a method that provides agricultural crops with a sufficient amount of required resources for that particular duration [3]. For example, traditional irrigation process has a typical time-based watering practice in which farmer irrigates the crop after a certain amount of time [4]. But the problem with this approach is that sometimes that crop doesn't need water so early, so ultimately that leads to wastage of water and affects the crop. Traditional systems provide generalized recommendations without considering the in-field variations [5]. Considering the above problems, we intend to develop an intelligent system which on deployment will provide detailed information about the suitability of the crop according to the farm parameters, real-time situation of the crop via analysis of the data collected through sensors employed in the farm.

2 Literature Review

2.1 *Literature Related to Existing Systems*

Soil Health Card

SHC indicates fertilizer recommendations and soil amendment required for the farm. SHC evaluates the quality of the soil considering the characteristics which are biological properties, water, etc. It is a tool to help the farmer in monitoring the field on their own experiences and knowledge of the soil.

Soil characteristics change after every cycle, a continuous cycle of testing should happen to ensure the requirements are fulfilled in a commensurate amount [6].

Flybird Innovations

Flybird provides irrigation-related services. Even though it provides irrigation-related solutions, it does not eradicate the ground problem. The farmer needs to be aware of the changes happening in the farm, i.e., the advantages of employing the product. Also, sensor-related data can be used in many different ways such as outlier analysis which is the answer for in-field variability, suitability recommendation based on soil type, storing it for further use, etc. [7].

To conclude, current systems like Soil Health card make a prediction based on soil factors which easily change with each harvest. It does not take into account the in-field variability. It takes one sample from 10-ha land leading to over-generalization. Also, the process consumes a lot of time. Other Systems such as Flybird use sensors only to schedule and control irrigation when different sensors have the capacity to monitor various aspects of a crop. They do not store and use data from sensors for anything which is a major wastage of resources. Thus, there is a need to create a system which stores and utilizes data gathered from various sensors to understand the properties and potential of a farm.

2.2 Literature Related to Methodology/Approaches

Following are various algorithms which can be used to create the model:

Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. As a predictive analysis, the multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical [8].

Support Vector Regression

Support Vector networks are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. In addition to performing linear classification, SVMs efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces [9].

K-means algorithm for clustering

K-means clustering is one of the simplest and most popular unsupervised machine learning algorithms.

Typically, unsupervised algorithms make inferences from datasets using only input vectors without referring to known, or labeled, outcomes [10].

2.3 Cloud Support

IoT

Platforms provide easy and secure connect, manage and collect data from multiple devices. This is combined with storage service to provide efficient storage and with analytics tools to provide valuable insights [11].

Database

Based on the requirement, the type of database can be chosen. For example, if the requirement is complex queries, transactions, etc. A relational database can be used. If the requirement is streaming, Mobile, Web, IoT, offline use then a NoSQL can be used. These services are provided by Cloud Computing platforms

Machine Learning

Machine learning (ML) uses algorithms to learn from data and prepares models by analyzing the data. Because of the learning-system, the overhead of ML is huge.

Since the data is continuously pushed from the sensors, the amount of data rises. The training of models requires high processing power which is provided by the cloud.

Cloud Storage

Cloud Storage is a unified object storage solution that allows worldwide storage and retrieval of any amount of data at any time [12].

2.4 Sensors to Measure Data

Some sensors that can be used to measure data are:

Ultrasonic sensors

It is used to monitor overwatering or underwatering of crops. Both the reasons damage the crops and hence affect the yield.

RTD (Resistance Temperature Detector)

Temperature sensors that contains a resistor that changes the resistance value when temperature changes.

Used because provides good range ($-200\text{--}800$) Celsius **pH sensor**

Major effects of extremes in pH levels include gaps in nutrient availability and the presence of high concentrations of minerals that are harmful to plants.

Soil moisture sensor

Extreme cases affect the pore size which is used for gas and water transfer

Humidity sensor

It measures humidity of air. High humidity can lead to soft growth, nutrient deficiency, edema etc. whereas low humidity leads to wilting, stunted plant, leaf curl etc.

3 Proposed Methodology

Internet of things (IoT) is a technology that allows real-time communication and data exchange between information devices [13]. It monitors, evaluates and analyzes data which is critical for decision-making. Sunlight, humidity, temperature, soil moisture, checking water-logging, pH are certain parameters of which data can be collected by deploying different sensors in the farm [14]. For example, when excessive water-logging occurs in some part of a field that can potentially damage the crop, the ultrasonic sensors sense it and the farmer is alerted via the mobile app. Thus, further damage is reduced due to real-time alerts.

Farmers often try to grow crops which are in high demand and can fetch them good prices. But finding suitability of the crop to be grown can be a challenge. Various parameters need to be taken into consideration and accurately measured. Inaccuracy in determining suitability can lead to poor growth of crops. Parameters which are required to determine suitability such as soil moisture, soil temperature can be measured using sensors in a time series fashion. Thus, we can use this data collected and use machine learning algorithms to easily and accurately determine suitability of various crops in an instant.

About 20% of the yield is destroyed every year with crop disease being the main culprit. This happens due to the fact that it is not possible to monitor the entire field and look for signs of disease being contracted. A plant contracts a disease in certain environments [15]. Sensors in the field can detect these conditions. Using data from the field and machine learning algorithms, we can predict the probability of contracting a disease. Thus, it gives a farmer to act before a disease is contracted and reduce potential damage in the future.

Further the data collected by the sensors can be used for various analysis. One application of analysis of the collected data can be using it to find variability in yield of a field. Suppose it has been observed that there exists inconsistency in the yield. This problem can be solved using outlier analysis. For example, if for a parameter the values range between 'X' and 'Y', if the value sensed is 'Z', then the reason for the difference in the yield might be attributed to the value 'Z'. Therefore, if the farmer has the knowledge of the inconsistency, necessary action can be taken.

3.1 *Features of Proposed System*

Suitability Prediction

By mapping data aggregated by various sensors to an available dataset you can find if a crop is suitable for a given agricultural land.

Monitoring the field via sensors reduces human effort, error and makes the whole process easier as well as more effective.

Crop Monitoring

Using various sensors in field, crop conditions can be monitored and it can be decided if human intervention is needed. Analysis of sensor data may provide useful insights related to yield

Disease Prediction

Based on real-time data from sensor one can predict if crops are in danger of contracting a disease. Different sensors collect real-time data of environmental parameters, utilizes this data to predict diseases using machine learning algorithms. Then it notifies farmers via text message or Web browser.

Data aggregation and storage

Monitored data is stored and analyzed. Analysis for an individual patch can be done to find any anomalies. Data is aggregated for future use in suitability prediction. Being deployed on cloud makes storage and processing simple.

3.2 Proposed System Architecture

The sensors employed in the field continuously sense and push the data in the Cloud for storage through Raspberry Pi. When an anomaly is detected in the data, the system sends a notification to the farmer for intervention. For example, if the water level for a particular crop is found higher or lower than required, the farmer is suggested to take necessary actions. Irrespective of anomaly detection, aggregation is performed on the data.

Machine learning algorithms are used for analysis on the data using which features like crop suitability and disease prediction are provided to the farmer.

The app is the primary source of interaction between the crop and the farmer. It is through this application that the farmer control and monitor the condition of the plant from anywhere in the world. It can be used to access the recommendation provided for suitability, checking the data and different services provided by the proposed system.

Sub-system 1: Uploading sensor data to cloud

Step 1. *Connecting the sensor and transmitting sensor data to Raspberry Pi* Based on the pin configuration of the sensor, it is connected to the Raspberry Pi. The input and output PINs on the Raspberry Pi should be noted and used in the program using which the data will be sent to the Raspberry Pi. In the program, the time interval is provided between which input data is sent to the Pi. On executing the program, the input will be received within the interval specified in the format specified in the program.

Step 2. Configuring Raspberry Pi to cloud

Registering your IoT device on the cloud service will present you with certain credentials that will be required to forward data from the Pi to the cloud.

Few modifications should be made in the program which includes adding the cloud service, your credentials related to the service and adding the table name provided at the cloud service in the table section which will allow you to output the sensor data to Raspberry Pi which will forward it to the cloud service.

Step 3. *At the cloud service*

A structure needs to be designed which consists of creating a primary key, adding the column names in which the incoming data will be stored.

Sub-system 2: Suitability prediction

The outcome of this sub-system is recommendation to the farmers for suitable crops.

The steps involved are:

Step 1. Aggregating sensor data

The sensor data is being collected continuously. It can provide accurate information about the farm conditions. This data is more accurate and can act as better input for the machine learning model providing precise predictions.

Step 2. Prediction

Machine learning models of various crops are run based on inputs from the above steps. Suitability of each crop with respect to a specific farm can be predicted. Hence the farmer has a good idea of variety of crops that can be grown on his field. Based on these suggestions and other factors like amount of care required, investment and return on investment, he can choose crops.

Sub-system 3: Crop Monitoring

Crop monitoring is an important aspect of farm management. If done incorrectly, it can have harmful effects on the yield and quality of the yield. In India, human intervention required in farming is quite high. During agricultural seasons, farmers stay at their respective farms to satisfy their farm requirements. Human involvement cannot be eliminated but it can be reduced to an extent where it is less hectic by automating the evaluation procedure.

In the proposed system, data stored in the cloud is checked for anomaly so that the farmer can be informed about it thus ensuring farmers to take necessary action.

Using the cloud services in which the farmer is added as the end device which will receive a notification, a SQL query is fired every time a data is pushed from the Raspberry Pi. If the output is between the set constraints, the farmer isn't notified else based on service approved by the farmer, the notification service proceeds.

Sub-system 4: Disease prediction

Step 1. Taking Inputs

- a. Meteorological data is taken as input
- b. Data from various sensors is taken as input
- a. Based on the parameters, probability of occurrence of disease is predicted
- b. The inputs are given to the ML model which produces the probability of disease

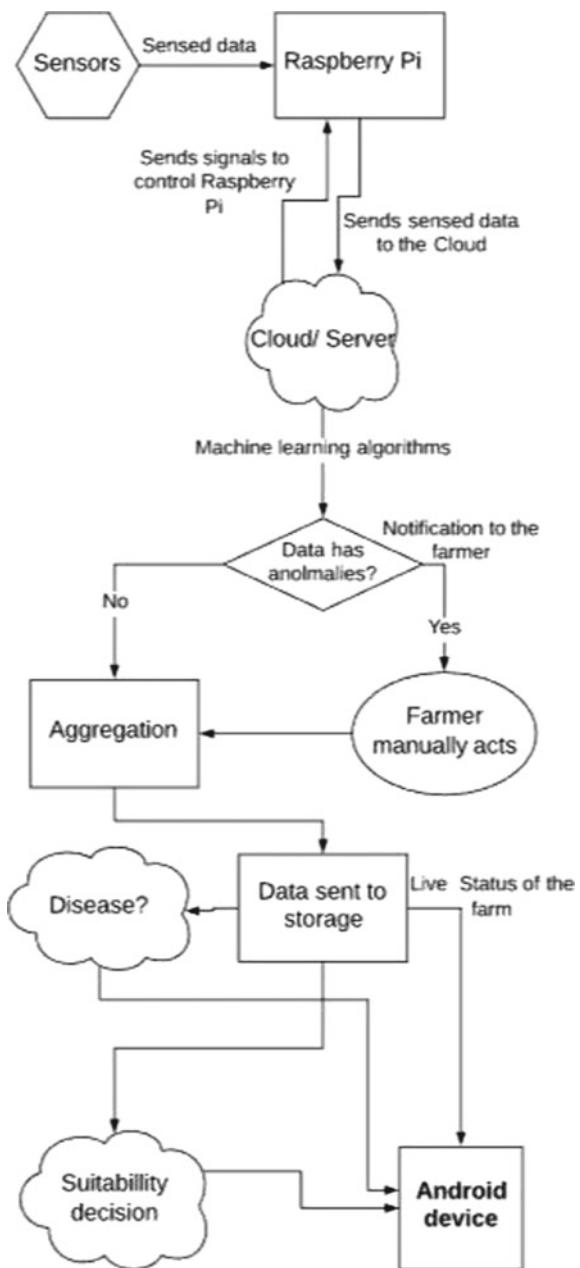
STEP 3. Alert

- a. The predicted value of occurrence of disease is sent to a cloud service.
 - b. If value is above threshold, it will send a message to the user
1. Camera and additional functionalities can be added to the system so that by using image processing techniques the amount of fertilizer required can be detected, disease prediction. Further, sending it to an expert can provide the farmer with a possible solution.
 2. A surfeit of data is collected using which in-field variability can be detected based on which suitable actions can be taken
 3. New parameters can be added to the system using which accuracy of algorithms can be increased thus providing better recommendations
 4. New sensors can be added to provide additional data about certain parameters, e.g., sunlight's intensity etc. (Fig. 1)

4 Conclusion

In this paper, we proposed an intelligent farming system which intends to solve the relevant problems of the farmers. Using data attributed to a single farm ensures highly accurate recommendation whose precision will increase as the amount of data collected increases. The damage caused to the crops due to varying diseases can be predicted and notified to the farmer which allows the farmer to take necessary action. This project aims to reduce human intervention by providing crop monitoring and taking action only if problems exist with crop which will be informed to the farmer. Storing the data over cloud provides control over data and secured storage. Thus, we have proposed a data centric solution for better crop yield.

Fig. 1 Proposed system architecture



References

1. Shirasath R, Khadke N, More D, Patil P, Patil H (2017) Agriculture decision support system using data mining. In: 2017 International conference on intelligent computing and control (I2C2), Coimbatore, pp 1–5
2. <https://www.bodhijournals.com>
3. Dholu M, Ghodinde KA (2018) Internet of things (IoT) for precision agriculture application. In: 2018 2nd International conference on trends in electronics and informatics (ICOEI)
4. Arshia Bhattacharjee A, Das P, Basu D, Roy D (2017) Smart farming using IoT. In: 2017 8th IEEE annual information technology, electronics and mobile communication conference (IEMCON)
5. Srinivasulu P, Sarath Babu M, Venkat R, Rajesh K (2017) Cloud service oriented architecture (CSoA) for agriculture through the internet of things (IoT) and big data. In: 2017 IEEE international conference on electrical, instrumentation and communication engineering (ICEICE)
6. <http://soilhealth.dac.gov.in/>
7. <http://flybirdinnovations.com/>
8. <https://www.statisticssolutions.com/multiple-linear-regression/>
9. <https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88>
10. <https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1>
11. Dagar R, Som S, Khatri SK (2018) Smart farming—IoT in agriculture. In: 2018 International conference on inventive research in computing applications (ICIRCA) Coimbatore, pp 1052–1056. <https://doi.org/10.1109/circa.2018.8597264>
12. Mekala MS, Viswanathan P (2017) A Survey: smart agriculture IoT with cloud computing. In: 2017 International conference on microelectronic devices, circuits and systems (ICMDCS)
13. Serikul P, Nakpong N, Nakjuatong N (2018) smart farm monitoring via the blynk IoT platform : case study: humidity monitoring and data recording. In: 2018 16th International conference on ICT and knowledge engineering (ICT&KE), Bangkok, 2018, pp 1–6
14. <https://www.link-labs.com/blog/iot-agriculture>
15. Shinde SS, Kulkarni M (2017) Review paper on prediction of crop disease using IoT and machine learning. In: 2017 International conference on transforming engineering education (ICTEE)

Solar-Powered Smart Agriculture and Irrigation Monitoring/Control System over Cloud—An Efficient and Eco-friendly Method for Effective Crop Production by Farmers in Rural India



Syed Musthak Ahmed, B. Kovela, and Vinit Kumar Gunjan

Abstract Agriculture is the basic profession of farmers in rural India, and cultivation is very important for their day-to-day living and survival. Crops like paddy, wheat, and vegetables require watering and regular monitoring to make effective production. Watering and monitoring are crucial and time consuming. Hence, the farmer has to spend most of his time to look over the crops leaving all his works behind. Hence, to make farmers time effective, an automated plant watering system through cloud is proposed. Here, the farmer need not stay at the fields all the time to fetch water to the crops or stay back at the field due to power cut or any other reason what so ever. The farmer can monitor the field from outside besides attending to his regular activities. Solar power supports agriculture farming during unseasonal and during power failure conditions. In the proposed work, monitoring of the field is done using ThingSpeak, an IoT platform that provides a user-friendly graphical representation of environmental parameters, and thereby, the field can be monitored from anywhere using IoT over the Internet. The project consists of an on-field LCD display for keeping track of water level in tank and pump status. This project has two power supply modes, one using AC mains supply and other using solar panels in case of power cut, so that field monitoring will never be interrupted.

Keywords Arduino UNO · NodeMCU · DHT11 · L298N · Solar panels · Lead-acid battery · Soil moisture sensor · Light intensity sensor · Raindrop sensor · LDR · LCD

S. M. Ahmed (✉) · B. Kovela
S R Engineering College, Warangal, Telangana State, India
e-mail: syedmusthak_gce@rediffmail.com

V. K. Gunjan
CMR Institute of Technology, Hyderabad, India

1 Introduction

Farmers play an important role in the existence of human lives, because they are the ones who provide us food to eat and clothes to wear. They utilize natural resources like water and manure to accomplish the task of growing good crop. A farmer's main goal is to produce enough food in order to feed the fast-growing population. Farmers always work outdoor in extreme weather conditions. They work so hard to supply animal products and crops to the market. Without farmers, the world would slowly die. Farmers have good knowledge on planting dates, breading cycles, and harvesting time to crop. Despite of all this, they should also have a good knowledge on mechanics and latest technology [1] which may help them in further good production of crop [2–7] without costing their lives.

The advent of IoT has brought tremendous changes in various fields, and one such is in agriculture [8, 9]. Here, we came up with a IoT-based smart agriculture irrigation monitoring system which can withstand any weather conditions and can update the farmers about weather and crop conditions and water requirement around the field, on the mobile phone/computers. Most of the time, villagers suffer from long power cuts. But to keep irrigation process uninterrupted due to power cut or power failure, solar powering facility is provided. Thus, unaltered monitoring and operation of field can be carried out despite power failures. This implementation contains bunch of environmental sensors like temperature, humidity, rainfall, soil moisture, light intensity for monitoring all environmental parameters around the field. The system also has a water level indicator for monitoring water level in the soil available for the crop. Whenever the crop soil gets dry, the water pump automatically gets ON. The system also has an on-field LCD display for displaying water level and pump status. Thus, this IoT-based system finds solution to much of the human-dependent farmers' problems in the coming days [10–12].

2 System Design

The proposed system consists of Arduino UNO microcontroller and ESP8266 Wi-Fi board for IoT operation and various environment monitoring sensors. The system block of implementation is shown in Fig. 1.

The DHT11 temperature sensor, rain sensor, light sensor, soil moisture sensor, and water level detector are connected to Arduino UNO. The water pump is controlled through a L298N motor driver. All sensors data are collected by Arduino and is sent to ESP8266 through four-wire communication protocol using 5 to 3.3 V logic level converter. Also, a 16X2 LCD display is connected to Arduino for displaying water level and water pump status. ESP8266 uploads received data to ThingSpeak IoT platform which gives a graphical representation of data in the form of graphs.

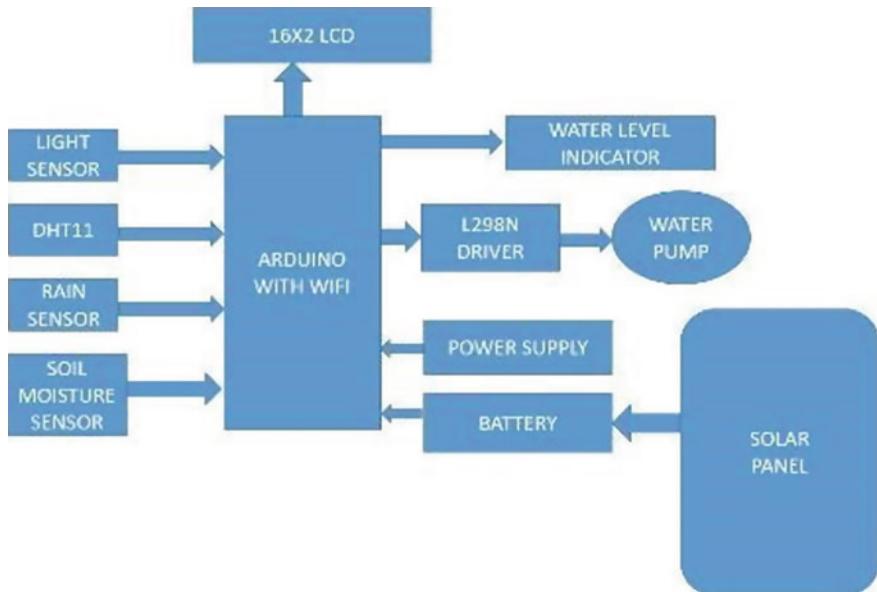


Fig. 1 Block diagram representation of system design

3 Hardware Requirements in Implementing the System

The following hardware's are made use of in the project implementation.

- NodeMCU:** It is an open-source platform having firmware [12] and hardware (based on the ESP-12 module). The “NodeMCU” has several features such as programmable Wi-Fi module, Arduino-like (software defined) hardware IO, PCB antenna, Wi-Fi networking, event-driven API. The NodeMCU board is shown in Fig. 2.
- Arduino UNO:** It is a microcontroller, based on ATmega328, an open-source platform [12], having 14 pins digital I/O, 16 MHz crystal oscillator, a USB connection, a power jacket, a ICSP header, and a reset button. This is shown in Fig. 3.
- Soil Moisture Sensor:** The moisture content in soil can be detected by this sensor. Farmers can manage their crops more effectively and efficiently by measuring the soil moisture content. Hence, farmers monitor the soil moisture [12] content by incorporating this sensor. Such a sensor is shown in Fig. 4.
- DHT11 Sensor:** It is an ultra-low-cost digital temperature and humidity sensor. It operates between 3 and 5 V power supply and I/O with humidity readings accuracy of 5%, temperature with ± 2 °C accuracy. It measures the surrounding air humidity and temperature and gives out a digital signal. The structure of DTH11 sensor is shown in Fig. 5.



Fig. 2 NodeMCU module

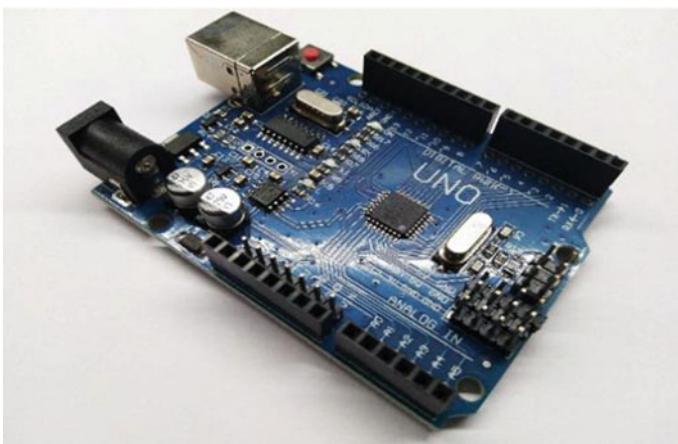


Fig. 3 Arduino UNO module

- E. **Water Pump:** Water pump is used for controlling the supply of water to the field. The operation of the pump is controlled by a motor driver to water the crop. Whenever the field gets dried up, i.e., moisture content gets reduced, the system makes the driver to operate which intern switches ON the motor till the required soil conditions is reached. Once the desired level for the selective crop is reached, the motor switched OFF via the motor driver. The water pump and relay module for controlling water power supply are shown in Figs. 6 and 7, respectively.

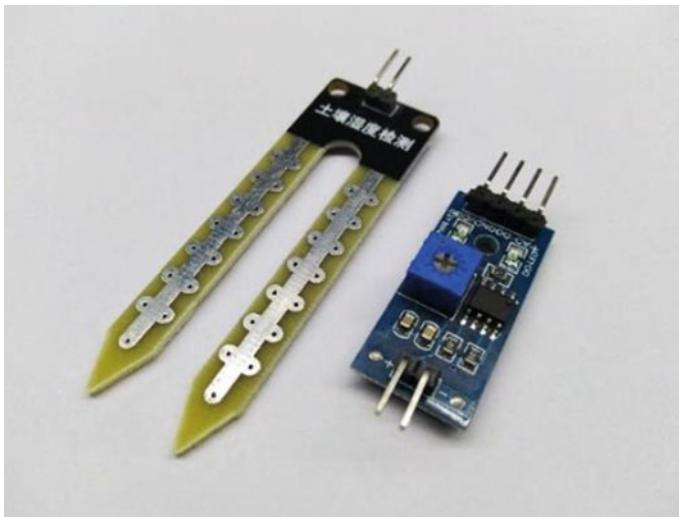


Fig. 4 Soil moisture sensor

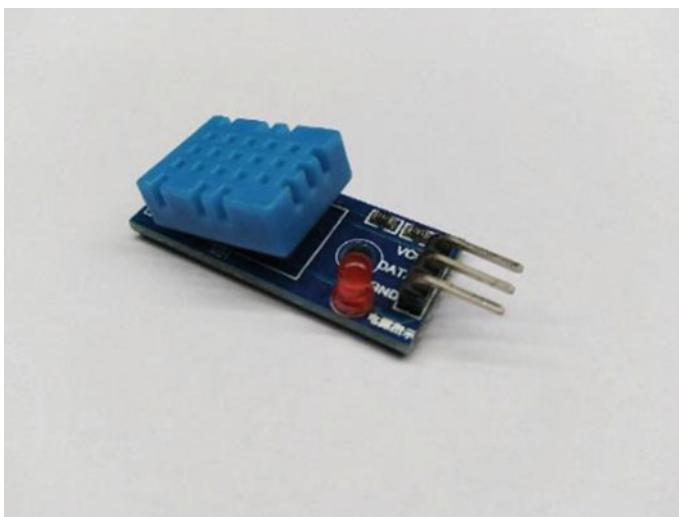


Fig. 5 DHT11 sensor

F. **Logic Level Converter:** The logic level convertor is used to step up or step down the voltage levels as required for the circuit. Here SparkFun bi-directional logic level converter is used in our application. This device simultaneously performs step-up and step-down operations from 5 to 3.3 V and vice versa. This is shown in Fig. 8.



Fig. 6 Water pump system

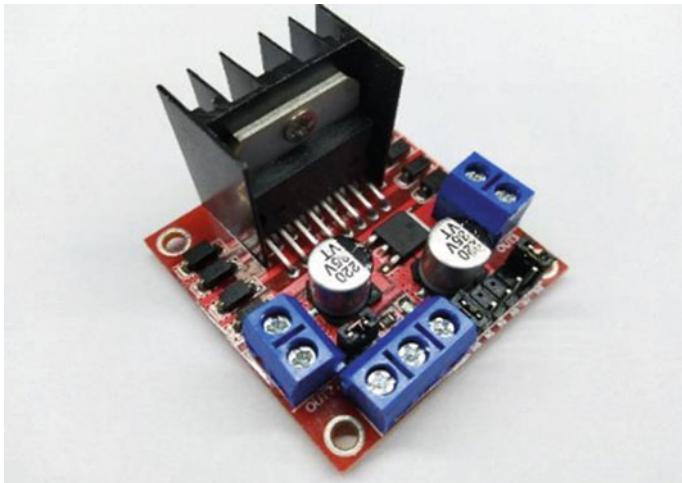


Fig. 7 Relay module for switching motor

- G. **LDR Module:** The light-dependent resistor (LDR) is a photoresistor whose resistance varies depending upon the light intensity falling on it [12]. It exhibits the property of photoconductivity and finds large applications in diverse fields of engineering. This is shown in Fig. 9.
- H. **Raindrop Module:** The rain sensor module is shown in Fig. 10. The function of this sensor is to shut down the system in the event of rainfall. It consists of a water conservation device to control the water supply to the fields during rainfall. It causes the irrigation system to automatically shut down in the event of rain, thereby saving power in system and time of farmers. This is shown in Fig. 10.

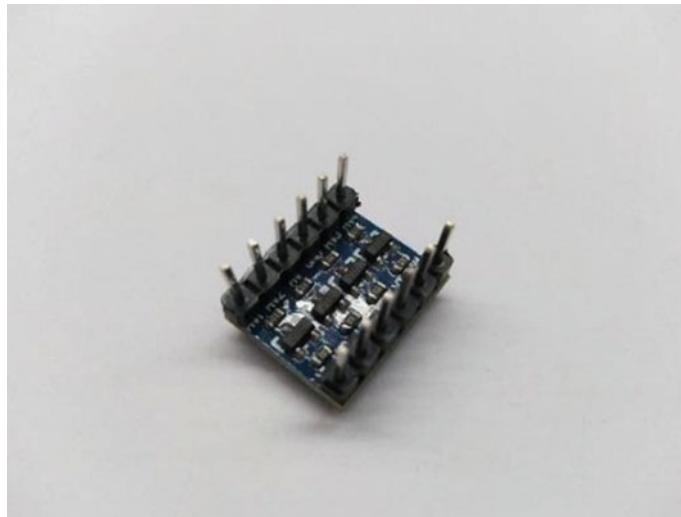


Fig. 8 Logic level converter

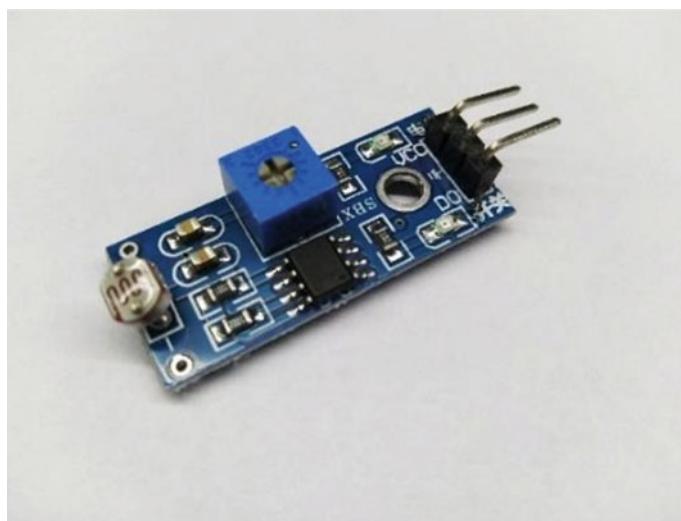


Fig. 9 LDR module to detect light sensitivity

- I. **Solar Panel:** Solar panels are the collection of photodiodes which convert light energy into electricity. For this project, we are using 12 V, 5 W solar panel which is used to charge a 12 V, 1.3AH lead-acid battery for seamless operation during power supply failure condition. Figures 11 and 12 show the solar panel and battery used in our implementation.

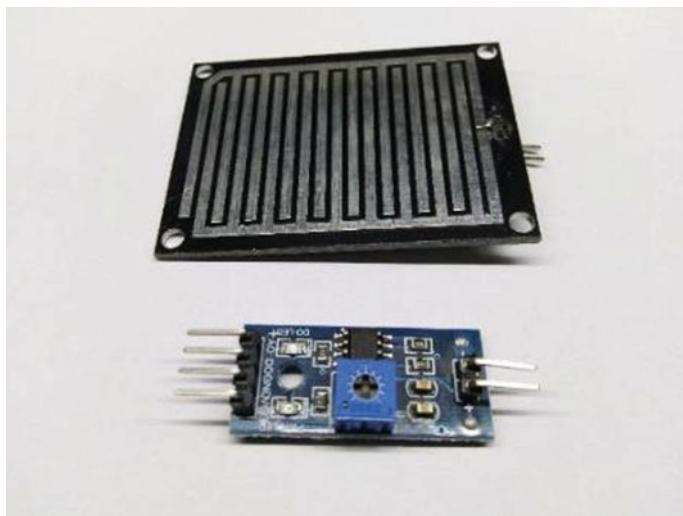


Fig. 10 Raindrop sensor circuit



Fig. 11 12 V, 5 W solar Panel

- J. **16X2 LCD Module:** To know the water level in tank and water pump status, an on-field LCD display is incorporated to display the status of water pump and water level content in tank. This is shown in Fig. 13.



Fig. 12 12 V, 1.3 AH battery

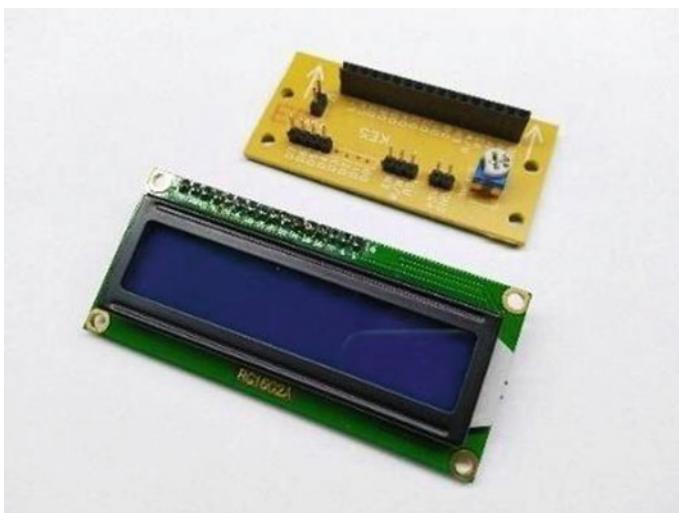


Fig. 13 16X2 LCD module

4 Software Requirement in Implementing the System

ThingSpeak is an open-data platform and API for IoT which helps to collect data, store data, analyze data, visualize data from various sensors and act accordingly. The data can be send from the device to ThingSpeak, create instant visualization of the

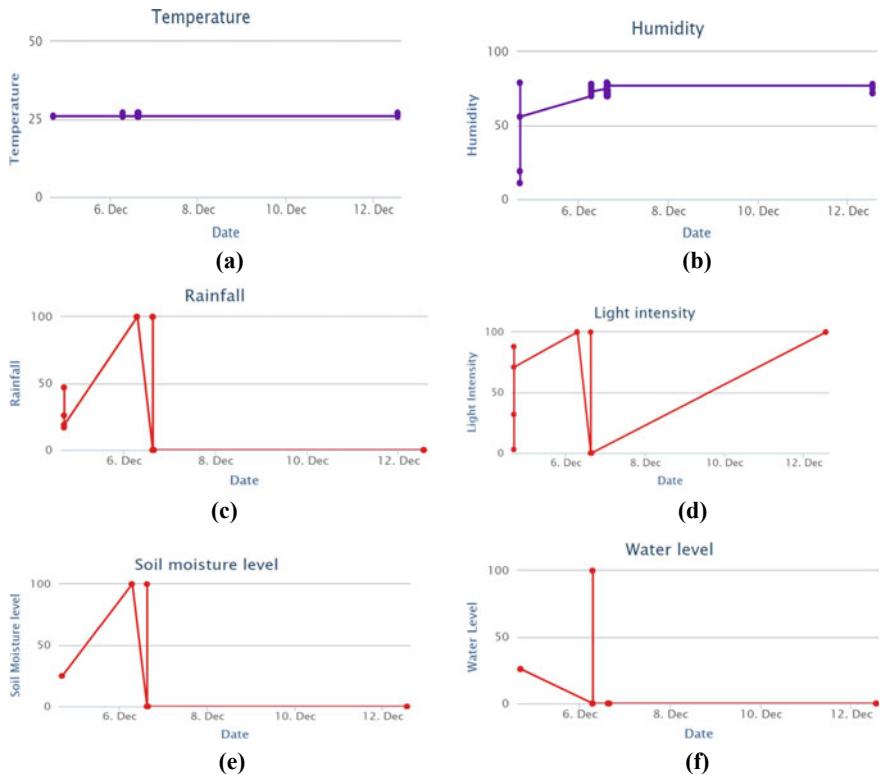


Fig. 14 ThingSpeak IoT platform showing graphical representation of **a** temperature, **b** humidity, **c** rainfall, **d** light intensity, **e** soil moisture, **f** water level

live data, and send alerts. ThingSpeak website displaying graphical representation of soil moisture, humidity, temperature, rainfall and light intensity, water level, and pump status is shown in Fig. 14.

5 Results

The implementation of complete agriculture monitoring system is shown in Fig. 15.

The project uploads the data of all sensors to ThingSpeak IoT website. Beforehand, we need to create an account in ThingSpeak using Gmail. Later, a channel ID and an API key are provided for secured operation which will be used in programming. After successful account creation, we will be promoted to field creation for graphical representation. Here we have created six graphical fields for temperature, humidity, rain, light, soil moisture, and water level. System always checks for sensor data for every 5 s, and whenever soil gets dry, the system checks for water content in tank.

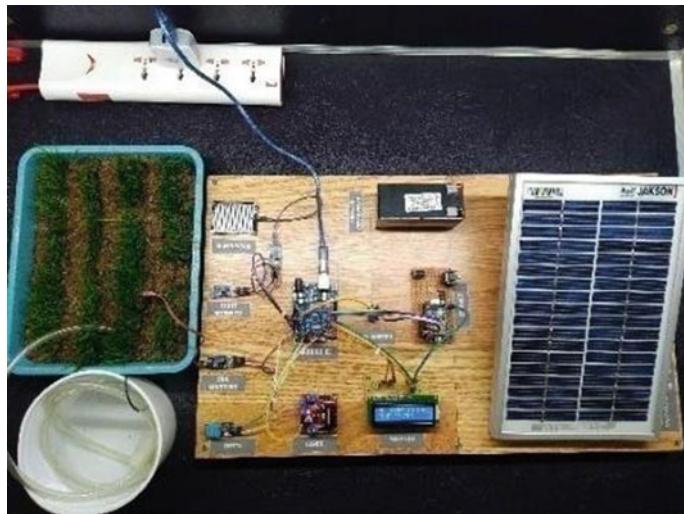


Fig. 15 Setup for testing

If water is present in the tank, then pump gets ON automatically. Whole system has dual power support with AC supply and with solar panels. On-field LCD is also provided for water level monitoring and pump status.

6 Conclusions

Thus, the solar-powered agriculture monitoring/control system over cloud has been implemented. ThingSpeak collects data from all sensors and uploads on to the ThingSpeak IoT server. The action will be performed depending on the moisture content present in the soil. If the soil is dry, it sends a high signal to Arduino which intern switches ON the water pump via the motor driver. If the soil gets enough wet, then the sensor sends a low signal to Arduino; hence, the water pump gets turned OFF automatically. The whole system is solar powered. The ThingSpeak website will get updated for every 5 s. Thus, the complete system is tested. The system developed is more efficient with affordable cost for implementation in agricultural fields and can be extended for large-scale cultivation of farmers.

References

1. Sandeep CH, Naresh KS, Pramod KP (2018) Security challenges and issues of the IoT system. Indian J Public Health Res Dev (IJPHRD) 9(11):748–753
2. Tani FH, Barrington S (2005) Zinc and copper uptake by plants under two transpiration rates. Part I. Wheat (*Triticum aestivum* L.). Environ Pollut 138:538–547
3. Gutierrez J, Villa-Medina JF, Nieto-Garibay A, Porta-Gándara MA (2013) Automated irrigation system using a wireless sensor network and GPRS module. IEEE
4. Patil SS, Malvijay AV (2014) Review for ARM based agriculture field monitoring system. Int J Sci Res Publ 4(2). Nagothu SK (2016) Weather based smart watering system using soil sensor and GSM. In: 2016 World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave). IEEE Xplore
5. Jury WA, Vaux HJ (2007) The emerging global water crisis: managing scarcity and conflict between water users. Adv Agron 95:1–76
6. Arun C, Lakshmi Sudha K (2012) Agricultural management using wireless sensor networks—a survey. In: 2nd international conference on environment science and biotechnology (IPCBEE), vol 48. IACSIT Press, Singapore
7. Nagothu SK, Anitha G, Annapantula S (2014) Navigation aid for people (joggers and runners) in unfamiliar urban environment using inertial navigation. In: 2014 sixth international conference on advanced computing (ICoAC), pp 216–219
8. Ahmed SM, Chandu TS, Rohit U, Naveen G, Naveen S (2019) IoT based garbage disposer for educating rural India. In: International conference on data sciences, machine learning and applications (DSMLA 2019), 29–30 Mar 2019. Springer
9. Patil SS, Malvijay AV (2014) Review for arm based agriculture field monitoring system. Int J Sci Res Publ 4(2)
10. Yuan G, Luo Y, Sun X, Tang D (2004) Evaluation of a crop water stress index for detecting water stress in winter wheat in the North China Plain. Agric Water Manag 64(1):29–40
11. Nagothu SK (2016) Weather based smart watering system using soil sensor and GSM. In: 2016 world conference on futuristic trends in research and innovation for social welfare (startup conclave). IEEE Xplore
12. Ahmed SM, Kovela B, Gunjan VK (2019) IoT based automatic watering system through soil moisture sensing—a technique to support farmers' cultivation in rural India. In: International conference on cybernetics, cognition and machine learning applications (ICCCMLA 2019), 16–17 Mar 2019 Springer

Exploration of Classification Algorithms for Divorce Prediction



Danussvar Jayanthi Narendran, R. Abilash, and B. S. Charulatha

Abstract Marital life is very important for any individual. Some marriages are successful, but nowadays many are unsuccessful. Divorce petition is filed in the family court due to various reasons. The couple faces lots of emotional and mental stress during the process. In addition, divorce gives wrong impact on the couple. The married couple ends up in divorce. Prediction is better than reaction. To avoid this situation, the comfort that will prevail among the bride and the groom is analyzed, and the success is predicted before the actual marriage. Using this prediction, the unnecessary formalities, expenses, stress can be devoid of. To analyze, the dataset is collected about the pre-marital status of the couple, which enables to predict if a marriage would be successful or otherwise before getting married. In this paper, the conclusion is drawn based on the performance of multiple classification algorithms for divorce prediction dataset. Evaluation was based on several criteria like k-fold cross-validation, mapping accuracy, sensitivity to dataset size and noise. The classification algorithms considered for the study are random forest, decision tree, XGBoost, bagging, and voting classifier.

Keywords Classification algorithms · Divorce · Ensemble learning · Machine learning · Prediction

D. J. Narendran
Intern, ASN Developers, Chennai, India
e-mail: danubmdjs@gmail.com

R. Abilash (✉)
Department of IT, Jawahar Engineering College, Chennai, India
e-mail: abilashr.86@gmail.com

B. S. Charulatha (✉)
Department of CSE, Rajalakshmi Engineering College, Chennai, India
e-mail: charu2303@yahoo.co.in

1 Introduction

Divorce, also known as the separation of marriage, is a method of breakup of marriage or civil partnership. Divorce usually involves the cancelation or reorganization of the legal obligations and obligations of marriage, thereby dissolving the matrimonial bonds between a married couple under the rule of law of a particular country or state. Divorce laws vary widely around the world, but in most countries, divorce requires the sanction of a court or other authority in a legal process that may involve issues, such as property distribution, child custody, child support, child visitation/access, parenting time, child support, and debt-sharing. In most countries, monogamy is required by law, and divorce requires each former partner to marry another. For those who are in a common-law partnership breakup, it is considered a breakup, not a divorce.

John and Julie Gottman presented us with four key predictors of divorce. The four primary predictors, the “Four Horsemen of the Apocalypse,” have been referred to as criticism contempt defensiveness and stonewalling.

For both partners, there are many social, physical, and emotional effects of divorce. Not only that, there are also many consequences of divorce on the children. There are many common after-effects of divorce on men, women, and children. The whole process of choosing to get a divorce and finally getting one can be very socially, physically, and financially exhausting for all the parties involved. The fear of these long-drawn-out and difficult times is what stops other people from going through the process. Nevertheless, for those who are working through their choices, there are certain after-effects that they have experienced. Not only do women feel the after-effects of divorce. It is just as difficult for men, not to mention children, even those who are older. The consequences of divorce include depression, anger, anxiety, social isolation, cynicism. Therefore, in order to avoid these challenges, technology takes the lead in determining the essence of the married life of two people before they get married. The purpose of this paper is to provide a way to prevent these circumstances. There are rumors that family courts and lawyers are mostly in Chennai [1].

2 Literature Survey

Many researchers have worked and are working on this dataset and the different classification algorithms for the prediction. This study aims at listing few works that have been done.

Boosting an ensemble method is one of the most successful classification techniques, but boosting algorithms are sensitive to noise. Row subsampling, Shrinkage parameter, Column subsampling, Regularization term in the objective function is used to avoid the overfitting [2].

Another ensemble method random forest is derived from decision tree. In random forest, the best attribute at each node in a decision tree is decided from a randomly

selected number of features. This random selection of features helps random forest to scale well, in reducing the interdependence between the feature's attributes [3] and is thus less vulnerable to inherent noise in the data. As mentioned by the author [4], the number of random features m selected per decision node in a tree decides the error rate of the forest classification.

In [5], the authors used naive Bayes and random forest for the prediction of occurrence of diabetes using the daily routine of the individuals.

Bagging, proposed in [6], improves the classification accuracy over an individual classifier or the approximation error in regression problems. The underlying idea is to perturb the training data by creating a number of bootstrap replicates of the training set, train a classifier on each bootstrap replicate, and aggregate their predictions. This allows reducing the variance component of the classification or estimation error [7]. Indeed, bagging has shown to be particularly successful when applied to “unstable” classifiers like decision trees and neural networks. In [8], it was argued that bagging equalizes the influence of training samples; namely, it reduces the influence of outlier samples in training data.

CART can be used for both classification and regression, depending on the available information of the dataset. Classification trees are used when we know the class of each instance. In CART, Gini impurity is used as the information measure. The selection of the feature is based on the sum of squared errors, and the best feature is the one with minimal sum of square error [9].

Another ensemble solution is by the voting classifier which is a wrapper for a set of different ones that are trained and evaluated in parallel in order to exploit the different peculiarities of each algorithm.

The final decision on a prediction is taken by majority vote according to two methods, namely

Hard voting is where a model is selected from an ensemble to make the final prediction by a simple majority vote for accuracy.

Soft voting can only be done when all your classifiers can calculate probabilities for the outcomes [10].

In [11], the authors discussed random forests and implemented for analyzing the determinants of divorce with SOEP data for German women 1984–2015. The algorithm is able to classify divorce determinants according to their importance, highlighting the most powerful attribute.

The application consists in achieving the prediction of the divorce general rate level of married population for 2005, for any district of Romania, using the ID3. The attributes used for the current application are the following: the number of marriages, the net medium nominal monthly income, the unemployment rate, the medium age at marriage, the education level index, and, the target attribute, the married population divorce general rate [12].

The dataset is taken from How Couples Meet and Stay Together; methods applied are Gaussian naive Bayes, support vector machine with linear kernel, k-nearest neighbors, decision tree, ridge regression, L1 regularized logistic regression, and AdaBoost using decision tree. Each classifier model was trained using fivefold cross-validation and is evaluated using a 70–30 training–testing split to avoid overfitting issues [13].

3 Experimental Setup

Divorce prediction models created and seen across the Internet are predicted with the help of the data collected about the post-marital status, and it cannot be used to predict if a couple would have a successful marriage or not, before getting married. In order to achieve this, a survey is conducted on the pre-marital status for couples who are currently married, still in marriage life, or broken marriage. The marriage life considered for the study is between 0 and 10 years. Using this data, research is conducted to predict the compatibility of a couple, which can be used to decide whether they will have a successful marriage or not. Four machine learning algorithms are implemented, so that it would enable us to select the model which gives the best results for this data. So, the research is proof to real world that the prediction of a successful marriage is possible before marriage itself.

4 Dataset Description

For this experiment, the authors collected dataset by creating a survey using a Google form, which was shared among their friends, family, and professional circle. This form was further shared by others to people they knew. The 226 samples were collected from a diverse range of people from different cultures and traditions. This survey consists of sixty questions which are grouped under the following categories: appearance, financial status, behavior, habits, culture, lifestyle, and relationships. And these questions are verified by a psychologist, whether they are valid or not. Then, these questions are used as fifty-nine input attributes and one output attribute.

4.1 Preprocessing

The multivariate data samples collected have missing values. The missing values are filled using the statistical imputation method. Mean replacement is pursued to fill the missing values. The missing values are replaced by the mean value of the corresponding variable in this process.

4.2 Feature Selection

The selection of the features for training the model is based on the univariate selection method called as chi-squared. The aim is to remove the features if it provides little or no additional information than the other features. Since the empirical score for all

Table 1 Feature selection scores of the dataset

Specs	Score	Specs	Score
Number of friends bride has	62.45	Number of credit cards owned by bride	33.9
Number of past relationships for bride	53.54	Dress sense	33.66
Having a stable job or not	52.02	Wearing accessories while going out	33.66
Chain smoker or not	48.65	Number of credit cards owned by groom	32.93
Alcoholic or not	47.92	Dress sense while going out	30.42

the 59 attributes are relatively good, none of the features are discarded. The top ten features of the dataset are represented in (Table 1).

4.3 Construction of Training and Testing Datasets

For the classification to develop the model, 70% of the data collected were used for training and the rest for testing the model.

The data consist of 60 attributes and 226 instances. The data are split into training data of 158 and testing data of 68 instances. Among these 60 attributes, 60th attribute is the class label. This is a two-class problem being in married life or divorced. The training and testing data are selected randomly from the dataset using the k-fold cross-validation, where $k = 10$.

4.4 Random Forest Classifier

The random forest classifier is an ensemble of the decision trees principle. Due to their flexibility between various types of data and their high precision of confirmation, they are very useful when processing complex data while avoiding overfitting to the noise in the data due to their existence of bootstrapping. When training a random forest model, the key parameters are the number of independent trees to evolve and the number of randomly sampled characteristics used in each decision node.

4.5 XGBoosting

XGBoost is a decision tree-based ensemble machine learning algorithm that applies the principle of boosting weak learners using the gradient descent architecture. XGBoost uses more regularized model formalization to manage overfitting, which gives it better performance. In fact, it has the capability to perform parallel computation on a single machine.

4.6 Decision Tree

The decision tree is a decision support method that uses a tree-like decision pattern. It is a flowchart like a system in which each internal node represents a test on the variable, each branch represents the test result, and each leaf node represents a class name. In addition, it can perform a multi-class dataset classification. The decision tree algorithm performs well even if its premises are somewhat broken by the true model from which the data were generated. In this study, a version of the decision tree classifier is called classification and regression trees (CART).

4.7 Bagging Classifier

A bagging classifier is an ensemble process that fits the base classifiers into the random subsets of the initial dataset and then aggregates their individual predictions to form the final prediction. This is a general approach used to reduce the variance of the foundation classifier. The basic parameters for the bagging classifier algorithm when training are the base classifier, the number of base classifiers, and the random state.

4.8 Voting Classifier

Voting classifier is not an actual classifier, but a wrapper for a set of different classification algorithms that are trained and evaluated in parallel. It has two forms of hard and soft voting. Hard voting is also referred to as majority voting. In this case, the forecast class with the highest number of votes will be chosen as the final predicted class. In the case of soft voting, the probability vector for each predicted class is summed up and averaged. The winning class is the one that corresponds to the highest value. While this algorithm is being trained, the key parameters are the list containing the artifacts of the classification algorithms and the method of voting. In this analysis, XGBoosting, decision tree, bagging classifier are the classification algorithms used for wrapping and hard as the form of voting type.

Table 2 Maximum depth versus accuracy for decision tree

Maximum depth	Accuracy score	Maximum depth	Accuracy score
1	0.81	15	0.8737
3	0.86	20	0.8737
5	0.869	25	0.8737
7	0.8604	30	0.8737
10	0.8737		

5 Experimental Results

5.1 Model Implementation

The results of the four models are summarized and compared in this section. It was observed that, the random forest classifier has more validation accuracy among the four machine learning algorithms for this data.

For all the processes conducted in this study, that is, training and testing the models, the dataset has a total of 60 features. The total number of 158 and 68 instances is used for training and testing the models, respectively, and the same training and testing data are used for all the models to compare the results more accurately.

5.2 Evaluation

The aim of this study is to create a best model for the divorce prediction dataset and to find which model would have the highest performance for the unseen data, so that the model would be reliable. This is achieved by comparing the accuracy score of different algorithms.

5.2.1 Decision Tree

Table 2 shows the effect of increasing the maximum depth to that of accuracy for decision tree. The same is shown as a graphical representation in Fig. 1.

5.2.2 Bagging Classifier

In order to evaluate the performance of the bagging classifier, the number of trees was started with five with accuracy of 90%. The accuracy of 93% was achieved when the number trees was 1000 (Fig. 2).

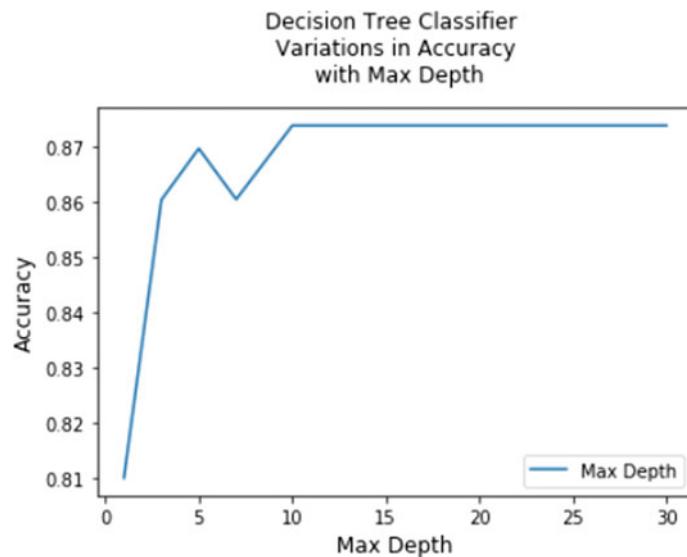


Fig. 1 Graphical representation of Table 2 for decision tree

Table 3 Number of trees versus accuracy score for bagging

Number of trees	Accuracy score	Number of trees	Accuracy score
5	0.9	400	0.9234
10	0.9	500	0.9368
20	0.91	700	0.9368
100	0.9234	900	0.9323
200	0.9279	1000	0.9323

5.2.3 XGBoost

Two experiments are performed in XGBoost to find the best model.

With the number of trees as three, the acquired accuracy is 83.84%, and accuracy of 93.68% was achieved when the number of trees is equal to 100. But the accuracy dropped when the number of trees was increased from 100. This is represented in Table 4.

In the rerun, the maximum depth was varied. The maximum depth ranged from one with accuracy of 92.3%. When the depth was increased to seven, the maximum accuracy of 93.69% was achieved, and it stayed constant beyond seven (Figs. 3 and 4).

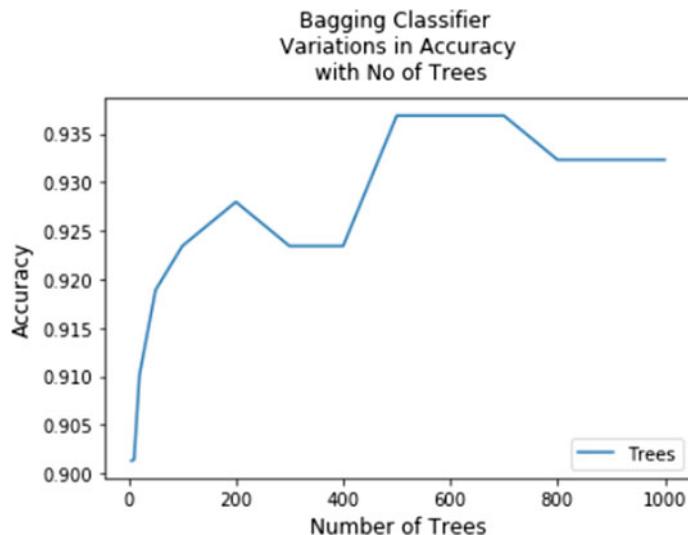


Fig. 2 Graphical representation of Table 3 for bagging classifier

Table 4 Number of trees versus accuracy score of XGBoost algorithm

Number of trees	Accuracy score	Number of trees	Accuracy score
3	0.8384	25	0.9
5	0.8294	50	0.928
7	0.8694	100	0.9368
10	0.8784	200	0.9277
20	0.8874	500	0.9277

Table 5 For maximum depth versus accuracy of XGBoost algorithm

Maximum depth	Accuracy score	Maximum depth	Accuracy score
1	0.923	15	0.9369
3	0.9368	20	0.9369
5	0.9324	25	0.9369
7	0.9369	30	0.9369
10	0.9369		

5.2.4 Voting Classifier

In voting classifier, the ensemble of decision tree, bagging classifier, and XGBoost is performed and acquired an overall accuracy of 94.14%.

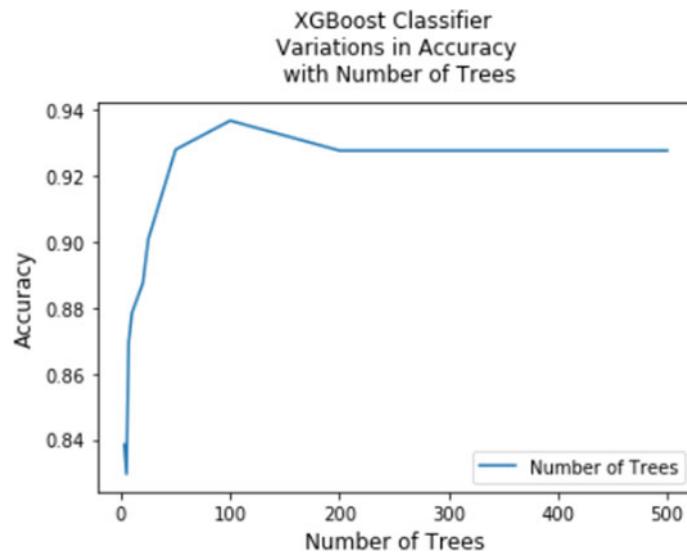


Fig. 3 Graphical representation of Table 4

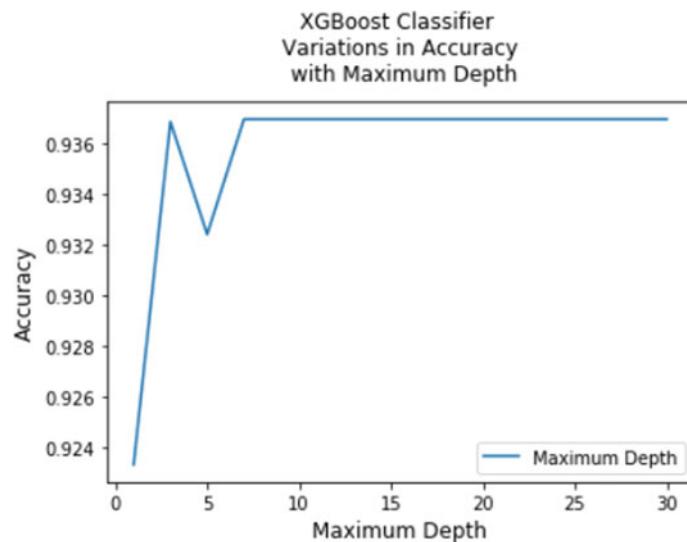


Fig. 4 Graphical representation of Table 5

Table 6 Random forest performance for variation in tree with accuracy

Number of trees	Accuracy score	Number of trees	Accuracy score
3	0.8827	20	0.937
5	0.8828	25	0.94
7	0.9367	50	0.9368
10	0.9323	100	0.9414

Table 7 Random forest performance for variation in maximum depth with accuracy

Maximum depth	Accuracy score	Maximum depth	Accuracy score
1	0.8	15	0.923
3	0.9	20	0.9145
5	0.92	25	0.9414
7	0.9368	30	0.9414
10	0.9367		

5.2.5 Random Forest

Again, two experiments are performed to find the best random forest model, one for varying the depth and the other with the variation in the trees. Analysis is presented in Tables 6 and 7.

First, we started with number of trees as three and acquired an accuracy of 88.27%, and as we increased the number of trees, we reached a maximum accuracy of 94% when the number of trees is equal to 100. Maximum depth in this experiment is none. If none, then the nodes expand until all leaves contain less than the minimum number of samples required to split an internal node. The minimum number of samples required to split an internal node in this experiment is two.

In the second experiment, the maximum depth was varied from one and acquired an accuracy of 80%. With the depth of 25, a maximum accuracy of 94.14% was achieved, and it stayed constant when increased beyond 25. The constant number of trees in this experiment is ten trees (Figs. 5 and 6).

6 Conclusion

Divorce is a major problem in our society. Many lives have been affected due to divorce. The one who has been affected the most due to divorce is the children, and the struggle they go through mentally is beyond words. Also, the couples lose a lot. This study would help a lot to prevent these things. This prediction model would help people to decide people to marry or not by giving their chances at their compatibility to have a successful marriage. Both random forest and the ensemble

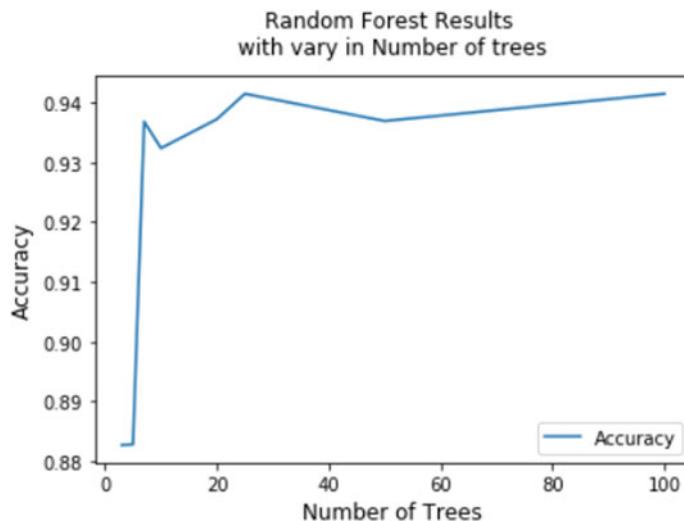


Fig. 5 Graphical representation of Table 6 for random forest

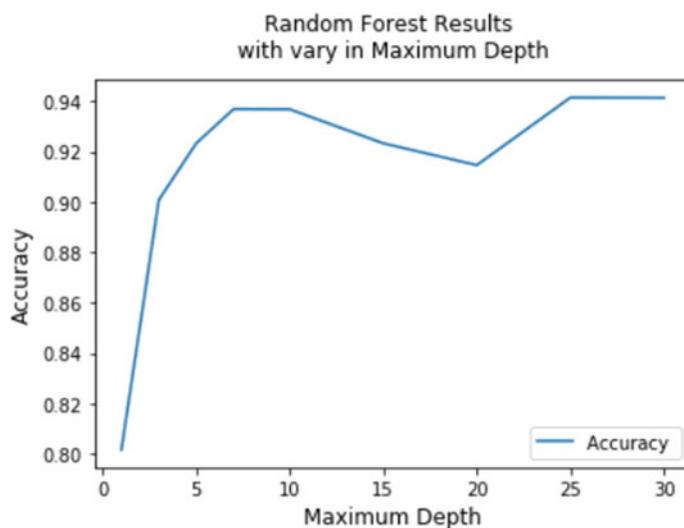


Fig. 6 Graphical representation of Table 7 for random Forest

model of decision tree, bagging classifier, and XGBoost have equal accuracy of 94.14% for this dataset which has outperformed decision tree (maximum accuracy = 87.37%), bagging classifier (maximum accuracy = 93.69%), XGBoost (maximum accuracy = 93.69%).

7 Future Enhancement

In this paper, the performances of four different machine learning classification algorithms are used to predict and study the collected dataset for divorce prediction. Reasons for divorce do vary with different regions and the cultural impact of the region the couples live in. Even though this data provided good insights, a dataset region specific and may increase the number of features according to a specific region and also use a different machine learning implementation to create a more accurate divorce prediction model.

References

1. <https://www.bonobology.com/what-are-the-after-effects-of-a-divorce-in-india/>
2. Gómez-Ríos A, Luengo J, Herrera F (2017) A study on the noise label influence in boosting algorithms: AdaBoost, GBM and XGBoost. Springer, Cham
3. Criminisi A, Shotton J, Konukoglu E (2012) Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Found Trends Comput Graph Vis* 7(2–3):81–227
4. Breiman L (2001) Random forests. *Mach Learn* 45(1):532. <https://doi.org/10.1023/A:1010933404324>
5. Abilash R, Charulatha BS (2020) Early detection of diabetes from daily routine activities: predictive modeling based on machine learning techniques. In: Intelligence in big data technologies—beyond the hype. Proceedings of 3rd ICBDCC 19
6. Breiman L (1996) Bagging predictors. *Mach Learn* 24(2):123–140
7. Biggio B, Corona I, Fumera G, Giacinto G, Roli F (2011) Bagging classifiers for fighting poisoning attacks in adversarial classification tasks. LNCS, vol 6713. Springer, Berlin, Heidelberg, pp 350–359
8. Grandvalet Y (2004) Bagging equalizes influence. *Mach Learn* 55:251–270
9. Che D, Liu Q, Rasheed K, Tao X (2011) Decision tree and ensemble learning algorithms with their applications in bioinformatics software tools and algorithms for biological systems. Advances in experimental medicine and biology, vol 696. Springer, New York. https://doi.org/10.1007/978-1-4419-7046-6_19
10. Bonacorso G (2018) Machine learning algorithms, 2nd edn. Packt Publishing, Birmingham
11. Arpino B, Le Moglie M, Mencarini L (2018) Machine-learning techniques for family demography: an application of random forests to the analysis of divorce determinants in Germany, a Creative Commons Attribution 4.0 International License
12. Cărbureanu M (2007) The divorce rate prediction using data mining techniques. *Seria Matematică-Informatică-Fizică*, vol LIX No. 2/2007, pp 37–42
13. Liu Y, Labiak C, Kliemann M, Srivastava K, Xiao Y (2014) Identification of promising couples using machine learning

Pronunciation Similarity Matching Using Deep Learning



**Ashish Upadhyay, Bhupendra Kumar Sonwani, Vimal Anand Baghel,
Yash Kirti Sinha, Ashish Singh Patel, and Muneendra Ojha**

Abstract The correct pronunciation of English language words poses a significant challenge for children in non-English speaking countries, which leads to incorrect interpretation and creates hurdles for effective communication. This work deals with Computer-Assisted Language Learning (CALL) concerning English language word pronunciation. In this work, a system is developed, which first teaches the correct pronunciation and scores the correctness of pronunciation produced by the user of a word. Feedback is provided when a person pronounces a particular word indicating how close is the pronunciation to the ideal pronunciation, thus serving as a metric of correctness. The proposed approach uses the Mel Frequency Cepstral Coefficients (MFCCs) of the spoken word and deep neural networks. Our results show that the proposed method is beneficial in identifying the accurately pronounced words and thus can be used for improving pronunciation. This work can readily be extended to other languages for pronunciation matching.

Keywords Pronunciation matching · MFCC · CNN · CALL

A. Upadhyay (✉) · B. K. Sonwani · V. A. Baghel · Y. K. Sinha · A. S. Patel · M. Ojha
Dr. SPM-International Institute of Information Technology Naya Raipur, Atal Nagar C.G, India
e-mail: ashish15100@alumni.iiitnr.ac.in

B. K. Sonwani
e-mail: bhupendra15100@alumni.iiitnr.ac.in

V. A. Baghel
e-mail: vimal15100@alumni.iiitnr.ac.in

Y. K. Sinha
e-mail: yash15100@alumni.iiitnr.ac.in

A. S. Patel
e-mail: ashish@iiitnr.edu.in

M. Ojha
e-mail: muneendra@iiitnr.edu.in

1 Introduction

The people who study and use English as a second language have trouble pronouncing the utterances of many words correctly. They often stress on particular syllables that don't need to be stressed too much otherwise may pronounce entire syllables incorrectly. For second language learners, it is easy to learn a second language in its written form without a human teacher. All it requires is dedication and a textbook. It is not necessarily an advantage to receive input from a teacher, depending on learning style and preferences. However, in the case of learning to speak a second language: It is not possible to learn correct pronunciation from a book. While a learner can use sound samples as a reference and repeat them, but feedback about the correction of repetition of the learner is not provided. Furthermore, even if the utterance sounds accurate to the speaker, but it may be due to the lack of awareness of certain aspects of the spoken language. Thus, it is a requirement to get feedback on the utterance. Usually, this means that a teacher must be available to listen to the learner. In many cases, this takes place in a classroom environment, leaving little time for individual feedback. The best way to improve the pronunciation of a learner is in one-on-one sessions with a teacher.

There have been several attempts in literature wherein researchers have tried CALL systems [1], but still the problem of learning to pronounce correctly remains unsolved.

This work facilitates people having English as their second language to improve their pronunciation. The proposed solution is implemented in two necessary steps. The first step is to provide correct utterances of every word being learned, thus eliminating the role of human teachers, which helps in reducing human error. The second step is to provide feedback. Therefore, when a person listens to the correct pronunciation of a word and tries to reproduce the same, they will receive a feedback percentage indicating how similar their pronunciation is to the ideal pronunciation of the word. This feedback will help them to judge and improve upon their pronunciation and also works as an indicative metric for any improvement. The techniques from speech recognition have been leveraged to create an automatic pronunciation checker for language learning software. Second language learners receive feedback based on their utterance. A long-term goal of this technology is to replace individual input from a human teacher with a language learning software.

The automatic pronunciation checker is devised by adapting the pattern matching algorithm, usually applied in speech recognition. Since the classic implementation of pattern matching is speaker-dependent between speech signals in most cases, a neural network that is trained to be speaker-independent is used as a distance metric [2]. There are two parts of the developed algorithm: first, MFCC Feature Extraction from the wave signal, and second, Deep Neural Network (CNN/LSTM) for learning the pattern from the signal.

The manuscript organization is as follows: In Sect. 2, we discuss some of the related works that have been done in the literature. Then, in Sect. 3, we propose our method to solve the problem. We discuss the experimental setup and dataset used

in Sect. 4 after which we move to Sect. 5 discusses the results obtained from the experiments performed. Finally, in Sect. 6 we conclude our work with some possible future works that can be done to develop this project further.

2 Related Work

Although speech recognition technology is gaining advancement in the research field but is still very far from other problems like image recognition. Computer Engineering and Computer Laboratory, ETH Zurich [2] is working to develop an automatic pronunciation checker for the issuance of Swiss visas and permanent residency (PR). The Government of the Netherlands is also working to develop a similar kind of software system to test their visa applicant's language capabilities [2–4]. Initial works used the approach of taking Euclidian Distance as a distance metric between wave signals. The model is speaker-dependent. New methods make use of neural networks as a distance metric between wave signals. Here the model is speaker-independent. Cucchiari et al. [5] developed a speech recognition technique to test the foreign speaker's proficiency in the Dutch language. A continuous speech recognizer (CSR) was introduced in this paper. CSR uses Context Independent Hidden Markov Model, language models (Unigram and Bigram), and a lexicon. Also, an automatic pronunciation checker is developed and compared using four approaches. In first, the acoustic-phoneme classifier is proposed using Linear Discriminant Analysis (LDA), the second classifier is based on cepstral coefficients in combination with LDA, and the third classifier implies a confidence measure technique. The fourth approach based on acoustic-phoneme and LDA gives the best results. Another framework is proposed by Strik et al. [6] for computer-assisted language learning systems in which the error can be automatically detected. Gao et al. [7] proposed an approach in which speech recognition of the spoken responses is done using Support Vector Machine (SVM) and Deep Neural Network model. Pocket sphinx library for extracting features from voices is used. They have used the phoneme approach instead of MFCC vector.

Authors in the work [8] introduced a method based on deep learning to assess the quality of a CALL system. This method has four different steps: in the first step, a third-party automated speech recognition (ASR) software is used for speech recognition which is then processed to identify 27 lexicon features and 22 grammatical features using different RNN-based language models in second and third step; which is then fed into a deep neural network with 49 input nodes for classification into acceptable or non-acceptable classes. The system was presented in the 2017 SLaTE challenge in Sweden and which gave a reasonable performance on various datasets. Similar work was proposed in [9] for automatic speech evaluation in CALL, where authors employed a multi-modal sparse autoencoder (MSAE) to make use of acoustic and lexical features which is then fed into a recurrent autoencoder (RAE) to utilize the temporal features. Finally, the obtained features are fed into an attention-based multi-scale bi-directional LSTM to generate the final score on speech expression.

The authors claim to have achieved a human level prediction ability with an acceptance rate of 70.4%. Other examples of similar works in the area of CALL can be found in works [10, 11] wherein [10] authors presented a system for automatic proficiency evaluation combining different Hidden Markov Models (HMM) and deep neural networks (DNN). The work proposed a novel reference-free error rate (RER) to evaluate the English proficiency of Japanese speakers. Li et al. [11] proposed three novel methods based on deep neural networks and Bi-LSTMs to improve the mispronunciation detection in the Mandarin language for non-native learners.

3 Proposed Work

The task of mapping words to their pronunciations seems daunting at first glance. It requires making the system intelligent. Thus, require a system based on basic concepts of pattern matching and speech recognition. In pattern matching, each word of the vocabulary that needs to be detected requires several recorded reference samples. The test sample is compared with each of the reference samples, and the best match is selected as the recognized word. Speech recognition is the process of turning spoken language into text. A division is made between two fundamentally different approaches to identify spoken language. The proposed model is a combination of MFCC for vectorization and Convolution Neural Network (CNN) architecture for the speaker-independent speech recognition model.

3.1 Vectorization Using MFCC

MFCC is the most advanced of the lot and produces the best results. The MFCC is used in speech detection to represent a signal. The data is analyzed at certain time intervals, where each time an MFCC vector is generated. The dimension of such a vector can be chosen depending on the application; also, the first and second derivatives can be included. A sequence of MFCC vectors represents a spoken signal and can be used for direct comparison or statistical analysis. The properties of MFCC are similar to the Discrete Fourier Transform Cepstrum, but research from psychoacoustics is incorporated into MFCC [2]. Its goal is to achieve resembling MFCC sequences for signals that are perceived to sound similar by humans. For example, instead of measuring the pitch of a tone by its frequency in hertz, the so-called Mel scale is used. After producing the vectorized output from the sound file, there is a need to find patterns in these sets of generated numbers.

3.2 Deep Neural Network Model Based Implementation

Neural Networks (NNs) are an essential tool in Machine learning [4]. A deep neural network is used to create a speaker-independent model for learning patterns and providing the desired mapping and hence achieving speech recognition. In this work, a multilayer perceptron (MLP) is used to create a new distance metric for the comparison of two MFCC vectors, replacing the Euclidean distance. The Euclidean distance is not speaker-independent. The main goal of using NN instead is to eliminate the component of the speaker from the distance metric [5]. For each word of the vocabulary that needs to be detected, one or several reference samples are recorded. The test sample is compared with each of the reference samples resulting in the selection of the best word. This approach primarily works for single words or (short) predefined sentences. The traditional method to calculate the similarity of two samples, i.e. the Euclidean distance between two MFCC, is speaker-dependent requiring both the reference and the test sample to be recorded by the same person to achieve accurate results. To resolve this issue, there have been efforts to create a speaker-independent distance metric using neural networks (NN) [3]. A diagrammatic representation of the workflow is shown in Fig. 1.

The steps followed for implementation are as follows:

1. Trim the recordings to a specific duration of seconds. The usual choice in this project is two seconds. This is done to remove any lingering noise after the utterance. It also ensures that all recordings are of equal length.
2. Vectorize these two-second clips to a matrix of numbers. We use Mel Frequency Cepstral Coefficients that was explained above for achieving the same. We, in this approach, have used an MFCC vector of dimensions 150×30 to convert sound waves to numerical values. Earlier, we were using an MFCC vector of size 20×11 , which is uneven to fully capture the information contained in the clip.
3. Build the architecture of CNN. The proposed architecture has been diagrammatically shown in Fig. 2.

The network can give predictions for any new words that it may encounter once it is trained on enough data.

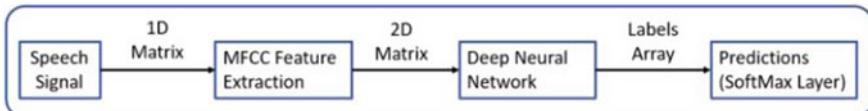


Fig. 1 Workflow of our proposed approach

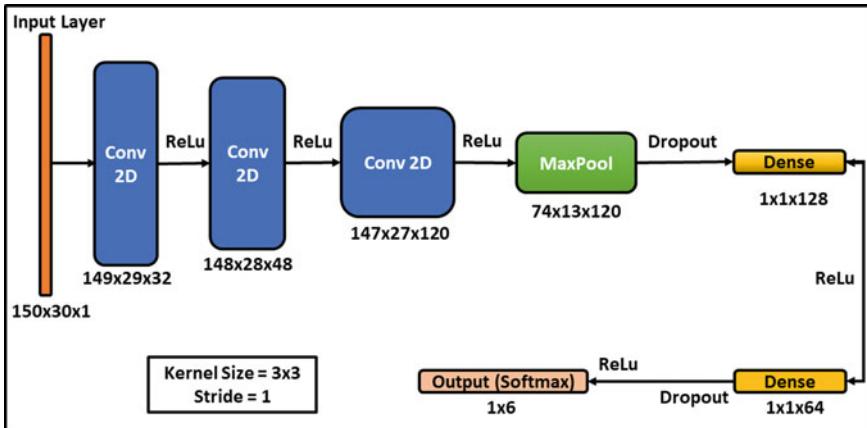


Fig. 2 Model architecture

4 Experimental Setup

The experiments are performed on two Nvidia GeForce GTX 1080 \times GPUs each with a memory of 11 GB. Python 3.5 is used as the programming language, Keras is used as the deep learning framework running on top of the Tensorflow. Python’s Librosa library is used to extract the MFCC vector and dealing with other speech-related problems.

4.1 Dataset

For this work, the data is provided by the National Informatics Centre (NIC). We also gathered the data by recording the audio of the university students specifically. Our dataset is described as follows:

- No. of Speakers: 45
- No. of Words: 5
- No. of Utterance/word by a speaker: 10
- Total data: $450 \text{ samples per word} \times \text{total 5 words}$.

We have collected a total of 2250 words for training the model. The hold-one out method is used for cross-validation in 60:40 ratios. Thus, 1350 data samples for training and 900 data samples for validation.

This dataset consists of audio files. The folder names correspond to the labels assigned to each sound file. For example, a folder named “Table” will contain utterances of the word Table. All these sound files are in *.wav format that consists of the sound waves. These sound waves cannot be identified by the system. The system

recognizes only numbers creating a need for vectorization, for which MFCC features are used.

4.2 Hyperparameters

In this work, there are two main algorithms for which we need to tune the hyperparameters in order to receive good performance.

First, the size of the MFCC vector, for which we experimented with different combinations ranging from 12 cepstral coefficients to 35 and 20 frequency intervals to 150 intervals. With different results, we found out 30 cepstral coefficient and 150 frequency intervals to be best and thus giving us an MFCC feature vector of size 150×30 .

Second, for the hyperparameters of a deep neural network model, we first started with simple network architecture and gradually increasing the depth while taking the overfitting in mind. With three convolutional layers and one maxpooling layer we ran grid search cross-validation to obtain the learning rate of our optimizer function, in this case, ada-delta. Four different learning rates were considered, [1, 0.1, 0.01 and 0.001] out of which 0.01 is found to be the best. To prevent overfitting, we also use dropout regularization on in our architecture, whose value is kept to 0.35 obtained from grid search CV.

5 Results

We ran the different combinations of experiments on the dataset obtained using the proposed method. We used two main classes of deep learning architectures, namely Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) with different combinations of classes for classification.

After training, the model on ideal pronunciations of each word's sufficient number of utterances, we achieved an accuracy of 90.5% test accuracy. The MFCC vector is shown in Fig. 3 and accuracy in Fig. 4a for size 150×30 .

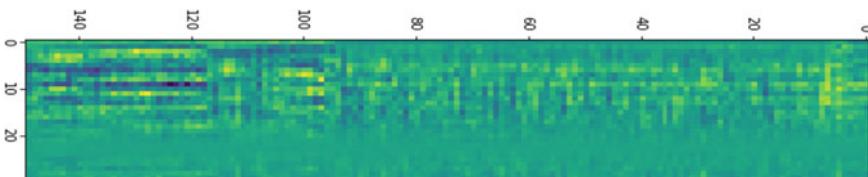


Fig. 3 Colour map representing MFCC vector for size 150×30 (the shown matrix is transposed)

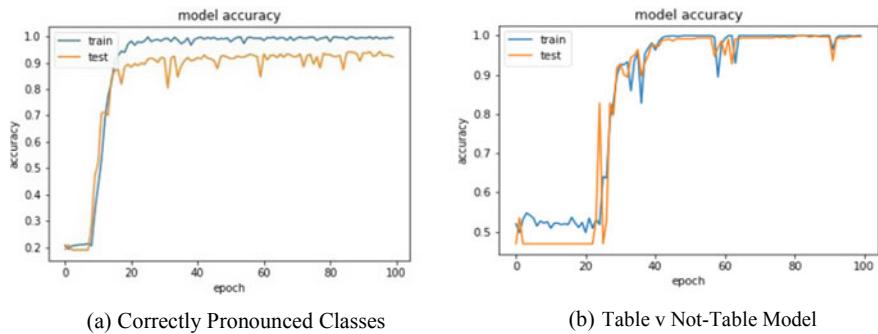


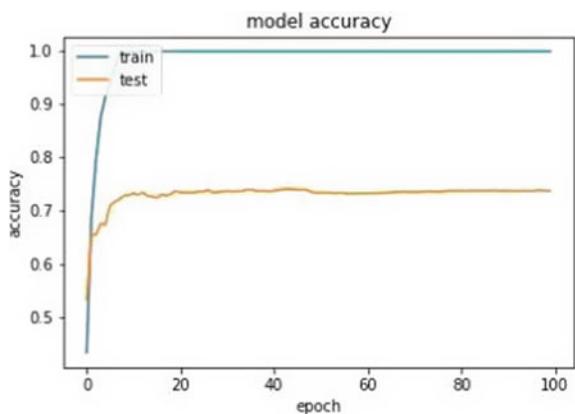
Fig. 4 Accuracy for different models with different class combinations

The proposed model can also provide a relative score to indicate if a word is partially mispronounced. Thus, we recorded false data for the word “Table” and found 450 partially incorrect utterances of the word. Then we made a CNN model for a single word “Table” having two output classes namely “Table” and “Not Table”. The train and cross-validation accuracy was 99%, signifying robust learning. The training and validation accuracy can be seen in Fig. 4b.

Long Short-Term Memory (LSTM) model is also explored, but the CNN model has given better results as shown in Fig. 5. This can be attributed to the fact that LSTM is particularly good at remembering sequences such as entire sentences, but CNN is better at recognizing individual words since they can be trusted as an image with a definite pattern.

A model is then constructed which outputs labels for all words. Two classes for word *Table*, *Table-yes*, and *Table-maybe* as partially incorrect utterances are also collected. Thus, the model had the final output classes as Australian, Japanese, Secretary, Saturday, Table-yes, Table-maybe. The accuracy of the model is shown in Fig. 6a and loss in Fig. 6b. Lastly, regularization techniques are used to improve accuracy.

Fig. 5 LSTM accuracy



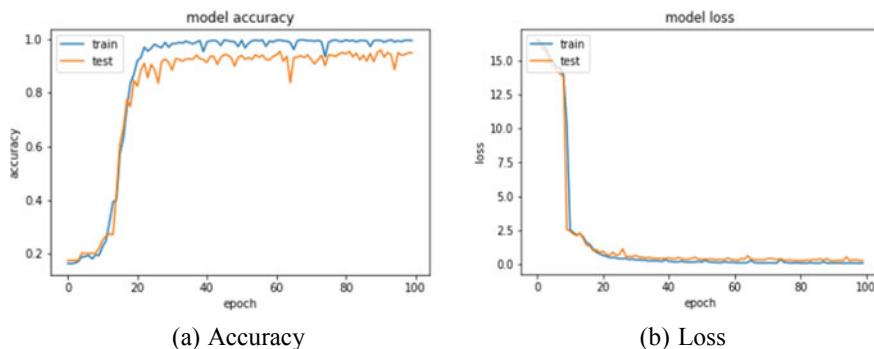


Fig. 6 Accuracy and loss for the model with six classes (Australian, Japanese, Secretary, Saturday, Table-yes, Table-maybe)

6 Conclusion

In this proposed work, a system is developed and demonstrated. The model takes a sound wave as an input and gives a certain metric of similarity as the output. The concepts of “Speech Recognition” and “Pattern Matching” are used to create a pronunciation matching tool. This system can be used to enhance the pronunciation skills of the English language for people having English as their second language. The tool matches the similarity of an utterance of a word by a speaker to the ideal pronunciation and gives a percentage similarity or a metric to judge the pronunciation similarity. The first requirement was to handle sound files that contain waves and are continuous in nature which is handled by MFCC. The second requirement was to make the system speaker-independent and to provide a model to match words to their utterances. This requirement was satisfied by using Convolutional Neural Networks. The proposed network model can be used to give predictions and ultimately give a percentage score which is our aim to achieve.

The possible future works can be adding more data to create a dataset with all words having two classes, one for the correct pronunciation and another for the incorrect pronunciation. Also, this model was very huge in size, around 70 MB which can also be reduced to accommodate a mobile app.

References

1. Krasnova E, Bulgakova E (2014) The use of speech technology in computer assisted language learning systems. Springer, Cham, pp 459–466
 2. Naghibi T, Jeisy K, Pfister B (2015) Automatic pronunciation checker. MS thesis, TIK-ETH Zurich, Switzerland
 3. Müller S (2008) Speech processing technologies for computer-assisted speech training. Diploma thesis, Inst. Tech. Informatics Commun. Networks, ETH Zürich, DA-2008-0

4. Gerber M, Kaufmann T, Pfister B (2007) Perceptron-based class verification. In: Chetouani M, Hussain A, Gas B, Milgram M, Zarader JL (eds) Advances in nonlinear speech processing. NOLISP 2007. Lecture notes in computer science, vol 4885. Springer, Berlin, Heidelberg, pp 124–131
5. Cucchiarini C, Strik H, Boves L (1997) Automatic evaluation of Dutch pronunciation by using speech recognition technology. In: 1997 IEEE workshop on automatic speech recognition and understanding proceedings, pp 622–629
6. Strik H, Truong KP, de Wet F, Cucchiarini C (2007) Comparing classifiers for pronunciation error detection. In: 8th annual conference of the international speech communication association, Antwerp, Belgium, pp 1837–1840
7. Gao Y, Lal Srivastava BM, Salsman J (2018) Spoken English intelligibility remediation with Pocketsphinx alignment and feature extraction improves substantially over the state of the art. In: 2018 2nd IEEE advanced information management, communicates, electronic and automation control conference (IMCEC), pp 924–927
8. Oh YR, Jeon HB, Song HJ, Kang BO, Lee YK, Park JG, Lee YK (2017) Deep-learning based automatic spontaneous speech assessment in a data-driven approach for the 2017 SLaTE CALL shared challenge. In SLaTE, pp 103–108
9. Li J, Wu Z, Li R, Xu M, Lei K, Cai L (2018) Multi-modal multi-scale speech expression evaluation in computer-assisted language learning. In: International conference on AI and mobile services. Springer, Cham, pp 16–28
10. Fu J, Chiba Y, Nose T, Ito A (2020) Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models. Speech Commun 116:86–97
11. Li W, Chen NF, Siniscalchi SM, Lee CH (2019) Improving mispronunciation detection of mandarin tones for non-native learners with soft-target tone labels and BLSTM-based deep tone models. IEEE/ACM Trans Audio Speech Lang Process 27(12):2012–2024

Noise Reduction in SAR Images with Variable Mode CT



R. Durga Bhavani, A. Ravi, and N. Mounika

Abstract Multiplicative noise of an image destructs the pixel association and leads to entropy loss. Somehow filters managed to reduce the noise but were not able to reform the image this leads to blurring of the image. To avoid this transform were measured, this results in a satisfactory image reconstruction with less part of information recovery. For this reason, enhancing image was considered, enhancing the image leads to improve entropy value. So, previously redundant and fusing methods were applied to Satellite Aperture Radar (SAR) images. Here we are providing a novel approach of fusing decomposition techniques i.e., Redundant curvelet transform (RFDCT) with variational Mode Decomposition (VMD). This results in the improvement of 11 parametric values and comparing with existing simulations of RFDCT, RFDCT with Empirical Mode Decomposition (EMD).

Keywords Multiplicative noise · Pixel association · Entropy · Blurring · SAR · Fusion · RFDCT · EMD · VMD · IMF

1 Introduction

RADAR images have an innate existence of SPECKLE noise, this will cause by utilizing equipment or by radiation of different sources. The impact of speckle noise in pictures is as per the random displacement of prominently bright or dark from Atlantis Scientific Inc [1, 2]. For the intention of noise reduction prior to different strategies were proposed some of them are Lee, Frost, refined map and map. The multiplicative or speckle noise decimates the image, in view of natural conditions. To affect this noise a noise estimator ‘Variance’ represents by σ^2 , is considering with a

R. D. Bhavani · A. Ravi (✉) · N. Mounika
PSCMR College of Engineering and Technology, Kothapeta, Vijayawada 520001, Andhra Pradesh, India
e-mail: ravigate117@gmail.com

value of 0.2, 0.5, 0.7 and 1. These values were comparative and just aides in resultant simulations and values can be changed. If the estimation crosses the boundary it will results in extra noise in the image and there is no probability of recovering the image to typical condition.

The portrayal of the speckle noise $b(x, y)$ is specified in Eq. (1). Input image is represented by $a(x, y)$.

$$b(x, y) = a(x, y) + \sqrt{12 * \sigma^2} * a(x, y) * (E(a) - 0.5) \quad (1)$$

Here $E(a)$ representing random function for size of input image ($a(x, y)$) is completely depending on Theorem 1.

Theorem 1 For a random variable $A : \Omega \rightarrow \mathbb{R}$, $E(A) = \sum_{r \in \text{range}(A)} P(A = r).r$

Proof Expectation is given by $E(A) = \sum_{s \in \Omega} P(s).A(s)$. But $\text{range}(A)$ If $\sum P(s)A(s) \exists A(s) = r$. Then $P(A = r) \cdot r$. Therefore, the finale outcome is $E(A) = \sum_{r \in \text{range}(A)} P(A = r).r$, this results fewer value outcomes from a broaden values.

To diminish the inter-pixel group and AWGN noise a multiplicative noise, SUBSPACE technique is utilized. To improve the intensity of the image in pre-processing phase the SAR noised image is experienced to Cholesky factorization approach. But the drawback of the framework in this methodology is it features the edges and there will be no enhancement found in PSNR, MSE parametric values. For edge enhancement, we can access this method proposed by Yahya et al. [3].

Konstantin Dragomiretskiy and Dominique Zosso proposing algorithm reduce the composite computation part of EMD and accomplishes band limited intrinsic mode function. This provides the mathematical proof of reducing noise and is considering as our main technique [4].

In Sect. 2 the method of RFDCT and mathematical proof of reducing the noise is analyzed. In Sect. 3 the combinational results of RFDCT and EMD were explained with algorithm proof. Section 4 deals with the brand-new algorithm proposing, i.e. RFDCT in combination with VMD.

2 RFDCT

RFDCT possess curvelet transformation characteristics, it includes the decomposition of the image depending on scale values ' j '. Here j is the scaling factor, for curvelet 2^j to 2^{-j} is restricted as scaling constant and helps mostly in localizing the decomposed values. This is more extended to DT CWT (Dual Tree Complex Wavelet Transform), the combination of wavelet transforms and Fast Fourier transform helps in generalizing the factors and decomposing the values based on rotation angle and scaling factor, Curvelet transformation is representing by ϕ [5].

All the coordinates were transformed into the frequency domain and polar coordinates (angular frequency, radius) were measured by the help of localizing values of ξ . Therefore, basic curvelet transform technique is given by

$$\hat{\phi}_{j,0,0}(r, w) = 2^{-\frac{3j}{4}} W(2^{-j} r) \tilde{V}_{N_j}(\omega) \quad (2)$$

where r and ω are polar coordinates and the limitation can be considered as $r \geq 0$ and $\omega \in [0, 2\pi]$, j is equivalent to total number of scales and ξ (ξ_1, ξ_2). Therefore, $r = \sqrt{\xi_1 + \xi_2}$ and $\omega = \tan \frac{\xi_1}{\xi_2}$ were used for measuring polar coordinates.

Now, for RFDCT the basic curvelet transform technique presented in Eq. (2) has been converted into the following value

$$\hat{\phi}_{j,0,0}(r, w) = 2^{-\frac{3j}{4}} W(2^{-j/2} r) \tilde{V}_{N_{j/2}}(\omega) + 2^{-\frac{3j}{4}} W(2^{j/2} r) \tilde{V}_{N_{j/2}}(\omega) \quad (3)$$

3 RFDCT with EMD

EMD is a time-frequency data determined by the adaptive decomposition mode of a signal. This acquires shifting algorithm assembles to AM/FM modulation. It has 3 stages (1) detecting local minima and maxima values (E_{\min}, E_{\max}), (2) detecting the edges of the images; (3) enumerate the fine points, and equalize the output to IMF's. Fine points will be calculated as

$$D = I - \left(\frac{1}{2} \right) * (E_{\min} + E_{\max}) \quad (4)$$

After applying EMD on the speckle noised image the boundaries of the images were enhanced and the decomposed image is dispensed to RFDCT by using Eq. (5). This helps in complete reduction of multiplicative noise from the image and mathematically it is representing as

$$\begin{aligned} \text{RFDCT}(\hat{\phi}_{j,0,0}(r, w)) \\ = a - \left(\frac{1}{2} \right) * \left(2^{-\frac{3j}{4}} W(2^{-j/2} r) \tilde{V}_{N_{j/2}}(\omega) + 2^{-\frac{3j}{4}} W(2^{j/2} r) \tilde{V}_{N_{j/2}}(\omega) \right) \end{aligned} \quad (5)$$

4 RFDCT with VMD

Variational mode decay decomposes signals in different modes or intrinsic mode capacities by analytics of variety. Each mode of the signal will be expected, require

conventional repetition aid approximately a central frequency. VMD will figure out central frequencies and inalienable mode works focused ahead the individuals. Frequencies which use a simultaneously streamlining method called (Alternate heading system for Multipliers) ADMM. Those first detailing of the streamlining issue will be constant in time Web-domain. Those constrained detailings is as [4]

$$\frac{\sum_k \|u_k^{n+1} - u_k^n\|_2^2}{\|u_k^n\|_2^2} < \epsilon \forall \omega \geq 0 \quad (6)$$

Equation 6 is indicating the optimized version of variable mode decomposition and this is more overly known as band limited intrinsic mode function and u_k is the intrinsic mode function. After applying VMD on the speckle noised image the edges of the images were enhanced and the decomposed image is subjected to RFDCT by using Eq. (7). This helps in the reduction of multiplicative noise from the image and mathematically it is represented as in equation. The maximum and minimum values are replaced with the decomposed scaled values of RFDCT.

$$\begin{aligned} \text{RFDCT}\left(\hat{\phi}_{j,0,0}(r, w)\right) := \\ \left\| \left(2^{-\frac{3j}{4}} W\left(2^{-\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) + 2^{-\frac{3j}{4}} W\left(2^{\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) \right)_k^{n+1} \right. \\ \left. - \left(2^{-\frac{3j}{4}} W\left(2^{-\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) + 2^{-\frac{3j}{4}} W\left(2^{\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) \right)_k^n \right\|_2^2 \\ \sum_k \frac{\left\| \left(2^{-\frac{3j}{4}} W\left(2^{-\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) + 2^{-\frac{3j}{4}} W\left(2^{\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) \right)_k^n \right\|_2^2}{\left\| \left(2^{-\frac{3j}{4}} W\left(2^{-\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) + 2^{-\frac{3j}{4}} W\left(2^{\frac{j}{2}} r\right) \tilde{V}_{N_{\frac{j}{2}}}(\omega) \right)_k^n \right\|_2^2} \end{aligned} \quad (7)$$

Task: Speckle Denoising of SAR Images by Integrated Redundant Curvelet Transformation Variational Mode Decomposition

KEY PARAMETER: Quantity of resolution levels I, minimum block size is B_{\min}
Calculate curvelet with I scales to obtain stable wavelet sub-bands W_j .

1. Let $\hat{I} = I$.

Locate Wedge minimum value

For iterator = 1 to size (I) do

1.1. Segment the sub-band W_j with blocks of side length.

1.2. Relate the digital ridgelet transform to all blocks to acquire the stable curvelet coefficients.

1.3. Identify the major stabilized curvelet coefficients to obtain M.

1.4. If length of the wedge is minimum the double the size of the wedge

1.5. Else equalize the same wedge with the new one

1.6. end if

1.7. Fusing the image repeatedly with same out image to enhance

1.8. Continue step 1 to 6

1.9. End For

2. Coefficients in Eq. (3) were obtained.
3. Interpolate the decomposed
4. Fine points ‘U’ has to be extracted from the envelope by Eq. (6).
5. Replicate 2-4 till U becomes the IMF of the U.
6. Apply inverse curvelet transform.

Reconstruction: Relate the HSD iterative reconstruction by the curvelet multi-resolution support to get the final estimate.

5 Results

Results in tabular form represent different referenced parameters to measure the efficient output and the mathematical expressions are considered from [6].

Figure 2 illustrates the comparison work of RCT, RCT with EMD and RCT with VMD. As a part of theoretical analysis VMD represents much more smoothing performed by EMD and this is practical results displays in Fig. 1. Figure 1 with tiles name as A, B, C and D, respectively known as Original image, RCT output, RCT + EMD and VMRCT. Tabular values from Table 1 suggest the efficient PSNR is obtained with VMRCT and Error representing values were also much more efficient than any else.

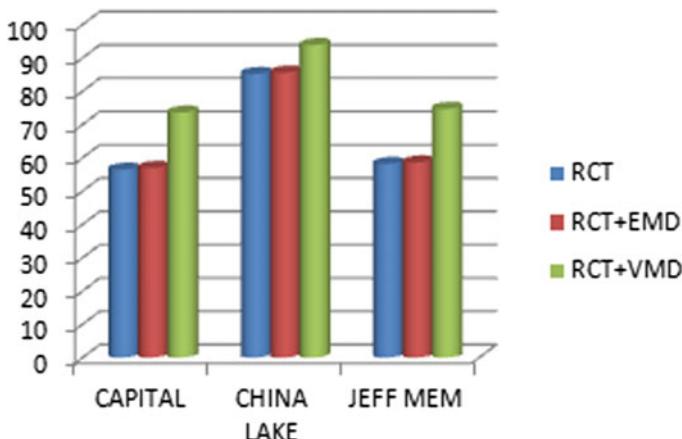


Fig. 1 PSNR comparison values for various SAR images

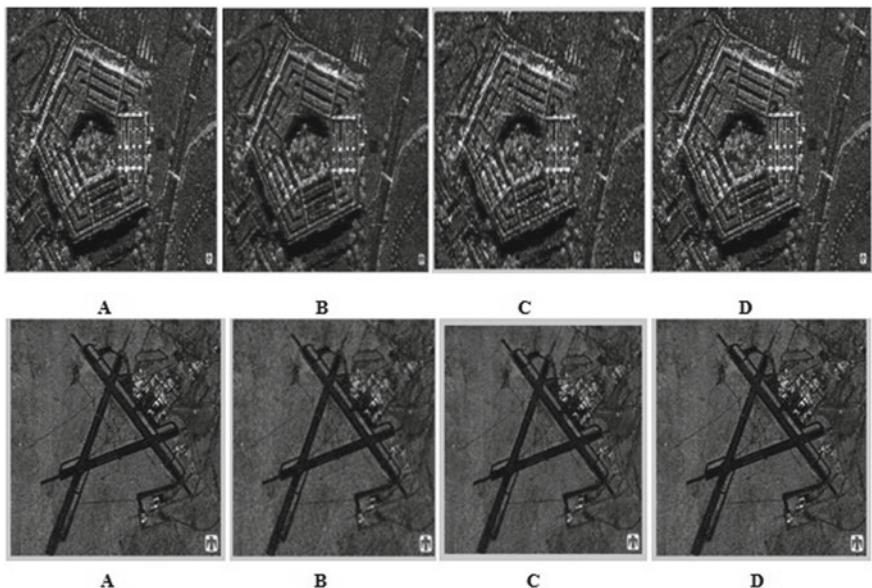


Fig. 2 Results for capital and china lakes SAR images

Table 1 Integrated redundant curvelet transform with empirical mode decomposition

	Redundant curvelet + variational mode decomposition		
	Capital	China lake	Jeff Mem
MSE	0.013	0.0012	0.005
PSNR	73.57	93.83	94.67
SNR	7.7992	13.605	8.65
Entropy	7.3249	7.8364	8.02
Corr	1.654	0.9876	0.9277
Std	0.25	0.29	0.5296

6 Conclusion

A novel method was implemented in this paper results in an efficient scheme of denoising and improving the quality parameters of an image. This is possible by the smoothing operation performed by VMD and the improvement of information carried using redundant curvelet transform. Here, curvelet coefficients were enhanced and this leads to the improvement of information. At last, a comparison a graph plotted between different images and quality measure were considered for different images.

References

1. Atlantis Scientific Inc. http://www.geo.uzh.ch/~fpaul/sar_theory.html
2. Zhang Y, Liu J, Li M, Guo Z (2014) Joint image denoising using adaptive principal component analysis and self-similarity. *Inf Sci* 259
3. Yahya N, Kamel NS, Malik AS (2014) Subspace-based technique for speckle noise reduction in SAR images. *IEEE Trans GEO-SCIENCE Remote Sens* 52(10):6257–6271
4. Dragomiretskiy K, Zosso D (2014) Variational mode decomposition. *IEEE Trans Signal Process* 62(3):531–544
5. Buades A, Coll B, Morel JM (2005) A non-local algorithm for image denoising. In: IEEE Computer Society conference on computer vision and pattern recognition, pp 20–26
6. Galbally J, Marcel S (2014) Image quality assessment for fake biometrics detection: application to Iris, Finger print and Face recognition. *IEEE Trans Image Process* 23(2)

Modeling IoT Based Automotive Collision Detection System Using Support Vector Machine



Nikhil Kumar, Debopam Acharya, and Divya Lohani

Abstract Due to the rise in automotive accidents across the globe, a cheap and reliable system is needed that can be retrofitted in any type of vehicle and can monitor road collision events. This research is aimed to develop an IoT system that uses contemporary smartphone's intrinsic sensors to accurately report vehicle collision accidents on the road. Absolute linear acceleration (ALA) and the speed of the vehicle have been used to train and test our Support Vector Machine (SVM) based collision detection model. During the testing, the accuracy of the model was found to be very high with a MAPE (mean absolute percentage error) of 0.6%.

Keywords Vehicle collision detection · Internet of things · Support vector machine · Absolute linear acceleration

1 Introduction

According to the World Health Organization (WHO), every year, nearly 1.35 million people die in road accidents, 3700 deaths a day on average. Moreover, 50 million people are injured or disabled every year [1]. More than 90% of road fatalities occur in low- and middle-income countries, which have less than half of the world's vehicles [1]. Most of these accidents are caused by vehicle collisions in which the car collides with other cars, road barriers, animals, pedestrians, or other fixed objects such as a pillar, tree, or building. Road accidents are often caused by road conditions, road

N. Kumar (✉) · D. Lohani

Department of Computer Science and Engineering, Shiv Nadar University, Gautam Buddha Nagar, UP, India

e-mail: nk438@snu.edu.in

D. Lohani

e-mail: divya.lohani@snu.edu.in

D. Acharya

School of Computing, DIT University, Dehradun, Uttarakhand, India

e-mail: dr.debopamacharya@dituniversity.edu.in

design, poor driving skills, and impairment due to intoxication (alcohol or drugs) and driver behavior, especially over speed, distracted driving, and road racing. In such cases, in the absence of a system that can identify and report the accident to the police, ambulance services, or family members, victims are often deprived of immediate medical attention. As per the golden hour principle (the relationship between mortality rate and delay in first aid after an accident), timely notification decreases the delay of medical treatment after an incident that can help reduce the mortality rate [2]. The types of road vehicle accidents can be categorized as a head-on collision, side collision, rear-end collision, rollovers, and fall-off. In recent past moves, information and communication systems have been used to reduce accident rescue response times [3, 4]. Knowledge of the type of accident is indicative of the type of injury suffered by the victims and is extremely useful in the planning and execution of the rescue operation.

Most of the researchers' collision detection systems built in recent years are expensive and restricted to high-end vehicles. These systems cannot be retrofitted to all vehicle types. In addition, such systems' ability to detect collisions is very limited. Work on designing solutions for road accident identification, management, and recovery have focused solely on improving reliability, assessing frequency, or reducing rescue after accident detection.

In this research, a vehicle collision detection system based on the Internet of Things (IoT) is proposed. The system not only detects but also records the kind of accident so that the victims can avail of the right kind of rescue and medical assistance timely. The proposed solution is a low cost, efficient, and can easily be retrofitted in any type of vehicle.

Contribution of this work. We have developed a smartphone-based IoT system that can accurately detect the vehicle collision event on the road. The proposed architecture of the system uses the accelerometer to measure the absolute linear acceleration and the GPS to measure the vehicle speed. In this work, the SVM algorithm with quadratic kernel has been used for training and testing the system to detect the vehicle collision events.

2 Architecture, Hardware Setup, and Software Setup

2.1 Architecture of the Proposed IoT System

We have proposed a context-aware IoT system to detect vehicle collision incidents with the help of an Android smartphone and a pre-built sensor system, Sensordrone (if necessary). Equipped with many in-built sensors such as accelerometer, gyroscope, compass, and GPS, etc., current smartphones can measure many parameters related to vehicle speed, maneuver, and orientation. Sensordrone is a multifunctional small sensor platform that contains 11 on-board ambient sensing sensors [5].

As shown in Fig. 1, the sensing system is deployed in a vehicle and placed on the

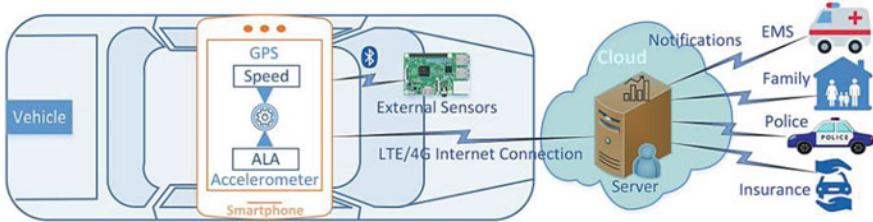


Fig. 1 Architecture of the system

flat armrest of the vehicle. The smartphone acts as an information sink and processor of the information. Most of the processing is done inside the smartphone, so it eventually reduces the information being sent to the server and saves the important internet bandwidth. Smartphone sends accident details, such as location, time, and vehicle ID, to the IoT server in the cloud using 4G/LTE internet. After aggregating and analyzing the sensors' data the IoT server delivers the incident alert to the intended recipients, including EMS (Emergency Medical Services), relatives, police, insurers, blood bank, and others.

2.2 Hardware Setup

In this study, the *Samsung Galaxy S8* android smartphone's intrinsic accelerometer (measurement range: ± 16 g) and intrinsic GPS sensor have been used to calculate the vehicle's ALA and velocity, respectively. To create a real-life collision scenario, a 1:12 scaled toy RC (Radio Controlled) car [6], which is made-up of high-performance ABS material, is used because real car tests are neither practical nor cost-effective. The maximum speed and operating range of this RC car is 53 km/h and 80 m, respectively. The hardware setup is shown in the inset of Fig. 3.

2.3 Software Setup

An android application called SNUSense is designed to collect the data from smartphone and Sensordrone sensors (Fig. 2 (left)). SNUSense pre-processes the information and reports the incident to the cloud-based IoT server (e.g., Google Firebase [7]). After analyzing the data, the IoT server sends the collision alert to the intended recipients. The application is designed to monitor more than 15 parameters for future research prospects so that it can handle multiple crash situations such as collision, rollover, and fall-off. Normally, the accelerometer produces a signal at a very high frequency of more than 2000 Hz. It is very difficult to record and process the deceleration fluctuation generated for 1–2 ms during the collision. In order to take account of

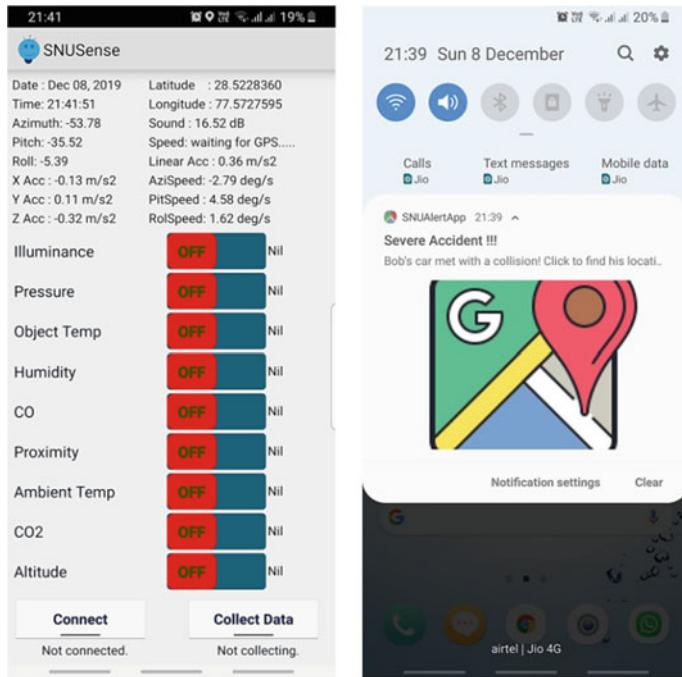


Fig. 2 (left) Screenshot of the SNUSense app, (right) screenshot of SNUAlertApp app

the peak value, the system is using a “*Ten Millisecond Moving Maximum*” approach in which the maximum value of every 10-millisecond duration is recorded.

We have created another android application, *SNUAlertApp*, to accept the collision alert notification, as shown in Fig. 2 (right). The notification incorporates the location of the collision, identity of the driver, and type of the accident. A simple touch on the notification area brings the user to *Google Maps*, where a place marker helps the user to identify the incident spot.

3 Parameters Used in the System

3.1 Speed, and Absolute Linear Acceleration (ALA)

The proposed solution uses GPS to measure vehicle speed and an accelerometer to measure absolute linear acceleration (ALA). ALA is a square root of the sum of squares of linear decelerations on the X , Y , and Z -axis of the accelerometer. ALA is a positive quantity, which is used to observe the force of impact, regardless of the direction of impact and the speed of the vehicle. GPS receives NMEA (National

Marine Electronics Association) [8] sentences which include several information such as date, time, longitude, latitude, etc. along with the *velocity*.

3.2 Dealing with False Positives

Driving activities such as sudden braking may produce an unintended false alarm. Therefore, to assess the maximum deceleration generated by the brakes and to differentiate the collision and brakes, we have used Maruti Suzuki Baleno car, as it is equipped with an anti-lock braking system [9]. It is observed that ALA does not cross 6.23 m/s^2 (i.e., 0.64 g) after the abrupt application of the brakes at different speeds ($30, 40, 50, 60 \text{ km/h}$).

4 Experimental Setup, and Data Collection

Smartphone and Sensordrone have been tied up close to the RC car's CG (center of gravity) (as shown in the inset in Fig. 3) to obtain an accurate dataset. All the tests have been performed at a speed of 35 km/h . We have used a plastic road barricading as a barrier (Fig. 3). Collision data has been collected by colliding the RC car setup into the barrier. This test has been performed 45 times. Test data is stored in a Comma Separated File (CSV) file in the storage of the smartphone.

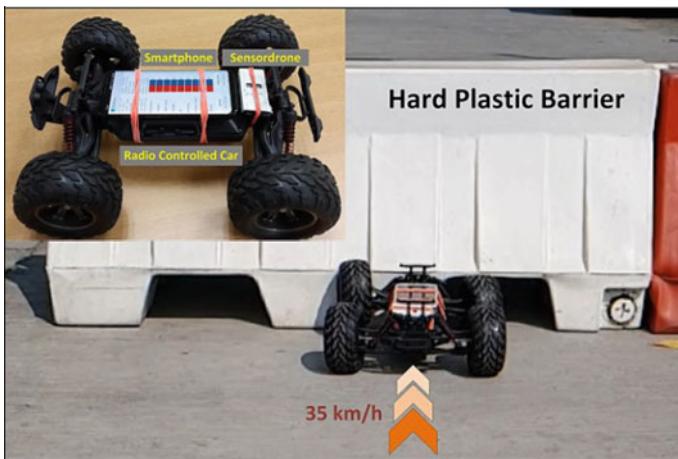


Fig. 3 Experimental setup

5 Mathematical Model for Accident Detection

This research examines support vector machines (SVM) to model the occurrence of the collision of a vehicle. The proposed model involves inputs of two specific parameters namely ALA, and speed of the vehicle, to detect the occurrence of a collision. The output of the proposed predicting model is dualistic, i.e., there can be only two potential outcomes of the proposed collision detection model; either collision occurs, or collision does not occur. Therefore, it is deduced as a binary classification problem in which the samples are represented as points in an n -dimensional space and mapped in such a way that the two category samples are segregated by a gap called hyperplane, which must be as wide as possible.

The SVM works on the rule of fitting a boundary to an area of points that all belong to the same class. Once a boundary has been fitted for the training data set, it is to classify whether the test sample lies within the border or not. It is the advantage of an SVM that if the boundary established once, most of the data points become redundant. Therefore, SVM requires a set of points called support vectors that can help to recognize and decide the boundary. The hyperplane is oriented in such a way that it is positioned as far as possible from the support vector sets. This orientation of hyperplane helps SVM to generalize the unknown cases more precisely in comparison to other classifiers such as neural network, which tries to minimize the training errors [10]. Another advantage of SVM classifier is that it requires less number of training data points for training in comparison to other statistical classifiers such as maximum likelihood classifiers [11]. With a large number of training samples, the K-Nearest Neighbor, Naïve Bayes, and Random Forest algorithms are more accurate and work faster, but our training data set is small, so SVM is best suited for our problem [12, 13].

The hyperplane is described in the “Eq. 1”, where W is the normal vector to the hyperplane, X is a data-point lying on the hyperplane, and B (i.e., bias) is defined as the position of the relative region to the coordinate center.

$$W \cdot X + B = 0 \quad (1)$$

The linearly separable case defines the two separate hyperplane classes as $W \cdot X_i + B \geq 1$ (for $Y_i = +1$) and $W \cdot X_i + B \leq -1$ (for $Y_i = -1$). Eventually, the margin condition would become

$$Y_i(W \cdot X_i + B) \geq 1 \quad \forall i \quad (2)$$

The training sample points on the two hyperplanes, $W \cdot X_i + B = \pm 1$, that are parallel to the optimal separate hyperplanes, are the support vectors. The margin on each side is $1/\|W\|$, so the total margin between the planes would become $2/\|W\|$. We want to maximize the margin, which is achieved by the constrained optimization problem under the inequality constraints of “Eq. 2”. Maximizing the margin between the planes is equivalent to minimizing the norm of W . i.e.

$$\min \left\{ \frac{1}{2} \|W\|^2 \right\} \quad (3)$$

If training data is noisy then to restrict the lower and upper bounds of input, slack variables $\{\xi_i \geq 0\}$ are introduced (distance by which it can violate the margin). So due to the soft-margin condition, “Eq. 2” becomes

$$Y(W \cdot X_i + B) \geq 1 - \xi_i \quad (4)$$

Due to the goal of maximizing the margin, while also minimizing the sum of slacks $C \sum_{i=1}^r \xi_i$, the primal objective for soft-margin SVM can be written as

$$\min \left[\frac{\|W\|^2}{2} + C \sum_{i=1}^r \xi_i \right] \quad (5)$$

Training data X is represented as $\phi(X)$ and mapped into a high-dimensional space H through a mapping function ϕ , which enables non-linear decision-making surfaces. Since the computation of $(\phi(X) \cdot \phi(X_i))$ is complex, it is simplified in high-dimensional space by using a quadratic kernel function such that

$$(\phi(X) \cdot \phi(X_i)) = k(X, X_i) \quad (6)$$

The decision function $f(x)$ is obtained in the form of the following equation.

$$f(x) = \text{sgn} \left(\sum_{i=1}^r \alpha_i Y_i k(X, X_i) + B \right) \quad (7)$$

Here, α_i denotes the Lagrange's multiplier.

6 Results and Discussion

The findings have been obtained from the proposed collision detection using the dataset created by SNUSense for training and testing. In our SVM model, the output function $f(x)$ uses the SVR (support vector regression) model with a cost of 2, 25 support vectors, quadratic kernel, and an epsilon value of 0.2 to develop our statistical model. Out of 900 experimental observations (data points), 550 observations of the collision event have been used for training and the rest 350 have been used for testing the model. If p is the likelihood of incidence, binary 1 is the occurrence and binary 0 is the non-occurrence of the collision. A scatter plot of test data of the SVM model is displaying in Fig. 4 (left) where blue dots are showing the occurrence while red dots are showing non-occurrence of the collision events. Red and blue cross signs are representing the false positive and false negative results of collision events.

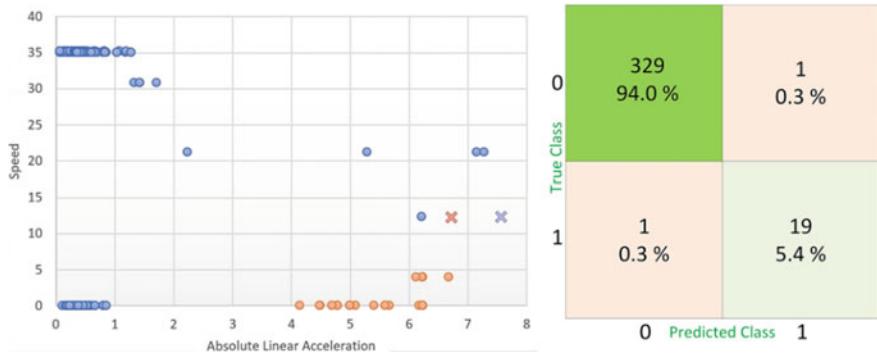


Fig. 4 (left) Scatter plot of collision testing, (right) confusion matrix for testing data

The accuracy of the collision detection system has been assessed using the mean absolute percentage error (MAPE) [14]. The MAPE is defined as

$$\text{MAPE} = \frac{1}{N} \sum_{i=0}^N \left| \frac{y_i - f(x_i)}{y_i} \right| * 100 \quad (8)$$

where N is the number of observations, y_i is the actual event that occurred at i th observation (0 or 1), x_i is the input vector (time), and f is the forecast model.

In Fig. 4 (right), a confusion matrix is showing that the MAPE for our SNU dataset was found to be 0.6, i.e., the system is capable of detecting 348 out of 350 test sample occurrences. Hence, it is concluded that our collision detection system has an accuracy of 99.4 percent and is capable of effectively identify automotive collision incidents.

7 Conclusion and Future Work

In this work, an IoT system, based on smartphone sensors, has been developed to accurately report the vehicle collision incidents to the intended recipients such as EMS, police, and family. SVM-based detection model has been used to detect the occurrence of the vehicle collision event. The proposed system is highly accurate to detect the vehicle collision occurrences with a MAPE of 0.6%.

As future work, we are intended to further refine the model to send the detailed collision notification with the degree of severity of the collision. The proposed system will be able to further classify the collision incident as mild, moderate, or severe.

References

1. GSRRS (2018) Global status report on road safety 2018: summary. World Health Organization, Geneva. https://www.who.int/violence_injury_prevention/road_safety_status/2018/English-Summary-GSRRS2018.pdf. Accessed 7 Dec 2019
2. Iyoda M, Trisdale T, Sherony R, Mikat D et al (2016) Event Data Recorder (EDR) developed by Toyota Motor Corporation. SAE Int J Trans Saf 4(1):187–201
3. Liyanage Y, Zois D-S, Chelmis C (2018) Quickest freeway accident detection under unknown post-accident conditions. In: 6th IEEE global conference on signal and information processing (GlobalSIP), Anaheim, CA, Nov 26–29
4. Dar BK, Shah MA, Islam SU, Maple C, Mussadiq S, Khan S (2019) Delay-aware accident detection and response system using fog computing. IEEE Access 7:70975–70985
5. Lohani D, Acharya D (2016) Real time in-vehicle air quality monitoring using mobile sensing. In: 2016 IEEE annual india conference (INDICON), Bangalore, 2016, pp 1–6
6. S911 (2020) GP TOYS Foxx S911 RC Truck. <https://g-p.hk/gptoys-foxx-s911.html>. Accessed 7 Dec 2019
7. Google Firebase (2019) <http://firebase.google.com>. Accessed 7 Dec 2019
8. NMEA (2019) National Marine Electronics Association, Severna Park, MD. <http://www.gps-information.org/dale/nmea.htm>. Accessed 7 Dec 2019
9. Gobbi M, Mastinu Gi, Prevati G (2014) The effect of mass properties on road accident reconstruction. Int J Crashworthiness 19(1):71–88
10. Utkin LV, Chekh AI, Zhuk YA (2016) Binary classification SVM-based algorithms with interval-valued training data using triangular and Epanechnikov kernels. Neural Netw 80:53–66. ISSN 0893-6080
11. Mathur A, Foody GM (2008) Multiclass and binary SVM classification: implications for training and classification users. IEEE Geosci Remote Sens Lett 5(2):241–245
12. Hmeidi I, Hawashin B, El-Qawasmeh E (2008) Performance of KNN and SVM classifiers on full word Arabic articles. Adv Eng Inform 22(1):106–111
13. Madzarov G, Gjorgjevikj D, Chorbev I (2009) A multi-class SVM classifier utilizing binary decision tree. Informatica 33:233–241
14. de Myttenaere A, Golden B, Le Grand B, Rossi F (2016) Mean Absolute Percentage Error for regression models. Neurocomputing 192:38–48. ISSN 0925-2312

Optimization-based Resource Allocation for Cloud Computing Environment



M. Chidambaram and R. Shanmugam

Abstract Cloud computing has become another age innovation that has colossal possibilities in undertakings and markets. Clouds can make it conceivable to get to applications and related information from anyplace. Organizations can lease resources from cloud for capacity and other computational purposes with the goal that their foundation cost can be diminished fundamentally. Further they can utilize expansive access to applications, in light of pay-more only as costs arise model. Henceforth, there is no requirement for getting licenses for singular items. Be that as it may, one of the significant traps in cloud computing is identified with upgrading the resources being dispensed. In view of the uniqueness of the model, resource allocation is performed with the goal of limiting the expenses related with it. Different difficulties of resource allocation are satisfying client needs and application necessities. In this paper, an optimization-based energy aware resource allocation is proposed for the allotting the resources in an improved way. An energy-aware particle swarm optimization method is proposed for the resource allocation in the cloud condition.

Keywords Cloud computing · Resource allocation · Particle swarm optimization · Optimization technique · Energy-aware resource allocation · Virtual machine

M. Chidambaram

Department of Computer Science, Rajah Serfoji Government College, Thanjavur, Tamil Nadu, India

e-mail: chidsuba@gmail.com

R. Shanmugam (✉)

Department of Computer Science, Arignar Anna Govt. Arts and Science College, Karaikal, Puducherry, India

e-mail: shanmugamrkk@gmail.com

1 Introduction

Cloud computing develops as another computing worldview which means to give solid, modified, and Quality of Service (QoS) ensured computing dynamic situations for end-clients [1]. Grid computing, parallel processing, and distributed processing together rose as cloud computing. The essential rule of cloud computing is that client data isn't put away locally however is put away in the data center of web. The organizations which give cloud computing service could oversee and keep up the activity of these data centers. The clients can get to the put away data whenever by utilizing Application Programming Interface (API) gave by cloud suppliers through any terminal gear associated with the Web.

Not exclusively are storage services gave yet additionally software and hardware services are available to the general open and business markets. The services gave by service suppliers can be everything, from the software or platform, infrastructure, resources. Each such service is individually called Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [2]. There are various advantages of cloud computing, and the most basic ones being lower costs, re-provisioning of resources, and remote accessibility. Cloud computing brings down cost by avoiding the capital consumption by the company in leasing the physical infrastructure from an outsider supplier. Because of the adaptable nature of cloud computing, we can rapidly access more resources from cloud suppliers when we have to expand our business. The remote accessibility enables us to access the cloud services from anywhere at any time. To gain the maximum level of the above-referenced advantages, the services offered as far as resources ought to be allocated optimally to the applications running in the cloud.

1.1 Significance of Resource Allocation

In cloud computing, Resource Allocation (RA) is the way toward assigning available resources to the required cloud applications over the web. Resource allocation starves services if the allocation is not managed correctly. Resource provisioning tackles that issue by allowing the service suppliers to manage the resources for each individual module.

Resource Allocation Strategy (RAS) is all about integrating cloud supplier activities for using and allocating scarce resources inside the farthest point of cloud condition to address the issues of the cloud application. It requires the sort and amount of resources required by each application so as to finish a client work. The request and time of allocation of resources are also a contribution for an optimal RAS. An optimal RAS ought to avoid the accompanying criteria as pursues:

- ***Under-provisioning*** of resources happens when the application is assigned with less quantities of resources than the demand.

- **Fragmentation of Resource:** situation arises when the resources are isolated. [There will be sufficient resources yet not able to allocate to the required application.]
- **Contention of Resource:** at the same time, when two applications are attempted to access the same resources.
- **Over-provisioning:** of resources arises when the application gets surplus resources than the demanded one.
- **Scarcity of resources:** due to the only availability of limited resources.

2 Related Works

Valliyammai and Mythreyi [3] proposed particle swarm optimization strategy with migration advances the allocation procedure utilizing computation and system-based parameters. Migration productively eliminates the issues of over-utilization of resources. The clustering of virtual machines has also been investigated in two measurements namely resource clustering and inactive clustering to increase the utilization of resources.

Anithakumari and Chandrasekaran [4] proposed a system of interoperability-based adaptable resource management. Initially, the SLA templates of private and open cloud are mapped utilizing the soft TF-IDF metric with case-based reasoning (CBR) approach. At that point, based on the mapped SLAs, various groups of cloud suppliers are framed with the assistance of k-means clustering system.

Alahmadi et al. [5] proposed an architecture that integrates VC with metro mist hubs and the central cloud to guarantee service congruity. We tackle the issue of energy effective resource allocation in this architecture by building up a Mixed Integer Linear Programming (MILP) model to limit control utilization by streamlining the assignment of various tasks to the available resources in this architecture.

Ficco et al. [6] proposed a meta-heuristic approach for cloud resource allocation based on the bio-roused coral-reefs optimization paradigm to display cloud elasticity in a cloud-data center, and on the classic game theory to improve the resource reallocation schema regarding cloud supplier's optimization destinations, as well as client necessities, communicated through service-level agreements formalized by utilizing a fuzzy linguistic technique.

Meng et al. [7] proposed the joint optimization strategy to enhance the quality of versatile cloud service. The authors formulate the computing resource allocation and remote bandwidth model as a triple-stage Stackelberg game, and understand it by utilizing backward technique. In addition, the interplays of triple-stage game are talked about and the subgame optimal harmony for each stage is analyzed. An iterative algorithm is proposed to obtain Stackelberg harmony.

Nayak et al. [8] demonstrated the current backfilling algorithm for booking deadline-based task utilizing Petri-Net. The paper displays the plan model of the

current backfilling algorithm. The model indicates real-time challenges of backfilling algorithm utilizing Petri-Net. The work also approaches with some structure issues of backfilling algorithm utilizing Petri-Net.

Guerrero et al. [9] addresses the optimization of document locality, record availability, and replica migration cost in a Hadoop architecture. Our optimization algorithm is based on the non-dominated sorting genetic algorithm-II and it simultaneously decides record block placement, with a variable replication factor, and MapReduce work scheduling.

3 Problem Statement

The information parameters to resource allocation strategy and the way of resource allocation vary based on the services, infrastructure, and the nature of applications which demand resources. Coming up next are the issues considered the resource allocation in the cloud condition.

- **Policy:** Since centralized client and resource management lacks in scalable management of clients, resources, and organization-level security policy.
- **Execution time:** estimating the execution time for a vocation is a hard task for error and client.
- **Virtual Machine:** The framework made out of a virtual network of virtual machines capable of live migration across physical infrastructure of multi-domain.

4 Optimization-Based Resource Allocation Algorithms

In cloud computing, various resources are furnished to the clients with the assistance of dynamic resource allocation. Resource allocation is an integral part of Infrastructure-as-a-Service model. Resource allocation is one of the major issues in cloud computing. Resource allocation is the way toward allocating the resources to the customers according to their need. There are various algorithms which are being utilized for resource allocation in cloud computing. These algorithms help in planning virtual machines on the server at various data centers. A portion of these algorithms are priority-based algorithm [10], preemptive and non-preemptive scheduling [10], bee algorithm [11], Choco-Based algorithm [12], bin-packing algorithm [10], and ant colony optimization [13]. These algorithms are utilized for productive resource allocation in cloud computing.

- **Ant Colony Optimization algorithm:** ACO is based on the behavior of the ants gathering nourishment [13].
- **Bee's Algorithm:** This algorithm is based on the action of honey bees to get their nourishment. In this algorithm, a metascheduler gets a new line of work with most reduced memory, input-yield, and processor necessity [11].

- **Priority Algorithm:** The ideas behind the dynamic resource allocation for seize able occupations in cloud is allocation of resources to the clients according to their demands; priority-based planning algorithm performs superior to the cloud min-min booking algorithm [10].
- **Bin-Packing Algorithm:** Bin-packing problem (BPP) includes the packing of the objects of given size into bins of given capacity [10].

5 Proposed Optimization-Based Resource Allocation for Cloud Computing Environment

PSO is an optimization method utilized for the energy conservation. In this paper, selected PSO as an optimization system to obtain the optimal ability and should be executed easily. PSO is selected with the goal that path finding is leisure and smarter to discover the VM energy and position on any host. Table 1 depicts the PSO used parameters and its description.

Algorithm 1 Particle Swarm Optimization

Step 1: Initialization Process

Step 1.1: Set constants k_{max} , c_1 , c_2 . Randomly initialize particle's position and velocity.

Step 2: Process of Optimization

Step 2.1: Evaluate function values f_k^i using design space coordinates x_k^i

Step 2.2: If $f_k^i \leq f_{best}^i$, then $f_{best}^i = f_k^i$, $p_k^i = x_k^i$

Step 2.3: If $f_k^i \leq f_{best}^g$ then $f_{best}^g = f_k^i$, $p_k^g = x_k^i$

Step 2.4: If terminating condition is satisfied, then go to third step

Step 2.5: Update particle velocities by the following formula:

$$v_{k+1}^i = v_k^i + c_1 r_1 (p_k^i - x_k^i) + c_2 r_2 (p_k^g - x_k^g)$$

Step 2.6: Increment k.

Step 2.7: Go to the first step.

Table 1 Terms and description used in PSO

Terms	Descriptions
C_1, C_2	Social and cognitive components
r_1, r_2	Random numbers among 0 and 1
x_k^i	Particle position
v_k^i	Particle velocity
p_k^i	best individual particle position
p_k^g	best global value

In this proposed work, PSO algorithm is utilized for the virtual machine (VM) allocation. For allocating the VM, the migration of VM is done. The overloading host is distinguished for the migration of VM. A local relapse strategy finishes the overloaded host detection. The migrations of VMs take place when the recognition of overloaded host is one, and the same VMs are allocated for the different hosts. In the proposed work, there are VM demands originating by the client and the demand handler of VM handles the solicitations. This solicitation handler allocates the solicitations to the hosts which will be random allocation. Each host's energy screening is done with other module. The PSO strategy is used to obtain the energy optimization. According to constraints of energy, the allocation of VMs to the hosts is done. Algorithm 2 used to check the overloaded host. The allocation of VMs takes place, when the host is not overloaded. The constraints in the allocation of VM are based on:

$$x_{ij} = \begin{cases} 1 & \text{where No. of request} > \text{No. of allotment on VMs} \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

When the allocation of VM allocation is finished, every host and overloaded load power are computed referenced in Algorithm 3. Every host power is noted in a rundown which assistance in computation of wellness function. When energy and position of VM on host are attained, then the PSO is implemented. Using these values, the computation of solitary host energy best value with various VMs, while is the host energy best value with all various hosts takes place. The computation of values takes places for the total quantity of hosts. In the final stage, with the minimal energy consumption, the allocation of VM takes place.

Algorithm 2 Detection of Overloaded Host

Input: Number of hosts, Number of VMs

Step 1: Find the overloaded host

Step 1.1: If () then

Step 1.2: VMs allocate to that host

Step 1.3: Else

Step 1.4: The host is overloaded host

Step 1.5: end if

Step 2: Calculate the power of host using:

Step 2.1: () ()

Step 3: The power of all the hosts are stored in the list.

Step 4: Exit

Algorithm 3 Particle Swarm Optimization based Resource Allocation

Step 1: Get, these values are obtained from the detection of overloaded host.

Step 2: Fetch

Step 3: Calculate()

Step 4: Select and

Step 5: Repeat {It is to be repeated up to n number of hosts}

Step 6: Updating of using Eqs. (1) and (2)

Step 7: Repeat Step 3.

Step 8: Update values of and

Step 9: Exit {Stopping condition is till the total number of hosts}

Output: Power on host having the lowest energy consumption.

6 Result and Discussion

This section illustrates the experimental setup and the results obtained with the proposed optimization-based resource allocation strategy. The proposed strategy is assessed in the Java CloudSim. The performance of the proposed optimization-based resource allocation strategy is compared with the existing resource allocation techniques like bin-packing, ant colony optimization, and priority-based algorithm. It is compared against the increasing size of the data center with the request size for the resource in the cloud environment. The convergence time is calculated in seconds for the proposed optimization-based resource allocation strategy, and the existing techniques. Table 2 depicts the convergence time in seconds by the bin-packing, ACO, priority-based resource allocation and proposed optimization-based resource allocation strategy for the resource request size is 150. Figure 1 represents the graphical representation of the performance analysis of the existing resource allocation techniques like bin-packing, ACO, and priority-based and the proposed technique

Table 2 Convergence time in seconds by the bin-packing, ACO, priority, proposed technique with increasing size of the data center for request size 150

Size of the data center	Convergence time in seconds			
	Bin-packing	ACO	Priority	Proposed technique
100	10	9	8	5
200	12	11	11	8
300	14	11	12	8
400	16	12	13	9
500	18	13	12	10
600	19	14	15	11
700	21	16	17	12
800	23	17	18	13
900	27	21	22	15
1000	28	22	26	17

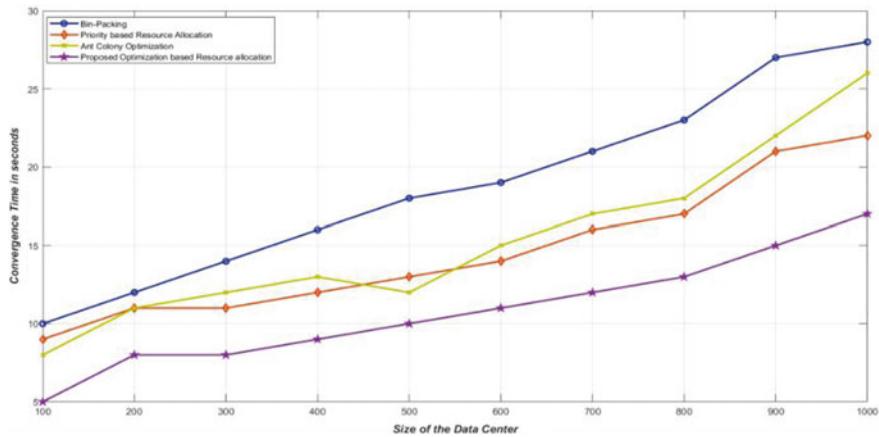


Fig. 1 Graphical representation of the convergence time in seconds by the bin-packing, ACO, priority-based and proposed optimization-based resource allocation with increasing size of the data center for resource request size of 150

of the convergence time in seconds with increasing number of data centers for the resource request size of 150. From Table 2 and Fig. 1, it is clear that the proposed technique performs in the least convergence time with the increasing size of the data center than the existing techniques for the resource request size of 150.

Table 3 depicts the convergence time in seconds by the bin-packing, ACO, priority-based resource allocation and proposed optimization-based resource allocation strategy for the resource request size is 300. Figure 2 represents the graphical representation of the performance analysis of the existing resource allocation techniques like bin-packing, ACO, and priority-based and the proposed technique of the

Table 3 Convergence time in seconds by the bin-packing, ACO, priority, proposed technique with increasing size of the data center for request size 300

Size of the data center	Convergence time in seconds			
	Bin-packing	ACO	Priority	Proposed technique
100	30	29	28	15
200	35	34	32	18
300	39	34	33	19
400	41	35	36	21
500	45	37	38	24
600	51	41	42	28
700	55	44	43	30
800	58	45	44	31
900	59	47	46	33
1000	61	49	49	36

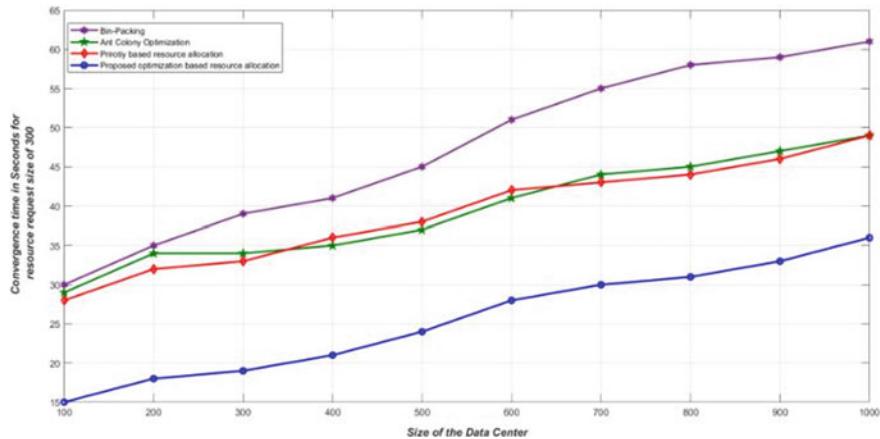


Fig. 2 Graphical representation of the convergence time in seconds by the bin-packing, ACO, priority-based and proposed optimization-based allocation with increasing size of the data center for resource request size of 300

convergence time in seconds with increasing number of data centers for the resource request size of 300. From Table 3 and Fig. 2, it is clear that the proposed technique performs in the least convergence time with the increasing size of the data center than the existing techniques for the resource request size of 300.

Table 4 depicts the convergence time in seconds by the bin-packing, ACO, priority-based resource allocation and proposed optimization-based resource allocation strategy for the resource request size is 450. Figure 3 represents the graphical representation of the performance analysis of the existing resource allocation techniques like bin-packing, ACO, and priority-based and the proposed technique of the

Table 4 Convergence time in seconds by the bin-packing, ACO, priority, proposed technique with increasing size of the data center for request size 450

Size of the data center	Convergence time in seconds			
	Bin-Packing	ACO	Priority	Proposed technique
100	76	61	63	41
200	79	63	62	42
300	85	65	69	47
400	88	67	70	49
500	91	72	73	51
600	94	73	74	52
700	97	74	74	53
800	99	75	76	55
900	102	78	79	57
1000	104	81	80	61

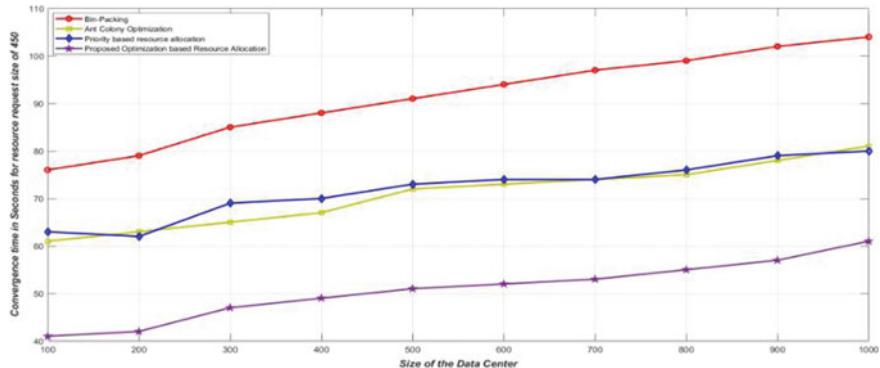


Fig. 3 Graphical representation of the convergence time in seconds by the bin-packing, ACO, priority-based and proposed optimization-based resource allocation with increasing size of the data center for resource request size of 450

convergence time in seconds with increasing number of data centers for the resource request size of 450. From Table 4 and Fig. 3, it is clear that the proposed technique performs in the least convergence time with the increasing size of the data center than the existing techniques for the resource request size of 450.

Table 5 depicts the convergence time in seconds by the bin-packing, ACO, priority-based resource allocation and proposed optimization-based resource allocation strategy for the resource request size is 600. Figure 4 represents the graphical representation of the performance analysis of the existing resource allocation techniques like bin-packing, ACO, and priority-based and the proposed technique of the convergence time in seconds with increasing number of data centers for the resource request size of 600. From Table 5 and Fig. 4, it is clear that the proposed technique

Table 5 Convergence time in seconds by the bin-packing, ACO, priority, proposed technique with increasing size of the data center for request size 600

Size of the data center	Convergence time in seconds			
	Bin-packing	ACO	Priority	Proposed technique
100	112	87	88	75
200	115	89	89	76
300	119	91	90	78
400	122	93	94	79
500	124	95	96	81
600	127	97	98	83
700	129	99	98	85
800	130	102	104	87
900	132	105	106	88
1000	133	107	108	90

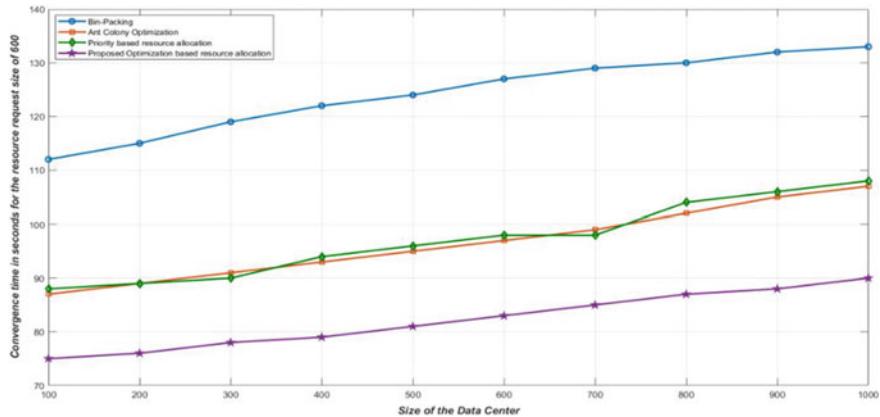


Fig. 4 Graphical representation of the convergence time in seconds by the bin-packing, ACO, priority-based and proposed optimization-based resource allocation with increasing size of the data center for resource request size of 600

performs in the least convergence time with the increasing size of the data center than the existing techniques for the resource request size of 600.

7 Conclusion

In this paper, we proposed an optimization-based resource allocation strategy for the energy productive resource allocation for the classical IaaS clouds. We formulate the problem of energy proficient resource allocation as the optimization model. This model is VM based and gives on-demand resource allocation. We proposed an algorithm for the overloaded host which is based on PSO. To deal with the dynamic resource consolidation, the proposed strategy for dynamic VM allocation is finished.

The proposed algorithm is compared with the current resource allocation approaches like ACO, priority-based resource allocation, and bin-packing with the increasing size of the data center and resource demand size.

References

1. Wang L, Tao J, Kunze M, Castellanos AC, Kramer D, Karl W (2008) High performance computing and communications. In: IEEE international conference HPCC, 2008, pp 825–830
2. Carroll M, Van Der Merwe A, Kotze P (2011) Secure cloud computing: benefits, risks and controls. In: 2011 information security for South Africa. IEEE
3. Valliyammai C, Mythreyi R (2019) A dynamic resource allocation strategy to minimize the operational cost in cloud. Emerging technologies in data mining and information security. Springer, Singapore, pp 309–317

4. Anithakumari S, Chandrasekaran K (2019) Adaptive resource allocation in interoperable cloud services. Advances in computer communication and computational sciences. Springer, Singapore, pp 229–240
5. Alahmadi AA et al (2019) Energy efficient resource allocation in vehicular cloud based architecture. arXiv preprint [arXiv:1904.12893](https://arxiv.org/abs/1904.12893)
6. Ficco M et al (2018) A coral-reefs and game theory-based approach for optimizing elastic cloud resource allocation. Future Gener Comput Syst 78:343–352
7. Meng S et al (2018) Joint optimization of wireless bandwidth and computing resource in cloudlet-based mobile cloud computing environment. Peer-to-Peer Netw Appl 11(3):462–472
8. Nayak SC et al (2018) Modeling of task scheduling algorithm using Petri-Net in cloud computing. Progress in advanced computing and intelligent engineering. Springer, Singapore, pp 633–643
9. Guerrero C, Lera I, Juiz C (2018) Migration-aware genetic optimization for MapReduce scheduling and replica placement in hadoop. J Grid Comput 16(2):265–284
10. Gokilavani M, Selvi S, Udhayakumar C (2013) A survey on resource allocation and task scheduling algorithms in cloud environment
11. Pradeep KR (2012) Resource scheduling in cloud using bee algorithm for heterogeneous environment
12. Lin W, Peng B, Liang C, Liu B (2013) Novel resource allocation model & algorithm for cloud computing
13. Ventresca M, Ombuki BM (2004) Ant colony optimization for job scheduling problem

Hardware Trojan Detection Using Deep Learning-Deep Stacked Auto Encoder



R. Vishnupriya and M. Nirmala Devi

Abstract Due to the outsourcing of integrated circuit manufacturing to third party vendors. The chances of malicious hardware insertion also got increased. Untrusted foundries can always add Hardware Trojan circuits, which can alter the behavior or working of the integrated circuits. Since new hardware threats are emerging day by day, a generalized solution for threat detection should be defined. Machine learning based threat detection which is widely popular nowadays, it requires handcrafted features. On the contrary, the Deep Learning (DL) class of ML can choose the relevant features and learn, which has been proposed in this paper. Raw circuit features are extracted and fed to the DL model (Deep Stacked Auto Encoder), which could extract features that can aid in Trojan detection. 95% average TPR and 75% average TNR is obtained and which is higher than previously discussed works.

Keywords Hardware Trojans · Deep learning · Deep stacked autoencoders · Features

1 Introduction

Outsourcing the fabrication to third party foundries due to an increase in manufacturing cost is the major trend seen in semi-conductor manufacturing companies [1]. These non-trusted foundries add HT circuits that are potentially harmful, hard to detect, or stealthy and can alter the circuit functionality. Since security is a critical issue, early threat detection is of utmost importance and the process of detection can be done at [2] design time, run time or test time. Of these, Machine learning based detection is widely popular, because of its effectiveness. In this, handcrafted features

R. Vishnupriya (✉) · M. Nirmala Devi

Department of Electronics and Communication Engineering, Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham, Coimbatore, India

e-mail: vishnu.priya735@gmail.com

M. Nirmala Devi

e-mail: m_nirmala@cb.amrita.edu

are given as the input to the algorithm, which is used to classify the affected and unaffected circuit. Most of the detection methods are Trojan nature dependent. The Trojan dependent features can vary according to the nature of them and generalized detection methods should be proposed, which is applicable for all kinds of malicious insertions that can be injected into a digital circuit. Recently, Deep Learning (DL) has become popular due to its generalization capability, it is the advanced model of neural network. [3] deals with raw and unstructured data and decisions are taken by analyzing the changes that occur in the parameters [4]. One of the main uses of DL is, it is widely used in feature selection for classification based problems. It produces superior results in many unsolved problems, which makes it more attractive these days. Here in this proposed work, generalized circuit features (net-based features) which are completely Trojan independent, are extracted using the synthesis tools. These are given to DL algorithm—DSAE, which analyses the parametric changes in Trojan affected and un-affected nets and selects the best features which show great variance in their value. Final decisions are taken by the softmax layer, which is attached at the end of the model. It outputs the probability corresponding to each label, which can be analyzed to detect the presence of Trojan. Most of the recent Trojans are moreover untraceable, guessing their behavior and applying detection methods accordingly is impractical and time-consuming, in such cases a generalized methodology which does not require much knowledge on Trojan will make detection easy and reliable. Paper gives an overview of existing Trojan detection schemes and deep learning based detection.

2 Overview

Hardware security means the protection against the malicious insertions (HTs) that can occur in a circuit. HTs can be added at any phase of IC design flow. They can be classified according to their characteristics like physical, action, and activation. Detection techniques can vary according to their characteristics and behavior [5]. Functional testing, side-channel parameter analysis, destructive approaches are some of the common detection schemes available. In side-channel analysis [6, 7], parameters like power, electromagnetic radiation, noise, leakage current, etc. are analyzed to detect the Trojan existence [8, 9]. In testing-based approach, automatic test patterns are applied to excite the Trojan nets and thus by increasing their visibility, which increases their detection. All the methods described exhibit a dependency on the parameters chosen for the detection. It constraints the system and hence there is no generalized solution available for different types of Trojans. Hence a procedure that handles similar HTs is developed using a machine learning algorithm [10]. ML is a field in which the model learns from its previous experience and do further computations, it is a supervised learning approach which requires learning examples and handcrafted features. It is used in clustering, anomaly detection, association mining, and dimensionality reduction [11]. Paper describes trojan detection by comparing the DFT power waveform of affected and non affected circuit and classification of DFT

waveform using SVM classifier [12–14], SVM, Random forest, neural networks like ML models were used in HT detection, these work uses handpicked trojan net features like fan-in, fan-out, number of gates, number of flipflops away from certain levels, etc. [15] paper describes the trojan detection in IoT chip using machine learning techniques, feature extraction of the net is carried out and a scoring mechanism, XGBoost is utilized in discarding the irrelevant feature.

Reshma et al. [16] discuss deep learning based Trojan detection which uses autoencoder based neural network for feature reduction and k means clustering for classification, with hand-picked features like controllability and transition probability, etc. In machine learning based trojan detection, the feature set should be manually selected according to the prior circuit knowledge, ML algorithms cannot handle more number of features or examples. Though it has more efficiency and convenience than other detection schemes we need a more generalized solution for the trojan detection problem. So that we move towards deep learning. This paper suggests a unique data extraction methodology in circuit domain. And also it exploits the benefits of Deep learning model in finding the malicious insertions in integrated circuits using the extracted parameters.

3 Methodology

A circuit independent and generalized hardware Trojan detection methodology, by exploiting all the benefits of deep learning is proposed.

Gate-level net-list is obtained from the design files and are given to the synthesis tools for the feature extraction. There are several circuit parameters which are common for all circuit and it can be either parameter related to overall circuit or parameters related to cell or parameters related to the net. In this work, raw and unstructured net features from the synthesis tools are directly given to the deep learning algorithm after pre-processing. The Deep stacked autoencoder selects the best possible features which are enough to classify Trojan affected and Trojan free circuits via the nets. Classification is done using the softmax layer attached at the end of the network. A threshold probability will be set the label which shows the higher probability is predicted.

3.1 Trojan Insertion

A Trojan circuit basically consist of trigger and payload [17], these malicious circuits are inserted into the rare nodes where the probability of detection is very less. The trigger circuit can be combinational or sequential. In this experiment we have used T000, T001, T002 combinational trojan circuits which can alter the internal signal values also T1000, T1100, T1200, Trojans inserted in RS232 circuit, In this, the trigger circuit is sequential comparator. These Trojans are internally conditionally

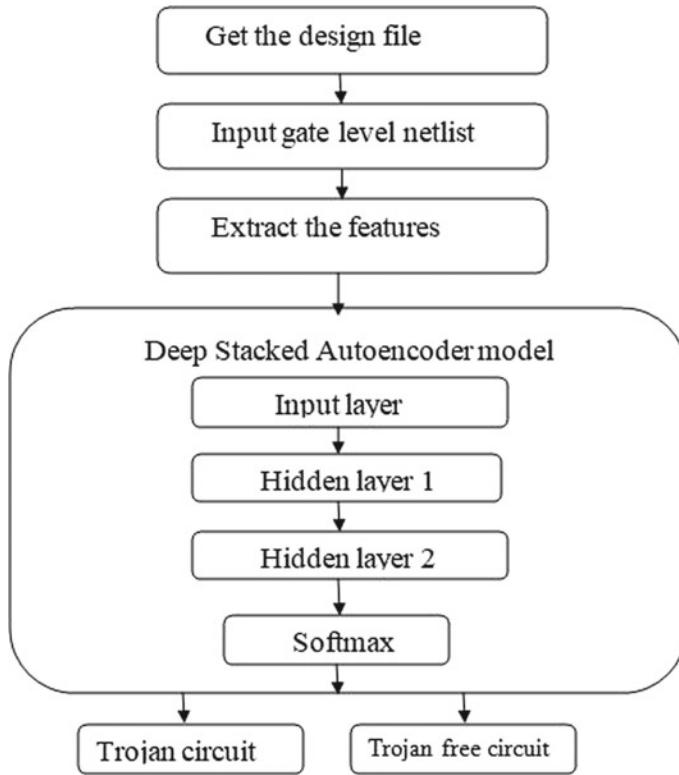


Fig. 1 Flow chart for Trojan detection

triggered. Most of the Trojan circuits inserted will be stealthy and hard to detect. Figure 1 describes the flow of the work.

3.2 Circuit Feature Extraction

In the process of fabrication, RTL file formation is the foremost step. The gate-level netlist is evolved from the RTL file using the synthesis tools like the design compiler. Net-list is the representation of the design file in terms of gates and nets. These files are dumped to the synthesis tools like Synopsys design compiler and ic compiler for the feature extraction. Here the experiments are done using combinational circuits and RS232 circuits. The basic circuit features like fan-in, fan-out, load values, resistance, capacitance, switching power, etc. like 16 features are obtained from the tools. The key factor in our paper is, it considers all generic features which could get altered on Trojan insertion, these features are independent of one another. Dataset prepared contains the parameters associated with the circuit nets which are obtained from

the synthesis tools. The Trojan affected nets are labeled as ‘1’ and unaffected nets are labeled as ‘0’. The overall dataset is splitted in 80:20 ratio. DL training always require labeled data and testing requires unlabelled data.

3.3 Deep Learning

DL is a subsection of ML which is ruled by a set of algorithms that mimic human brain activities [18], these algorithms train the layers involved in the DL which will enable them to take decisions in critical problems from the training examples. Auto-encoders are neural networks, which are used in representing the higher dimension input in a lower dimension. They contain an input layer, intermediate layer, and an output layer. The network includes an encoding part and decoding part (reconstructing part). Let X be the input ‘ H ’ be the hidden layer, then encoded input can be represented as $H = F(X)$, let the reconstructed output be Y , it can be represented as $Y = F(H)$. The upcoming encoders [19] make use of the probabilistic equations for encoding and decoding which reduces the error and noise during the process. Auto-encoders when stacked together form deep-stacked autoencoders which are the deep version of encoders.

3.4 Deep Stacked Auto Encoders

They are deep neural networks formed by [20] stacking multiple autoencoders, for better classification accuracy and feature reduction. Output of each auto encoder is connected to the input layer of the successive layer. Here the encoding is done between the output layer of the previous auto encoder and input of successive auto encoder, the information is contained in the deepest encoder layer. The decoding step is provided by running each autoencoder’s decoding stack in reverse order. The encoder layer gives us input representation in terms of higher-order characteristics. A deep network is trained in different steps, like pretraining (which can also be called as a feature learning step) [21], fine-tuning (stacking the decoding layer with encoder layer). These steps increase the potency of the network. Most of the DL networks use softmax as a classification layer, here it classifies based on the predefined threshold which is set.

4 Model Description

Deep stacked autoencoder used in this work consists of five layers, in which the first layer consists of 16 nodes, which represent 16 features. The hidden layer node numbers can be fixed by hyperparameter tuning, the node numbers which give minimum loss and maximum accuracy are considered. Here 16 features are broken

down to 10, which means the 16 features can now be represented by 10 features, at the end a softmax layer is attached for classification, which classifies the Trojan affected and unaffected nets according to probability. During the hyper-parameter tuning activation function, loss function, optimizers, node numbers, etc. are selected according to minimum loss and maximum accuracy (Figs. 2 and 3).

- Taking several auto-encoder layers
- Training the encoders by reducing reconstruction error.
- Output of previous AE is given as input to next AE.
- Repeat the steps for remaining AE layers.
- Save the obtained weights.
- Add the decoder layers.
- Compute loss using Back propagation, by making use of saved weights.
- Repeat the procedure till minimum loss is obtained.
- Adding softmax layer at the output, for the classification.

Pre-training

Fine-tuning

Classification

Fig. 2 DL training

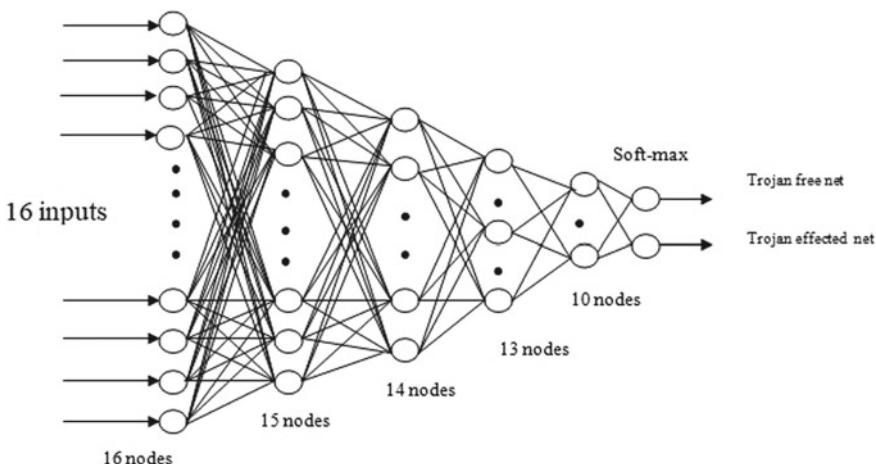


Fig. 3 Deep stacked auto encoder for Trojan detection

Table 1 Analysis of ISCAS'85 Benchmark circuit *TPR TNR in %

Circuit	TPR	TNR	TP	FP	FN	TN
C432-T0	100	64	138	0	6	11
C2670-T0	99	58	913	7	7	10
C2670-T1	98	54	909	11	11	13
C6288-T0	99	53	2446	3	8	9
C6288-T2	99	59	2444	5	7	10

Table 2 Analysis of TRUST-HUB circuits *TPR TNR in %

Circuit	TPR [14]	TPR ours	TNR [14]	TNR ours
RS232-T1000	100	94	24	64
RS232-T1100	78	97	25	60
RS232-T1200	91	99	55	92
RS232-T1300	86	99	65	71
RS232-T1400	100	98	15	94

Table 3 Analysis of TRUST-HUB circuit %

Circuit	TP	FP	FN	TN
RS232-T1000	252	14	16	28
RS232-T1100	265	7	17	25
RS232-T1200	271	1	3	39
RS232-T1300	275	2	9	22
RS232-T1400	258	4	3	45

5 Result

The experiment was conducted on ISCAS' 85 benchmark circuit and RS232 circuit. Hasegawa et al. [14] evaluate the performance on the basis of TPR and TNR hence these two parameters are analyzed. Average TPR of 95% and TNR of 75% is obtained, means Deep stacked autoencoder shows higher classification (Tables 1, 2, and 3).

6 Conclusion

Paper describes a hardware Trojan detection method using deep-stacked autoencoders. The results obtained show that this method shows better TNR of 95% and TPR of 75% when compared to existing methods. The problem is attempted on

combinational circuits and RS232 Trojan circuits the same can be extended to sequential circuits. Our future work will be concentrated on improving the TNR and thus reducing the miss classification. Outcomes which describes the benefits of deep learning over machine learning will be analyzed and discussed in our future work.

References

1. Garg S (2017) Inspiring trust in outsourced integrated circuit fabrication. In: Proceedings of the conference on design, automation test in Europe. European Design and Automation Association
2. Francq J, Frick F (2015) Introduction to hardware Trojan detection methods. In: Proceedings of the 2015 design, automation test in Europe conference exhibition. EDA consortium
3. Fan J, Ma C, Zhong Y (2019) A selective overview of deep learning. arXiv preprint [arXiv:1904.05526](https://arxiv.org/abs/1904.05526)
4. Charmisha KS, Sowmya V, Soman KP (2018) Dimensionally reduced features for hyperspectral image classification using deep learning. In: International conference on communications and cyber physical engineering 2018. Springer, Singapore
5. Rostami M, Koushanfar F, Karri R (2014) A primer on hardware security: models, methods, and metrics. Proc IEEE 102(8):1283–1295
6. Agrawal D et al (2007) Trojan detection using IC fingerprinting. In: 2007 IEEE symposium on security and privacy (SP'07). IEEE
7. Rosenfeld K (2011) Hardware Trojan detection solutions and design-for-trust challenges
8. Chakraborty RS et al (2009) MERO: a statistical approach for hardware Trojan detection. In: International workshop on cryptographic hardware and embedded systems. Springer, Berlin, Heidelberg
9. Nourian MA, Fazeli M, Hely D (2018) Hardware Trojan detection using an advised genetic algorithm based logic testing. J Electron Test 34(4):461–470
10. Elnaggar R, Chakrabarty K (2018) Machine learning for hardware security: opportunities and risks. J Electron Test 34(2):183–201
11. Iwase T et al (2015) Detection technique for hardware Trojans using machine learning in frequency domain. In: 2015 IEEE 4th global conference on consumer electronics (GCCE). IEEE
12. Inoue T et al (2017) Designing hardware trojans and their detection based on a SVM-based approach. In: 2017 IEEE 12th international conference on ASIC (ASICON). IEEE
13. Hasegawa K, Yanagisawa M, Togawa N (2017) Trojan-feature extraction at gate-level netlists and its application to hardware-Trojan detection using random forest classifier. In: 2017 IEEE international symposium on circuits and systems (ISCAS). IEEE
14. Hasegawa K, Yanagisawa M, Togawa N (2017) Hardware Trojans classification for gate-level netlists using multi-layer neural networks. In: 2017 IEEE 23rd international symposium on on-line testing and robust system design (IOLTS). IEEE
15. Dong C et al (2019) A machine-learning-based hardware-Trojan detection approach for chips in the Internet of Things. Int J Distrib Sens Netw 15(12). <https://doi.org/10.1177/1550147719888098>
16. Reshma K, Priyatharishini M, Nirmala Devi M (2019) Hardware Trojan detection using deep learning technique. Soft computing and signal processing. Springer, Singapore, pp 671–680
17. Salmani H, Tehranipoor M, Karri R (2013) On design vulnerability analysis and trust benchmarks development. In: 2013 IEEE 31st international conference on computer design (ICCD). IEEE
18. Chandra B, Sharma RK (2015) Exploring autoencoders for unsupervised feature selection. In: 2015 international joint conference on neural networks (IJCNN). IEEE
19. Sadati N et al (2019) Representation learning with autoencoders for electronic health records: a comparative study. arXiv preprint [arXiv:1908.09174](https://arxiv.org/abs/1908.09174)

20. Vincent P et al (2010) Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. *J Mach Learn Res* 11:3371–3408
21. Kading C et al (2016) Fine-tuning deep neural networks in continuous learning scenarios. In: Asian conference on computer vision. Springer, Cham

IOT and Intelligent Asthmatics Monitoring Sensors—A Literature Survey



Aditya Bothra, Saumya Bansal, and Surender Dhiman

Abstract Internet of things is taking the technology world by storm and healthcare was the latest one to join. Engineers and Medicine have had very close relationship in determining the response time in the field of Health Care. With technological advancements, response time has been reduced and success of treatments has increased. But there still remains a problem in the accuracy of detection of symptoms that can be concrete, and enough promptness of those readings to assist physicians in determining the remedy. One such prevalent and common disease in Indian society and the world as a whole is Asthma. This paper deals with the factors responsible in triggering Asthmatic symptoms and Sensors that are used to measure them as part of Smart Asthma monitoring. It also highlights the common IOT domain electronic sensors pertaining to this field of healthcare and factors that these sensors are capable of recording. Though better sensors that are of commercial use are available but our motive is to compile the sensors used as a part of IOT projects throughout the world. Our aim, ultimately, is to determine the relevance of all these sensors combined, for this field of Medical science, in improving the condition of such patients.

Keywords Smart asthmatics system · Ozone sensor · Sensor drone · Spirometry · Pulse-oximeter · Peak flow meter · Allergens · Propeller

A. Bothra (✉) · S. Bansal · S. Dhiman
DADGITM, affiliated GGSIPU, Delhi, India
e-mail: adityabothra@gmail.com

S. Bansal
e-mail: sauumzz19@gmail.com

S. Dhiman
e-mail: surender.dhiman@adgitmdelhi.ac.in; surender.dhiman@gmail.com

1 Introduction

In November of 2019 Delhi NCR regions of India, clinched second spot in the world's worst polluted cities and the cities with air quality index [1] in poorest category while Kolkata was ranked fourteenth. AQI for Delhi NCR as worse as 202 which is constantly inquisitive to when will pollution checks will be implemented and air quality improvement measures be implemented. Such deteriorated levels are hazardous to everyone from an infant to an old age person, from the healthy to the diseased. One particular chronic respiratory disease, which is specifically prone to pollution and deteriorated Air Quality Index is asthma. According to an estimate 15–20 million people in India suffer from Asthma [2]. 90% of childhood asthma and 50% of adult asthma is caused due to environmental allergens and other pollutants. We hence make an effort to survey symptoms, prediction, prevention and detection of Asthma using the new age technology of IOT.

Asthma is a condition in which a person's airways become inflamed, narrow and swollen producing extra mucus, which makes it hard to breathe [3]. According to The Global Asthma Report 2014, more than 334 million people are suffering from this disease globally [4]. Asthma or an asthmatic attack is a major discomfort caused by air pollution that can originate from diverse sources—some are human-made while others are naturally occurring. Air pollutants include gaseous emissions from industries, smoke from fires, volcanic ash, fuel combustions and dust particles. Research exhibits that pollution arising in air can lead to deterioration in asthma symptoms. A study on young group of people with mild to severe Asthma showed, they were 40% more likely to have acute asthmatic tendencies on highly polluted summer noon than on days with mediocre pollution level. Another study among comparatively older adults, showed, they had more chances of visiting the emergency room for breathlessness related problems when summer aerial pollutant concentration was higher. This suggests the difference in severity due to age group but even then, it is dangerous for both the sections and is even more hazardous for infants and children with such problems.

For ages asthma and asthmatic attacks have been monitored less vigilantly and thus rendering more such people vulnerable to hardships. Lack of proper diagnosis, reaction time, expensive routine consultancy and more factors have induced these problems. With the dawn of IOT, more and more people have started noticing the plight of such folks and dived into study to develop practices to get the prevention, monitoring, remedy and consultation of physicians at every patient's doorstep. To sense Asthma and related problems, we first need to look out for factors and how can those factors be measured. Though, many have worked upon this domain and have highly scholarly experimental and theoretical results to show, with this paper we aim at incorporating a review of all such factors and their measuring sensor, devices or systems in one place.

Out of all the macro and micro factors here are the following key factors affecting Asthma patients:

- (1) **Gases/Pollutants:** Among the poisonous gases and pollutants present in air, the most significant particles causing havoc in Asthma condition are:
- (1.1) **Carbon Dioxide**—If the concentration of CO₂ is anyway above 380 ppm that area is dangerous for asthmatic patients and ventilation or fresh air is required. Apart from causing affects like heavy breathing, itching in nasal tract etc. research suggests that carbon dioxide emissions from automobiles and factories are causing flora and molds in the region to elevate pollen and spore production. More pollen aerial concentration is likely to increase trouble in cases of allergic diseases such as asthma [5].
 - (1.2) **Hydrocarbons**—If the concentrations of hydrocarbons like carbon monoxide reaches a level greater than 35 ppm, it can be attention worthy [6]. Exposure to particulate matter less than 2.5 μm (PM2.5), a common air pollutant, has been linked to asthma exacerbations, the development of new wheeze, and decreased lung function. More recently, exposure to individual chemical components of PM2.5 has been associated with the development of asthma in young children, including the diesel soot fraction of PM2.5 (black carbon or elemental carbon), transition metals (e.g., nickel and vanadium),^{9,10} and polycyclic aromatic hydrocarbons (PAHs) [7].
 - (1.3) **Ozone**—Epidemiologic studies suggest an extremely high correlation between ozone exposure and the condition of an asthma patient's lung function. Ozone exposure can cause itching in wind tract and also cause a burning sensation. When ozone levels are high, people with asthma tend to experience: Lung function decrements, Increased respiratory symptoms like shortness of breath, increased medication usage than regular need [8].
 - (1.4) **Dust particles/allergen**—If dust particles/allergen concentration is above 0.1 ppm, the area needs attention. Allergens and dry air are widely accepted as major stimulating agents in case of Asthma [2]. Dry air and lead to nasal tract to get swollen thus increasing mucus production and in all trigger asthmatic symptoms. According to some serial analysis of reports of 63,000 patients between the years 2013 to 2017, allergens like diverse types of dust particles propagated by cockroaches have come out to be one of the major factor in 60 percent of the cases.
- (2) **Temperature and Humidity:** Heat and humidity are usually a combined occurring process. So when there is difficulty in breathing air, your body temperature can rise, leading to extra sweat. This in consequence causes dehydration, which causes faster breathing. All of these factors in combination can trigger fiery asthmatic symptoms [9]. When the temperature goes down, your asthma symptoms may experience a worse. Chilly winds can dry out the tissues in your airways, increasing their sensitivity and likely leading them to close up [3].
- (3) **Lack of Oxygen/Oxygen saturation levels in body:** Although there exists no evidence to support the use of oxygen in acute asthma, but, it is known to be

Table 1 Factors and sensors used

Factor to be measured	Sensors/combination used to achieve the goal
Lack of oxygen	Pulse-oximeter
Over exertion/extraneous activity	ECG sensor, blood pressure sensor, Gyro sensor, accelerometer, GPS
Ozone	Ozone sensor
Pollutants, gaseous concentration level, dust particles and allergens	Sensor drone (consisting of various sensors for environmental sensing)
Temperature and vapor concentration (external)	Temperature and humidity sensor

effective and should be administered when oxygen saturation levels fall below 94% in all cases of acute asthma [10]. An isolated pulse oximetry reading at triage is unpredictable in most cases (with some noteworthy exception of severe attacks that, often times, are self-evident on visual inspection), vigilant monitoring of pulse oximetry status can provide subtler evidence for or against the need for hospital admission [11]. Measuring SpO₂, or saturation of oxygen levels gives the values of oxygenation of blood and thus determining if there is any link in between lack of energy synthesis and shortness of breath.

- (4) **Extraneous Activity:** Exercised-induced asthma is a narrowing of the air flow paths in the lungs initiated by strenuous exercise. This leads to breathlessness, cough, wheezing and other symptoms during or after exercise. The preferred term for this condition is exercise-induced bronchoconstriction (brong-koh-kun-STRIK-shun) [12]. Theforesaid term is more precise because the exercise is the contributing factor in narrowing of airways (bronchoconstriction) but this isn't the basic cause of asthma. Among people with asthma, exercise is likely one of the several factors that may trigger breathing difficulties, decrease in lung capacity and getting restless which all combined is not suitable for Asthma patients.

For each of the factors mentioned above, our goal is to detect them via a sensor. There are a lot of possibilities of using “N” number of combinations of different sensors to achieve the same goal, but this paper deals with fairly popular and the coherent ones considering cohesiveness of the platform used. Following is a tabular representation of the factors to be measured and some of the sensors utilized to accomplish the objective (Table 1).

2 Related Work

The concept of wireless sensing as intelligent asthmatics is advantageous on grounds of accuracy, speed, power consumption, manual effort, promptness and decision making by incorporation of sensors and devices as a system on whole. Routine

monitoring and regular visits to the hospital is not affordable to everyone and also sometimes traveling to the hospital is a bigger problem than economic reasons. If a system/device that can give you prior indication of your condition and thus, enabling you to make a choice to commute or treat it at home using medications mentioned. Considering one such hypothetical system that will have the capacity of measuring all such factors together the description following this, will deal with all such sensors needed, brief information regarding sensors and their applications.

2.1 Pulse-Oximeter

Aileni and Pașca [13] Measuring saturation of peripheral oxygen, or SpO₂, is a method of finding hemoglobin saturation which can be measured non-invasively, for example, with an electronic clip device on the finger or ear. The sensor more than often employs a pair of small LEDs facing a photodiode that measures the amount of light passing through the skin. The sensor is then put around your finger or earlobe. TI AFE4490 Pulse Oximetry sensor is one of the leading sensors in this criteria. Pulse oximetry is advantageous in a set up where oxygenation of the patient shows unstable values, including ICU (Intensive care Unit), operation procedure, healing and recovery, emergency rooms and hospital wards, unpressurized aircraft (pilot driven), for evaluation of oxygenation of any random patient, and drawing out the effectiveness of or requirement for any supplemental oxygen. Although a pulse oximeter is used to monitor oxygenation levels, it is incapable of determining oxygen metabolism, or the amount of oxygen being consumed by a patient. Hence, for this purpose, it is required to also gauge carbon dioxide (CO₂) levels. Another possibility is that it can also be employed in detection of anomalies in ventilation. However, on similar grounds employment of pulse oximeter in detection of hypoventilation is impaired with the use of similar supplemental oxygen. This is based on the finding that only when the person breathes room air, that, anomalies in respiration functionality can be identified reliably. The biggest utilization of this pulse oximeter is its help in determining sleep deprivedness, insufficiency of energy and thus measuring oxygen levels to determine symptoms of acute Asthma. In combination with gyro sensors and accelerometer it can measure patient's activity and can help determine accurately the change in BP and Heartbeat if is due to over exertion or sudden induced mental stress.

2.2 ECG Sensor

ECG [14] (electrocardiography) is a method of collecting electrical signals generated by the heart. This enables our understanding of the levels of physiological arousal that someone is going through, but it can also be deployed to effectively understand somebody's psychological condition. ECG takes into account the electrical activity

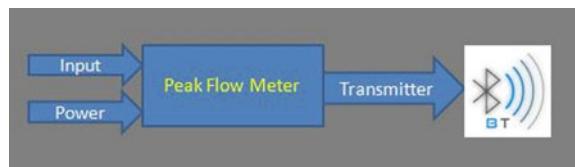
generated by heart muscle's depolarization and consequent repolarization and termed it as one cardiac cycle, which travels in pulsating electrical waves in the direction of the skin. A graph of voltage versus time—gathered by recording the electrical activity of the heart is displayed or produced on a screen or as a printable sheet. ECG uses electrodes, which are typically wet sensors, needing the use of a conductive gel like substance to increase conductivity between skin and these electrodes. ECG's are very noisy which is based on the reason that we are measuring muscular activation. The farther these sensor pads are placed from the heart; the more muscle noise you will observe often termed as "Motion Artifacts." AD8232, TI ADS1292R 24-bit sensors are the most common and cost efficient sensors in IOT-based health monitoring systems. These are good for estimative indication over a range but as Asthma is a critical deal, an improvement of these sensors can be used to obtain better results. Also ECG sensor is not solely capable of making observations to indicate/identify any parameters of Asthma but on combination with few other sensors it can suggest striking results.

2.3 Blood Pressure Sensor

Vernier Software and Technology [15] the Blood Pressure Sensor is a non-invasive sensor whose purpose is to measure human blood pressure. It measures systolic, diastolic and mean arterial pressure utilizing the oscillometric technique [16]. The oscillo-metric technique has been used with a very high rate of success in ambulatory B.P. monitors and household monitors. Pulse rate is also reported along with BP. Now this pulse rate reporting feature is an advantage here and helps to reduce a pulse sensor from the sensory network. A transducer AG3-03 assists in the process of measuring Blood pressure. The Transducer converts the pulsatile signal into a mechanical pressure signal. This pressure sensor comprises of a silicon piezo-resistive pressure sensing chip and an Integrated circuit for signal conditioning [17]. Blood Pressure sensor is crucial to determination of the amount of strain/stress on arteries and heart pumps during any arousal of asthmatics symptoms in combination with human's activity detection using Gyro and Accelerometer sensors. The main con of this sensor is that such recorders do not operate well during physical activities when there are chances of considerable movement artifacts. Blood pressure sensors have been fairly readily available with some variants now being developed to be clubbed with Arduino/Pi board etc. for IOT research.

2.4 Spirometry and PEF/PFM

[3] Spirometry, an effort-based procedure to diagnose asthma, has been in use for a long time to diagnose asthma. This method uses the Peak Flow Meter, also known as peak expiratory flow (PEF), which gauges the maximum level of a forced exhalation

Fig. 1 IOT-based PFM [18]

air flow rate parameter. This device is electro-mechanical based in operating and is used for the diagnosis of the asthma level of the person under observation; it requires a cautious handling to ensure that the patient level of asthma is accurately measured [18]. In such patients, the PEF percentage predicted aligns well with the percentage predicted value for the forced expiratory volume per second (FEV1) and also provides a measure of airflow limitation when spirometry is not available. These are not just sensors but are sensor systems. Nowadays, PFM has incorporated a part of IOT progress and we have thus started to fetch its data from patient and using that data by transmitting it via Bluetooth (as shown) or Internet and thus, making the data available to the Physician. Figure 1 illustrates this approach of spirometer these days.

2.5 Ozone Sensor

Ozone sensors are typically based upon either electrochemical or metal-oxide sensing procedures [19]. Consuming relatively Less power and exhibiting relatively high specificity is the biggest advantage of electrochemical sensors. One of the widely cited ozone sensor is MICS2614, SGX SensorTech (IS) Ltd., Essex, UK. Another ozone sensor readily available for prototyping using Arduino platform is MQ131. The MQ131 ozone gas sensor is capable of sensing ozone air with O_3 concentration levels between 10 parts per million(ppm) and 1000 ppm. An internal preheater inside the sensor helps in achieving the ideal sensing conditions of $20\text{ }^{\circ}\text{C} \pm 2\text{ }^{\circ}\text{C}$ at $65\% \pm 5\%$ humidity, but the manufacture provided datasheet recommends over 24 h (1 + day) for preheating to achieve optimum accuracy. Another sensor MICS2614 has a sufficiently quick response time of approximately 30 s, but similar to MQ131, there is a relatively longer duration of time required for the initial preheat of the sensor, once switched on. Its complete use is given in the works of Low power wearable system's ambient sensing [20]. So ozone sensors since are as important to Asthma detection as are any VOCs, such sensors which can achieve optimal results in lesser preheat period can be hypothesized.

2.6 Sensor Drone

Sensor drone is a generic term for sensor network incorporated on a drone. In this context it is an Unmanned autonomous vehicle (UAV) having the capacity to sense and record several environmental factors simultaneously. It can consist of a combination of sensors which to this domain should consist of sensors such as DHT22 (temperature & humidity sensor), LM393 (carbon dioxide sensor), UL-2034 (carbon monoxide sensor), PM2.5 GP2Y1010AU0F (Dust Smoke Particle sensor) etc. These sensors have individual existence as well but to compact the circuitry a single mobile sensory system such as a UAV (Drone) can prove advantageous in collecting data remotely. All such data is collected and retrieved in a database or desired report format. If utilized based upon correct algorithm as these factors affect the asthmatic condition, one can actually derive striking correlation in between these sensor node data. This data can thus serve to determine the conditions of the surroundings around a patient and thus assist in preventing such triggers. Drones, UAV, as popularly known is widely used throughout the globe for different factors from surveillance to leisure. Ours reference might be a new application of drone in conducting regular environmental survey but there are no claims to it. Also such mobile drones can be programmed to periodic data collection and thus help achieve better learning of information, by the system, regarding the environmental conditions.

Presently, serial monitoring device and systems for such patients include machines that look over heart rate and breathing rate. With the GPS incorporation in such system, the patient's exposure to such hazardous/polluted areas can be lowered. Some consumer products in beginning stages for tracking physiological parameters are available but none of these incorporates the environmental and physiological sensors, specifically to track the impact of ozone and volatile organic compounds on health. A system with such assembled parts or made by combination of the above solutions listed, becomes bulky, difficult to synchronize, highly power consuming, and hard to use along with daily routine hence increasing the rejection chances by the consumer. Moreover, the inconvenience caused due to the need of frequent charging and thus requiring high power sources suggests the need for a system operating on lower power which in practice, incorporates energy synthesis methods in order to make these wearable devices a part of our/patient's daily routines.

Some of the earlier works with similar motives to predict and prevent such asthma related symptoms/attacks are there in planning stage either as a publication or prototypical but not many are of commercial availability or catering to direct patients. Also, few of these systems need un-automated/manual observations and recordings along with regularized visits to the nearest hospitals for routine checkups by the specialist doctor. For reference a device called Propeller is an application that allows patients and doctors to track efficiently the time and place patients use their inhalers [21]. Propeller is a smart system that learns about user's breathing pattern over a brief period of time. The mobile application learns and identifies about your irregular flare-ups and routine medication usage by just connecting your inhalers and other devices via app and can assist the patient in managing their symptoms and

detecting their triggers like an expert. Another device SAM [6] reminds asthma sufferers to take regular medications and informs asthma triggers such as dangerous pollen levels depending upon the locations etc. These devices fall in the category of remedy/reaction and prevention and are on pace to become possible systems for Asthma prevention.

3 Conclusion

In this paper, we have summarized all factors that affect the condition of patient suffering from Asthma and their measuring sensors with mentions of the most popular in IOT prototyping domain. Through this paper I believe the consensus of a system considering all elements yet being compact can be planned and implemented. We invite people to develop systems and make those resources available to help the mankind as a whole. Advancement in the field of wearable sensors that can have efficiency and consume lesser power can help build a system that can promptly detect Asthma and its related symptoms.

Acknowledgements We would like to thank the researchers as well as the publishers for making their resources available for study and our teachers for their valuable guidance. Authors are also thankful to the reviewer and mentor for their valuable time and informative suggestions. We would also like to thank the authorities of our institution ADGITM for providing us with the required infrastructure and aid. We would also like to thank our family and friends for the moral support needed during the course of research.

References

1. IQAir (2019) Air quality index US AQI. Retrieved from <https://www.airvisual.com/world-air-quality-ranking> on date: 12 Oct 2019
2. DTE Staff-Asthma affected population in India. Retrieved from <https://www.downtoearth.org.in/news/health/one-in-10-asthma-patients-in-the-world-are-in-india-60376> on date 12 Oct 2019
3. Google,webmd.comm() - Internet search engine/websit references
4. The Global Asthma Network (2014) The Global Asthma Report. Auckland, New Zealand
5. Zwillich T (2019) High CO₂ levels may up the rate of asthma. Retrieved from <https://www.webmd.com/asthma/news/20040429/high-carbon-dioxide-levels-may-up-asthma-rate#1> on date 12 Oct 2019
6. Isaac N, Sampath N, Gay V SAM smart asthma monitoring: focus on air quality data and Internet of Things (IoT). University of Technology Sydney
7. Jung KH, Yan B, Moors K, Chillrud SN, Perzanowski MS, Perera FP, Miller RL Repeated exposure to polycyclic aromatic hydrocarbons and asthma: effect of seroatopy
8. U.S. EPA (2019) Ozone pollution and your patients health. Retrieved from <https://www.epa.gov/ozone-pollution-and-your-patients-health/health-effects-ozone-patients-asthma-and-other-chronic> on date 12 Oct 2019
9. AAFA community services (2019) Affect of humidity on asthma. Retrieved from <https://community.aafa.org/blog/3-ways-humidity-affects-asthma> on date 12/10/2019

10. Foster S, Chavasse R, Paton JY, Wilson M Acute asthma and other recurrent wheezing disorders in children-A Publication of BMJ Publishing Group
11. Morris MJ (2019) Pulse oximetry in determining severity of acute asthma. Retrieved from <https://www.medscape.com/answers/296301-7995/how-is-pulse-oximetry-used-to-determine-the-severity-of-acute-asthma-in-children> on date 12 Oct 2019
12. Mayo clinic staff (2019) Bronchoconstriction and exercising as a potential danger for asthma patients. Retrieved from <https://www.mayoclinic.org/diseases-conditions/exercise-induced-asthma/symptoms-causes/syc-20372300> on date 12 Oct 2019
13. Aileni RM, Paşa S (2019) E-health monitoring by smart pulse oximeter systems integrated in SDU. In: The 11th international symposium on advanced topics in electrical engineering. Bucharest, Romania
14. Farnsworth B (2019) What is ECG. Retrieved from <https://imotions.com/blog/what-is-ecg/> on date 12 Oct 2019
15. Vernier Software and Technology (2019) Blood pressure sensor- definition, specs. Retrieved from <https://www.vernier.com/products/sensors/blood-pressure-sensors/bps-bta/> on date 12 Oct 2019
16. Ball-llovera A, Del Rey R, Ruso R, Ramos J, Batista O, Niubo I (2003) An experience in implementing the oscillometric algorithm for the noninvasive determination of human blood pressure. In IEEE conference 2003
17. Naveen, Sharma RK, Nair AR (2019) IoT-based Secure healthcare monitoring system. NIT, Kurukshetra
18. Kassem A, Hamad M, El-Moucary C, Neghawi E, Bou Jaoude G, Merhej C (2013) Asthma Care Apps. In: 2013 2nd international conference on advances in biomedical engineering. Notre Dame University-Louaize, Lebanon
19. Arshak K, Moore E, Lyons GM, Harris J, Clifford S Gas sensors employed in electronic nose applications
20. Dieffenderfer J, Goodell H, Mills S, McKnight M, Yao S, Lin F, Beppler E, Bent B, Lee B, Misra V, Zhu Y, Oralkan O, Strohmaier J, Muth J, Peden D, Bozkurt A Low power wearable systems for continuous monitoring of environment and health for chronic respiratory diseases
21. Propeller health (2019) Propeller. Our solution. Retrieved from <http://propellerhealth.com> on date 12 Oct 2019

Exploring in the Context of Development of Smart Cities in India



Smita Bharne and Suryakant Patil

Abstract In the year 2014, “smart cities” in India started through an initiative known as “Smart Cities Mission” by the Indian Government. The key idea behind this mission is to develop technology-driven cities in India. The urbanization rate in India is rapidly increasing. Due to fast urban growth, citizens suffer from problems like heavy traffic jams, poor waste management, inadequate water supply and many more. To overcome these issues, smart city terminology comes into existence. This paper provides the inclusive improvement of smart cities in India, in terms of market scale, development growth and industry standards. The challenges while making the “smart cities” in India are also discussed. The research also reviews the main themes constituents of the “smart cities” like “smart governance”, “smart education”, “smart living”, “smart healthcare”, “smart citizens”.

Keywords Smart city · Smart city development · Social challenges · Smart solutions · Technology · Urban

1 Introduction

In the last 20 years, the world economy shows fast progress towards globalization and urbanization around the world [1]. Due to growing population all over the world, there is pressure for sustainability on the global cities. The urban population growth rate in 1950 was 30%. According to the United Nation reports on the world’s population, around 56% of people were living in the urban areas in 2015 [2]. The growth rate of the urban population will be increased to 55% till 2018, and it is expected to increase by nearly 69% till 2050 [3]. Figure 1 shows the urbanization rate in India. The urban population in India will reach 40% by 2026 [4]. Thus, the cities around

S. Bharne (✉)
Ramrao Adik Institute of Technology, Navi Mumbai, India
e-mail: smita146@gmail.com

S. Bharne · S. Patil
Sandip University, Nashik, India

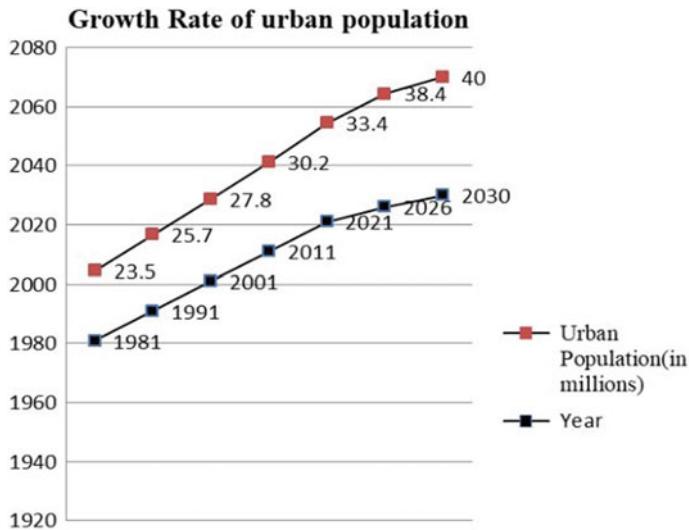


Fig. 1 Urbanization rate in India

the world will face the challenges for maintaining the high-quality life of citizens. It also makes difficult to offer smart, sustainable, economical and efficient services towards the growing urban population.

A smart city has been emerged to provide an efficient, sustainable life to citizens for economic growth. The expanding scope of the citizen demands the efficient management of urban planning. From the perspective of globalization, a “smart city” is important and considered as a part of e-governance-driven architecture. A smart city aims to provide citizenship management by providing basic infrastructure to citizens, green environment, ease of transportation and sustainable development. The best way to meet these requirements is the use of information technology with sufficient infrastructure. This notion is broadly recognized as smart cities (SC) [5].

1.1 Major Constituents of Smart City

The fundamental constitutes of “smart cities” are listed here:“smart governance”, “smart education”, “smart living”, “smart mobility”, “smart environment”, “smart energy”, “smart healthcare”, “smart citizens” [5–8]. Smart cities can grow more smartly using technologies like IoT (Internet of Things), artificial intelligence and big data analytics [8–11]. All key components are described in the following section (Table 1).

Table 1 Major constituents of the smart city

Key areas	The task associated with areas	Citations
Smart governance	Online citizens portals	[4, 9, 16, 21]
	Efficient and fast public services	
	Effective resource management	
	Innovative planning approaches	
	Public asset management	
Smart education	E-services, connecting people through social media	
	Smart infrastructure with CCTV surveillance	[4, 9, 16, 21]
	GPS tracking of school buses	
	Smart learning through video conferencing lectures	
	Teacher–students management solutions	
Smart living	Virtual labs	
	Public security tools	[5, 4, 8, 9, 16, 21]
	Safety alarms at public places at panic situations	
	Community network management	
Smart mobility	Safety of senior citizens	
	A smart toll collection system	[4, 5, 8, 9, 16, 21]
	Community carpooling system	
	Charging point for electric vehicles	
Smart environment	Smart parking system	
	Traffic management	[4, 5, 8, 9, 16, 21]
	Vehicle monitoring	
	Water quality management	
	Air quality management	
	Smart water storage and purification system	
	Wastewater management	
	Pollution sensors	
Smart energy	Disaster management	
	Green and clean environment	
	Intelligent smart metres	[4, 9, 21, 16]
Smart healthcare	Efficient utilization of energy subsystem	
	Energy distribution through sensors	
	E-health records	[4, 9, 21, 16]
Smart citizens	Diagnostic analytics portals	
	Emergency medical services	
	Privacy and security of citizens	[4, 5, 8, 9, 16, 21]
	Social engagement of the people	
	Raising awareness of smart solutions	

(continued)

Table 1 (continued)

Key areas	The task associated with areas	Citations
	Community interactions	

2 Review of the Literature

In 2008, IBM launched the “smarter planet and smart city concept” all over the world, in a few selected cities based on information communication technology (ICT). Japan, Singapore, China already started building smart cities with the help of ICT [1]. From the past decade, different characteristics related to the smart cities are evolving with ICT. Before 2008, very few literature studies are having the key terminology as a smart city. The articles having a major keyword as “smart city” have been increased from the last 10 years [5]. There are many definitions of “smart city” available in the literature. SC will be having different insinuations in India as this concept varies from cities: countries depending on development and resources [12].

After analysing the different definitions relevant to different domains, Elvira Ismagilova et al. finally concluded the “smart cities use an information system (IS)-centric approach to the intelligent use of ICT within an interactive infrastructure to provide advanced and innovative services to its citizens, impacting the quality of life and sustainable management of natural resources” [5].

These definitions are also categorized according to the different domains. The word smart city is having relevance with the digital city, intelligent city, information city, according to the information communication technology domain. Creative city and entrepreneurial city, knowledge city are categorized into the learning- and knowledge-based development, sustainable city; eco-city is based on sustainable development [5, 4, 6, 13]. Francesco Paolo et al. summarized the definitions of SC by classifying three components, namely physical structure, the potential eminence of life to citizens and improvement in ecosystems [9].

Bhattacharya et al. [14] proposed the reference framework for India’s Smart Cities Mission. The author provided a detailed understanding of the progress of SC at a global level: a benefit of “Smart Cities Mission” to development urban India. The details of the study approach include the evolution of the SC concept; review of SC concept from research and an academic point of view; corporate sectors and government sector. The smart city’s reference framework is based on the four guiding principles—comfort with security, equity, effectiveness and foresight [4].

Sharma and Ghosh [15] describe India’s Smart Cities Mission through the social innovation perspective, to enhance the citywide services. Its focuses are on water supply management, transport and green spaces. One of the major goals is to connect the citizens to local officials through Internet connectivity. Citizen’s participation makes much more transparency in urban planning and management using information communication technology.

Rehena and Janssen [4] focus their work on smart city Pune after the initiative of “Smart Cities Mission” in India. The author identifies the current challenges to make “Pune” as a smart city and address these challenges. India has lots of difficulties to meet the requirements of smart cities, like lack of infrastructure, inefficient traffic management, inadequate water supply. After identifying the potential long-term challenges, the establishment of powerful planning between the authorities is needed to achieve the urban and its surrounding development.

Ismagilova et al. [5] summarized the SC research in the view of the information system perspective. This paper is based on key aspects—(1) exploring research related to smart cities over the past few decades, (2) discussion and analysis of themes of smart cities and (3) identifying potential future growth areas for further research. The author has summarized the different characteristics of SC like “smart living”, “smart environment”, “smart citizens”, “smart mobility” through the information system perspective [5, 11].

Das and Misra [2] described the connections between the management of SC and e-governance within India. “Smart Cities Mission” in India aims to provide urban services to the people through smart e-governance. The author proposed the framework to connect the various services such as public information, grievance redressal without making too many changes between a service provider and service user.

Sarkheyli and Sarkheyli [15] present the work-related to the various issues and challenges of the smart cities. The author focuses on how a smart city uses information smartly. Various cities use the label of “smart city” to describe the environmentally ecological development, cost-effective growth and superior quality life of peoples. In a smart city, information gathering from various sources and its efficient management are a quite difficult task. Smart cities can face various challenges such as social cohesion, controlled transitions in the market due to the automated systems. A major focus is given on exploring the different extents of the smart city like “smart economy”, “smart mobility”, “smart environment”, “smart people”, “smart living” and “smart governance” connected through technology, people and community. Generation and management of data in SC are described through various categories such as infrastructure, sustainability, health, commerce, citizen, safety, energy and electricity. Since the gathered information is too large, security and privacy are the critical part of this architecture. These aspects of privacy-related issues and security attacks in SC are also covered in this paper. Security and privacy protections of data in SC have also become one of the key challenges and growing area of research.

3 General Situation of Smart City in India

In the year June 2015, the Government of India started “Smart City Mission”. The objective of this mission to deliver a good infrastructure, quality life to people, clean and green environment with smart solutions. Smart City Mission was started on 25 June 2015 for developing 100 smart cities. Initially, 20 cities are selected in the list

Table 2 Top ten emerging cities in India around the world

Rank	City	States	Average annual growth (in %)
1	Surat	Gujrat	9.17
2	Agra	Uttar Pradesh	8.58
3	Bengaluru	Karnataka	8.50
4	Hyderabad	Telangana	8.47
5	Nagpur	Maharashtra	8.41
6	Tiruppur	Tamil Nadu	8.36
7	Rajkot	Gujrat	8.33
8	Tiruchirappalli	Tamil Nadu	8.29
9	Chennai	Tamil Nadu	8.17
10	Vijaywada	Andhra Pradesh	8.16

of Smart City Mission in 2015. In the second and third round of selection, 13 and 27 cities are added, respectively, in the year 2016. During the fourth round, 39 cities are selected as of Jan 2018. The selection criteria for the cities included in Smart India Mission has at minimum two smart cities in 29 states, state having an urban population between 1 million and 5 million [12]. The categorization of cities is based on city sizes such as megacities, mid-size cities and small cities. The population of more than 5 million was included in megacities. The population of between one and five million was included in mega mid-size cities. The population of less than one million was included in small cities. Amongst the first 20 selected cities as on 29 January 2016, Bhubaneshwar and Pune have got first and second rank, respectively [16].

In 2018, Oxford Economics had released a list of top ten emerging cities in India all over the world along with an average annual growth percentage rate. Table 2 shows the top ten emerging cities in India around the world [17].

A smart city selection criterion is through the two-stage competition as follows. Phase 1: Maintaining a past track record of under Jawaharlal Nehru National Urban Renewal Mission, its service levels and financial strength [4].

Phase 2: Financial control of smart city planning including citizen's involvement [4].

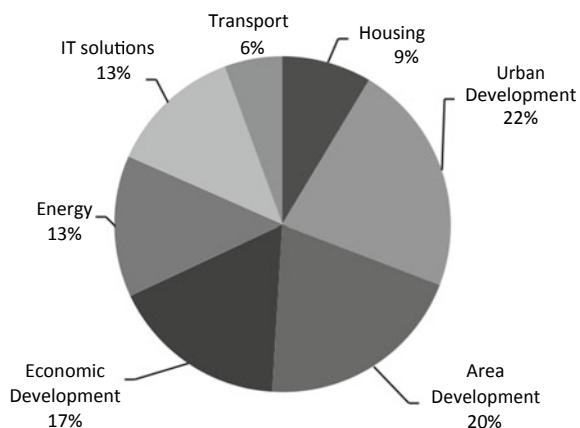
Between January 2016 and January 2018, the government selected the 99 cities through the competitive process. The updated status of project completion, smart city mission as on 31 March 2018, is summarized in Table 3. (Source: Ministry of Housing and Urban Affairs, National Institute of Urban Affairs) [18].

In the starting year of Smart City Mission, growth rate of completed projects is slower. First two and half years of "Smart City Mission" merely 5.2% of the total recognized projects are completed. In the year of March 2018 to June 2019, the number of projects completed had a rise of 182% compared to earlier years. Table 3 shows the number of the updated status of completed projects in Smart City Mission during the year March 2018 to June 2019. A large number of projects included in the

Table 3 Updated status of Smart City Mission as on 31 March 2018

Attributes of the project selection/completion	Values
Number of selected cities	99
Cities that appointed project management consultants	69
Total number of population affected	99.5 million
Town planner working on the scheme	5500
Number of afforded housing projects	73
Cost of affordable projects	Rs. 17.036 Crore
Total investment proposed in 4500 projects by 99 cities	Rs. 203,979 Crore

Fig. 2 India smart city sector-wise investment structure 2018 [18]



Smart City Mission with key component areas are listed in Table 2. An investment of Rs. 203,979 is made on 4500 projects in 99 cities. Figure 2 shows the sector-wise investment growth in major areas of constituents of a smart city. These statistics show the growth of smart cities in later years is fast compared to earlier years.

4 Challenges of Smart Cities in India

India has a bright prospect of SC with the support of the Government of India. In the foundation year, efficient planning for designing SC in various cities and well-defined policies give support for the successful execution of “Smart Cities Mission”. SC projects can have a good impact on the quality of life of the citizens. Some of the major challenges towards completion of Smart City Mission are discussing below [4, 19].

The accurate background model for smart city: To make the progress of SC in India, the same background framework with well-defined indicators must be applied within all cities.

Societal adequacy and lack of awareness about ICT: Smart citizens are playing an important role in using smart solutions within smart cities. To make the urban citizen's comfortable life the majority of components of a smart city fuse with ICT innovation. Unfortunately, there is a lack of awareness of how ICT is used within smart cities amongst the people.

Difficulty in upgrading the old cities as smart cities: The upgradation of the old cities requires more effort than building new cities. The work is always under process. **Sustainable growth:** It is based on utilizing natural resources in smart cities. It includes less pollution, tree plantation and lower energy consumption to maintain a clean environment. Thus, it is difficult to maintain the sustainable growth of smart cities.

A requirement of efficient coordination: Development of smart cities requires good coordination between various departments of government.

Efficient governance: To maintain the proper growth of the smart city's projects, various systems are implemented by Smart Cities Mission. It requires control monitoring. This becomes one of the big challenges.

The requirement of funds: Urban infrastructure needs lots of funds to fulfil the requirement of smart cities. Seven lakhs are needed in the next 20 years to build the smart city infrastructure [4, 20]. The most appropriate solution to this problem is to raise the funds from the public-private partnerships.

5 Smart Solutions for Smart Cities in India

The government is taking initiatives to provide smart solutions in the major components of smart cities. It will help to drive the various innovative solutions that incorporated to make cities "smart".

Innovation in transportation: To provide better mobility services, the Government of India launch many metro rail projects in various cities. It will cover the 47 lakh kilometres of the road span. Metro cities like Mumbai will have ropeway transportation to connect the different sides of the city.

Digital solutions: Making all government initiates online will lift the transparency in the system. It also provides ease of access to citizens to reduce manual efforts. **Eco-friendly homes:** The houses having solar panels and wind generators can contribute towards the increases in renewable energy sources.

Enhanced surveillance system: To monitor every corner of the city, a good quality of surveillance system is required. Data can also be saved on the centralized server to track the road accidents.

Smart homes: Technologies such as automation in security and surveillance kits can help in monitoring activities around the homes to provide more security to the citizens.

6 Conclusion

The government is playing a vital role during the progress of a smart city. The growth of smart cities is good in later years compared to earlier years. The current growth of smart cities will strengthen the assimilation of resources. In a country like India, to make existing cities as smart cities are quite challenging. The challenges faced during the constructions of smart cities can be overcome by providing a good economic structure and with good governance. Smart cities can provide a sustainable environment with innovations using information communication technology such as big data platforms, artificial intelligence and the Internet of Things (IoT) make cities smarter in a much better way.

References

1. Guo M, Liu Y, Yu H, Hu B, Sang Z (2016) An overview of smart city in China. Communications 13(5):203–211. <https://doi.org/10.1109/CC.2016.7489987>
2. Das RK, Misra H (2017) Smart city and E-Governance: exploring the connect in the context of local development in India. In: Fourth international conference on eDemocracy & eGovernment (ICEDEG), Quito, pp 232–233. <https://doi.org/10.1109/icedeg.2017.7962540>
3. Sang Z, Li K (2019) ITU-T standardization activities on smart sustainable cities. In: IET smart cities. 1(1):3–9. <https://doi.org/10.1049/iet-smc.2019.0023> www.ietdl.org
4. Rehena Z, Janssen M (2019) The smart city of Pune. J Smart City Emergence 2019:261–282. <https://doi.org/10.1016/B978-0-12-816169-2.00012-2>
5. Ismagilova E, Hughes L, Dwivedi YK et al (2019) Smart cities: Advances in research—an information systems perspective. Int J Inf Manag 47:88–100. <https://doi.org/10.1016/j.ijinfo@mgt.2019.01.004>
6. Cocchia A (2014) Smart and digital city: a systematic literature review. In: Dameri R, Rosenthal-Sabroux C (eds) Smart city. Progress in IS. Springer, Cham
7. Anthopoulos LG, Reddick CG (2016) Smart city and smart government: synonymous or complementary? In: Proceedings of the 25th international conference companion on World Wide Web. International World Wide Web Conferences Steering Committee, April 2016, pp 351–355
8. Vinod Kumar TM, Dahiya B (2017) Smart economy in smart cities. In: Book: Smart economy in smart cities edition: 1st chapter: Springer, Berlin. https://doi.org/10.1007/978-981-10-1610-3_1
9. Appio FP, Lima M, Paroutis S (2019) Understanding smart cities: innovation ecosystems, technological advancements, and societal challenges. Technol Forecast Soc Change 142:1–14. <https://doi.org/10.1016/j.techfore.2018.12.018>
10. Anthopoulos LG, Reddick CG (2016) Smart city and smart government: synonymous or complementary? In: Proceedings of the 25th international conference companion on World Wide Web (WWW '16 Companion). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp 351–355. <https://doi.org/10.1145/2872518.2888615>
11. Vinod Kumar TM (2019) Smart metropolitan regional development. In: Book, chapter advances in 21st century human settlements book series (ACHS). Springer, Berlin. <https://doi.org/10.1007/978-981-10-8588-8>
12. Smart City Government of India Guidelines Information. [http://smartcities.gov.in/upload/uploads/files/SmartCityGuidelines\(1\).pdf](http://smartcities.gov.in/upload/uploads/files/SmartCityGuidelines(1).pdf)

13. Praharaj S, Han H (2019) Cutting through the clutter of smart city definitions: a reading into the smart city perceptions in India. City Culture Soc. <https://doi.org/10.1016/j.ccs.2019.05.005>
14. Bhattacharya S, Rathi S et al (2015) Reconceptualising smart cities: a reference framework for India. <https://niti.gov.in> › files › CSTEP Report Smart Cities Framework
15. Sharma M, Ghosh A (2015) Imagining smart cities in India. In: Book chapter. What does India think? ISBN 978-1-910118-45-0
16. Growing trend of urban population in India. http://www.governoruk.gov.in/files/Smart_City_Dashboard.pdf
17. Oxford Source Economics—<https://www.bloomberg.com/news/articles/2018-12-05/india-claims-top-ten-in-list-of-world-s-fastest-growing-cities>
18. Ministry of housing and urban affairs, National Institute of Urban Affairs <http://mohua.gov.in/>
19. Sarkheyli A, Sarkheyli E (2019) Smart megaprojects in smart cities, dimensions, and challenges. In: Chapter 19-smart cities cybersecurity and privacy. Elsevier, pp 269–277. <https://doi.org/10.1016/B978-0-12-815032-0.00019-6>
20. Kandpal V, Kaur H, Tyagi V (2017) Smart city projects in India: issues and challenges. SSRN <https://ssrn.com/abstract=2926260>. <http://dx.doi.org/10.2139/ssrn.2926260>
21. Yigitcanlar T, Kamruzzaman M, Foth M, Sabatini-Marques J, da Costa E, Ioppolo G (2019) Can cities become smart without being sustainable? A systematic review of the literature Sustain Cities Soc 45. <https://doi.org/10.1016/j.scs.2018.11.033>

Microservices and DevOps for Optimal Benefits from IoT in Manufacturing



Anurag Choudhry and Anshu Premchand

Abstract Internet of Things (IoT) is not a pipe dream anymore. The trajectory of growth in the Internet of Things (IoT) space has been nothing short of phenomenal in the last few years. It is envisaged to keep a similar pace of growth for the next few years. This paper looks at Microservices and DevOps as relevant levers for faster adoption of IoT as well as for following the path of least resistance keeping in mind factors like future growth, pervasiveness, miniaturization, economics, etc. This paper proposes that usage of DevOps and Microservices is essential for eliciting maximum benefits from IoT and suggests a framework for the same. It also discusses the combined effects and benefits of IoT, Microservices, and DevOps.

Keywords Internet of Things · IoT · Microservices · DevOps · Manufacturing

1 Overview of IoT, Micro Services and DevOps

Today, Internet of Things (IoT) has become an inescapable reality. The interest in Internet of Things is increasing because IoT has societal as well as industrial uses. The evolution of Internet of Things (IoT) will continue for some time to come. The IoT uses computer networks to connect physical objects or things with information technology systems. According to a report [1]:

- The Industrial Internet of Things, also referred to as IIoT, is likely to have a major impact on the manufacturing industry as well as on the global economy. By the

A. Choudhry (✉)

Solution Architect, Research & Innovation, Manufacturing & Utilities Business Group, Tata Consultancy Services, Delhi, India
e-mail: Anurag.choudhry@tcs.com

A. Premchand

Agile & DevOps Practice Lead, Manufacturing & Utilities Business Group, Tata Consultancy Services, Chennai, India
e-mail: Anshu.premchand@tcs.com

year 2030, it is projected that the industrial Internet of Things (IIoT) will create \$15 trillion in terms of the global GDP.

- Global spend on the Internet of Things is likely to be around \$772 billion by the year 2018. Also, by the year 2020, it is predicted to cross \$1 trillion. The projected expense of \$189 billion on IoT by the manufacturing industry in 2018 is the largest amount from any industry.

International Organization for Standardization (ISO) has defined IoT as “an infrastructure of interconnected physical objects or entities, information resources, and systems along with a number of services which are intelligent and can process as well as react to information of both the worlds—virtual and physical. This IoT infrastructure can also make an impact on actual tasks or activities in the physical world” [2].

The concept of creating a network by connecting objects to each other and to the internet has been around for long, then, why has Internet of Things (IoT) suddenly become all-pervasive now? Convergence of several technologies and industry trends makes it possible to interconnect more and more devices inexpensively. The following are some industry trends and technologies which are the main drivers of IoT [3].

- Pervasive Connectivity—Low-cost, high-speed network connectivity, and wireless technologies and services make almost everything connectable.
- Prevalent Adoption of IP Based Networking—IP based networking has become new normal for connecting devices.
- Computing Economics—Research and Innovations continue to deliver greater computing power at lower cost and lower power consumption.
- Miniaturization—Small and low-cost sensor devices drive many IoT applications and this could have been possible due to advancement in communication technologies and manufacturing.
- Data Analytics—one of the key requirements of IoT is to extract information and knowledge from large and dynamic datasets. Evolution of data store, new algorithms, computing power, and cloud services enable data aggregation, correlation, and analysis to extract information and knowledge from datasets.
- Cloud Services—Cloud services powering IoT solutions by allowing small and distributed devices to connect with high computing enabled back-end analytics.

From the definition of IoT mentioned above, we can note that IoT is an infrastructure of connected things and services are enablers of IoT. IoT solutions can be implemented using Service-Oriented Architecture (SOA). However, SOA-based solutions are not sufficient for IoT. IoT solution consists of a large number of single-function distributed devices. In recent years, we have seen that most of the organizations are implementing IoT solutions based on Microservices because Microservices characteristics are matched with IoT solutions. Let us first look at the definition of Microservices followed by characteristics of Microservices.

Lewis and Fowler [4] defined Microservices as “an approach for the development of a single application in the form of a suite of small-sized service(s) such that each one runs in its own process and communicates using light-weight mechanism”,

which could be, for example, HTTP resource API. The Microservices are generally created in alignment with business capabilities. They are deployed independently of each other by utilizing a (potentially) fully automated deployment mechanism. There is no necessity to centrally manage Microservices. Also, the Microservices can be written in a variety of programming languages as well as utilize different data storing methods and technologies.

Microservices are a methodology to build an application as a set of small but loosely coupled services that can communicate with each other using light-weight mechanisms like HTTP resource APIs. Microservices are nearly always modeled around business bound context(s) and have their own life cycle, which is independent of other Microservices. These services are independently deployable by fully automated deployment processes. Microservices can be built using different languages/frameworks/tools.

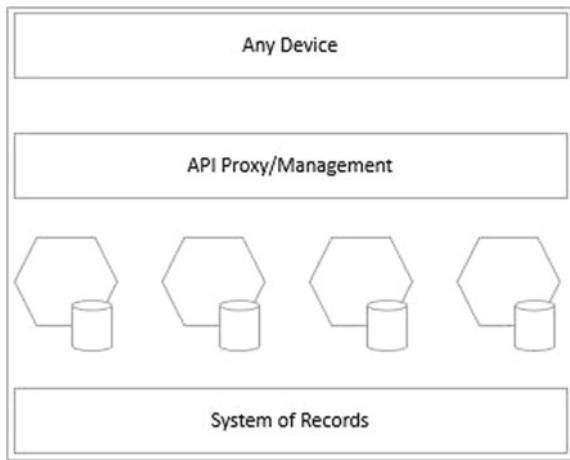
A microservice has the following key characteristics [5]:

- Independence of Microservices: A microservice is inherently independent of other Microservices. Each microservice is developed, deployed, and grow independently.
- Evolutionary Nature—Microservices can be developed along with existing monolithic applications providing a bridge to a future state.
- Resilient and Fault-Tolerant—Isolating Microservices that are down would not affect the other functions. This enables ease of operational maintenance and also reduces overall downtime across functions.
- Technology independent—Each microservice can be built on its own technology and be exposed through API gateways. This enables to move away from any technology at any point in time and adapt to newer models.
- Flexible Deployment—A microservice has its own lifecycle, and can be built (or changed), tested, and deployed on its own.
- Flexibility of Scale—Dynamic scaling of Microservices leads to better cost control.
- Reusability—Microservice is highly reusable and can be composed of other services.
- Decoupled Architecture—Microservices are independent functions and thus decoupled by their inherent nature. Clients are not aware of the implementation details and also two Microservices are not aware of each others functions. Decoupling is high in the Microservices environment.

We have seen so far that Microservices are relevant to IoT solution implementation. We have not discussed one important aspect of the solution implementation which is DevOps. We cannot even think about Microservices-based solution without DevOps. Let us look at DevOps concept and how this is important for Microservice-based IoT solutions (Fig.1).

- DevOps provides the process, practices, and tools for Microservices delivery enabling faster time to market.

Fig. 1 Micro service architecture



- For successful microservice architecture (MSA) based IoT solution and to achieve faster time to market, it must adopt development-to-delivery practices including the DevOps philosophy.
- Microservices and DevOps enable IoT solution delivery to target the same set of common objectives—speed of delivery, business value, and cost-benefit by faster time to market.
- DevOps achieves the goal of continuous delivery, which is required for an IoT solution, through its three pillars approach:
 - Automation
 - Standardization
 - Frequent releases of code.

Figure 2 shows DevOps framework for continuous service delivery.

Details of each aspect of DevOps framework are covered in Sect. 4. Figure 3 shows Microservices-based IoT solution architecture. The Fig. 3 shows that Microservices can be used in IoT solution implementations.

2 Combined Effect and Benefits of IoT, Micro Services and DevOps

In many ways, Microservices are best fit for IoT solutions and without DevOps it would be hard to envisage MSA-based solutions [6]. Following are some of the areas which support this resemblance:



Fig. 2 DevOps framework for continuous services delivery

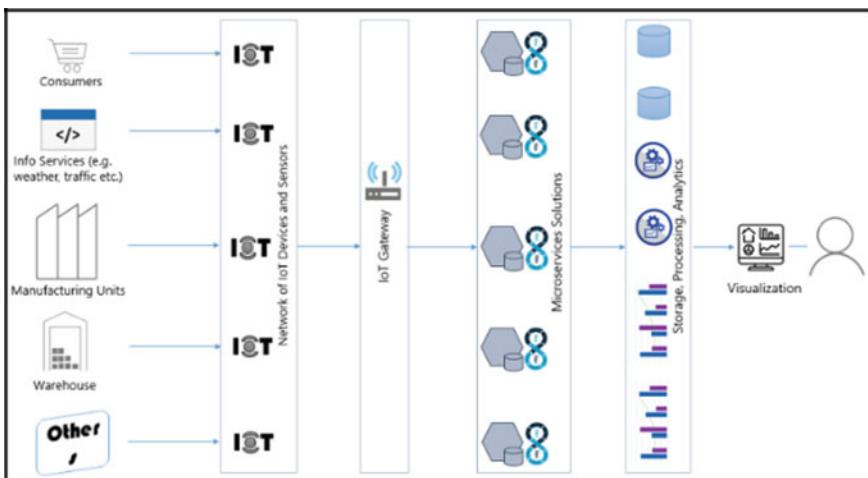


Fig. 3 Microservices-based IoT Solution architecture

- Diverse Networks

IoT networks are inherently heterogeneous. They also evolve continuously. These characteristics of independence and decoupling make Microservices the best fit for such network of devices.

- Distributed Nature of Data Management and Governance

Deployment of services in heterogeneous networks are performed by independent parties. And, deployment of components and services are done by different vendors as and when required. Centralized governance of such architecture would be extremely difficult. Independent nature of Microservices makes decentralized governance possible at a very granular level. In a solution, Microservices can be built using different programming languages.

- Designed for Failure

IoT networks are composed of a large number of components that cannot afford to be dependent on the health and status of a single or small number of these components. Independence characteristic of Microservices provides an architecture that is robust in the event of failure. If a microservice fails it may be replaced by the new instantiation of that service and any failed IoT device may be replaced by other device or data may be extrapolated generated by other devices.

- Independence

One of the key characteristics of Microservices-based architecture is independence and this differentiates Microservices Architecture style from Service-Oriented Architecture. IoT network characteristics such as varied networks, resiliency, distributed governance, etc. are better aided by Microservices-based architecture where services are inherently independent.

- Evolution at the atomic level

IoT networks consists of many devices, components and services, and constantly evolving business needs require changes in IoT architecture. Hence, architecture needs to be designed to enable this change to happen at an atomic level. Microservices allows this flexibility by adding, replacing, and deploying services independently.

- Smart endpoints and dumb pipes

Deploying new functionality in smart networks becomes expensive and complicated because of tight coupling between components. Shifting the intelligence and domain logic into many endpoints of an IoT network makes the network highly decoupled. And, Microservices are aimed to be decoupled and as cohesive as possible [6].

3 Micro Service Patterns Applicable to IoT Solutions

Following are some of the key patterns of Microservices which are applicable to IoT solution implementation [6].

- Interpolation

Problem Statement: Information will be lost if an individual sensor fails in a grid of network-enabled sensors when try to read the information.

Solution: Information provided by nearby sensors can be interpolated and used in place of missing information.

- Sensor Facade

Problem Statement: A device acting as an IoT sensor does not implement its services in a way that is compatible with the desired consumers.

Solution: Use a façade service in front of the IoT device that provides compatible data to the desired consumer.

- Cache

Problem Statement: Sensor data may be required on a regular basis by more consumers than a sensor can handle.

Solution: Sensor data may be stored in the cache and this will return the most current sensor value until the cache entry expires. This leads to a single direct sensor read to refresh the cache entry.

- Gateway

Problem Statement: Grid of IoT devices has limitations such as protocol transformation, security, services enhancements, etc.

Solution: IoT devices must be accessed through Microservices-based gateway. This enables inter-service communication and with the IoT nodes.

- Sensor Aggregator

Problem Statement: Analysis of data is required which is collected from many sensors. Analysis result is required to take desired action.

Solution: Data is collected from sensors and after analysis, the result is exposed as a microservice operation.

- Multicast

Problem Statement: In some scenarios, consumer needs to be notified when an event occurs at an IoT device.

Solution: Multicast pattern can be used in such scenarios. All consumers who are registered as subscriber notified when an event has occurred at an IoT device.

4 DevOps for Microservice-Based Architecture (MSA) and IoT

This section lists the details of stages of DevOps framework for continuous services delivery.

Continuous Planning—Release Planning, Feedback, Analytics

- Scope the team intends to deliver by a given deadline.
- Feedback from Operations team in release planning.
- Analytics based on user behavior pattern

Continuous Build—Dedicated Environment, Private Build, Automated Integration, Build Label, Minimal Dependencies

- Dedicated build servers to isolate from other activities.
- Using the existing build configuration to compile the code that is not yet checked in.
- Trigger a build on check-in and deploy the source to a continuous integration server.
- Every build to be uniquely identified across the environments *Continuous Integration*—Automated Testing, Integration Testing, Continuous Inspection, Thresholds, Non-Functional Test
- Automate the process for initiating various testing.
- Integration testing validates the microservice in the system for dependencies with other services.
- Frequent inspections to raise early quality issues.
- Include non-functional tests such as performance, security. *Continuous Testing*—Test Automation, Test Data, Defects and Feedback
- Test automation to control the execution of tests and comparison of the actual outcomes with predicted ones.
- Defects to be recorded and informed to concerned people. *Continuous Deployment*—Deployment Script, Rollback Script, Deployment Test, Unified Deployment
- Scripts to take care of the environment variables and other aspects related to the deployment of applications in various servers.
- Automated testing such as Smoke test using Selenium to be performed post-deployment.
- The environment not being deployed appropriately should be scrapped.

Continuous Monitoring—Proactive Incident Management, Incident, and Problem Management

- To analyze the metrics intelligently and identify reasons for the potential problems. Also to take any action to avoid the incident.

Continuous Feedback—Analytics

- To proactively seek the information about the application and analyze the behavior of the same.

5 Use Case Example

Following are some of the use cases where Microservices-based architecture can be used to develop IoT solutions:

- Prognostic Maintenance—IoT Devices, Sensors, and Data Analytics enable us to detect when equipment will fail before it does.

Advantage: Longer life of the equipment, increased plant safety, and reduced accidents with negative environmental impact.

- Asset and Material Tracking—Enables organizations in locating and monitoring of key assets such as raw materials, final products, and containers. *Advantage:* Helps in optimizing logistics, maintain inventory levels, prevent quality issues, and detect theft.
- Connected Operations—Enables in connecting disparate silos of operational data such as manufacturing, supplier, and logistics into unified, real-time visibility across heterogeneous systems.

Advantage: Helps in making faster and better decisions and improve operational performance.

- Unified Key Performance Indicators—Generate KPIs by aggregating and contextualizing data from isolated manufacturing systems.

Advantage: Helps to diagnose problems more quickly and improve performance.

- Operations Management Improvements—This is to accelerate smart factory and ‘Industry 4.0’ initiatives by extending existing equipment and ERP/MES systems with connectivity, interoperability, mobility, and intelligence.

Advantage: Helps in improving process stages, i.e., monitored, managed, and optimized.

6 Combined Effect and Benefits of IoT, Micro Services and DevOps

The paper looks at an overview of IoT and how Microservices-based architecture is the best fit for IoT solution implementation. It also describes how DevOps is the necessity of Microservices-based solutions. Further, the paper discusses the combined effect and benefits of IoT, Microservices and DevOps followed by some of the applicable Microservices patterns to IoT solutions and IoT use cases. In conclusion, the researchers recommend that IoT implementation and adoption should strongly consider the proposed usage of Microservices and DevOps for realizing optimal benefits.

The biggest challenge in an IoT solution is the lack of/less communication amongst sensors, components, services, and other system applications. IoT networks are heterogeneously composed of many sensors that are widely distributed. It is obvious that monolithic application architecture cannot handle the data generated and consumed by IoT network devices. Hence, we need an architecture approach to handle this complexity which can allow applications to work independently. Microservice architectural style is best fit in such scenarios since the independence, distributed governance, and resiliency inherent in Microservices-based architecture.

References

1. Microsoft (2019) 2019 manufacturing trends. <http://info.microsoft.com/rs/157-GQE-382/images/EN-US-CNTNT-Report-2019-Manufacturing-Trends.pdf>
2. ISO/IEC 30141 Internet of Things (IoT)—Reference Architecture
3. ISOC-The Internet of Things: an overview. <https://www.internetsociety.org/wp-content/uploads/2017/08/ISOC-IoT-Overview-20151221-en.pdf>
4. Lewis J, Fowler M Microservices—a definition of new architectural term. <https://martinfowler.com/articles/microservices.html>
5. The Open Group (2016) Microservices architecture, White Paper (W169). <https://www.opengroup.org/library/w169>
6. The Open Group (2018) Microservices architecture for the Internet of Things (MSA-IoT)- The open group guide. <https://publications.opengroup.org/g187>

Semantic Interoperability for IoT Agriculture Framework with Heterogeneous Devices



P. Salma Khatoon and Muqeem Ahmed

Abstract Interoperability is a challenging issue faced by IoT developers all over the world. This is due to the fact that the IoT devices currently being used are utilizing a variety of data formats, protocols, and technologies for operation. As currently there are no standardized rules framed for IoT applications, interoperability tools remain limited. This paper focuses on the concept of developing a framework on the principles of interoperability for agriculture-related IoT devices. The proposed framework enables interoperability among heterogeneous devices. The data gathered from different sensors in the farms is semantically annotated and presented in a user-friendly manner. A lightweight semantic annotation model is used to annotate the data. Resource Description Framework (RDF) is used to provide semantic functionality to the data. The proposed framework helps in providing interoperability to the heterogeneous data gathered from IoT devices.

Keywords Interoperability · IoT · Resource description framework · Heterogeneous devices

1 Introduction

The precision agriculture [1] is enabled by developing very precise embedded sensors that measure the environmental concept in the farms in order to enhance the yield and to increase the production as well as profitability. Moreover, the environmental footprint is reduced by implementing the methods like effective irrigation, targeted, more accurate usage of fertilizers as well as pesticides for crops, as well as food and antibiotics for animals, the environmental footprint. The concept of smart farming is enabled by the precision agriculture that implements the collection, processing,

P. S. Khatoon (✉) · M. Ahmed
Maulana Azad National Urdu University, Hyderabad, India
e-mail: salmakhatoon537@gmail.com

M. Ahmed
e-mail: muqeem.ahmed@gmail.com

and analysis of real-time data, in addition to the automated techniques upon the farming approaches. The enhancement and management of the complete farming activities are improved by it and provides additional information for the farmers in decision making. Since farming is highly dependent on weather and environmental conditions like rain, temperature, humidity, hail, etc., unpredictable events such as animal diseases, and pests, as well as instability of prices within agricultural markets, therefore it is greatly unpredictable. The farmers are advised regarding the condition of their farms and proper countermeasures are suggested and their crops or livestock along with complete production can be protected by the implementation of wide-ranging agriculture-related frameworks by using sensors or automated technologies. As a result of high interoperability, scalability, pervasiveness, and inclusiveness, an absolute equivalent for smart farming is the Internet of Things (IoT) [2]. An impulsion is gained by IoT in the agriculture sector due to the realization of the huge potentiality of IoT technologies for smart farming.

For instance, at present, affixing the passive RFID ear tags to the cattle is compulsory for the overall Australian farmers and their activities in the farms must be reported to an online national database [3]. The outdated farming methods are revolutionized by the progression of the IoT-based environment for smart farming. IoT frameworks targeting smart farming is implemented in order to achieve this and the frameworks are implemented in various related areas like smart cities, healthcare as well as smart homes [4–8]. With the help of open standards such as 6LoWPAN, ZigBee, CoAP, ISOBUS, SigFox, REST, MQTT, XMPP, etc., and semantics like RDF, OWL, SWE, SensorML between the components, devices, processes as well as platforms in order to bring seamless connections and advanced compatibility. Several advantages are provided by the IoT frameworks for smart farming which reduces vendor lock-in problems by adopting equipment in addition to sensing or automated systems for several companies. Moreover, it is compatible with the entire smart system of the farms and the information can be exchanged easily amongst various heterogeneous components and internet standards are implemented by this in order to attain enhanced automation with minimum efforts.

A high-adaptive online platform is proposed for IoT-based innovative data analytic solutions by the motivation provided by the IoT advantages as well as potentiality regarding the smart farming that considers the wide-ranging deficiency, reliability, and well-established as well as bulletproof solutions in addition to frameworks in this area. Depending upon the real-time data streams containing various sources like sensory systems, surveillance cameras, hyperspectral images from drones [9], online weather forecasting services, social media streams for fast identification of events (e.g., hazards, earthquakes, floods), information, warning and alerts from governmental organizations, such as the Ministry of Agriculture, etc., enables the wide-ranging data processing, investigation as well as automatic reasoning.

The evaluation and combining capability of the data streams like the above benefits which help the farmers for decision making in the adjacent real-time and rapid-response for the variations and the unexpected incidents are provided by Agri-IoT. For instance, the sensory data regarding the soil fertility and web services on behalf of the weather forecasting is combined in order to provide optimum solutions with

highly accurate irrigation in addition to the crop's fertilization. This paper explains the Agri-IoT (Sect. 3) operation and the performances are evaluated with the help of two demanding smart farming scenarios (Sect. 4). Moreover, it discusses the provision of the novel potentials within the farming industry (Sect. 5) using the IoT frameworks.

2 Related Work

The IoT frameworks within the related fields mainly smart cities are involved in this section as the IoT frameworks as well as platforms are always non-existent or deficient for agriculture. The smart city-built IoT frameworks are mainly considered excluding the additional domains like healthcare, smart homes, sports, inclusive sensing and so on due to the smart farming necessities in common with the urban environments such as scalability, heterogeneous data streams assistance, numerous performers or operators, real-time investigation as well as the conclusion, decision support in addition to realistic services from embedded sensors.

On the basis of IoT, a management information system architecture on behalf of intelligent agriculture is proposed in [10] concerning the IoT frameworks within the farming sector. Even if the data system is described in this paper, the details regarding the integration of the IoT is not provided. An absolute interoperable IoT-based smart farming ecosystem, semantics of information, and ontologies are designed that involves the description of the relation of the data.

The alerting conditions within the farm are defined and the data within the stream management system is published by the implementation of the semantic web technologies by Taylor et al. [3]. Depending upon the Semantic Sensor Network (SSN) ontology [11], the translation of the sensor data summarization is done towards the RDF by implementing a Global Sensor Network (GSN) [12] middleware. Though the existing methods are optimistic, they are still immature and extracts frequently.

The most extensive, powerful, and scalable frameworks are identified by concentrating upon smart cities and provide enhanced efficiency as well as scalability. The ODAA platform [4] is included in this framework which provides a direct data accession for the aggregated IoT data from the City of Aarhus, and Spitfire [5], where a uniform way is provided for the searching, interpreting and transforming the sensed data by using semantic techniques. The real-time IoT data analytics are supported by IBM Star City [6] for detecting the event with reference to the traffic domain. A service-oriented method and a complicated event processing for on-demand discovery and integration of urban data streams are combined in CityPulse [7]. The semantically annotating sensory data streams are discovered by the software platforms in OpenIoT [13] and IoT-A [14]. The complicated detection of an event within the large heterogeneous systems is processed by providing an event-driven middleware by PLAY [8].

The existing systems are allocated for urban atmospheres, intended for city-specific events as well as notification patterns, devices in addition to service

discovery, fault tolerance, and reusability even though several features and services for semantic-based sensory data processing and analysis are provided by them.

Lastly, a powerful set of APIs is provided by the FIWARE platform [15] that reduces the smart application's development in various vertical sections. The proposed method performs efficiently that brings the IoT within the agricultural sector by assuming similar directions as [3, 10]. Moreover, depending upon the IoT standards and semantic web technologies, this paper considers the developments made in the smart city frameworks [7, 13–17] which reuse certain precise software features and proposes the Agri-IoT as a complete, interoperable, flexible and adaptable, large-scale data analytics framework for smart farming.

3 Agri-IoT Framework

A prefilled Fig. 1 represents the Agri-IoT data analytics platform in which several layers such as low layers (device, communication planes), intermediate layers (data, data analytics) as well as high layers (application, end-user planes) are included. Concerning data achievement, modeling, investigation, or visualization, certain activities are performed by multiple software components at every single layer. The smart farming requirements and scenarios that are discussed in the previous section are covered and the re-using of the components is considered here rather than a reinvention of the wheel in accordance with the specific requirements of smart farming due to the presence of specific related software features that are developed in IoT and smart city-related projects [7, 13–15, 18–21].

Using the individual open API, every single software component operates by means of a particular entity and on the basis of their specific needs, a flexible distributed architecture is provided in which the components from various layers are integrated by the applications. Therefore, the components are ready to use and are implemented optionally in accordance with the necessities of specific agricultural applications. Hence, on the basis of the presented ecosystem shown in Fig. 1, the selection and combination of numerous software components are done which are suitable for the smart agriculture domain that includes the Agri-IoT framework as shown in Fig. 2.

A wide-ranging cross-domain streaming data source can be integrated, manipulated, and processed using Agri-IoT easily and standardized methods are used in order to acquire the data following IoT principles, employing semantics.

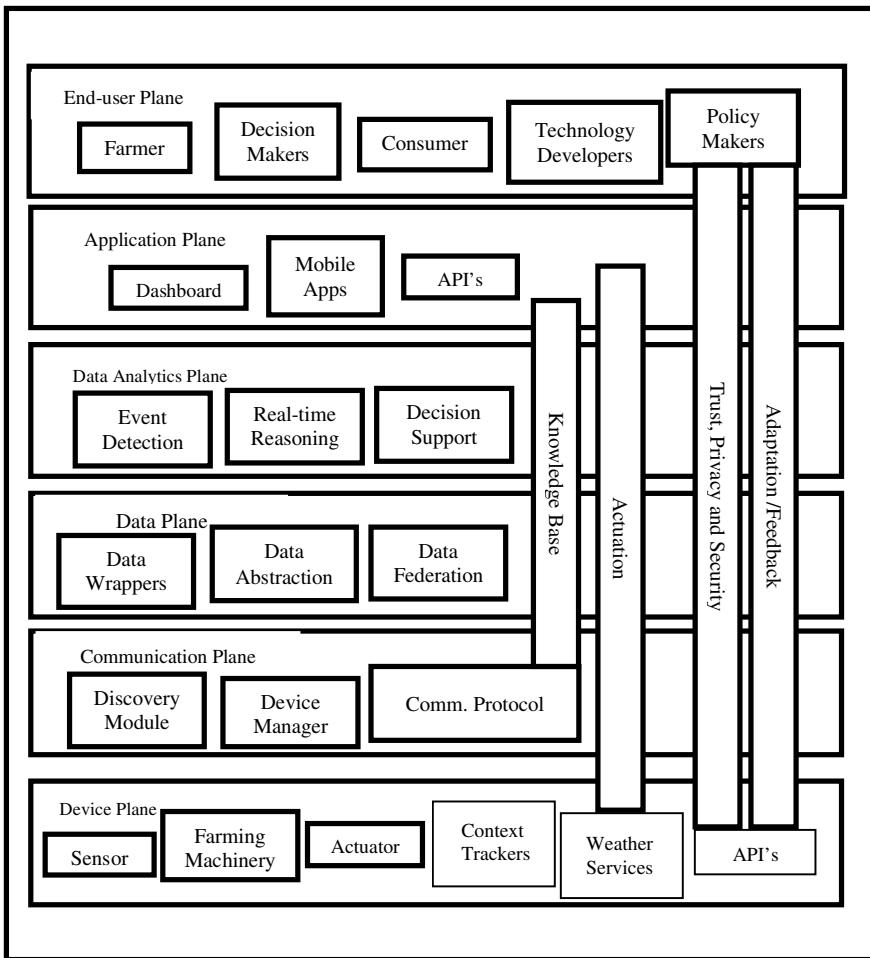


Fig. 1 The Agri-IoT ecosystem layered architecture

3.1 Main Components

Figure 2 given below represents the main components of the Agri-IoT framework.

- (a) Data wrapper: With the help of sensory metadata, the features of the sensors are discussed generally regarding the data stream. The parsed sensory data is annotated using a semantic annotation component.
- (b) Device manager: The IoT devices are managed inevitably by excluding the human operator's necessity and provides essential equipment to manage the autonomic procedures and for enforcing the decisions finally. By considering the data stream's accuracy and fault recovery, the identity and authorization of the device are managed.

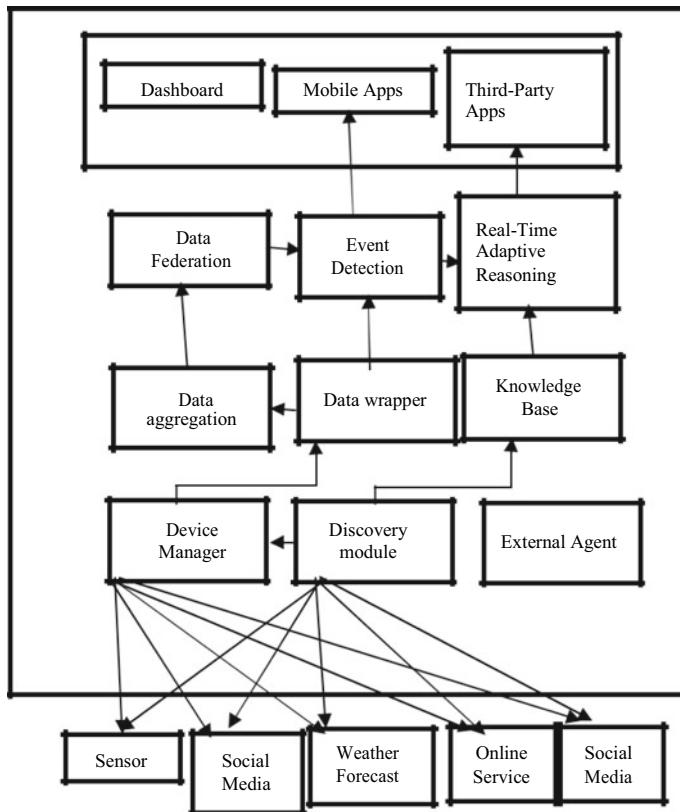


Fig. 2 Agri-IoT architecture

- (c) Discovery module: The scalable registration as well as IoT devices and service's detection is ensured using plug-and-play in reality. The location of the devices can be at similar places or accessible distantly over the internet or web.
- (d) Data aggregation: The raw sensory observation's dimension which is supplied through the data wrappers is reduced by the huge amounts of data along with time series analysis as well as data compression methods.
- (e) Data federation: The queries of the users are answered, for instance, the quantity of the fertilizer essential for applying over a particular region. In accordance with the necessities listed in the request, the related streams are found using this feature. Later, the requests from the users are translated into RDF Stream Processing (RSP) queries and the results are obtained by evaluating the queries. It is clear that appropriate technology is RSP since rapid variations in the data from real-world sensors and online services, as well as real-time processing and analytics depending upon the semantics, is involved in the IoT-based smart farming. CSPARQL [22] and CQELS [23] support the RDF-based reasoning

- those are the RDF query languages which manage the data streams that are unceasing.
- (f) Event detection: The farm events like irrigation necessity, sick animals, or pest identification in crops, is obtained here by providing the essential equipment for the data streams that are aggregated and for handling the annotated.
 - (g) Real-time adaptive reasoning: The optimum decision support in real-time is provided by considering the preferences of the farmers in addition to the dynamic contextual farm-related information (in the form of real-time events). Depending upon the condition of the farms, the reliable, accurate as well as rapid decision making for the farmer is provided as measured using the smart sensing procedure that is present.
 - (h) External agent: For the virtualization of the objects, facilities, approaches, and procedures, by assuming the identity of the user as well as authorization, a significant role is played by addressing the interoperability, device heterogeneity, data processing as well as protocol adaptation.
 - (i) Dashboard: The instant and intuitive visual accession is provided for the processing outcomes and for the investigation of the data and events.
 - (j) Mobile Apps: It is constructed upon the top of the other elements just like the dashboard. Many facilities are offered by using the APIs for the mobile users and moreover for the farmers on behalf of the real-time data in rapid decision making. Moreover, the consumers and transport agents are offered for more transparency at the sales points.
 - (k) Knowledge base: For the sensor or data stream detection, the service metadata is provided here.

3.2 Semantic Annotation

The data streams are annotated semantically by implementing the lightweight data patterns by the Agri-IoT which are on the basis of significant patterns like the SSN [11] and OWL-S [24] ontologies. The streams pertaining to the sensors which are deployed within the farm by the Stream Annotation Ontology1 are described and the events related to the farm are identified by the Complex Event Ontology2. AGROVOC [25] and the Agricultural Ontology Service (AOS) [26] are popular vocabularies as well as ontologies, in relation to agriculture. An additional ontology targeting farming constructed by the terms of the agriculture as well as lifecycles which includes seeds, grains, transportation, storage as well as consumption is AgOnt [27].

The data streams of the real-time sensors in addition to the metadata are described by the previous ontologies with the help of AgOnt for the agricultural domain. The probable events that occur at the farm are referred by the agricultural products of the metadata. The heterogeneous data streams can be easily processed by using Stream Processing (RSP) techniques, i.e., CSPARQL [22] and CQELS [23] whenever the annotation of the real-time data streams is done semantically. The query patterns over

static or dynamic information are evaluated using SPARQL-like query languages by the RSP and the on-demand stream discovery is facilitated.

Various sensors placed at the field will start sending the data to the cloud. This cloud would give various facilities to the data, such as semantic annotations, web storage through interoperability by semantic as well as the big data services of statistics. The cloud of intelligent health would extract the keywords from the data of sensors and this recommends the necessary action to the farmer. This model that has been suggested has two primary constituents that are UI, Semantic interoperability along the services of big data that are distributed. The farmer is in contact only with IoT devices in UI (User Interface). He would monitor the sensors data irrespective of the place and time with any hardware which has some relevance. The devices of IoT of heterogeneous would take diverse readings in segment of UI and then transmit to subsequent semantic interoperability along with services of big data section that has been distributed. Further, this segment has been separated into three subsections that are cloud services, big data analytics as well as semantic interoperability. The primary working of the suggested model has been operated in this segment.

The service of the intelligent cloud has been presented as an online provision of storage for the gathered data. This collected data during devices of heterogeneous IoT comprise the raw statistics along with several keywords that have been saved in the cloud system. Openly, Semantic interoperability communicates with the UI section on the cloud of intelligent. Interoperability amongst devices of IoT from several other traders is a big challenge. No standards that are identified globally for interoperability along with devices of heterogeneous IoT. Semantic interoperability is the swapping of data by means of important and apparent meetings. This consists of semantics within the data “self-described” addition bundles of data. IoT devices capitulate the information from UI and later include the semantic annotations along with semantic interoperability on the cloud system for making it crucial with distributed terminologies. Analytics of data is the scientific process of data that would pull the evocative inferences from raw facts. It would fetch money-saving and fast decision making help for processing the data. Data analytics would reveal patterns that have been hidden in the conception of data for making healthier and sensible decisions in trade. SPARQL query has utilized to discover the hidden patterns from data that have been semantically annotated. Data analytics method has been applied on pooled data from IoT devices and later applies annotations of semantic for making it more descriptive and price effective as is displayed in Fig. 3.

Data semantic annotations utilizing IoT devices of Heterogeneous Intelligent cloud gives semantic interoperability to the statistics gathered from the devices of IoT. The annotations of RDF data have been required for resolving the semantics interoperability challenge in devices of heterogeneous IoT. IoT devices are associating with the sensors. And the sensor has been utilized the network API for monitoring the data in the section of the sensor network. And later, this data would transmit to SaaS (Software as a Service) section of cloud. SaaS is a model of software allocation that would give the third-party submissions existing with the farmers over the internet. SaaS has been utilized for receiving the data of farms from IoT network of sensors and save on the cloud system would be obtainable for additional procedures. Such particular data

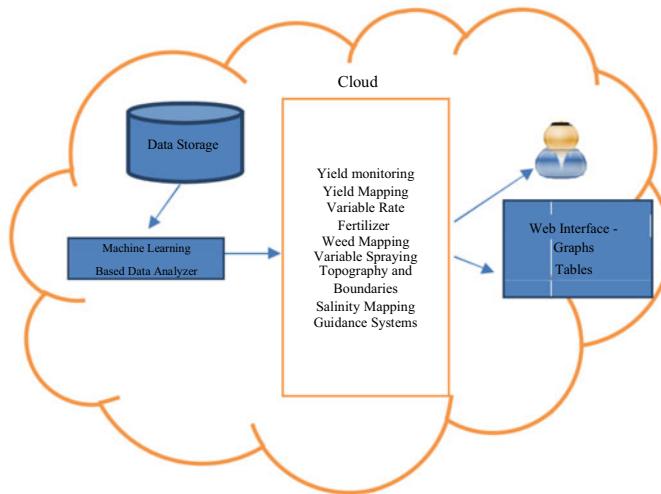


Fig. 3 IoT cloud structure for proposed Agro IoT framework

comprises values gathered from the sensors of IoT of heterogeneous over the cloud system. The statistics comprise of parameters like temperature, humidity, soil moisture, etc. which utilized as the keywords. This dataset will be transmitted to big data semantic of the interoperability segment. All the keywords have been obtained from the section of cloud that has been categorized in accordance with received parameters. The standard metadata model RDF that has utilized URIs' (Uniform Resource Identifier) for semantically related elements. The data which has been annotated was passed to the unit of big data for analysis about concluding as well as analysis of data. The triple store is utilized for storage space and triples' retrieval by semantic queries. SPARQL has been operated for extracting the particulars semantically from the triple store regarding the crops.

4 Experiments and Results

Please make SPARQL query has been utilized for discovering the patterns that have been hidden in a huge graphical database volume. The SPARQL query 1 has been utilized to take out the exclusive records from the graphical record.

Fig. 4 Sensor data

1. Air Temperature
2. Humidity and Pressure
3. Soil Temperature / Moisture
4. Leaf Wetness
5. Solar Radiation - PAR
6. pH Sensor
7. Wind Vane
8. Pluviometer (precipitation)
9. Luminosity
10. Ultrasound (distance measurement)

```
select ?parameter?value where
```

```
{ { ?parameter?value
```

} } SPARQL Query 1

The query output has been displayed in Fig. 4, where attributes are taken out from graphical database of RDF. SPARQL query has been utilized for extracting those characteristics that are authenticated.

5 Conclusion and Future Work

We suggested a model for semantic interoperability amongst heterogeneous devices of IoT in agriculture in this article. The primary objective of the suggested representation is to grant the semantic interoperability in big data amongst heterogeneous devices of IoT by data model of RDF. The data from the sensors is explained through RDF schema to readable of semantically. IoT devices acquire diverse physical parameters from the farm and after applying big data analytics to annotated dataset; it would suggest the action to be taken by the farmer. Farmers can monitor the farm, anytime utilizing any hardware. Then, the annotated details express to intelligent cloud that has been recommended actions has prescribed. The offerings of RDF graph parameters catalog in triples that are semantically comprehensible and also utilizing SPARQL queries we possibly might acquire any figures regarding the farm. Farmers could ask from the device of IoT any time regarding the available circumstance of the farm from database distantly. The end users would never concern regarding the distance, time, and hardware. The recommended form could be developed in upcoming days for providing the syntactic interoperability amongst the heterogeneous devices of IoT. Syntactic interoperability would provide the syntactic planning of the specified guidance.

References

1. Bongiovanni R, Lowenberg-DeBoer J (2004) Precision agriculture and sustainability. *Preci Agricu* 5(4):359–387
2. Gershenfeld N, Krikorian R, Cohen D (2004) The Internet of Things. *Sci Am* 291(4):76–81
3. Taylor K et al (2013) Farming the web of things. *Intell Syst* 28(6):12–19
4. Open Data Aarhus (2016) <http://www.odaa.dk>
5. Pfisterer D et al (2011) SPITFIRE: toward a semantic web of things. *IEEE Commun Mag* 49(11):40–48
6. Lecué F et al (2014) Smart traffic analytics in the semantic web with STAR-CITY: scenarios, system and lessons learned in Dublin City. *Web Semant Sci Serv Agents World Wide Web* 27:26–33
7. CityPulse EU FP7 Project (2016) <http://www.ict-citypulse.eu/page/>
8. Stuhmer R et al (2013) PLAY: Semantics-based event marketplace. *Collaborative Syst Reindustrialization* 699–707
9. senseFly SA (2016) <https://www.sensefly.com/>
10. Yan-e D (2011) Design of intelligent agriculture management information system based on IoT. In: Proceedings of ICICTA, pp 1045–1049
11. Compton M et al (2012) The SSN ontology of the W3C semantic sensor network incubator group. *Web Seman Sci Serv Agents World Wide Web* 17(1):25–32
12. Aberer K, Hauswirth M, Salehi A (2006) A middleware for fast and flexible sensor network deployment. In: Proceedings of VLDB. VLDB Endowment, pp 1199–1202
13. OpenIoT EU FP7 Project (2016) <https://github.com/OpenIotOrg>
14. IoT-A EU FP7 Project (2016) <http://www.iot-a.eu/public>
15. FIWARE. The FIWARE Catalogue (2016) <http://catalogue.ifiware.org/>
16. Kamilaris A et al (2011) Bridging the mobile web and the Web of Things in urban environments. In: Proceedings of IoT 2010, first workshop on the urban Internet of Things, Tokyo, Japan
17. Kamilaris A et al (2012) Integrating web-enabled energy-aware smart homes to the smart grid. *Int J Adv Intell Syst* 5(1)
18. ThingSpeak (2016) <https://thingspeak.com/>
19. freeBoard (2016) <https://freeboard.io/>
20. Map your meal Europe Aid funded project (2016) <http://www.mapyourmeal.org/>
21. FoodLoop GmbH. FoodLoop (2016) <https://www.foodloop.net/en/>
22. Barbieri DF et al (2009) C-SPARQL: SPARQL for continuous querying. In: Proceedings of WWW. ACM, pp 1061–1062
23. Le-Phuoc D et al (2011) A native and adaptive approach for unified processing of linked streams and linked data. In: Proceedings of ISWC. Springer, Berlin, pp 370–388
24. W3C. OWL-S: Semantic Markup for Web Services (2016) <https://www.w3.org/Submission/OWL-S/>
25. FAO. AGROVOC Thesaurus (2009) <http://www.taxobank.org/content/agrovoc-thesaurus>
26. Lauser B et al (2006) From AGROVOC to the agricultural ontology service/concept server. An OWL model for creating ontologies in the agricultural domain. In: Dublin Core conference
27. Hu S et al (2010) AgOnt: ontology for agriculture internet of things. In: Computer and computing technologies in agriculture IV. Springer, Berlin, pp 131–137

Boosting Approach for Multiclass Fake News Detection



Rajkamal Kareddula and Pradeep Singh

Abstract In the modern era of information, data integrity is of utmost priority. With the rapid development in the field of Artificial Intelligence, one who has credible data owns the key to build a reliable future. But with the breakneck development of communication over social media the reliability of data is no more guaranteed. “Fake News” is data that doesn’t have any real-world significance (or) a fact which has been modified by some middleman over the chain of communication. Spreading of such fake news affects humanity in various unacceptable perspectives. As a solution, in this paper, a machine learning approach is proposed to verify the trustworthiness of news. Instead of just classifying the data as true or fake, various degrees of truth and falsehood are also explored. The proposed methodology has been applied to “Liar, Liar Pants on Fire”, a benchmark data set for fake news detection. The proposed approach with 41.1% accuracy, outperforms the baseline approaches.

Keywords Fake news detection · Machine learning · XGBoost · TF-IDF

1 Introduction

The modern world is completely data-driven. Spreading fake news may lead to some unacceptable situations which might include taking down a whole business, a threat to people living, damage the well-balanced social-economic life cycle, and many more. Fake news in any form is a serious threat to humanity. The advancements in social media are acting as a catalyst in spreading fake news making it a critical issue of the highest priority and for which an immediate solution is required.

The early and prominent work on fake news detection was done by Jin et al. [1], they made use of a hierarchical propagation model for credibility evaluation.

R. Kareddula (✉) · P. Singh

Computer Science and Engineering, National Institute of Technology, Raipur, India
e-mail: rajkamalk99@gmail.com

P. Singh
e-mail: psingh.cs@nitrr.ac.in

The further studies have been conducted by Conroy et al. [2], they made a study on automatic deception detection for detecting fake news and proposed various linguistic cue and network analysis approaches for veracity assessment.

Fake news might include faulty data from political, educational, entertainment, industrial, or a combination of these and many other domains. In this paper, fake news detection in the political domain has been deeply explored. For this purpose, “Liar, Liar Pants on Fire” [3] a labelled benchmark data set for multiclass fake news detection has been used. This data set contains statements given by formal officials as well as statements that are taken from social media like Twitter. Along with each statement, the Metadata related to that statement such as the name of the speaker, designation of the speaker, subject(s) of the statement, speaker state information, speaker party affiliation, venue at which the statement was given, and the total credit history counts of the speaker are also given. The label has six fine-grained truthfulness ratings: pants-fire, false, barely-true, half-true, mostly-true, and true, this division of truthfulness is inspired by the work done by Rubin et al. [2]. Except for the pants-fire class, the overall data set is completely balanced.

As a solution addressing the above-stated problems of fake news, a machine learning model involving the eXtreme Gradient Boosting algorithm (XGBoost) [4] is proposed. The XGBoost algorithm uses boosting an ensemble learning technique in machine learning. To ensure the rationality of the proposed approach, Pipe-lining has been used to prevent data leakage during training and testing phases. The rest of the paper is organized as follows. Chapter 2 explains the currently existing work on fake news detection, chapter 3 explains the proposed methodology, chapter 4 presents the results obtained with our approach and its comparison with other existing approaches, chapter 5 contains the conclusion and references used.

2 Literature Overview

Various research works have been carried out for fake news detection from the past few years., The prominent among them are, Automated fake news detection in social networks by Tacchini et al. [5], they proposed two classification approaches one based on Logistic Regression and the other based on a novel adaptation of Boolean crowd-sourcing algorithms. Shu et al. [6], present a comprehensive review of detecting fake news on social media, fake news characterizations, and existing algorithms from a data mining perspective. Granik and Mesyura proposed an approach for fake news detection using the Naive Bayes Classifier in [7].

The dataset “Liar, Liar Pants on Fire” by Wang [3] is a labelled multiclass data set containing statements gathered from the PolitiFact site, this is one among the few publicly available benchmark data sets for multiclass fake news detection. A hybrid deep model approach by Ruchansky et al. [8], uses CSI (Capture, Score, and Integrate) methodology through deep learning for detecting fake news. A Deception Detection for news by Rubin et al. [2] provides a detailed study of three types of fake news. Multi-Source multiclass Fake news detection by Karimi et al. [9], proposes

an approach as a combination of automated feature extraction, multi-source fusion, and automated degrees of fake news detection. Sieving Fake News from Genuine: A Synopsis, by Alam and Ravshanbekov [10] gives a good review of fake news and the currently available approaches for fake news detection. Combating Fake News with Adversarial Domain Adaptation and Neural Models by Xu [11] proposes a deep learning approach for detecting the fake news. In their paper [12], Shu et al. explore a trio-relationship between publisher bias, news stance, and relevant user engagements and propose a Trio-Relation fake news detection framework, Thorne et al. [13] propose an approach involving stacking ensemble of classifiers for fake news detection.

3 Proposed Methodology

Three major challenges need to be solved while designing an end-to-end solution for fake news detection. They include being able to realize the importance of each word in the given news, most of the data is real-world data and it won't be available in a well-formatted way, choosing an appropriate model and tuning its hyperparameters to achieve maximum efficiency.

The designed approach solves all the three major problems stated above and gives an end-to-end solution for fake news detection. The proposed methodology can be divided into different segments as follows:

3.1 Data Set and Preprocessing

There are quite a few labelled and well-balanced multiclass data sets available currently and “Liar, Liar Pants on Fire” [3] being the major one among those. The data set proposed by Wang [3] is considered as a benchmark data set for fake news detection. This data set contains a total of twelve features and a multiclass target variable. The statements in this data set are gathered from the PolitiFact website and most of them are statements given by formal officials around the world and some being taken from the social media sites.

In the data preprocessing phase, initially, dropping of duplicate statements has been done, followed by dropping rows having null values. Since most of the data is real-world data it should be well-formatted which involves converting the text to string format and finally into Unicode format for further proceedings. In the next step, one-hot encoded vectors were generated for the target variable. This preprocessing phase also includes the tokenization of sentences, removal of stop words, position tagging, and filtering out only verbs, adjectives, and nouns. On the resulting tokens, lemmatization is performed to reduce each word to its base form. The scaling of numerical features is not required because the use of XGBoost [4], will automatically scale them. The Natural Language Toolkit [14] has been used for preprocessing

textual data, the NLTK library provides a very rich set of features to handle the textual data. The whole data set has been split into training, validation, and testing data having 80%, 10%, and 10% data, respectively.

3.2 TF-IDF Vectorizer

Ramos [15] uses TF-IDF to determine the word relevance in the text documents which will be very useful to find the importance of each word in the given news. The pre-processed data in the previous step is passed to the TF-IDF vectorizer available from sklearn API [16] to get the TF-IDF score of each word which is eventually used to form the TF-IDF vectors for each sentence. For generating the TF-IDF vector of entry the corresponding statement along with its Metadata is considered as a single document so that the Metadata related to that statement will also contribute to generating the TF-IDF vector of that statement. The TF-IDF score of a word signifies the importance of that word. The TF-IDF score is computed in two steps, first is term frequency, which is the number of times the term t appears in a document to the total number of terms in the document. Second, compute the Inverse Document Frequency which is the total number of documents to the number of documents in which the term t is present. Term frequency measures how frequently the term is appearing in that document and Inverse Document Frequency measures how frequently the term is appearing in all the documents. The final TF-IDF score is computed by multiplying the Term Frequency and Inverse Document Frequency.

$$\text{Term Frequency} = \frac{\text{No.of instances of term } t \text{ in the document}}{\text{Total No.of terms in the document}} \quad (1)$$

$$\text{Inverse Document Frequency} = \frac{\text{No.of documents}}{\text{No.of documents where } t \text{ appears}} \quad (2)$$

$$\text{Tf - IDF} = \text{Term Frequency} \times \text{Inverse Document Frequency} \quad (3)$$

3.3 Pipe-Lining

The use of pipelines ensures the efficient and safe usage of data. Pipelines are abstract structures that are generally used to automate the work-flow. Pipelines implement the fit method which in turn fits the input data sequentially to a series of transformers, in sklearn [16] convention a transformer is a class that implements both fit and transform methods. The output of one transformer is passed on as an input to the next transformer in the sequence and the output of the last transformer is the final

result. The main advantage of pipelines is that the data leakage won't happen, the training and testing phases are handled separately by different pipelines.

Pipelines for handling the text data and numeric data have been built separately. The individual text and numeric pipelines have been combined by using feature union which combines the outputs generated by multiple transformers (pipelines in this case). The resultant combined features are put in a pipeline along with the machine learning model, building a pipeline having nested pipelines for generating the required pre-processed data and the final model to take the generated data as input and perform the classification.

3.4 Model

In the proposed approach eXtreme Gradient Boosting model (XGBoost) [4] has been used. In recent days XGBoost has been gaining popularity due to its ease of use and efficiency. XGBoost has been successfully employed to solve the problem of classification as well as regression with state-of-art results [17]. XGBoost uses boosting which is an ensemble learning methodology in which a number of weak learners are trained sequentially with the main objective of reducing the variance. The weak learners are trained iteratively to build strong learners upon them. Weak learners are generally decision trees with each tree initialized with a different set of parameters. Boosting applies iteratively by adjusting the weights of the weak learners and learning from classification mistakes made in the previous iteration.

3.4.1 XGBoost

XGBoost [4] is a supervised ensemble model that implements the gradient boosting decision tree algorithm. XGBoost uses gradient descent technique to reduce the loss when adding new models to the existing sequence of weak learners during boosting. It uses the approximation on the split points so that it selects only the top splits and ignores the others. By the use of this approach, it runs almost ten times faster than the traditional gradient boosting algorithms. A binary tree has been chosen as a weak learner for boosting which works similarly to a decision tree. The problem of fake news detection is something that is related to working on real-world data which might be very noisy and may contain potential outliers. XGBoost works fine even in case of noisy data and can even take care of outliers, outliers in this case are TF-IDF scores which are too high or which are too low. Other algorithms using boosting techniques such as AdaBoost [18] and Gradient Boosting Trees can't withstand the noisy data since they try to fit out every single feature and will be largely affected by the outliers too. Taking these problems into consideration XGBoost has been chosen as the final model. In order to avoid the over-fitting of XGBoost, early stopping rounds have been used.

Grid search optimization has been performed to tune the hyperparameters of XGBoost. After tuning XGBoost on training data the following hyperparameters have been achieved: subsample: 0.8, learning_rate: 0.09, max_depth: 6, n_estimators: 80, scale_pos_weight: 1, min_child_weight: 12, max_delta_step: 2, gamma: 0, colsample_bytree: 0.9, colsample_bynode: 1, colsample_bylevel: 1, early_stopping_rounds: 5.

Figure 1 shows the complete context flow of the proposed approach. Making use of pipelines makes sure that the training and testing data are never merged, not even in the preprocessing phase, hence data leakage won't happen. The generated TF-IDF vectors are used as input features for XGBoost. The XGBoost model is trained with training data and validated on the validation data in each boosting round. Once the whole training and final validation phases are completed, predictions are made on the hold-out test set.

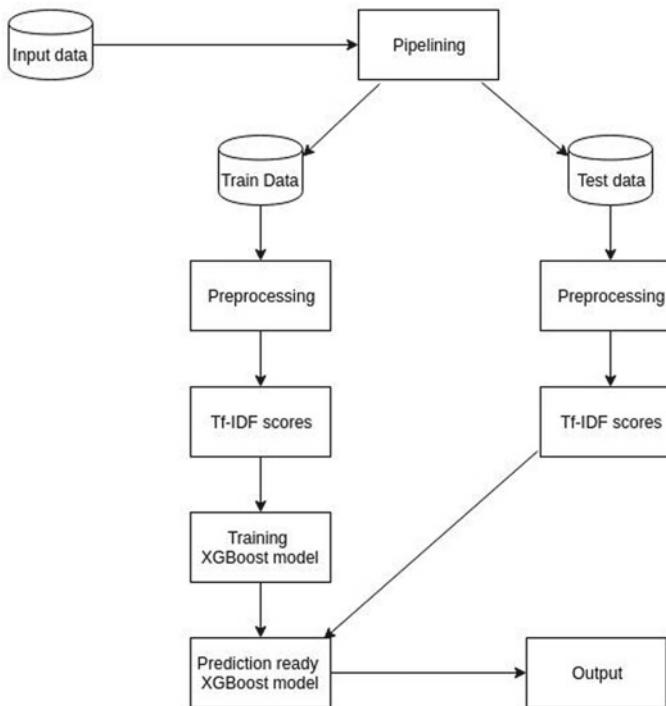


Fig. 1 The overall control flow of the proposed approach

Table 1 Comparison of results

Model	Accuracy
SVM [3]	25.5
Random forest [9]	27.6
Basic neural network [9]	28.1
Wang [3]	27.4
MMFD [9]	38.8
Logistic regression [3]	24.7
CNN [3]	27.0
Proposed approach	41.1

4 Results

For the evaluation of the proposed methodology and comparison with other available approaches, accuracy has been used as a metric. Accuracy is a fraction of number of correct predictions made to the total number of predictions. Accuracy is calculated using the True positive, True negative, False positive, and False negatives as follows:

where

$$TP = \text{Number of True Positives}$$

$$TN = \text{Number of True Negatives}$$

$$FP = \text{Number of False Positives}$$

$$FN = \text{Number of False Negatives}$$

$$\text{Accuracy} = \frac{TP + NP}{TP + TN + FP + FN} \quad (4)$$

To prove the effectiveness of the proposed approach, it has been compared with seven baseline approaches, which include Support Vector Machines (SVM) [3], Logistic Regression [3], Basic neural network [9], Wang hybrid model [3], Multi-Source **Multiclass** Fake News Detection model (MMFD) [9], Random Forest [9] and Convolution neural networks [3]. The results of the comparison are shown in Table 1. Similar to the other reported results and experiments conducted we also performed the comparison presented in Table 1 on the same pre-split test set having 10% of the dataset.

5 Conclusion

Due to the rapid development of communication over social media, the spreading of fake news has become a critical issue having the utmost priority. To address this problem, in this paper, an XGBoost based model for multiclass Fake News detection

has been proposed. The whole experimental work has been conducted using the publicly available “Liar, Liar Pants on Fire” [3] data set. The proposed approach has been compared with various existing works and it is evident from the results that the proposed model has significantly outperformed all other compared models when trained on the same data set. Future work for this paper includes employing advanced deep learning techniques which will preserve the contextual meaning of the text for even better reliability. In addition, nested ensemble architectures such as ensemble XGBoost, having XGBoost itself as a weak learner during boosting would be a great concept to explore.

References

1. Jin Z, Cao J, Jiang Y, Zhang Y (2014) News credibility evaluation on microblog with a hierarchical propagation model. In: 2014 IEEE international conference on data mining. IEEE, vol 2015, pp 230–239
2. Conroy NJ, Rubin VL, Chen Y (2015) Automatic deception detection: methods for finding fake news. Proc Assoc Inf Sci Technol 52(1):1–4
3. Wang WY (2017) Liar, liar pants on fire: a new benchmark dataset for fake news detection. In: Proceedings of the 16th international conference (long paper), ACL 2017—55th annual meeting of the association for computational, vol 2, pp 422–426
4. Chen T, Guestrin C (2016) Xgboost: a scalable tree boosting system. In: Proceedings of 22nd acm sigkdd international conference on knowledge discovery on data mining. ACM, vol 42, no 8, 2016, pp 665
5. Tacchini E, Ballarin G, Della Vedova ML, Moret S, de Alfaro L (2017) Some like it Hoax: automated fake news detection in social networks. In: CEUR workshop proceedings, vol 1960, pp 1–12
6. Shu K, Sliva A, Wang S, Tang J, Liu H (2017) Fake news detection on social media: a data mining perspective. In: ACM SIGKDD Explor Newsl 19(1):22–36
7. Granik M, Mesyura V (2017) Fake news detection using naive Bayes classifier. In: Proceedings of 2017 IEEE 1st Ukraine conference on electrical and computer engineering UKRCON 2017, pp 900–903
8. Ruchansky N, Seo S, Liu Y (2017) CSI: a hybrid deep model for fake news detection. In: International conference on information and knowledge management, vol Part F1318, pp 797–806
9. Karimi H, Chandan Roy P, Saba-Sadiya S, Tang J (2018) Multi-source multi-class fake news detection. In: Proceeding of 27th international conference on computational linguistics, pp 1546–1557
10. Alam S, Ravshanbekov A (2019) Sieving fake news from genuine: a synopsis. arXiv Prepr. arXiv1911.08516
11. Xu B (2019) Combating fake news with adversarial domain adaptation and neural models. Dissertation Massachusetts Institute Technology 2019, no. 2018
12. Shu K, Wang S, Liu H (2019) Exploiting tri-relationship for fake news detection. In: WSDM 2019—proceedings of 12th ACM international conference on web search and data mining, pp 312–320
13. Thorne J, Chen M, Myrianthous G, Pu J, Wang X, Vlachos A (2018) Fake news stance detection using stacked ensemble of classifiers. pp 80–83
14. Loper E, Bird S (2002) NLTK: the natural language toolkit. arXiv Prepr. cs/0205028 (2002), 2002
15. Ramos J (2003) Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning, vol 242, no. 1, pp 29–48

16. Buitinck L et al (2013) API design for machine learning software: experiences from the scikit-learn project. arXiv Prepr. arXiv1309.0238 (2013), pp 1–15
17. Friedman J (2001) Greedy Function approximation: a gradient boosting machine. Ann Stat 29(5):1189–1232
18. Dietterich TG (2000) Ensemble methods in machine learning. Lecture notes computer science (including subseries lecture notes in artificial intelligence in lecture notes bioinformatics), vol 1857 LNCS, pp 1–15

A Sentiment-Based Recommender System Framework for Social Media Big Data Using Open-Source Tech Stack



Shini Renjith, Mable Biju, and Monica Merin Mathew

Abstract The large volume of data getting generated on the internet has become an opportunity for data analysts to retrieve information and perform decision-making. However, the extraction of relevant information has turned out to be an extremely difficult task, especially when dealing with big data sources like social media. Intelligent approaches like recommendation systems came into existence to deal with such situations. These systems require the capability to deal with various big data sources like user-generated content—reviews, comments, ratings, likes, etc. This work proposes a four-stage recommender system architecture for big data processing using open-source technology stack. The key objective of our proposed architecture is to ensure an efficient and robust recommendation system with excellent efficiency.

Keywords Big data · Recommendation system · Open-source · Natural language processing · Machine learning

S. Renjith · M. Biju · M. M. Mathew

Department of Computer Science and Engineering, Mar Baselios College of Engineering, Thiruvananthapuram, Kerala 695015, India
e-mail: mablebj15@gmail.com

M. M. Mathew
e-mail: monica.m.mathew@gmail.com

S. Renjith (✉)
Department of Computer Applications, Cochin University of Science and Technology, Kochi, Kerala 682022, India
e-mail: shinirenjith@gmail.com

1 Introduction

Recommender system is an information filtering technique used for predicting the potential behavior of a user by processing a variety of information like past behavior traits, societal activities, contextual information, personal preferences, etc. The objective is to cater personalized and contextualized suggestions to enable better decision-making in various contexts like online shopping [1], travel planning [2, 3], customer retention [4, 5], fraud or anomaly detection [6], targeted advertisements [7], etc. Social media big data can be considered as a potential source of information in such scenarios, which provide avenues for their registered users to communicate with each other and share opinions on various by means of reviews, ratings, comments, suggestions, etc. thereby creating social media big data. Big data encompasses data in structured and unstructured forms. The typical definition of big data deals with five dimensions (5Vs) [8]. This paper proposes an open-source-based recommender system framework that is capable of handling inputs from big data sources like social media data. Section 2 of this paper analyses recent literature in this area and Sect. 3 details on the proposed architecture. Section 4 describes the implementation of information and Sect. 5 concludes the paper along with details of enhancements in consideration.

2 Related Works

There are many literatures existing on big data analysis for recommender systems using different technologies, approaches, and platforms [9]. Verma et al. [10] proposed a recommendation system build by making use of numerical data like ratings or ranks awarded to various products or services by users. Gandhi et al. [11] conducted a survey on collaborative filtering based big data recommender systems and attempted a personalized movie recommendation system based on past behavior of users. Nundlall et al. [12] came up with a hybrid recommendation approach combining collaborative and content-based filtering with sentiment analysis. Dwivedi et al. [13] proposed a recommendation system for education system domain using collaborative filtering on the grades obtained in other subjects. Ashraf [14] presented a new recommendation system for the e-commerce domain using a social network knowledge base. Amato et al. [15, 16] proposed a big data recommender system based on user interactions and generated multimedia content on social media channels. Coelho et al. [17] presented a model that exploits tweet features like URL count, favorites count, hash-tag count, count of user mentions, count of media attachments, tweet length, etc. to generate personalized travel recommendations. A feature-based dynamic system was proposed by Bafna et al. [18] for the summarization of customer reviews on online products and services. Rodavia et al. [19] presented AutoRec, a platform for automatic recommendations generation of items for its users. Lin et al.

[20] utilized customer surveys to build a review-driven recommender solution for e-commerce websites. Qu et al. [21] proposed a framework for recommending friends on social media platforms like Weibo based on Deep Graph-Based Neural Network. In their work on factoring personalization in social media recommendations, Ge et al. [22] differentiated generic and social media recommender systems based on various and derived five critical factors that influence the architecture of social media recommender systems.

3 Proposed Architecture

In this work, we are presenting an open-source-based recommender system framework capable of handling social media data as inputs. In this section, we will cover the functional architecture and the details of the proposed technology stack for implementation. We have four stages in the proposed recommender system architecture as portrayed in Fig. 1. The first stage of the architecture is designed specifically for information retrieval purposes. The next stage is called Natural Language Processing phase (NLP phase) and deals with data-pre-processing. In the third stage namely the machine

Learning phase, as the name implies, we carry out machine learning tasks like sentiment classification, popularity ranking, and recommendation generation. The final stage is mainly reserved for academic interest and various performance evaluation measures are calculated and visualized. [9, 23]. We propose to use open-source technology stack for each phase as well as for big data compatibility. We consider Apache Spark cluster computing framework [24] to provide big data compatibility and to deal with requirements of real-time analytics.

Fig. 1 Functional architecture of the proposed big data recommender system

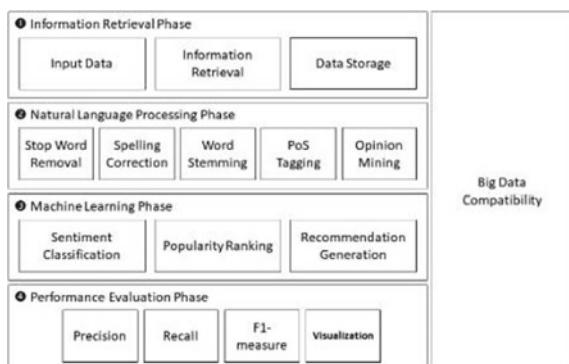




Fig. 2 Implementation architecture of information retrieval phase

4 Discussion on Implementation Aspects

In this section, we will cover the implementation aspects for building the proposed big data recommender system explained in this paper. We focus to cover the high volume of social media data available on the internet in various forms like ratings, opinions, reviews, comments, feedback, blog posts, etc. using Apache Spark Framework.

4.1 Information Retrieval Phase

The first stage of the framework covers a series of activities that represent the sequencing of data, from data entry to a document data model. The purpose is to capture a relevant subset of social media big data for generating better results through the recommendation system. Our implementation architecture for this phase is depicted in Fig. 2.

Input Data. Data entry to the framework can be from different social media channels, mostly as unstructured data. We use the Yelp dataset for validating the proposed system. **Information Retrieval.** Information Retrieval is an important phase in building big data recommender systems. Our proposed design is capable of receiving real-time streaming data using a message broker like RabbitMQ and MongoDB NoSQL database. **Data Storage.** Data preparation and data modeling are the two key activities performed prior to actual data storage into the database. During data preparation, data is cleansed and made ready for further analysis. Data Modeling is done by using MongoDB. Data is stored in the form of flexible JSON-like documents.

4.2 Natural Language Processing Phase

The purpose of the Natural Language Processing (NLP) Phase is to transform unstructured textual data to the tokenized and structured text format. However, utmost care is taken to ensure the quality and the originality of the raw data is preserved in the exact same way. The preprocessing steps are depicted in Fig. 3 and as discussed below.

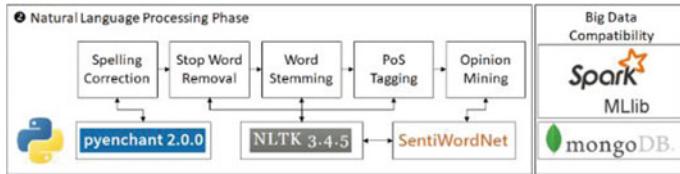


Fig. 3 Implementation architecture of NLP phase

Stop Word Removal. Stop words such as “the”, “is”, etc. do not contribute much to the meaning of the sentence. These words do not help in the text classification process. We leverage NLTK library of Python for stop word removal in our implementation.

Spelling Correction. Incorrect spellings are real prevention to concentrate on the importance of the content. Our framework utilizes the PyEnchant library, a free accessible spell checker that replaces the most appropriate right word in place of the miss-spelled one.

Word Stemming. Word stemming is a linguistic normalization method that uses morphological techniques. The process is primarily considered by removing the pre-fixes and suffixes from a word and identifying the root word.

PoS Tagging. Part of Speech tagging is a significant activity in our proposed framework. Using the NLTK module we parse each review text and sentences are extracted from it. Such sentences are further sliced and PoS tags are allocated for each word. The taggers mine out the nouns, verbs, and adjectives from the reviewers’ feedback.

Opinion Mining. Opinion mining comprehends genuine opinion inside any sentence. In our proposed framework, we utilize the NLTK interface for SentiWordNet for opinion mining. Opinion mining is performed as the last step in the NLP phase.

4.3 Machine Learning Phase

Machine Learning phase covers the key activities of sentiment classification, popularity ranking, and recommendation generation. We have populated a list of attributes that can be applied to a generic traveler. Figure 4. depicts various steps involved.

Sentiment Classification. In the Sentiment classification phase, we populate the user interest profile as a reflection of users’ sentiment or inclination towards certain attributes. The user interest profile is a matrix with each user representing one unique row in it and each item from reference attributes forming its column. The key assumption is that a user will provide feedback on an attribute only when he/she has an interest in it. **Popularity Ranking.** The results of opinion mining (polarity) are used for rank-ordering point of interest (POI) like attractions, hotels, restaurants, etc. We form a POI profile by representing each POI as a row and each attribute from reference

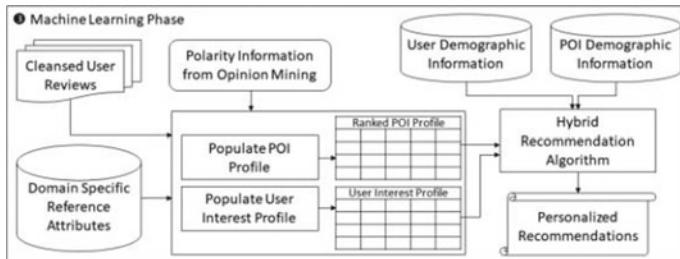


Fig. 4 Implementation architecture of machine learning phase

attributes as its column. For each sentence with a polarity, we increment/decrement the score against the relevant attribute in a weighted manner. After completing the formation for POI profile, we can rank order POIs against any of the items in the set of reference attributes. **Recommendation Generation.** The user interest profile and POI profile are the key inputs for recommendation. In addition to these two, we do leverage the demographic details of users as well as POIs. We apply a hybrid approach combining content-based, collaborative, and demographic algorithms to predict user behavior. The demographic approach is used to tackle the cold start problems in recommender systems.

4.4 Performance Evaluation Phase

The anticipated output is a list of top recommendations for a target user. The performance of a recommender system is evaluated by measuring the quality of recommendations made, by checking the closeness of estimated preferences against the actual preferences of the user. We use precision, recall, F1-measure, and accuracy as evaluation measures.

4.5 Big Data Compatibility

We leveraged the Apache Spark cluster computing framework to provide big data compatibility and to deal with real-time analytics requirements.

5 Conclusion

Social media big data has evolved as the most prominent information source for modern recommender systems. In this work, we propose a sentiment-based recommender system framework for social media text data. The architecture is highly scalable and can be easily extended to big data platforms. We propose end-to-end usage of open-source technology stacks so that researchers can easily adopt this framework without incurring licensing costs. We have successfully implemented a pilot version and it could successfully generate recommendations adopting this framework. As a subsequent step, we are planning to extend the recommendation algorithms to use dimensionality reduction to reduce computational complexity. In addition, we are exploring options to leverage deep learning approaches to generate recommendations from social media text content.

References

1. Kim K, Ahn H (2008) A recommender system using GA K-means clustering in an online shopping market. *Expert Syst Appl* 34:1200–1209. <https://doi.org/10.1016/J.ESWA.2006.12.025>
2. Renjith S, Anjali C (2013) A Personalized travel recommender model based on content-based prediction and collaborative recommendation. *Int J Comput Sci Mobile Comput ICMIC13*:66–73)
3. Renjith S, Anjali C (2014) A personalized mobile travel recommender system using hybrid algorithm. In: 2014 first international conference on computational systems and communications (ICCSC) (2014). <https://doi.org/10.1109/compsc.2014.7032612>
4. Renjith S (2015) An integrated framework to recommend personalized retention actions to control B2C E-commerce customer churn. *Int J Eng Trends Technol* 27(3):152–157. <https://doi.org/10.14445/22315381/IJETT-V27P227>
5. Renjith S (2017) B2C E-commerce customer churn management: churn detection using support vector machine and personalized retention using hybrid recommendations. *Int J Future Revolution Comput Sci Commun Eng* 3(11):34–39
6. Renjith S (2018) Detection of fraudulent sellers in online marketplaces using support vector machine approach. *Int J Eng Trends Technol* 57(1):48–53. <https://doi.org/10.14445/22315381/IJETT-V57P210>
7. Yang W, Dia J, Cheng H, Lin H (2006) Mining social networks for targeted advertising. In: Proceedings of the 39th annual Hawaii international conference on system sciences (HICSS'06). <https://doi.org/10.1109/hicss.2006.272>
8. Jagadish H, Gehrke J, Labrinidis A, Papakonstantinou Y, Patel J, Ramakrishnan R, Shahabi C (2014) Big data and its technical challenges. *Commun ACM* 57:86–94. <https://doi.org/10.1145/2611567>
9. Renjith S, Sreekumar A, Jathavedan M (2019) An extensive study on the evolution of context-aware personalized travel recommender systems. *Inf Process Manag* 102078. <https://doi.org/10.1016/j.ipm.2019.102078>
10. Verma J, Patel B, Patel A (2015) Big data analysis: recommendation system with hadoop framework. In: 2015 IEEE international conference on computational intelligence & communication technology. <https://doi.org/10.1109/cict.2015.86>
11. Gandhi SR, Gheewala J (2017) A survey on recommendation system with collaborative filtering using big data. In: 2017 international conference on innovative mechanisms for industry applications (ICIMIA). <https://doi.org/10.1109/icimia.2017.7975657>

12. Nundlall C, Sohun G, Nagowah SD (2018) A hybrid recommendation technique data systems. In: 2018 international conference on intelligent and innovative computing applications (ICONIC). <https://doi.org/10.1109/iconic.2018.8601282>
13. Dwivedi S, Roshni VSK (2017) Recommender system for big data in education. In: 2017 5th national conference on e-learning & e-learning technologies (EELTECH). <https://doi.org/10.1109/eeltech.2017.8074993>
14. Ashraf M, Choudhary D, Das R, Ghosal P (2014) An efficient and optimized recommendation system using social network knowledge base. In: 2014 international conference on advances in electrical engineering (ICAEE). <https://doi.org/10.1109/icaee.2014.6838561>
15. Amato F, Moscato V, Picariello A, Sperli G (2017) Recommendation in social media networks. In: 2017 IEEE third international conference on multimedia big data (BigMM). <https://doi.org/10.1109/bigmm.2017.55>
16. Amato F, Moscato V, Picariello A, Piccialli F (2019) SOS: a multimedia recommender system for online social networks. Future Gener Comput Syst 93:914–923. <https://doi.org/10.1016/J.FUTURE.2017.04.028>
17. Coelho J, Nitu P, Madiraju P (2018) A personalized travel recommendation system using social media analysis. In: 2018 IEEE international congress on big data (BigData Congress). <https://doi.org/10.1109/bigdatacongress.2018.00046>
18. Bafna K, Toshniwal D (2013) Feature based summarization of customers' reviews of online products. Procedia Comput Sci 22:142–151. <https://doi.org/10.1016/J.PROCS.2013.09.090>
19. Rodavia M, Ballera M, Clemente G, Ambat S (2017) AutoRec: a recommender system based on social media stream. In: 2017 international conference on platform technology and service (PlatCon) (2017). <https://doi.org/10.1109/platcon.2017.7883691>
20. Lin K, Shen C, Chang T, Chang T (2017) A consumer review-driven recommender service for web E-commerce. In: 2017 IEEE 10th conference on service-oriented computing and applications (SOCA). <https://doi.org/10.1109/soca.2017.35>
21. Qu Z, Li B, Wang X, Yin S, Zheng S (2018) An efficient recommendation framework on social media platforms based on deep learning. In: 2018 IEEE international conference on big data and smart computing (BigComp). <https://doi.org/10.1109/bigcomp.2018.00104>
22. Ge M, Persia F (2019) Factoring personalization in social media recommendations. In: 2019 IEEE 13th international conference on semantic computing (ICSC). <https://doi.org/10.1109/icosc.2019.8665624>
23. Renjith S, Anjali C (2013) Fitness function in genetic algorithm based information filtering—a survey. Int J Comput Sci Mobile Comput ICMIC13:80–86
24. Zaharia M, Franklin M, Ghodsi A, Gonzalez J, Shenker S, Stoica I, Xin R, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S (2016) Apache spark: a unified engine for big data processing. Commun ACM 59:56–65. <https://doi.org/10.1145/2934664>

Hardware Trojan Detection Using Machine Learning Technique



Nikhila Shri Chockaiah, S. K. Swetha Kayal, J. Kavin Malar, P. Kirithika, and M. Nirmala Devi

Abstract As the internationalization of Integrated Circuit (IC) production increased, the inclusion of deliberately stealthy modification called hardware Trojans has also escalated. A hardware Trojan detection method that works at the gate-level using the netlist of the circuit under test is presented in this paper. The unsupervised machine learning algorithm, K-Means classification is used for categorization. Every net of the circuit is analyzed to determine if the net is genuine or is Trojan infected by the extraction of seven relevant features from every net. The technique has been validated on ISCAS'85 benchmark circuits and parameters like true positive (TP), false negative (FN) and recall (TPR) have been illustrated.

Keywords Hardware security · Hardware Trojan · Machine learning · K-means · Classification · Feature extraction

N. S. Chockaiah · S. K. S. Kayal (✉) · J. K. Malar · P. Kirithika · M. N. Devi

Department of Electronics and Communication Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India

e-mail: swethakayal@gmail.com

N. S. Chockaiah

e-mail: shrinikhila@gmail.com

J. K. Malar

e-mail: kavinmalar.j@gmail.com

P. Kirithika

e-mail: kirithikapalanisamy@gmail.com

M. N. Devi

e-mail: m_nirmala@cb.amrita.edu

1 Introduction

The demand for built-in and high-functioning hardware devices has been on the rise. The need to mass-produce Integrated Circuit (IC) products through internationalization for cost reduction thus arises. The design and fabrication stages of the hardware devices are often allocated to third-party vendors, where an intentional third-party can implant into ICs, deliberately stealthy modification called *hardware Trojan* (HT). This emphasizes the fact that it is crucial to be circumspect to the installation of hardware Trojans at any of the different stages of the chip production life-cycle. Due to the increase in the size and the complexity of ICs, detecting hardware Trojans has become very difficult. An HT can have various impacts on a chip which include information leakage, change in functionality, Denial of Service (DoS), and reduced statistical reliability.

Areas prone to the insertion of HTs are the design phase, fabrication phase, or testing phase. Within the design phase, various IP cores, libraries, tools, standard cells, and models are used by the designers. These can be obtained from untrusted sources and can thus contain malicious components. In the fabrication phase, HTs can be introduced by modifying the parameters of the involved processes or making changes in the layout. Within the testing and deployment phase, possible attacks are the placement of HTs or the obfuscation of Trojans which were inserted during previous stages from being detected. The cost and time involved in hardware Trojan detection in the fabrication and testing phase are comparatively more than that in the design phase. The significance of the necessity to identify hardware Trojans in the design step of IC production life-cycle can thus be inferred.

In a few HT identification methods such as logic testing, side-channel analysis, and post verification, HTs are identified by observing the variations from the golden IC, where golden ICs are HT-free ICs. For instance, in side-channel analysis, a comparison of involved parameters like power consumption, delay, etc., is used to identify the existence of HTs. In a practical situation, it is not possible to assume the availability of *Golden netlists* or prior information related to the hardware Trojan that has been inserted. This highlights the need for golden IC-free HT detection methods. The aim of the proposed detection method is to get as input the gate-level netlist, extract relevant features, and identify the presence of HT in the given netlist.

Techniques used for this purpose can be broadly sorted into pre-silicon and post-silicon techniques. The former techniques analyze gate-level netlists of the circuits to establish gates or nets that are a part of the Trojan by using the values of extracted features. In contrast, post-silicon techniques target the identification of HTs in ICs after the fabrication. Detection in this phase is challenging in the context that process variations (PVs) exist. Side-channel analysis, one of the widely used detection schemes requires a golden chip. However, the invasion might be required to obtain a golden chip which might make it quite an expensive process too. Even in the presence of a golden chip, the above methods do not accurately identify the presence of HTs. One of the causes of the non-detectability is that the inserted HTs are usually very small in terms of the physical area occupied or the size of code

involved. The detection of hardware Trojans in the design step (pre-silicon) of IC production life-cycle is thus required.

It has been expected that with the development of advanced HT detection methods, the design of HTs that cannot be identified by those methods will be faster. Such HTs have been developed in the past, a hardware Trojans called DeTrust [1] was proposed immediately post the development of FANCI [2], which is a static HT detection method. To identify an effective solution to handle such a scenario, machine learning (ML) can be utilized to detect unknown hardware Trojans. It forms the basis for the proposed work here. It is executed using an ML classifier where the learning phase involves only data that is fed to it and machine learning has been used to find out malicious behaviors. Machine learning is advantageous as the classification of hardware Trojans is possible without actually simulating the considered circuit. The focus of this paper is on IC design step where hardware Trojan is placed at gate-level netlists.

In this proposed work, the hardware Trojan classification method based on a static machine learning algorithm is used to identify if HT exists in the given circuit. It is implemented to work with gate-level netlists in the absence of golden netlists.

In our approach, HTs themselves are not detected but nets where HTs are included, the *Trojan nets* are detected. In the first place, seven features of each net in the gate-level netlist having the ability to help in identifying the presence of a hardware Trojan is extracted in our method. Then machine learning is applied to these seven features of each and every given net in the netlists. Secondly, those features are represented in a vector form and learned by using a K-Means classifier. Eventually, a group of nets of the netlist can be identified as a Trojan-infected or Trojan-free net by the trained classifier.

This method works in the absence of pre-information of the HT inserted and only makes use of information that is acquired from the gate-level netlist and identifies if the given netlist is free from HT or not. Nevertheless, whether the circuit under test is HT-free or HT inserted will be recognized.

The remainder of the paper is organized hence: Sect. 2 briefs the prior works dedicated to this very issue. Section 3 primes the proposed methodology of the detection of HT. The experimental results have been presented in Sect. 4 succeeded by the conclusion in Sect. 5.

2 Related Works

To understand the motivation behind this study, the literature which discusses the classification techniques for identification and detection of HTs using a machine learning-based algorithm is presented.

2.1 Identification of HTs Through Classification

Initial attempts of using classification techniques for the identification of HTs include methods where the HTs are classified into five attributes: design stage, abstraction juncture, trigger of Trojan, effect of Trojan, and physical location [3]. This diverse set of observed HTs is then compiled and classified based on these five attributes. In [3], previous works related to the classification of HTs have been mentioned in detail. HT classification has been established using the two categories of activation mechanism, trigger, and payload of the Trojan. In later works, the classification categories are: physical structure, activation method, and consequence created. Despite the addition of two more classes to the previously proposed taxonomy, this consignment is not associated with the entire IC production life-cycle. In [4], a thorough categorization has been established into five levels: insertion level, abstraction stage, activation methodology, effects, and physical location. The categorization evaluates the life-cycle of the chip and the location of target. However, the physical characteristics of the Trojans are not considered. The taxonomy used in [4] analyzes the nature of Trojans and identifies the attributes corresponding to that particular Trojan.

A comprehensive classification based on 33 attributes which are further classified into eight major categories was introduced in [4]. The logic type of an HT, which is significant to predict the effect of an HT, is included. Thus, deciding the relevant detection methodology and diminution techniques become easier. In [4], a new approach has

been exhibited to study HT attributes and the relationship that exists between them. It is versatile and can be used with any HT classification methods. Newly discovered attributes can be easily fit in with and categorized into the proposed taxonomy.

In [5], the initial dataset of HTs is analyzed by developing a vulnerability analysis flow and detectability metric. The HTs are implemented based on the hard to detect areas that are determined by the vulnerability analysis flow. Another classification technique called score-based classification is developed in [6] to identify HT-free or HT-infected circuit without using a golden netlist. Two types of class: weak and strong are developed for score-based technique. This score-based technique asserts that it can detect all HTs in some selected benchmark circuits which are obtained from TrustHUB compared to the UCI and the VeriTrust [7] techniques which were developed earlier.

In [8], for a classification technique, SVM is used to differentiate normal nets from those affected by HTs in a set of given gate-level netlists. These extracted features are learned using SVM based on three conditions: no weighting, static weighting, and dynamic weighting. The accuracy of identifying the HTs are 80% or higher with dynamic weighting. An exhaustive classification of hardware Trojans is detailed in [9]. In [10], a side-channel analysis that uses ML for the purpose of classification is implemented.

2.2 *Classifying Using ML for the Detection of HTs*

Compared to the identification of HTs, there is more research literature on machine learning-based classification for HT detection. In [11], an SVM-based approach was developed to assist reverse engineering (RE) in detecting HTs. This approach eliminated the last two steps: annotation and schematic creation in RE. The features were extracted from the first three steps of RE without labels. To solve this, one class v-SVM is used as the class for this training sample. This type of SVM has values between ‘0’ and ‘1’ and these values were determined by the decision boundary that closely surrounds the training sample. Also in [11] reverse engineering is utilized to detect Trojan-free ICs by reformulating the problem of HT detection as an ML-based clustering problem (K-means). Images from RE/golden layout are preprocessed to obtain feature vectors. K-means clustering is performed on these features. A collation between Support Vector Machine and K-Means clustering approach to classification implied that the latter has lower parameter dependence.

Different types of hardware Trojans which differ by their trigger circuit are designed and injected into normal circuits in [12]. These inserted Trojans are then detected using a machine-learning-based technique that employs a neural network. The designed neural network comprises 11 neurons and 2 neurons in the input layer and output layer respectively. The developed algorithm is then validated on benchmark circuits. The usage of neural networks for the said purpose has been detailed in [12, 13], whereas [14] presents the use of random forest classifier and [15] an SVM classifier instead.

In [16] a comprehensive analysis of how machine learning and side-channel power analysis are integrated into the mechanism of HT recognition is presented.

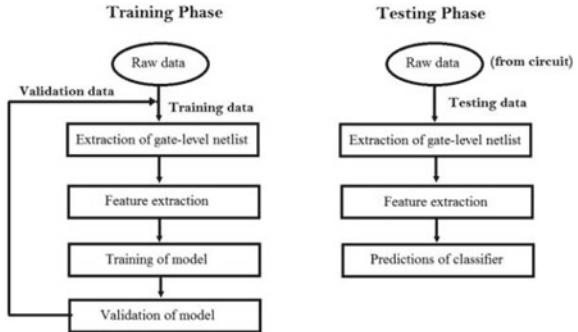
Gleaned on the above literature, it is evident that there is still lack of machine learning-based algorithms for identifying the HTs compared to detecting HTs. However, the utilization of classification methodology which involves machine learning for HT detection is complex as it is tailored to the detection techniques used.

3 Proposed Methodology

The two main processes for an ideal Trojan detector would be the careful scrutinization of features to be selected and the algorithm that would model and segregate between Trojan-infected and Trojan-free nets. For the former decisive-factor, we have intuited to gather seven features.

Benchmark circuits are formulated in Verilog-HDL have been used for analysis. Synopsys Design Compiler was then used for synthesizing by GSCLib_3.0 in 90 nm technology. Netlists have been generated from the given Trojan-infected circuits. The features pertaining to these circuits have been extricated from the said netlists. The

Fig. 1 Unsupervised ML algorithm



features are drawn out from the gate-level netlists using Python 3. These lineaments assist in the shaping of the ML model.

ML algorithms can be categorized as supervised and unsupervised algorithms. While supervised algorithms require a set of predefined input-target pairs for sculpting the model, unsupervised models do not have that obligation. This gives supervised algorithms a slight advantage in terms of accuracy as it has examples to work with. When one does not have at hand the predefined datasets, unsupervised learning serves.

In this work, K-Means classifier has been used to typecast Trojan-infected and Trojan-free nets. The model, as shown in Fig. 1, has a training phase along with a testing phase.

4 Experimental Results

The machine learning section of the proposed flow has been implemented. The K-Means classifier is realized in Python 3.7 using the scikit-learn Python library for ML. The outcome of the hardware Trojan classification technique that employs the previously obtained seven gate-level features of every net, will be demonstrated in this section.

The method has been validated on ISCAS ‘85 benchmark circuits [17]. Table 1 depicts the circuits on which the test is performed and the corresponding number of gates, total count of nets involved, the total of Trojans nets actually present, and the count of nets that are predicted as Trojan-infected by the K-Means classifier that has been employed.

Figure 2 shows the results obtained from the clustering performed by the K-Means classifier on the benchmark circuits. Figure 3 showcases the histogram version of the same. Table 2 encapsulates the overall result of the work. It is observed from Table 2 that the features chosen can be refined, so as to improve the TPR on all possible test circuits.

Table 1 K -means clustering results

Test data	Number of gates	Number of nets	No. of Trojan nets	Predicted No. of Trojan nets
c17	6	17	2	5
c432	160	432	3	2
c2670	1269	2670	12	7
c3540	1669	3540	24	17
c3540a	1668	3540	15	17
c5315	2307	5315	16	33

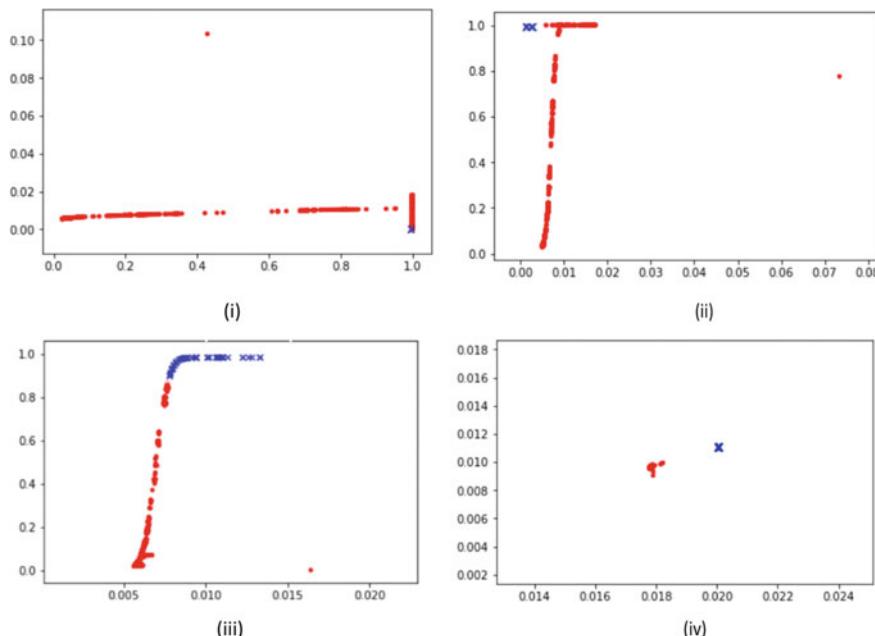


Fig. 2 Clustering results: (i) c3540a (ii) c2670 (iii) c5315 (iv) c432—plotted with feature 1 and feature 2 in x-axis and y-axis, respectively

5 Conclusion

In this work, a hardware Trojan identification technique by adopting ML to label the nets comprised in the circuit under test as Trojan nets or ordinary net has been proposed. At the outset, seven feature values are extracted from every net of the gate-level netlists. The obtained features are then fed to the classifier. Thereafter, the unknown netlists (of circuit under test) have been classified and identified to have hardware Trojans by employing the trained K-Means classifier.

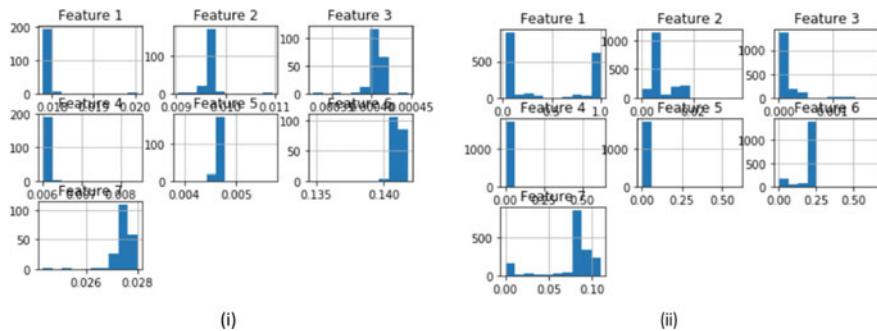


Fig. 3 Histograms: (i) c432 (ii) c3540a—plotted with the value of the feature in x -axis and its frequency in y -axis for each extracted feature

Table 2 Classification results

Data	TP	FN	TPR (%)
c17	2	0	100
c432	2	1	66.67
c2670	7	5	58.33
c3540	17	7	70.83
c3540a	15	0	100
c5315	16	0	100

The results of the experiment indicate that the sensitivity (TPR) of the method proposed is up to 100% on selected circuits. Despite the fact that our proposed method is a golden-free method we are able to obtain results that are at par or better than the previously proposed methods. On comparing the obtained results with [18], it has been observed that the TPR achieved by our method is higher in certain benchmarks considered, but the latter has a lower FPR.

Future scope of our proposed method includes improving the method so as to obtain a higher TPR and decrease the FPR so that every part of the hardware Trojan can be extracted from a given netlist.

References

1. Zhang J, Yuan F, Xu Q (2014) DeTrust: defeating hardware trust verification with stealthy implicitly-triggered hardware trojans. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp 153–166
2. Waksman A, Suozzo M, Sethumadhavan S (2013) FANCI: identification of stealthy malicious logic using boolean functional analysis. In: Proceedings of the ACM conference on computer and communications security, pp 697–708. <https://doi.org/10.1145/2508859.2516654>
3. Moein S, Gulliver TA, Gebali F, Alkandari A (2016) A new characterization of hardware trojans. IEEE Access 4:2721–2731

4. Moein S, Khan S, Gulliver TA, Gebali F, El-Kharashi MW (2015) An attribute based classification of hardware trojans. In: 2015 tenth international conference on computer engineering & systems (ICCES), Cairo, pp 351–356
5. Salmani H, Tehranipoor M, Karri R (2013) On design vulnerability analysis and trust benchmarks development. In: 2013 IEEE 31st international conference on computer design (ICCD), Asheville, NC, pp 471–474
6. Oya M, Shi Y, Yanagisawa M, Togawa N (2015) A score-based classification method for identifying hardware-trojans at gate-level netlists. In: 2015 design automation & test in Europe conference & exhibition (DATE), pp 465–470
7. Zhang J, Yuan F, Wei L, Liu Y, Xu Q (2015) VeriTrust: verification for hardware trust. *IEEE Trans Comput Aided Des Integr Circ Syst* 34(7):1148–1161
8. Hasegawa K, Yanagisawa M, Togawa N (2017) A hardware-trojan classification method using machine learning at gate-level netlists based on trojan features. *IEICE Trans Fundam Electron Commun Comput Sci* E100.A(7):1427–1438
9. Karri R, Rajendran J, Rosenfeld K, Tehranipoor M (2010) Trustworthy hardware: identifying and classifying hardware trojans. *IEEE Comput* 43(10):39–46
10. Suresh Babu N, Mohankumar N (2019) Wire load variation-based hardware trojan detection using machine learning techniques. In: Soft computing and signal processing. Advances in intelligent systems and computing, vol 900. Springer, Singapore
11. Bao C, Forte D, Srivastava A (2016) On reverse engineering-based hardware trojan detection. *IEEE Trans Comput Aided Des Integr Circuits Syst* 35(1):49–57
12. Inoue T, Hasegawa K, Kobayashi Y, Yanagisawa M, Togawa N (2018) Designing subspecies of hardware trojans and their detection using neural network approach. In: Proceedings of IEEE 8th international conference on consumer electronics—Berlin (ICCE-Berlin)
13. Hasegawa K, Yanagisawa M, Togawa N (2017) Hardware trojans classification for gate-level netlists using multi-layer neural networks. In Proceedings of IEEE 23rd international symposium on on-line testing and robust system design (IOLTS), pp 227–232
14. Hasegawa K, Yanagisawa M, Togawa N (2017) Trojan-feature extraction at gate-level netlists and its application to hardware-trojan detection using random forest classifier. In: Proceedings of international symposium on circuits and systems
15. Inoue T, Hasegawa K, Yanagisawa M, Togawa N (2017) Designing hardware trojans and their detection based ON a SVM-based approach. In: Proceedings of international conference on ASIC (ASICON), pp 811–814
16. Zantout S (2018) Hardware trojan detection in FPGA through side-channel power analysis and machine learning. UC Irvine. ProQuest ID: Zantout_uci_0030M_14966. Merritt ID: ark:/13030/m5k40r5c. Retrieved from <https://escholarship.org/uc/item/7hk8x6rb>
17. Brglez F, Fujiwara H (1985) A neutral netlist of 10 combinational benchmark circuits. In: Proceedings of IEEE international symposium on circuits and systems, IEEE Press, Piscataway, N.J., pp 695–698; see also the ISCAS-85 benchmark directory at <http://www.cbl.ncsu.edu/benchmarks>
18. Reshma K, Priyatharshini M, Nirmala Devi M (2019) Hardware trojan detection using deep learning technique. In: Computing and signal processing. advances in intelligent systems and computing, vol 898. Springer, Singapore

An Application Suite: Effectiveness in Tracking and Monitoring of Skill Training Programs



Balu M. Menon, P. Aswathi, and Shekar Lekha

Abstract Conducting and monitoring skill training programs for rural population always involves a huge amount of administrative tasks and tracking processes. To ensure the effectiveness and efficiency of the program it is mandatory to implement a proper tracking and monitoring system. In this paper, we present a software application suite for monitoring and tracking of the PMKVY (Pradhan Mantri Kaushal Vikas Yojana) training programs in the rural settlements of India conducted by an NGO. We discuss the common challenges and the design considerations that were used while designing and implementing the application. Sentiment analysis was applied on the feedback received through the mobile application and it was observed that 66% of the feedback were of positive polarity. Also, the word frequency analysis on the same data revealed that the feedback was mostly about the students and that too on a daily basis. With these observations, the potential to apply NLP in the future is also discussed.

Keywords Real-world applications of text mining · Deployment · Skill training program · Sentiment analysis

1 Introduction

Monitoring and evaluation of the activities of any training program are crucial for its effective completion. In vocational training, technology has been useful in motivating the students toward better performance [1] and thereby proving to be more effective in learning and teaching [2]. Through the monitoring processes of such programs, we track the key indicators of progress over the course of a program on the basis of which the outcomes of the intervention are evaluated [3]. The main challenges faced in implementing a monitoring process concerning a training program are (a)

B. M. Menon · P. Aswathi (✉) · S. Lekha

Center for Women's Empowerment and Gender Equality, Amrita Vishwa Vidyapeetham, Amritapuri, India

e-mail: aswathi.p@ammachilabs.org

Identifying the goal that the training program is meant to achieve (b) Identify the key indicators that are to be used to monitor the goals set for the program. (c) Setting targets and milestones to check if the key performance indicators are met by the given timeline. (d) To implement a monitoring system to track progress.

With the services and smart mobile phone technology growing rapidly [4] there exists many software tools to support mobile app development [5]. Software suites of mobile and web application offers great support for or data collection, monitoring and report generation [6] irrespective of the application domain. In the health domain, it is not only used for collecting patient data, and reporting [7] but also toward disease diagnosis with a reference guide on antibiotics [8]. Behavioral modeling has been detrimental research in health [9, 10] and education where it is used as a medium to improve teaching effectiveness and student performance [11]. In social science, they have become a new frontier for advancing field experiment methodology [12].

Through this paper, we present the varying challenges in conducting such remote skill training programs in India, and also we present a case study on how these problems are effectively addressed by implementing a mobile application-based solution. The APAMA software suite presented in this paper consists of a mobile android app component, cloud storage, and a web application. The collective contribution of these components is data accumulation, event tracking, monitoring, reporting and data visualizations that are further explained in the sections mentioned below. The data thus accumulated enables the use of text analysis to automatically extract the insights from data there by providing easier means to summarize the information gathered from different centers.

2 Challenges in Tracking and Monitoring the Training Programs

The PMKVY project trains the youth on vocational courses like two-wheeler maintenance, tailoring, plumbing, and carpentry. Post-training the students will be assessed to certify the youth of India on industry-relevant vocational skills [13]. Often the unemployed population is from the rural and encouraging the trainers to join such training programs can be challenging. Owing to the fact that the rural communities are scattered across India and also considering the lack of population's exposure to such formal training programs, it becomes more challenging to organize and monitor. We need a proper event log and data log system to ensure effective implementation and tracking. AMMACHI Labs at Kollam, Kerala is an NGO organization that has been conducting several rural development programs. As part of this PMKVY program, the team conducts tailoring, two-wheeler maintenance, and plumbing skills to the youth in 10 rural centers. Even though technology outreach has far improved in rural India, being less educated and less exposed to the digital world, any programs that leverage on digitization needs good strategies at least for the initial phases.

The conventional method of reporting is often done through a shared document, e-mail, whatsapp, messages, image files, etc. The regional staff fills in data, consolidates it and shares it with the authorities. These multiple sources of data then need to be processed in order to prepare any report as it involves a lot of paperwork, data entry work and human work hours. The data reported by the regional staffs often tends to vary from the decided format according to their convenience. This scattered data sources and the lack of structure in data becomes the biggest bottleneck. Also in India in the rural areas it is difficult to get access to the internet all throughout the day and at a particular location and this might be an additional hindrance to add in and share the data by mail or by other means of shared documents.

3 Software Design Considerations

Explained below are the points that we considered while designing the software.

Accessibility: Making the application accessible 24×7 irrespective of the network coverage at both urban and rural settlements of the country.

Accountability: The APP is designed to make sure that the trainers and other actors are providing relevant data and are accountable for the information provided.

Improved turnaround time: Data entered into the mobile APP after every training session, inventory, maintenance, or meetings helps the supporting staff at the headquarters to look into any issues raised immediately.

Availability: The primary reason to choose a mobile application to collect data was to help in the increasing percentage of availability. The usage of smartphones is expected to rise further.

User roles and their responsibilities: The mobile APP is designed to facilitate the user to select his respective role to enter data. The advantage is when a user is assigned two or more roles during the project. The different roles that are part of the program are Facilitator, Hub Manager, Researcher, Trainer and Zonal Operations Manager.

Information Storage: The mobile APP acts as a platform to collect and store collection of activities. Activities can be defined as a set of actions or steps taken to accomplish a particular task. For example, the activity of starting a tailoring batch can be further broken down into a set of sub tasks like hiring tutor, student registration, equipment purchase and so on. These three steps will be stored under one activity named ‘starting a tailoring batch’.

Data services required: The whole purpose of collecting data is to produce meaningful reports/data services. Hashtags were included in the design so that common activities could be grouped together and tagged with a particular hashtag. This helps the user to retrieve all activities with common behavior under a particular hashtag.

4 APAMA: The Software Solution

4.1 System Architecture and the Mobile Application

The entire software architecture comprises of 3 components as shown in Fig. 1. The mobile application, cloud data storage and a web portal. The data accumulated through the mobile app is stored in the Google Firebase storage and from there it is consumed by the website portal. The main features of the mobile app are as given below.

User Authentication with Fire base Centralised cloud storage

Online and Offline mode of data entry

Data entry based on user's role in the activity Hash tagging of activities

Upload the related documents like reports, images or audio files. Linked to whatsapp for further communication

Automated generation of whatsapp text message

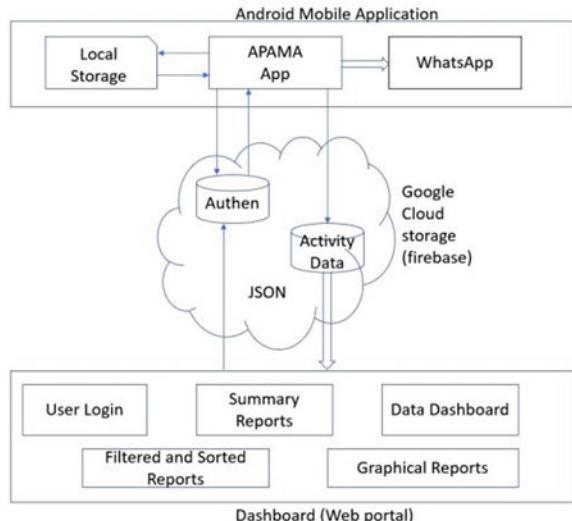
Activities can have sub-activities or other related activities The feedback of the trainer about the training session

The curriculum plan and topics covered for a day

The challenges faced by the students and trainers. Also what measure they took to handle those issues

Student performance and their feedback in the day.

Fig. 1 System architecture



4.2 Web Based Data Portal

The web application acts as an interface for the administrator to monitor and review the data sent from the mobile application. The overall status of the data accumulated is reported through the portal. Reports can also be generated through the portal as graphs and also in textual format which can be downloaded and saved as local copies. The various filter options based on the data parameters helps in better tuning of the reports. Utilizing the real-time cloud based data, the web application leverages the database to provide accurate reports.

5 Data Collection and Analysis

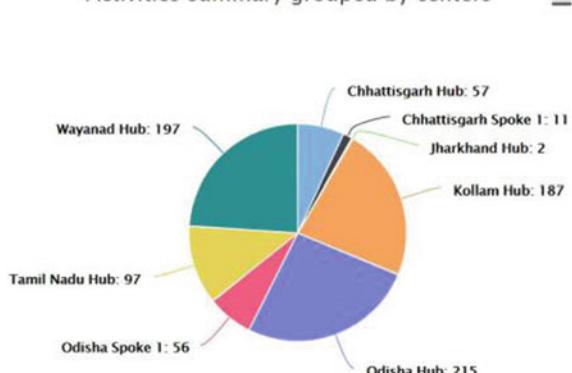
The number of training activities conducted so far across all the centers is over 1000+. The 42 users of the system have collectively entered the data. The Odisha center being the first has the most number of updates (as given in Table 1 and Fig. 2). We

Table 1 The PMKVY centers and corresponding data collected through the mobile application

Centers	#Updates
Chhattisgarh	57
Chattisgarh spoke 1	11
Jharkhand hub	2
Kollam hub	187
Odisha hub	215
Odisha spoke 1	56
Tamil Nadu hub	97
Wayanad hub	197

Fig. 2 Pie chart depicting the proportion of data acquisition from different centers

Activities summary grouped by centers



can observe that the centers are scattered geographically. Also, the training follows a hub-spoke model that would make it easier for us to scale geographically. The resources at Odisha hub is shared to the spokes at Odisha, making it easy for sharing the resources and monitoring. The most commonly used feature of the app is the message sharing feature. Even though there were multiple options for sharing the information, it is observed that it has become a standard to post the automatically generated activity summary to the official whatsapp group. This enabled the program monitors to respond to the issues or challenges raised by the remote staff without any delay. The details that get stored in the cloud can be accessed immediately which helps to understand the situation properly.

The trainers are given an entire set of questions to give feedback on the training. Based on the feedback the training monitoring team and the content creation team alters the training strategies as needed. But for the effectiveness of this approach, it is vital that the trainers provide sincere and detailed feedbacks. The methods in which we can encourage data entry to the very detail remains an open challenge. For example, the question “Explain the challenges that you faced today” doesn’t add value if not entered with proper reflection. Out of the whole data, 260 feedbacks were constructive for further introspection. But instead of gathering objective close ended questions we implemented open-ended questions to enable users to provide their perspectives in detail. Later these answers were subjected to text analysis techniques to automatically extract the insights from the entire data. Training multiple batches of students often gives varying experience to the tutors based on the challenges they face. Word frequency analysis and sentiment analysis provides an easier way to extract these batch specific or course specific insights from the tutor’s feedback.

5.1 *Text Analysis and Results*

As part of the questions, there are some open-ended questions like “General feedback on today’s class?”, “How was the student performance?”, “What all were the challenges and how did you handle them”. These questions were intentionally left as open-ended to make the users think and write in their original way and thus became eligible for text analysis through Natural Language Processing approach. But this led to some challenges too. The trainers who are generally hired from the nearby rural locality of the centers, often found it difficult to enter the data in English. The texts that they entered were in a combination of Hindi and English. Also, we could observe that there were instances where the data entered was in Hindi typed in as English. This challenge in itself becomes a good research problem in the domain of Computational Social Science and NLP.

As a basic analysis, we applied word frequency analysis and sentiment analysis on the data provided by the users as feedback to the training sessions. Figure 3 depicts the word frequency count applied to the user feedback data. The story formed by the most frequent words. Based on the user’s daily feedback data the frequency of the words is as Student> Today> Class> Good> Tailoring> Batch> Practical> Plumbing.

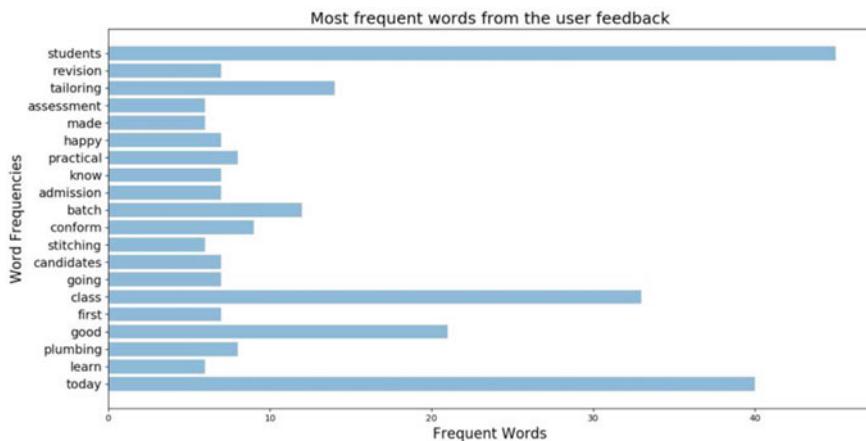


Fig. 3 Graph that demonstrates the most frequent words and their frequencies

These words give us an idea of the context of feedback in general. Students, today and class being the most frequent words we can say that the discussion is around a day to day class and the students that attend the class. We can also say that classes are generally discussed as being good. We can also see that tailoring and plumbing were often discussed. Table 2 shows the results of the sentiment analysis that we applied to the same feedback data. With 66% of positive polarity, we can observe that the feedback is generally satisfactory. But we noticed that several messages were Hindi typed in English and most reflect either an issue or a challenge that they faced on the day. On discussing with these respondents we understood that they were finding it difficult to express their complex view in the English language and hence used Hindi. For the sentiment, we avoided such feedback. But in the future, we need to either translate those messages to pure English or implement automatic text translation methods to process such texts.

Table 2 Results of the sentiment analysis obtained on user feedback data

Centers	Number of updates
Average polarity	0.3
Average subjectivity	0.35
Percentage of positive feedback	66
Percentage of negative feedback	13
Maximum value of positive polarity	0.447
Maximum value of negative polarity	0.316

6 Conclusion and Future Work

The paper describes a software application suite that comprises of a mobile android component and a web application to monitor and evaluate training programs to ensure its effective implementation. The mobile app is to collect data and the web application to support the monitoring, reporting, and visualization of data collected through dashboards. It aims at replacing the conventional method of data collection. From the 1000+ entries collected through the mobile application, shows that it could be effective in monitoring and reducing the response time needed to track down on issues. The challenges faced with the data were that as most of the training programs are located in the remote locations of rural India, the users are not fluent in English. This has lead to grammatical mistakes in the entered data and has created fluctuations in the words processing results like frequency, co-occurrence or dependencies. Another drawback of the users not being fluent in the English language is that they tend to mix language while entering the data. For instance, they might type Hindi along with an English statement. This challenge can leverage the advanced algorithms of NLP to process the multilingual texts and extract data for sentiment analysis. As an added feature to reduce the response time even further, we can send out notification in the form of an email or text message to the training coordinators or monitors for any possible issues addressed by the trainers through their feedback.

References

1. Bhavani B, Sheshadri S, Unnikrishnan R (2010) Vocational education technology: rural India. In: Proceedings of the 1st amrita ACM-W celebration on women in computing in India. ACM
2. Sachith KP et al (2017) Contextualizing ICT based vocational education for rural communities: addressing ethnographic issues and assessing design principles. In: IFIP conference on human-computer interaction. Springer, Cham
3. Khandker S, Koolwal GB, Samad H (2009) Handbook on impact evaluation: quantitative methods and practices. The World Bank, Washington, DC
4. Kang J-M, Seo S-S, Hong JW (2011) Usage pattern analysis of smartphones. In: 2011 13th Asia-Pacific network operations and management symposium. IEEE
5. Gandhewar N, Sheikh R (2010) Google android: an emerging software platform for mobile devices. Int J Comput Sci Eng 1(1):12–17
6. Nejmeh BA, Dean T (2010) The charms application suite: a community-based mobile data collection and alerting environment for HIV/AIDS orphan and vulnerable children in Zambia. Int J Comput ICT Res 4(2):46–63
7. Hamou A et al (2010) Data collection with iPhone web apps efficiently collecting patient data using mobile devices. In: The 12th IEEE international conference on e-health networking, applications and services. IEEE
8. Burdette SD, Herchline TE, Oehler R (2008) Practicing medicine in a technological age: using smartphones in clinical practice. Clin Infect Dis 47(1):117–122
9. Vaizman Y et al (2018) Extrasensory app: data collection in-the-wild with rich user interface to self-report behavior. In: Proceedings of the 2018 CHI conference on human factors in computing systems. ACM

10. Wang R et al (2014) StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In: Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing. ACM
11. Nedungadi P, Mulki K, Raman R (2018) Improving educational outcomes & reducing absenteeism at remote villages with mobile technology and WhatsApp: Findings from rural India. *Edu Inf Technol* 23(1):113–127
12. Zhang J et al Advantages and challenges in using mobile apps for field experiments: a systematic review and a case study. *Mobile Media Commun* 6(2): 179–196
13. Ministry of Skill Development and Entrepreneurship. Pradhan Mantri Kaushal Vikas Yojana (PMKVY) (2015). <http://pmkvyofficial.org/>

Automated Water Management System Using Internet of Things



Ritik Gupta, B. Shivalal Patro, and Manas Chandan Behera

Abstract This paper presents a framework for the realization of smart cities through the help of the Internet of things (IoT). It focuses on a method to save one of the natural resources, i.e., water and manage it at the same time. It focuses on the area of a normal human daily routine, i.e., the amount of water to be used per person in a house. This is designed so that the world do not faces crisis of water in future. This paper has an application of one of the highest-rated technology, i.e., Internet of things (IoT), on one of the most significant issues by utilizing technology.

Keywords Internet of things (IoT) · Smart home nodeMCU (ESP8266)

1 Introduction

Water management is defined as the activity of planning, developing, distributing and managing the optimum use of water resources. This impact of several key matters of human lives such as food production, water consumption, sewage treatment, generation of electricity etc. [1]. Everyone knows water means life. Everything is directly or indirectly dependent on the water whether it be the food we eat, the energy we get or anything else, and deep inside everyone also knows that the most important things are being misused. So, this paper has brought a solution to water management with the help of the Internet of things (IoT) [2, 3].

The IoT is a system of interrelated computing devices, mechanical and digital machines or objects that can provide transfer data over a network without having a

R. Gupta · B. Shivalal Patro (✉) · M. C. Behera

School of Electronics Engineering, KIIT Deemed to be University, Bhubaneswar, India
e-mail: shivalalpatro@gmail.com

R. Gupta
e-mail: ritikgupta8936@gmail.com

M. C. Behera
e-mail: manaschandan79@gmail.com

need of human to human or human to computer interaction, the interaction is only conducted but computer to computer in this technology. IoT has given rise to this era of technology. It is used many other fields such as in fracture management, industrial applications, medical purpose etc. [4, 5]. The rest of the paper consists the project description, its result analysis followed by conclusion.

2 Project Description

The project is proposed for water consumption of the smart cities based on a simple theory that the amount of water allowed in a particular home dependent on the number of members it has in the house. Based on the theory the proposed model does have the following stages of work. The various components with their detailed description of the specifications used in this project is listed in Table 1.

2.1 Hardware Setup

A person counter is made to install in the door frame of the house. The number of the person entering the house is counted through the counter and the relevant data is sent to the database which maintains the number of persons entering inside house. The process of sending relevant data to the database from the counter is achieved through the usage of nodeMCU(ESP8266). The counter has been connected to the nodeMCU where the microcontroller is responsible for sending the relevant data to the database. The devices being used are highly optimal and sensitive for environmental conditions which are calibrated up to the mark. With fast connectivity, the nodeMCU sends the data to database very fast and effectively [6, 7]. Figure 1 depicts the initial count to be 0 as the detection has not been started. The database allows the user to set the count value to a default value so that the number of permanent members can be

Table 1 Specifications of the components used in the prototype

Device	Model	Specification
nodeMCU	ESP8266	This small module allows various electronics equipment's and electrical equipment's to connect to itself and transmit and receive the data by creating its own WiFi field using hash style
Proximity sensor	Infrared type	A proximity sensor is designed to detect obstacles near it by radiating infrared rays which would strike obstacle and return back and the device properly monitors the radiating waves
Jumper wires	-	An electrical wire that is used to interconnect various electronics or electrical equipment's without soldering
Breadboard	-	A breadboard is a solder less device for temporary prototype with electronics and test circuit designs

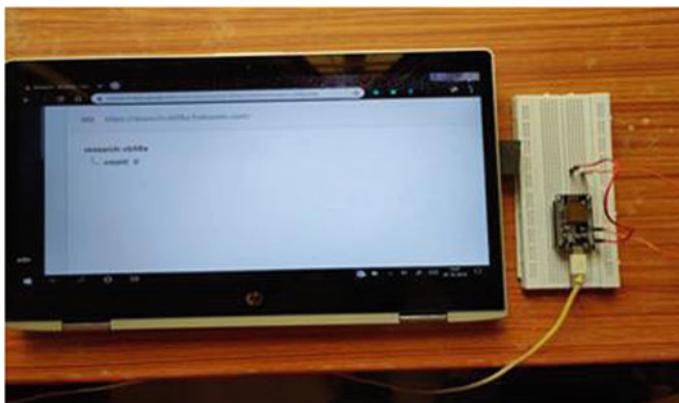


Fig. 1 The counter in the idle state, the counter is 0



Fig. 2 The counter in the working state, database changing in real time, i.e., led glowing

assigned to the database. Whenever the person counter counts any person the value gets updated in the database. The yellow mark in the variable (when updated to 3 as shown in Fig. 2) of the database depicts that the value has been changed recently.

2.2 Software Setup

The data that has been collected from the sensors by wireless sensor networks is updated and accessed so that based on that information the inflow of the water inside the house can be controlled [8–10]. This can be achieved with the help of the motor being controlled by the motor driver. The figure here doesn't contain any motor driver. But the proposed solution will contain. The data is retrieved by the second side of the model, where the motor driver is connected to the nodeMCU, based on the number of persons in the house decides the timing for which the motor would operate, i.e., the motor driver would receive positive input for a certain time only. This

would open the gate for a certain degree only. The degree of the gate is proportional to the number of persons inside the home. So, in the figure when the value is being retrieved by the nodeMCU. Then, the led glows for a certain time, which depicts that the motor driver is active for that particular certain time. This will increase the gate's degree. When the value retrieved is more than the amount of time has been increased, which increases the gate degree. From a single network modulator, both the microcontrollers could be handled. There is no physical connection between the two systems.

3 Working and Result Analysis

Figure 3 shows the detailed flow chart of the project implemented in this paper. The database in the paper refers to a changing real-time database that is present in a cloud which are used for official purpose. The model aim was fully justified, i.e., to have an automated water management system. It only focuses on the fact to dispense an adequate amount of water that a person needs in his daily chores. The prototype was operated on two different stages as it was meant to be one the person counter and the second one, the motor driver operating system. For each count, in person the relevant data is sent to database and gets updated (the yellow mark over variable shows that variable has been updated recently) and the motor driver runs the motor for 0.5 s that opens the gate of inflow pipe for a certain degree depending on the rpm of motor. The municipality (Government) rules of water supply would define the rpm or speed of the motor or contradictory to it the time of operation of motors could be varied by changing the script of nodeMCU on the motor driver system part. Thereby it could be stated as the total number of people/folks present in the house would directly be proportional to the amount of water delivered to that house.

4 Conclusions and Future Scope

With blistering development in the field of Internet of things (IoT) technology, this paper presents a compendious blueprint of developing smart homes with effective water management. So, using this proposed prototype or model water wastage can be minimized and can be used smartly and most importantly in automated manner with high efficiency. Furthermore, the study will be directed toward setting an individual system for a complete society coming under a particular municipality that would regulate the water supply in effective and optimal manner that would result in a more efficient management. The study would also concentrate to apply this prototype in various other management systems for the effective usage of available resources.

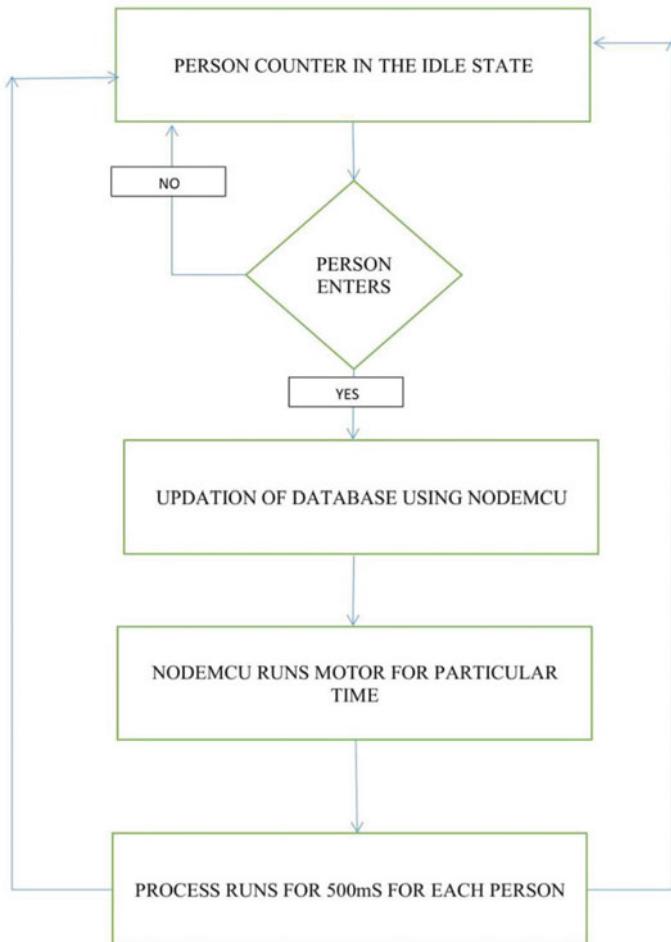


Fig. 3 Flow chart of the implemented project

References

1. Gupta K, Kulkarni M, Magdum M, Baldawa Y, Patil S (2018) Smart water management in housing societies using IoT. In: Second international conference on inventive communication and computational technologies (ICICCT) 2018 Apr 20, pp. 1609–1613 IEEE
2. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. IEEE 1(1)
3. Mesiti M, Valtolina S, Ferrari L, Dao MS, Zettsu K (2015) An editable live ETL system for ambient intelligence environments. In: Internet of the things (WF-IoT). In: IEEE 2nd world forum 2015
4. Mattern F, Floerkemeier C (2010) From the internet of computers to the internet of things. In: From active data management to event-based system and more

5. Alkhamisi A, Nazmudeen MS, Buhari SM (2016) A cross-layer framework for sensor data aggregation for IoT applications in smart cities
6. Atzori L, Lera A, Morabito G (2010) The internet of things: a survey. *Comput Netw* 54(15)
7. Cuff D, Hansen M, Kang J (2000) Urban sensing: out of the woods. *Commun ACM* 51(03)
8. Xu LD, He W, Li S (2014) Internet of things in industries: a survey. *IEEE Trans Ind Inform* 10(4)
9. Chen S, Xu H, Liu D, Hu B, Wang H (2014) A vision of IoT: applications, challenges, and opportunities with china perspective. *IEEE Internet Things J* 1(4)
10. Bellavista P, Cardone G, Corradi A, Foschini L (2013) Convergence of MANET and WSN in IoT urban scenarios. *IEEE Sens J* 13(10)
11. Beneditt M, Ioriatti L, Martinelli M, Viani F (2010) Wireless sensor network: a pervasive technology for earth observation. *IEEE J Sel Top Appl Earth Obs Remote Sens* 3(4)

Model-Based Observer Performance Study for Speed Estimation of Brushed DC Motor with Uncertain Contact Resistance



Sayantan Moulik and Biswajit Halder

Abstract In view of recent advancement in consumer and industrial devices, sensorless estimation has been accepted and applied widely. Speed measurement of brushed permanent magnet direct current (PMDC) motor is an integral part of controlling velocity servomechanism. The model-based observer design is simply achieved by following pole placement technique; however, in the parameter varying environment, the issue of robust performance has to be accounted for. The LMI technique-based observer is supposed to accommodate parameter variation but with relatively complex configuration. In this paper, a thorough performance comparison between rotor speed observer designs of armature controlled DC motor using pole placement technique with that of the LMI technique has been accomplished. The results generated through MATLAB simulation were interesting in the perspective of speed and current observation error.

Keywords Pole placement · LMI · Full-order observer · Uncertain contact resistance

1 Introduction

In the existing significant application area encompassing consumer and industrial products, the brushed DC motors are used despite the availability of different types of electrical drives. The utilization of brushed DC motors as described in [1] is majorly due to its simplicity in control and operation. In view of modern control applications in velocity and position servomechanism, measurement of rotational speed of the shaft is an essential part. Some of the critical applications constrain the installation of speed measuring device particularly in the area of robotics. This

S. Moulik · B. Halder (✉)
Narula Institute of Technology, Kolkata, India
e-mail: halder.biswajit@nit.ac.in

S. Moulik
e-mail: sayantanmoulik@gmail.com

increase in the cost of measuring devices, maintenance, and mounting motivates the use of sensorless speed measurement of brushed DC motors. There exist many potential methods of sensorless speed measurement as illustrated in the work [2–5]. In the present day, the model-based observers are being introduced greatly due to the advent of low-cost processor with adequate processing speed and commensuration with miniaturization requirement.

Although the model-based observers are generally designed for DC motor which is deterministic one with constant time-invariant parameters, in real time this model may vary due to the existence of sliding electrical contact resistance uncertainty. This uncertain resistance occurs during the conduction of armature current between the stationary and moving parts of a motor. The recent study in [6, 7] illustrates the critical issues like degradation of brushes, reduction condition of contact resistance uncertainty in view of magnitude of current, ambient temperature, and rotating speed. In the study [8], a method has been proposed for efficiently computing a brushed DC motor circuit model which accommodates the brush segment contact resistance. In the recent research [9, 10], the speed estimation in view of armature resistance variation with respect to thermal effect has been discussed.

The majority of the model-based observers that are designed with the philosophy of that of Luenberger observer [11] which has been the foundation of Utkin observer [12] and Kalman filter [13] dealing, respectively, with the robustness and stochastic problems. A numerous research has been accomplished in the past few decades in designing an observer for linear as well as nonlinear systems. The pole placement method of model-based observer design is a well-accepted method among them as found in [1] which is very simple with good observed error convergence property for LTI systems. Recently, linear matrix inequality (LMI) technique-based observer which exhibits robustness property is of great significance in the parameter varying environment as discussed in [14, 15] pertaining to the intuition described in [16].

In this paper, a performance comparison is made for the full-order observers designed with pole placement technique and LMI methodology for armature controlled brushed PMDC under the existence of uncertain contact resistance. Analysis of results has been obtained through the simulation in the MATLAB/SIMULINK environment. A comprehensive observation has been discussed through the variation of armature resistance which is a series combination of constant armature resistance and uncertain contact resistance.

In the next section, the pertaining formulations of the observers design have been discussed briefly. In Sect. 3, description of brushed DC motor taken as the system and corresponding observer is illustrated. In Sect. 4, detailed simulation result and its discussion over that has been illustrated are followed by essential conclusion in Sect. 5.

2 Observer Design Formulations

A continuous-time finite-dimensional linear time-invariant (FDLTI) model of a system may be represented by the following expression.

$$\left. \begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) \end{array} \right\} \quad (1)$$

where $x \rightarrow R^{(n \times 1)}$ represents the system states, $u \rightarrow R^{(m \times 1)}$ represents the system input, $y \rightarrow R^{(p \times 1)}$ represents the system output, $A \rightarrow R^{(n \times n)}$ is system matrix, $B \rightarrow R^{(n \times m)}$ is the input matrix and $C \rightarrow R^{(p \times n)}$ is the output matrix, n is the number of states, m is the number of inputs, p is the number of outputs with $p \geq m$.

The assumptions for the design are: B and C matrices are full rank matrices and (A, C) pair is observable. The model-based full-order state observer may be represented as given below.

$$\dot{\hat{x}}(t) = (A - K_e C)\hat{x}(t) + Bu(t) + K_e y(t) \quad (2)$$

where $\hat{x} \rightarrow R^{(n \times 1)}$ represents the system estimated states and $K_e \rightarrow R^{(n \times p)}$ represents the observer gain. The observer gain is chosen such that the error between the estimated and actual state asymptotically approach to zero following a desired transient characteristics. The error dynamics is given by the following expression.

$$\dot{e}(t) = \dot{x}(t) - \dot{\hat{x}}(t) = (A - K_e C)\{x(t) - \hat{x}(t)\} \quad (3)$$

where $e \rightarrow R^{(n \times 1)}$ represents the estimated error vector.

2.1 Pole Placement Method of Observer Design

The proper choice of K_e makes the matrix $(A - K_e C)$ asymptotically stable with all the eigenvalues of the matrix which are termed as the observer pole(s) having negative real parts. The design should be such that the pole(s) of the observer located far left half of s -plane than that of the system poles. The following steps may be followed for obtaining the desired result.

Step 1. Checking of the observability condition through the Kalman test. Once the system is found to be observable, next step is followed.

Step 2. The characteristic polynomial is found from the expression of $(A - K_e C)$ with unknown parameter K_e to determine the coefficient of the characteristic polynomial as given below.

$$|sI - (A - K_e C)| = s^n + a_1 s^{n-1} + a_2 s^{n-2} + \cdots + a_n \quad (4)$$

Step 3. The desired eigenvalues of the observer are chosen, and desired characteristic polynomial is hence determined as per the following expression.

$$(s - \mu_1)(s - \mu_2) \cdots (s - \mu_n) = s^n + \alpha_1 s^{n-1} + \alpha_2 s^{n-2} + \cdots + a_n \quad (5)$$

Through the comparison of Eqs (4) and (5), the value of K_e is determined.

2.2 LMI Method of Observer Design

The observer design by using LMI methodology may be illustrated by using the following simple plant dynamics as expressed below.

$$\left. \begin{array}{l} \dot{x}(t) = Ax(t) + Bu(t) + f(x) \\ y(t) = Cx(t) \end{array} \right\}. \quad (6)$$

where all the variables and parameters are similar as that of Eq. (1) except $f(x) \rightarrow R^{(n \times 1)}$ which represents the system nonlinearity and/or uncertainty. The model-based full-order state observer dynamics may be represented by

$$\dot{\hat{x}}(t) = (A - K_e C)\hat{x}(t) + Bu(t) + K_e y(t) + f\{\hat{x}(t)\}, \quad (7)$$

which is similar as that of Eq. (2).

The observer gain K_e is determined with the intuition of bounded Jacobian system such that the nonlinear function has bounded derivatives as

$$K_1 \leq \frac{\partial f}{\partial x} \leq K_2, \quad (8)$$

where \leq denotes element-wise inequality operator. It is to be noted that K_1 and K_2 are representing matrices, with $K_1(i, j)$ and $K_2(i, j)$ being the lower and upper bounds on $\frac{\partial f_i}{\partial x_j}$.

The Jacobian bounds are now considered as

$$\tilde{f}\{x(t), \hat{x}(t)\} = f\{x(t)\} - f\{\hat{x}(t)\}. \quad (9)$$

Using differential mean value theorem (DMVT) [17] and that of Eq. (8), it can be shown that

$$\left[\tilde{f}\{x(t), \hat{x}(t)\} - K_1 \tilde{x}(t) \right]^T \left[\tilde{f}\{x(t), \hat{x}(t)\} - K_2 \tilde{x}(t) \right] \leq 0, \quad (10)$$

with 0 represents the matrix with all zero elements. Equation (10) may be represented in matrix form as

$$\begin{bmatrix} \tilde{x} \\ \tilde{f} \end{bmatrix}^T \begin{bmatrix} \frac{K_1^T K_2 + K_2^T K_1}{2} & \frac{K_1^T + K_2^T}{2} \\ -\frac{K_1^T + K_2^T}{2} & I \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{f} \end{bmatrix} \leq 0, \quad (11)$$

Thus, the observer design is accomplished by finding K_e such that

$$\begin{bmatrix} (A - K_e C)^T P + P(A - K_e C) & P \\ P & 0 \end{bmatrix} - \begin{bmatrix} \frac{K_1^T K_2 + K_2^T K_1}{2} & \frac{K_1^T + K_2^T}{2} \\ -\frac{K_1^T + K_2^T}{2} & I \end{bmatrix} < 0, \quad (12)$$

where P constitutes the Lyapunov function candidate V such that

$$V\{x(t)\} = \hat{x}^T(t)Px(t), \text{ with } P > 0. \quad (13)$$

Equation (13) ensures asymptotic stability of the observer. By using LMI solver tools of MATLAB, the observer gain K_e has been obtained in this paper.

3 Brushed DC Motor System and Observer Description

The performance of observers designed with pole placement and LMI methods has been compared for a system as an armature controlled permanent magnet brushed DC motor having significant uncertain contact resistance in the armature circuit.

3.1 State Space Representation of PMDC

The observer performance has been found out by considering a fractional horsepower PMDC motor (Model-TM110). The rotor speed is estimated through observer by using armature current and applied armature voltage as the two known input to the observer. The load torque input, which is proportional to the armature current, is actually the third input of the observer.

The differential equations describing the dynamics of the PMDC motor are given by

$$\left. \begin{aligned} v_a(t) &= R_a i_a(t) + L_a \frac{di_a(t)}{dt} + e_b(t) \\ T(t) &= J \frac{d\omega(t)}{dt} + D\omega(t) + T_L(t) \end{aligned} \right\}, \quad (14)$$

where $v_a(t)$ is the applied armature voltage, $e_b(t) = K_v \omega(t)$ is the back emf., $i_a(t)$ is the armature current, $T(t) = K_t i_a(t)$ is the electromagnetic torque, $\omega(t)$ is the angular velocity of the rotor shaft, and $T_L(t)$ is the load torque. The electromechanical system parameters R_a , L_a , K_v , K_t , J and D , respectively, denote the armature resistance including uncertain brush contact resistance, the armature circuit self-inductance, back emf. constant, torque constant, moment of inertia and the damping constant.

Now Eq. (14) may be represented in vector matrix form as

$$\left[\begin{array}{c} \dot{x}_1(t) \\ \dot{x}_2(t) \end{array} \right] = \left[\begin{array}{cc} -\frac{R_a}{L_a} & -\frac{K_v}{L_a} \\ \frac{K_t}{J} & -\frac{D}{J} \end{array} \right] \left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right] + \left[\begin{array}{cc} \frac{1}{L_a} & 0 \\ 0 & -\frac{1}{J} \end{array} \right] \left[\begin{array}{c} u_1(t) \\ u_2(t) \end{array} \right],$$

$$y(t) = [1 \ 0] \left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right], \quad (15)$$

where $x_1(t)$ and $x_2(t)$ are the two state variables representing, respectively, $i_a(t)$ and $\omega(t)$ for the system described in Eq. (14) which is basically describing a velocity servomechanism. The variables $u_1(t)$ and $u_2(t)$ are the system inputs as $v_a(t)$ and $T(t)$, respectively, and the output $y(t)$ representing $i_a(t)$.

3.2 Performance Analysis of Observers Designed with PMDC

The parameter of the PMDC motor shown in Fig. 1 is described in Table 1 with the rated armature voltage of 200 V. The state Eq. (15) may be reduced to the following Eq. (16) with the real parameters as

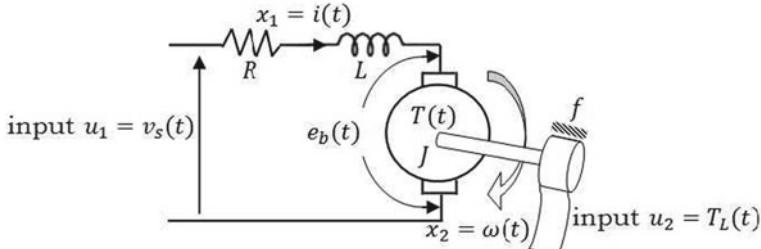


Fig. 1 Schematic of a PMDC motor

Table 1 PMDC motor parameters

Parameters	Values
Nominal armature resistance, R_a	16.5 Ω
Armature winding inductance, L_a	143 mH
Back emf. constant, K_b	0.816 Vrad $^{-1}$ s
Torque constant, K_t	0.816 Nm A $^{-1}$
Damping constant, D	1.23×10^{-3} Nm rad $^{-1}$ s
Moment of inertia of rotor, J	5.83×10^{-5} Kgm 2

$$\left. \begin{aligned} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} &= \begin{bmatrix} -115.38 & -5.706 \\ 13997 & -21.16 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 6.99 & 0 \\ 0 & -17152.6 \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \end{bmatrix}, \\ y(t) &= [1 \ 0] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \end{aligned} \right\}, \quad (16)$$

having pair of PMDC motor eigenvalues located at $-68 \pm j278.65$. In view of designing an observer with less noise sensitivity, the pole location must be appropriately chosen so that the observer is still exhibited faster response to follow the speed transient of the motor. In this problem, both the poles of the observer have been placed at 100 corresponding to pole placement method of design.

However, in case of LMI method-based observer design the issue of contact resistance uncertainty of PMDC motor brush is considered as the part of armature resistance R_a with known bounds of variations. Three types of bounds of variation, $\pm \Delta R_a$, are considered here which are $\pm 2.5 \Omega$, $\pm 5 \Omega$, and $\pm 7.5 \Omega$. Finally, observer response of two different designs has been compared in the next section.

4 Results and Discussions

The performance comparison has been accomplished with respect to a given PMDC motor in the MATLAB/SIMULINK environment and the system used for execution is Intel(R) Core i3-3217U CPU, 1.80 GHZ, and RAM of 2 GB. In the pole placement method of observer design, the desired pole position is taken at $[-100, -100]$ with the observer gain $K_e = [64 \ 12906]^T$. This design does not require any bounds information of the contact resistance uncertainty, but the nominal value of armature resistance as given in Table 1. The observer gain with LMI method is found to be $K_e = [6758 \ 26431]^T$, $K_e = [6797 \ 26428]^T$ and $K_e = [6715 \ 26435]^T$ corresponding to the contact resistance variation of $\pm 2.5 \Omega$, $\pm 5.0 \Omega$, and $\pm 7.5 \Omega$, respectively.

4.1 Observed Output Response Analysis

The observed output of the designed full-order observers yields two state variables, namely the rotor speed and armature current. The inputs of the observer are step change in armature voltage and load torque. The armature voltage step change has been conducted under no-load condition, whereas the load torque step change has been conducted under constant rated armature voltage condition.

The no-load speed profile in Fig. 2 indicating that the input voltage has been increased twice by 100 V at the time interval of 0.5 s from the starting of simulation. It has been observed that practically no significant difference lies in two different observer responses with the bound $|R_a|_{max} = 7.5 \Omega$, which are obviously clear from the zoomed portion of the figure. In Fig. 3, speed profile under load condition has been

Fig. 2 No-load speed profile for armature voltage variation to 100 V and 200 V at 0.5 and 1 s interval, respectively, with $|R_a|_{\max} = 7.5 \Omega$

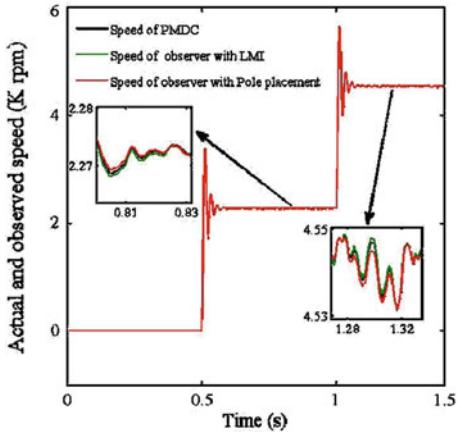
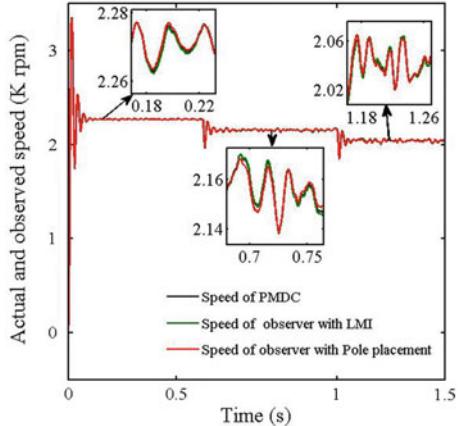


Fig. 3 Speed profile with constant armature voltage for load variation to 0.5 Nm and 1 Nm at 0.5 and 1 s interval, respectively, with $|R_a|_{\max} = 7.5 \Omega$



shown where load torque has been increased twice by 0.5 Nm at the time interval of 0.5 s from the starting of simulation. It has been observed that like Fig. 2 no significant difference practically observed in two different observer responses with the bound $|R_a|_{\max} = 7.5 \Omega$, as supported from the zoomed portion of the figure.

In Fig. 4, no-load current profile has been shown where step change of voltage fed to the system from the starting of simulation similar to that of Fig. 2. It has been found that both the observers generate similar current profile with little deviation. In Fig. 5, armature current profile under load condition has been shown where step change of load fed to the system from the starting of simulation similar to that of Fig. 3. Again similar current profile with little deviation has been noted.

Fig. 4 No-load current profile for armature voltage variation to 100 V and 200 V at 0.5 and 1 s interval, respectively, with $|R_a|_{\max} = 7.5 \Omega$

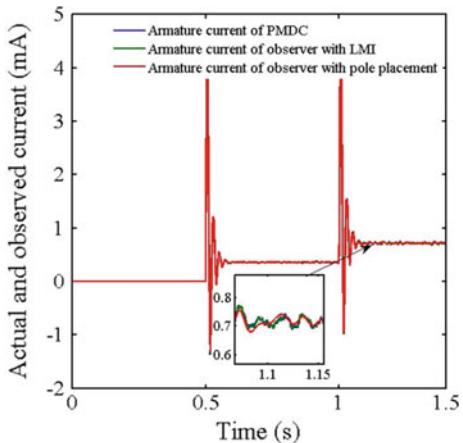
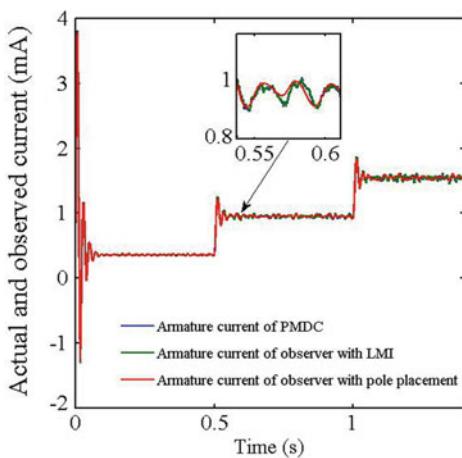


Fig. 5 Current profile with constant armature voltage for load variation to 0.5 Nm and 1 Nm at 0.5 and 1 s interval, respectively, with $|R_a|_{\max} = 7.5 \Omega$



4.2 Observed Error Response Analysis

A performance comparison for observed error response has been illustrated with three different LMI-based observer design against single-pole placement-based design. In Fig. 6, no-load speed estimation error for armature voltage variation with LMI method varies approximately from -1.6 to 1.5 rpm, and that of pole placement method varies from -3 to 9 rpm. In the case of speed, estimation error as shown in Fig. 7 with the variation of load under constant rated armature voltage varies from -5 to 3 rpm with LMI method and -5 to 7 rpm with that of pole placement method.

On the other hand, the no-load current estimation error profile as shown in Fig. 8 for variation armature voltage lies between -60 and 70 mA with LMI method and -60 – 100 mA with that of pole placement method. In the case of current, estimation

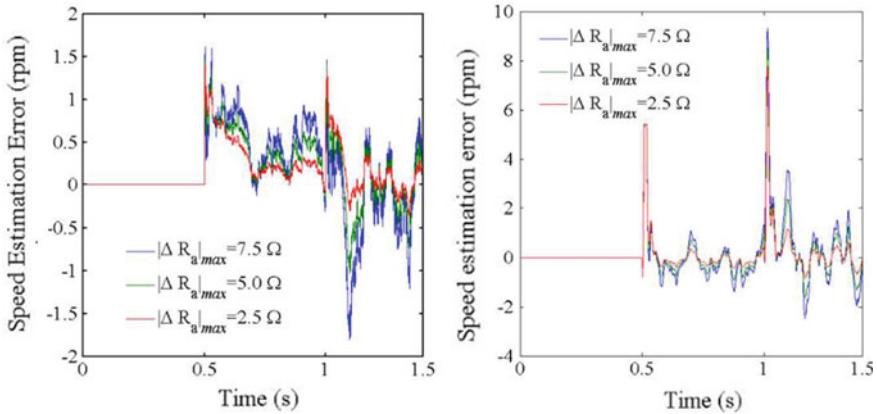


Fig. 6 No-load speed estimation error for armature voltage variation to 100 V and 200 V at 0.5 and 1 s interval, respectively, with LMI (left) and pole placement (right) method

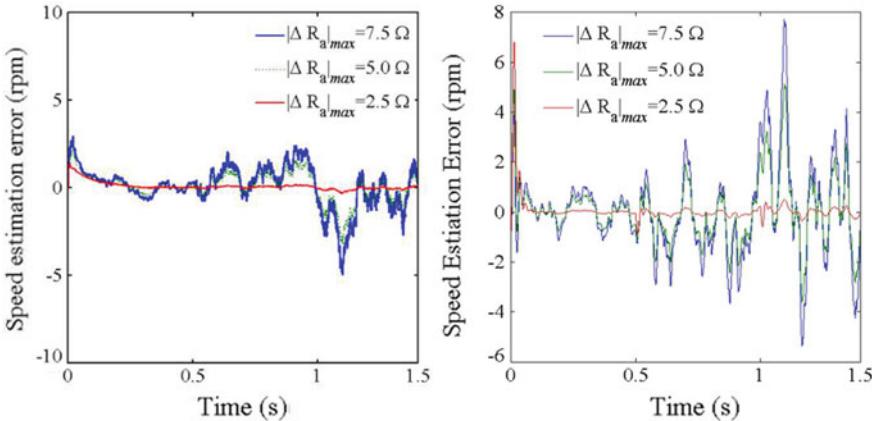


Fig. 7 Speed estimation error with constant armature voltage for load variation to 0.5 Nm and 1 Nm at 0.5 s and 1 s interval, respectively, with LMI (left) and pole placement (right) method

error as shown in Fig. 9 with variation of load under constant armature voltage varies from -40 to 20 rpm with LMI method and -60 – 50 rpm that of pole placement method.

5 Conclusion

The performance comparison of model-based observers, designed with LMI, and pole placement methods has been illustrated in this paper with respect to speed estimation problem of brushed DC motor. The estimated states and estimation errors have been

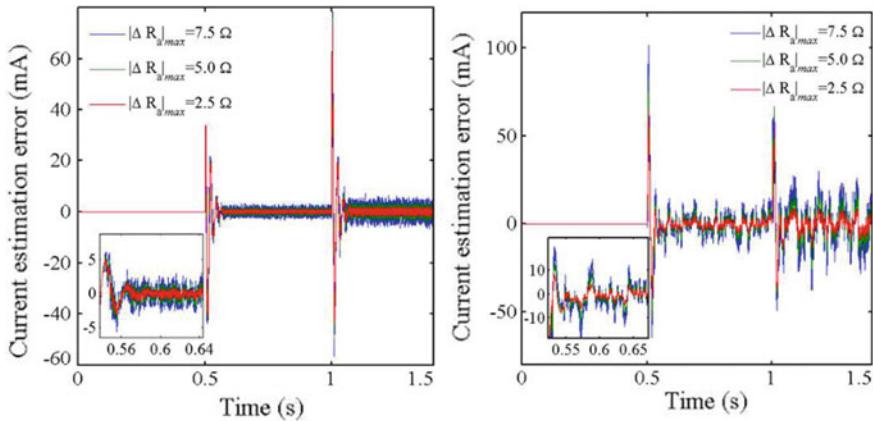


Fig. 8 No-load current estimation error for armature voltage variation to 100 V and 200 V at 0.5 s and 1 s interval, respectively, with LMI (left) and pole placement (right) method

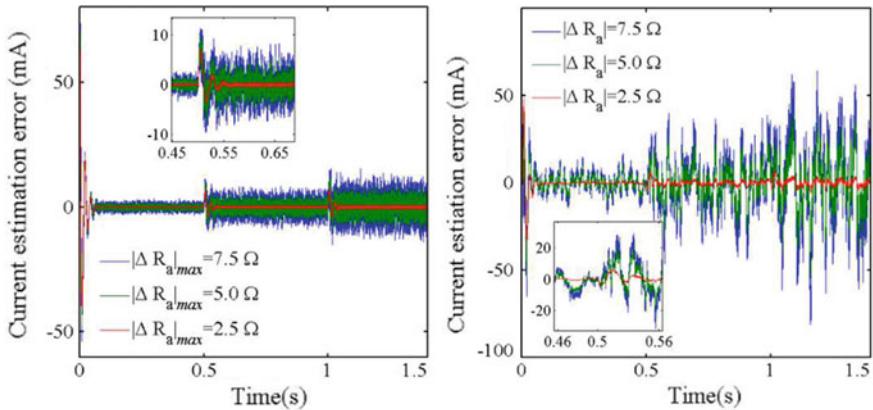


Fig. 9 Current estimation error with constant armature voltage for load variation to 0.5 Nm and 1 Nm at 0.5 s and 1 s interval, respectively, with LMI (left) and pole placement (right) method

analyzed in view of variation of motor parameter accommodating uncertain contact resistance due to brush. It has been shown that both the observers exhibit significantly close estimation of the states. Although LMI-based observer is found better with respect to estimation error, it is essential to know the Jacobian bound as an additional design requirement. In view of pole placement method-based observer design, only nominal parameter values of the system are sufficient. Thus, it is evident that for linear time-varying electromechanical systems, pole placement method-based observer design is an obvious choice when Jacobian bound is not available.

References

1. Rao TM, Ghosh M, Halder B (2016) Effect of pole placement of a full order state observer in sensorless speed estimation of brushed DC motor. In: Proceedings of 7th IEEE power India international conference (PIICON) conference. IEEE Press, Bikaner. <https://doi.org/10.1109/poweri.2016.8077440>
2. Radcliff PJ, Kumar D (2015) Sensorless speed measurement for brushed DC motors. IET Power Electron 8(11):2223–2228
3. Vazquez-Sanchez E, Gomez-Gil JC, Diez-Higuera JF (2012) A new method for sensorless estimation of the speed and position in brushed DC motors using support vector machines. IEEE Tran Ind Eletron 59(3):1397–1408
4. Saurav K, Polturi R (2013) Sensorless speed control of a permanent magnet DC motor by compensating the plant nonlinearities. In: IEEE international symposium on industrial electronics (ISIE). Taipei, Taiwan, pp 1–4
5. Chi CT, Yin SA (2012) Speed measurement of a general DC brushed motor based on sensorless method. In: Proceedings of 10th international power energy conference. Ho Chi Minh City, Vietnam, pp 332–337
6. Turel A, Slavic J, Boltezar M (2017) Electrical contact resistance and wear of a dynamically excited metal-graphite brush. Adv Mech Eng 9(3):1–8. <https://doi.org/10.1177/1687814017694801>
7. Shin WG, Lee SH (2010) An analysis of the main factors on the wear of brushes for automotive small brush type DC motor. J Mech Sci Technol 24(1):37–41. <https://doi.org/10.1007/s12206-009-1135-4>
8. Andreux R, Fontchastagner J, Takorabet N, Labbe N, Metral JS (2014) A general approach for brushed DC machines simulation using a dedicated field/circuit coupled method. Prog Electromagn Res 145:213–227. <https://doi.org/10.2528/pier14011402>
9. Mellah H, Hemsas KE, Taleb R, Cecati C (2018) Estimation of speed, armature temperature, and resistance in brushed DC machines using a CFNN based on BFGS BP. Turk J Elec Eng Comp Sci 26: 3181–3191. <https://doi.org/10.3906/elk-1711-330>
10. Mellah H, Hemsas KE, Taleb R (2016) Intelligent sensor based bayesian neural network for combined parameters and states estimation of a brushed DC motor. Int J Adv Comput Sci Appl 7(7):230–235
11. Luenberger DG (1971) An introduction to observers. IEEE Trans Autom Control 16(6):596–602
12. Utkin VI (1990) Method of separation of motions in observation problems. Automation and remote control 51(3):300–308
13. Kalman RE (1960) A new approach to linear filtering and prediction problems. Trans ASME J Basic Eng 82(Series D):35–45
14. Zemouche A, Rajamani R, Phanomchoeng G, Boulkroune B, Rafaralahy H, Zasadzinski M (2017) Circle criterion based H-infinity observer design for Lipschitz and monotonic nonlinear systems—Enhanced LMI conditions and constructive discussions. Automatica 85:412–425
15. Wang Y, Rajamani R, Bevly DM (2017) Observer design for parameter varying differentiable nonlinear systems. IEEE Trans Autom Control 62:1940–1945
16. Rajamani R (1998) Observer for lipschitz nonlinear systems. IEEE Trans Autom Control 43:397–401
17. Zemouche A, Boutayeb M, Bara GI (2005) Observer design for nonlinear systems: an approach based on the differential mean value theorem. In: Proceedings of 44th IEEE conference on decision and European control, pp 12–15 Seville, Spain

Severity Prediction of Software Vulnerabilities Using Textual Data



Ruchika Malhotra and Vidushi

Abstract Nowadays, all essential activities carried out by society are dependent on software systems. These systems with time have become more and more complex. Increase in complexity leads to introduction of vulnerabilities in the software system. Vulnerabilities are being reported every year and are increasing exponentially with time. These vulnerabilities required to be patched and fixed before getting exploited. To achieve this goal, textual data available on vulnerabilities is needed to be exploited to retrieve valuable information from it. Therefore, in this paper, we aim to develop a prediction model which will take textual description of Apache Tomcat vulnerabilities as input and will predict the severity of the vulnerabilities. Text mining techniques and machine learning algorithms are used to perform this task. The huge volume of textual data available on vulnerabilities has to be prioritized so that more severe vulnerabilities can be targeted first and limited resources can be put to its best use. Hence, it is essential to use text mining techniques for feature selection to reduce the volume of the data. The model being developed uses chi-square and information gain for feature selection, and among the machine learning algorithm, bagging technique, random forest, naïve Bayes, support vector machine are used for prediction. From results, it is observed that comparatively information gain gave better results among feature selection technique. Model based on naïve Bayes algorithm along with information gain as feature selection techniques performed good with the average values of 0.91, 90.90%, and 92.30%, respectively, for AUC, sensitivity, specificity, respectively.

Keywords Vulnerability · Text mining · Feature selection · Machine learning · Prediction model

R. Malhotra (✉) · Vidushi

Delhi Technological University, Shahbad Daulatpur, Main Bawana Road, Delhi 110042, India
e-mail: ruchikamalhotra@dtu.ac.in

1 Introduction

In the current scenario, as the dependency of almost all the major human activities is increasing on various software systems, it is becoming crucial for the developers to keep in mind the security aspects of these systems. Software systems currently being developed are huge and complex and are likely to have various issues regarding its security. In order to develop systems having zero security issues, during the development process, testing and debugging should be done thoroughly. Such type of softwares with zero security issues can be referred as high-quality software system. However, because of limited available resources, there is a high possibility of ending up with softwares with various security issues such as vulnerabilities. In general, vulnerability can be defined as ability to be easily attacked. In the field of computer science, vulnerability refers to a flaw or a weakness in a system which can be exploited by an attacker to do unauthorized activities in a system [1].

If the vulnerabilities are not fixed, it can lead to various serious impacts such as information leakage or elevate user privileges or grant otherwise unauthorized access [2]. With each passing year, huge amount of data is being added on vulnerabilities for a software system, which is in the form of text. This textual data is increasing exponentially with time. To perform appropriate actions effectively on the system in order to remove the vulnerabilities making it secure, this huge amount of data which is available on vulnerabilities should be prioritized as severe and non-severe so that the amount of data needed to be considered and to be focused on can be reduced. Afterward, according to the prioritization, these vulnerabilities can be looked upon and resolved.

Hence, prioritization of vulnerability greatly helps in reducing the amount of data to be focused on in the process of removing the vulnerability from the software systems. This reduction of data greatly supports in fast fixing of vulnerabilities, dealing with the problem of limited number of resources available. Therefore, a need arises to develop a model which can help in the prediction of severity of vulnerability so that it can be prioritized and solved accordingly.

In this paper, an approach is proposed to identify whether vulnerability is severe or not. The approach aimed at developing a model which can perform the task efficiently. The modeling procedures will be dependent on the exploitation of various text mining techniques and machine learning algorithms. Text mining techniques will be used to handle the data present in the form of raw text and convert it from unstructured form to structured data. There are various steps which are involved in order to achieve this. These steps are elaborated and explained in Sect. 3 of the paper. After generation of the structured data, machine learning techniques will be applied on the data, to predict the severity of the vulnerability.

This paper uses security vulnerabilities of Apache Tomcat of last 20 years. It contained lots of information such as ID, publish date, update date, vulnerability type, but for our work, the main concern was with vulnerability description and the common vulnerability scoring system (CVSS) score. It is a scoring system which is formulated to rate the vulnerabilities in the range of 0–10 [3]. To the extent of our

knowledge, the data set used in this paper is never been used as an input to any such model, and it is the first time that such analysis will be done on this data set.

At the end, evaluation is done of the models which are proposed in this study. These models are assessed and evaluated on the basis of various performance measures such as sensitivity, specificity, and AUC. Therefore, the following research questions are investigated in this study:

RQ1: What is the effectiveness of software vulnerability severity prediction (SVSP) model developed using chi-square feature selection technique for dimensionality reduction?

This question tries to find the effectiveness of the models which are based on chi-square feature selection technique. In this paper, effectiveness is measured in terms of sensitivity, specificity, and AUC, and according to the values of these parameters, performance of the models is analyzed.

RQ2: What is the effectiveness of SVSP model developed using information gain feature selection technique for dimensionality reduction?

As described above, here in this question, same task is done, i.e., finding the effectiveness of the models but which are using information gain technique. Effectiveness is judged on the basis of similar parameters which are stated in the above RQ, i.e., RQ1.

RQ3: In terms of dimensionality reduction, how the effectiveness of the model vary with the use of chi-square and information gain, respectively?

The answer of this RQ helps us to investigate all the models and find a fruitful result that which dimensionality reduction technique worked better when coupled with various machine learning algorithms.

The rest of the paper is structured as follows: In the next section, i.e., Sect. 2, the related research work is summarized. In Sect. 3, the approach proposed by the paper is presented. In Sect. 4, experimental results observed are documented in a tabular form, and finally in Sect. 5, the paper is concluded by providing key findings and the future work.

2 Related Research Work

This is the area which is not much explored by the researchers, and very limited amount of work is done exploiting the textual data available on vulnerabilities. In this section ahead, the related research work done on vulnerabilities by the researchers is discussed.

Bozorgi et al. [4] presented a methodology to classify the vulnerabilities. Open-source vulnerability database (OSVDB) and MITRE common vulnerabilities and exposures (CVE) database were used as the data source containing vulnerability reports. Classifier is built using SVM which classifies vulnerabilities into positive and negative examples. The error rate of the classifier developed is found to be 14%, a false negative rate of 9%, and a false positive rate of 5%.

A study by Hovsepyan, et al. [5] used Java files of 19 versions of the K9 mail client application of an android phone as data source. Text mining techniques are applied on the source code and converted those to feature vector. Afterward, SVM was used to classify these vectors as vulnerable or clean. Prediction model developed found to have average accuracy of 0.87, precision of 0.85, and recall of 0.88.

Sadeghi et al. [6] used HP fortify as the main static analysis tool providing built-in vulnerability detection rule. Probabilistic rule classifier was used further to rank the vulnerability on the basis of their occurrence in the analysis report generated by Fortify. Android applications belonging to different categories were used as data source. Probabilistic rule classifier applies conditional probability to find the likelihood of occurrence of each vulnerability in each category.

Zhang et al. [7] proposed an approach and coined a term VulPredictor to predict vulnerable files and labeled the files as vulnerable or not. VulPredictor used both software metrics as well as text features as input variables. Datasets from Drupal, phpMyAdmin, and Moodle were used as data source, and experimental results showed that VulPredictor can achieve F1 and effectiveness ratio at 20% scores of up to 0.683 and 75%, respectively.

Khazaei et al. [8] investigated the textual data available regarding vulnerability and developed a solution predictor for newly detected software vulnerability. Overlapping SOM algorithm (OSOM) algorithm was used to categorize the existing vulnerabilities and their solutions. After categorization, support vector machine (SVM) and random forest algorithms were used to predict the type of solution for newly reported vulnerabilities.

Pang et al. [9] tried to identify the vulnerable components of software by proposing a hybrid technique which uses SVM algorithm and ensemble learning strategy. Source code of five android applications is used in this study as data source. The average accuracy, precision, and recall values are found to be 75.98%, 58.75%, and 64.78%, respectively.

Gupta et al. [10] used text mining and pattern matching to label PHP source code files as yes if vulnerable and as no if non-vulnerable. Seven machine learning algorithms were used to do the prediction. Highest precision was obtained of 99.29 when JRip was used. On the other hand, bagging produced best result in terms of recall value of 86.87.

Dam et al. [11] tried to explore long short-term memory model to evaluate source code of 18 android applications on the basis of its semantic and syntactic features and joint features. Recall values were observed to be 0.87, 0.87, 0.86 for semantic and syntactic features and joint features, respectively.

3 Proposed Approach

The main aim of this work is to develop a prediction model which takes the textual data available on vulnerabilities as input, and after application of the text mining techniques and machine learning algorithms, producing an output predicting whether

the particular vulnerability of a given description is severe or not. A given description of vulnerability is considered to be severe or non-severe according to Eq. (1) which is defined as follows:

$$\text{Vulnerability} = \begin{cases} \text{severe,} & | \text{ CVSS score} \geq \text{threshold} \\ \text{non severe,} & | \text{ otherwise} \end{cases} \quad (1)$$

As mentioned above in Sect. 1, CVSS score rates the vulnerabilities in the range of 0–10. This severity score allows us to prioritize the resources to be able to solve more threatening vulnerabilities first. According to most recent version of CVSS, v3.0, score of 0.0 receives a “None” rating, 0.1–3.9 score gets a “Low” severity rating, score of 4–6.9 is a “Medium” rating, score of 7–8.9 is a “High” rating, and score of 9–10 is a “Critical” rating [3]. The framework of our proposed model is depicted in Fig. 1 which considered critical and high-rating vulnerabilities as severe and others as non-severe. Hence, the threshold is decided to be 6.9. Figure 1 presents that the entire approach can be divided broadly into two phases. The first phase is model building phase followed by the second phase, i.e., prediction phase. The focus of the first phase is to train the classifier being developed. To achieve this aim, the first step is to apply text mining techniques on the vulnerability textual data collected to conduct this experiment. In text mining, initially the training data is being preprocessed [12]. Preprocessing includes tokenization which means separation of text into words where every word is considered to be a token. Stop word removal is done after tokenization which means elimination of all the insignificant words from the textual data so that important and useful works containing the information can be focused. “Is,” “the,” “for” are some examples of stop words of English language. For the conduct of this experiment, the default list of stop words present in Python’s

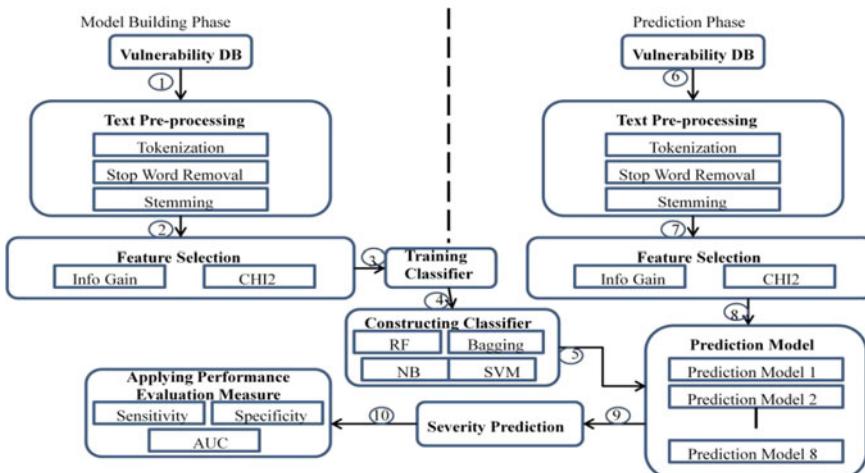


Fig. 1 Proposed approach

nlkt package was used along with few additions. After the removal of stop words, final step of preprocessing, i.e., stemming, is performed. It is a process in which all the words are reduced into their root form. For example, “likes,” “liking,” “liked” all are reduced into their root word, i.e., “like.” To achieve this, many algorithms have been developed by researchers, but the most widely used and accepted algorithm was Porter Stemmer algorithm [13] which is used in this study as well. By the end of this entire process of preprocessing, we are left with unique and meaningful tokens. These tokens are nothing but the attributes or the dimensions or the features around which our model will make the predictions.

These features representing vulnerabilities are converted into feature vectors. Feature vectors are constructed using “bag-of-words” formed after data preprocessing. The length of the feature vector is equivalent to the number of tokens in “bag-of-words,” and for a feature vector of a particular vulnerability, it contains the information about the frequency of each token in the textual description of the respective vulnerability. For further filtering of features based on their importance, weighing is done using $tf * idf$ approach [14]. It is a weight which is calculated for every token using its term frequency, i.e., tf and inverse document frequency, i.e., idf . Calculation of this weight is done according to Eq. (2).

$$\text{For a term } i \text{ in document } j : w_{i,j} = tf_{i,j} \cdot X \log\left(\frac{N}{df_i}\right) \quad (2)$$

$tf_{i,j}$ frequency if i in j

df_i number of description containing i

N total number of description.

Dimensionality reduction is done further which helps in the reduction of features by selecting the subset of features and removing the irrelevant features from the entire set. Irrelevant features are those which do not contribute in finding the prediction variable or the output. This technique of dimensionality reduction is known as feature selection. For this purpose, two techniques, namely chi-square and information gain, were used.

Chi-square [15] is a statistical test which is used by the researchers to find if there is dependence between two variables or not. The value of chi-square is calculated according to Eq. (3) listed below.

$$\text{CHI}(t_i, c_j) = \frac{N * (ps - qr)^2}{(p + q) * (q + r) * (p + q) * (r + s)} \quad (3)$$

In the above equation, N represents the total number of vulnerabilities, p is the number of vulnerabilities having t_i term and belonging to c_j category, q is the number of vulnerabilities having t_i term but does belong to c_j category, r is the number of vulnerabilities not having t_i term but belonging to c_j category, and s is the number of vulnerabilities neither having t_i term nor belonging to c_j category. Hence, finally Eq. (3) gives the chi-square score of term t_i and category c_j .

Larger the value of chi-square statistics indicates, higher the significance of the term t_i . Likewise, the score is calculated for all the terms, and top n terms having the highest score are selected after dimension reduction.

The second technique used for dimensionality reduction via feature selection is information gain. Information gain [16] aims at finding out those words that best simplifies the target concept. It tries to find the total number of bits required to encode arbitrary class distribution C is $H(C)$ as follows:

$$H(C) = - \sum_{c \in C} \frac{n(c)}{N} \log_2 p(c) \quad (4)$$

where

$$N = \sum_{c \in C} n(c) \quad (5)$$

The number of bits required to encode a class after observing an attribute A will be:

$$H(C|A) = - \sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2(p(c|a)) \quad (6)$$

The attribute having the highest information gain is the one with highest rank which is denoted by the symbol A_i

$$\text{Infogain}(A_i) = H(C) - H(C/A_i) \quad (7)$$

In a nutshell, it can be said that a series of methods were applied starting from tokenization, stop word removal, stemming, vectorization, $tf * idf$ weight calculation, info gain, or chi-square to do dimensionality reduction.

After getting the final list of attributes, the model is trained using the training data to do the prediction using machine learning algorithms. Among the various machine learning algorithms, decision tree, multinomial naïve Bayes, bagging technique, random forest, naïve Bayes, support vector machine were used to make the predictions. The details of these algorithms can be found in [17]. From the above-mentioned algorithms, the first two are not documented in this paper because of their poor performance. Hence, this lead to the development of eight prediction models by combining one feature selection technique with one machine learning algorithm. The performance of the model is tested on the basis of result of prediction which it gave over testing data. The aspects regarding the performance measurement of the prediction model is discussed in next section that is Sect. 4.

4 Result Analysis

Confusion matrix and the measures derived from its elements were used to analyze and access the performance of the prediction models. Confusion matrix has four basic components, namely true positives (tp), true negatives (tn), false positives (fp), and false negatives (fn). In context to our experiment, we can say tp is the number of those vulnerabilities having severity as 1 is classified as having severity 1 where as tn refers to vulnerabilities having severity as 0 and classified as well having 0 severity. Similarly, fp represents those vulnerabilities having 0 severities but predicted to have severity of 1 where as fn represents those vulnerabilities having severity as 1 but classified to be of 0 severity. 1 is the representation of severe, and 0 is the representation of non-severe. Using these basic components of confusion matrix, other parameters such as AUC, sensitivity, specificity were derived.

AUC [18] is an abbreviation for area under ROC curve, popularly also known as area under curve. Receiver operating characteristics is known to be ROC. It is a plot of sensitivity values and specificity values on x-axis and y-axis, respectively. Accuracy of a model is measured in terms of value of AUC. The value ranges between 0 and 1. Prediction model is considered to give good results if the value AUC is large and reaching to 1. Sensitivity and specificity [19] both are statistical measures to evaluate the performance of a prediction model. Sensitivity popularly also known as recall is nothing but true positive rate. It identifies the ratio of true positives that are correctly identified. It can be calculated according to the formula specified in Eq. (8)

$$\text{Sensitivity} = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (8)$$

Specificity on the other hand is also known as true negative rate. It calculates the ratio of actual negatives that are correctly identified. It can be calculated according to the formula specified in Eq. (9).

$$\text{Specificity} = \frac{\text{tn}}{\text{tn} + \text{fp}} \quad (9)$$

Major machine learning algorithms along with feature selection technique when applied on vulnerability dataset of Apache Tomcat gave the results as depicted in Tables 1 and 2. The feature selection technique used is chi-square and information gain, and the ML algorithms used in this experimental setup are bagging technique, random forest, naïve Bayes, and support vector machine.

Table 1 depicts the value of all performance measures for all four machine learning algorithms when applied along with chi-square as feature selection technique. Similarly, Table 2 depicts the value of all performance measures for all four machine learning algorithms when applied along with information gain as a feature selection technique. The values in the table are rounded off to two decimal places. Depending upon all these parameters which are discussed above, the RQs which are listed in Section I are addressed here in this section as follows:

Table 1 Results using chi-square

No. of terms		5	10	50	100
Bagging	AUC	0.62	0.65	0.78	0.79
	Sensitivity	1	1	0.95	0.82
	Specificity	0.23	0.31	0.62	0.77
RF	AUC	0.58	0.65	0.68	0.70
	Sensitivity	1	1	0.91	0.86
	Specificity	0.15	0.31	0.46	0.54
NB	AUC	0.62	0.58	0.73	0.75
	Sensitivity	1	1	1	0.95
	Specificity	0.23	0.15	0.46	0.54
SVM	AUC	0.58	0.65	0.69	0.75
	Sensitivity	1	1	1	0.95
	Specificity	0.15	0.31	0.38	0.54

Table 2 Results using information gain

Number of terms		5	10	50	100
Bagging	AUC	0.59	0.62	0.74	0.78
	Sensitivity	0.869	0.86	0.64	0.73
	Specificity	0.31	0.38	0.85	0.85
RF	AUC	0.55	0.56	0.69	0.78
	Sensitivity	0.86	0.90	0.68	0.95
	Specificity	0.23	0.23	0.69	0.62
NB	AUC	0.61	0.65	0.78	0.92
	Sensitivity	0.91	1	0.95	0.91
	Specificity	0.31	0.31	0.61	0.92
SVM	AUC	0.58	0.58	0.65	0.73
	Sensitivity	0.86	0.86	0.91	1
	Specificity	0.31	0.31	0.38	0.46

RQ1: What is the effectiveness of software vulnerability severity prediction (SVSP) model developed using chi-square feature selection technique for dimensionality reduction?

To answer this RQ, values of Table 1 can be referred. It gives the results in terms of AUC, sensitivity, specificity for top-5, top-10, top-50, top-100 terms selected by feature selection technique. As can be seen from Table 1, when chi-square is used as feature selection technique, model using bagging technique performed fairly enough even though in the case when the number of terms selected was as low as 5. The value of AUC was found to be 0.61, 0.65, 0.78, 0.79 for top-5, top-10, top-50, top-100 terms, respectively. It is also observed that the model performed really well

when the number of terms selected was large irrespective of the machine learning technique used. As can be seen from Table 1, model using RF technique gave the lowest value of 0.70 for AUC even for top-100 terms as compared with other models. The specificity results had values as 15.38% and 30.7% for top-5 and top-10 terms, respectively, which increased to 46.15% and 53.84% in case of top-50 and top-100 terms, respectively. Therefore, in general, it can be deduced that a model performance is dependent on the number of top n terms used as input, and it is found that model gave good performance for large number of top n terms.

After the complete analysis of the values documented in Table 1, we can state that the model using bagging technique and top-100 terms as input performed the best with the values of 0.79, 81%, and 76% for AUC, sensitivity and specificity, respectively. On the other hand, model using RF gave low performance with the values of 0.57, 0.65, 0.68, 0.70 of AUC for top-5, top-10, top-50, top-100 terms, respectively.

RQ2: What is the effectiveness of SVSP model developed using information gain feature selection technique for dimensionality reduction?

To answer this RQ, values of Table 2 can be referred. In case of information gain, the highest value was obtained by the model using naïve Bayes as machine learning algorithm. The value of AUC for top-100 terms was 0.91 with specificity value of 92.30%. The model gave an average performance with the values of 0.60, 0.65, 0.78, 0.91 for AUC for top-5, top-10, top-50, top-100 terms, respectively. As observed in Table 1, similar trend is seen in the result values of Table 2; i.e., models gave a good performance in case of top-100 terms, and the performance degraded as the number of terms reduced to top-5. When SVM was used to develop the model, it gave the lowest value of AUC of 0.73 as compared to other models. The specificity results had values as 30.77% and 30.77% for top-5 and top-10 terms, respectively, which increased to 38.46% and 46.15% in case of top-50 and top-100 terms, respectively. Hence, in a nutshell, it can be stated that, for model using information gain as feature selection technique, naïve Bayes gave the best results with the values of 0.91, 91%, and 92.30% for AUC, sensitivity and specificity, respectively, where as SVM gave the worst performance with the values of 0.58, 0.58, 0.64, 0.73 of AUC for top-5, top-10, top-50, top-100 terms, respectively.

RQ3: In terms of dimensionality reduction, how the effectiveness of the model vary with the use of chi-square and information gain, respectively?

If a comparison has to be made between the performances of feature selection techniques, then we can say, use of information gain resulted in better results overall. From the results depicted in Table 1, bagging technique came out to be the best technique. In case of both the models using bagging technique, the specificity value came out to be 76.92% for chi-square feature selection technique and came out to be 84.61% in case of information gain feature selection technique indicating information gain is a better technique for dimensionality reduction. Similarly, from result values of Table 2, naïve Bayes is deduced to be the best technique. In case of both the models using naïve Bayes technique, the specificity value came out to be 53.84% for chi-square feature selection technique and came out to be 92.30% in case of information gain feature selection technique, again indicating information gain is

a better technique for dimensionality reduction. The AUC values ranged between 0.70 and 0.79 in case of chi-square and 0.73–0.91 in case of information gains. The AUC values compared are for models using top-100 terms. Other values can also be analyzed vividly by comparing the values of Tables 1 and 2 giving the same deduction.

Thus, we can conclude that model based on naïve Bayes algorithm along with information gain feature selection techniques performed the best with the values of 0.91, 90.90%, and 92.30%, respectively, for AUC, sensitivity, specificity, respectively, for the model using top-100 terms.

5 Conclusion and Findings

Identification of vulnerabilities has become very crucial with increase in dependence on software systems for all activities. Major crucial activities are carried out on Internet which leads to the threat of exploitation of these vulnerabilities. Lots of textual data is available on vulnerabilities which is increasing exponentially with time and is not much explored by the researchers. Because of limited number of resources, it becomes quite essential for researchers to reduce the amount of data to be focused on first by finding the level of severity of vulnerability. Therefore, it becomes significant to develop a model which can identify the severity of vulnerabilities with the help of text mining and machine learning techniques.

As a result, in this paper, we aimed for developing a prediction model which will predict the severity of the vulnerabilities present in software. Machine learning algorithms like bagging technique, random forest, naïve Bayes, and support vector machine were applied along with text mining technique for feature selection. Chi-square and information gain are used as feature selection techniques along with a sequence of application of text mining techniques. This leads to eight combinations hence development of eight prediction models. These eight prediction models took security vulnerabilities of Apache Tomcat of last 20 years as input files. The performance of the models is tabulated in the paper selecting top-5, top-10, top-50, and top-100 terms. For gauging the performance of the models, AUC, sensitivity and specificity are the measures used in this paper. From the results, it is observed that overall information gain in combination with other machine learning techniques performed better in comparison with chi-square as feature selection technique. The best performance was given by the model using naïve Bayes and information gain as the combination of algorithms with values of 0.91, 90.90%, and 92.30%, respectively, for AUC, sensitivity, specificity, respectively.

In future, we plan to include more databases on vulnerabilities in addition to the one used in this study. As there is always a scope of improvement, we also intend to improve the performance of the model by developing a novel technique for feature selection. Apart from this, we can integrate a proper validation technique to validate the results of the model being developed.

References

1. Meunier P (2008) Classes of vulnerabilities and attacks. In: Wiley handbook of science and technology for homeland security
2. Winkler I, Gomes AI (2017) A cyberwarfare approach to implementing adaptive enterprise protection, detection, and reaction strategies, how to hack computers. In: Advanced Persistent Security, pp 41–46. <https://doi.org/10.1016/b978-0-12-809316-0.00005-1>
3. Zerkane S, Espes D, Le Parc P, Cuppens F (2017) Vulnerability analysis of software defined networking. In: International symposium on foundations and practice of security, pp 97–116. https://doi.org/10.1007/978-3-319-51966-1_7
4. Mehran B, Saul LK, Savage S, Voelker GM (2010) Beyond heuristics: learning to classify vulnerabilities and predict exploits. In: International conference on knowledge discovery and data mining, pp 105–114
5. Hovsepyan A, Scandariato R, Joosen W, Walden J (2012) Software vulnerability prediction using text analysis techniques. In: International workshop on Security measurements and metrics, pp 7–10. <https://doi.org/10.1145/2372225.2372230>
6. Sadeghi A, Esfahani N, Malek S (2014) Mining the categorized software repositories to improve the analysis of security vulnerabilities. In: International conference on fundamental approaches to software engineering, pp 155–169
7. Zhang Y, Lo D, Xia X, Xu B, Sun J, Li S (2015) Combining software metrics and text features for vulnerable file prediction. In: 20th international conferences on engineering of complex computer systems (ICECCS), pp 40–49. <https://doi.org/10.1109/iceccs.2015.15>
8. Khazaei A, Ghasemzadeh M, Meinel C (2016) Solution prediction for vulnerabilities using textual data. In: International conference applied computing
9. Pang Y, Xue X, Namin A S (2016) Early Identification of Vulnerable Software Components via Ensemble Learning. In: International Conference on Machine Learning and Applications (ICMLA), pp 476–481. <https://doi.org/10.1109/icmla.2016.0084>
10. Gupta MK, Govil MC, Singh G (2018) Text-mining and pattern-matching based prediction models for detecting vulnerable files in web applications. *J Web Eng* 17:028–044
11. Dam HK, Trany T, Phamy T, Ng SW, Grundyy J, Ghose A (2018) Automatic feature learning for vulnerability prediction In: IEEE Transactions on Software Engineering, pp 1–12. <https://doi.org/10.1109/tse.2018.2881961>
12. Vijayarani S, Ilamathi J, Nithya S (2015) Preprocessing techniques for text mining—an overview. *Int J Comput Sci Commun Netw* 5(1):7–16
13. Porter M (1980) An Algorithm for Suffix Stripping, Program, pp 30–137
14. Keefe TM, Koprinska I (2009) Feature selection and weighting methods in sentiment analysis. in: australasian document computing symposium
15. Yiming Y, Pederson JO (1997) A comparative study on feature selection in text categorization. *Int Conf Mach Learn* 97:412–420
16. Menzies T, Mracus A (2008) Automated severity assessment of software defect reports. In: IEEE international conference on software maintenance (ICSM)
17. Witten IH, Frank E, Hall MA (2011) Data mining: practical machine learning tools and techniques. Morgan Kaufmann Publishers, Burlington, MH
18. Jindal R, Malhotra R, Jain A (2015) Mining defect reports for predicting software maintenance effort. In: International conference on advances in computing, communications and informatics (ICACCI), pp 270–276. <https://doi.org/10.1109/icacci.2015.7275620>
19. Jiang Y, Cukic B, Ma Y (2008) Techniques for evaluating fault prediction models. *Empirical Software Eng* 13(15):561–595

Data Analysis of Cricket Score Prediction



Suyoga Srinivas, Naveen N. Bhat, and M. Revanasiddappa

Abstract Cricket, a game played on 22-yard strip among 22 players with a bat and ball, is one among the most popular sport in the world. Even though it is played by a lesser number of countries, the sport is followed all across the globe. People are not only interested in following this sport, they also try predicting the flow of the match. Predicting the flow of a cricket match has always been a strenuous task as a particular player may not perform the same way against every opposition nor will his performance be the same in every venue. Also, the way a player performs depends according to the dynamics of the game. While predicting the flow of the match, a majority of the people tend to give more weightage to the previous results. In this paper, we have come up with a model which not only takes previous results into consideration but also the opposition, the venue and the current state of the match such as, number of wickets fallen, number of overs remaining, the way the players have fared till that moment. We have developed various algorithms to find batting and bowling index of the players involved in the match. These indices, along with a special feature, RunFactor form the input parameters to our machine learning model. The generated output from this model is the number of runs that will be scored in the particular over. Compiling these, we estimate the final score scored by that team in a One Day International (ODI).

Keywords Data analysis · Machine learning · Multivariate linear regression · Score prediction · Support vector regression

S. Srinivas (✉) · N. N. Bhat · M. Revanasiddappa
PES University, Bangalore 560100, India
e-mail: suyoga06031999@gmail.com

N. N. Bhat
e-mail: naveennbhat22@gmail.com

M. Revanasiddappa
e-mail: revanasiddappam@pes.edu

1 Introduction

Cricket is the most viewed sport in the world after football [1]. It is played among eleven players from each side. Originated in the sixteenth century in the southeastern parts of England, the game is followed across the world. In this paper, we will be predicting the runs scored by a team in every over in ODI format.

The reason for choosing ODI over the test match is because it might not be feasible to wait for five days to get the result of our prediction. In a T20I, the entire game is played for 40 overs due to which we do not get enough data points to predict.

A 50 over World Cup is played once in four years. Started in the year 1975, the top teams in the world compete against each other to claim the championship. In this paper, we have used our model to predict the flow of all matches in the Cricket World Cup 2019 played at England and Wales. The 48 matches played in this World Cup had each team facing each other. A total of ten different venues hosted these matches. Hence, we had sufficient data on which we could predict the course of the match.

It was a challenging task to predict the total runs scored by a team as the highest runs scored was 397 and the lowest was 105 [2]. Also, the teams in this World Cup were not consistent. For example, India scored 352 against Australia [3], whereas the same batting line up could only manage 224 against Afghanistan [4].

Data analysis in sports is not something new. Researches have been conducted in other sports as well to predict the outcome of a match. Bhandari et al. [5] developed a tool that is now used by NBA teams to strategize their game. Luckner et al. [6] predicted the outcome of the 2006 FIFA World Cup. Similarly, Gartheepan et al. [7] have conducted their research in baseball. However, these models are specifically developed for their particular sports and cannot give an accurate result when applied to cricket.

Duckworth and Lewis [8] were the first in cricket to propose a model in cricket. This method or commonly known as the DL method is currently used in cricket, which allows a fair adjustment of targets when due to time lost in a limited over match. This method considers the resources remaining. The resources remaining constitute the number of wickets and the number of overs that were remaining before the interruption. This method fails to consider the strength of the batsman who is yet to come or the skills of the bowler who couldn't bowl due to the interruption. Apart from this, Lewis [9], Lemmer [10], Alsoppp and Clarke [11], and Beaudoin [12] have developed performance measures to rate teams. With the change in rules and the way in which the batsmen have dominated the sport recently, we felt there is a need for a fresh model to analyze the sport.

In our paper, we have used a weighted combination of existing mined data and current state as an input to our model to accurately predict the runs scored in the coming overs. In the upcoming parts of the paper, we have mentioned how we have implemented our model and the results we have obtained.

2 Methodology

2.1 Data Collection

The first step of a successful data analysis model is to collect data. For our research work, we have collected our data from the Web site www.espncricinfo.com [100]. The data consists of only completed (where a result was declared) One Day International matches played between the 2011 and the 2019 World Cups. Change of rules post the 2011 World Cup made cricket a batsman dominant sport. Hence, we have not considered data from matches played before the 2011 World Cup. Runs scored, Balls faced and Average (runs scored per dismissal) for a batsman, Wickets taken, Economy (Runs conceded per over) and Average (Runs conceded per dismissal) for a bowler were scraped. Ground stats such as average run rate and average runs per dismissal of venues which hosted at least a completed One Day International match was also recorded.

2.2 Algorithms

We have used these algorithms to find out the four independent variables that we pass to our machine learning models.

Algorithm 1 RunFactor:

```

Input: Avg_Score, Runs
Output: RunFact
1:   if Runs >= Avg_Score:
2:     Append Runs to L1
3:     c = 1
4:   else:
5:     Append Runs to L2
6:     c = 2
7:   Ratio = Average(L1) / Average(L2)
8:   if c == 1:
9:     R1 = R1 + Ratio
10:  else:
11:    R1 = R1 - Ratio
12:  if R1 > 1:
13:    R1 = 1
14:  If R1 < 0:
15:    R1 = 0
16:  R2 = 1-R1
17:  RunFact = (R1*Average(L1)
+ R2*Average(L2))
18:  Return RunFact

```

Algorithm 2 BatIndex:

```

Input: Avg_Sr, RunsPerOver, BallsFaced
Output: BatSr
1:   Sr = RunsPerOver/BallsFaced
2:   if Sr >= Avg_Sr:
3:     Append Sr to G1
4:     k = 1
5:   else:
6:     Append Sr to G2
7:     k = 2
8:   Ratio = Average(G1) / Average(G2)
9:   if k == 1:
10:    S1 = S1 + Ratio
11:  else:
12:    S1 = S1 - Ratio
13:  if R1 > 1:
14:    S1 = 1
15:  If R1 < 0:
16:    S1 = 0
17:  S2 = 1-S1
18:  BatSr = (S1*Average(G1)
+ S2*Average(G2))
19:  Return BatSr

```

Algorithm 1 RunFactor takes Avg_Score and Runs as input parameters, where Avg_Score is the average run rate (runs scored per over) in that particular ground taken across all completed One Day International matches. Runs are the number of runs scored in that particular over. L1 and L2 are the two lists we are using to classify the runs scored in every over. R1 and R2 are the corresponding weights for the lists L1 and L2. Avg_Score acts as a threshold value. If Runs is greater than the threshold value, it is appended to the list L1 else to L2. We calculate RunFact as shown in step 17 of Algorithm 1, RunFactor. The obtained value of RunFact is returned.

Algorithm 2 BatIndex takes Avg_Sr, RunsPerOver, and BallsFaced as input parameters, where Avg_Sr is the average strike rate (runs scored per 100 balls by a batsman) of a batsman taken across all completed One Day International Matches. RunsPerOver represents the number of runs scored in that particular over by that particular batsman. BallsFaced denotes the number of balls faced by that particular batsman in that particular over. Sr is calculated as the ratio between the RunsPerOver to BallsFaced. G1 and G2 are the two lists, and we are using to classify each batsman's strike rate every over. S1 and S2 are the corresponding weights for the lists G1 and G2. Avg_Sr acts as a threshold value. If Sr is greater than the threshold value, it is appended to the list G1 else to G2. We calculate BatSr as shown in step 18 of Algorithm 2, BatIndex. The obtained value of BatSr is returned.

Similarly, we have used an algorithm to find BowlEr by taking in average economy and runs given per over as inputs.

We now have BatSr of two batsmen at the crease, BowlEr of the Bowler bowling and RunFact. With these four parameters, we predict the runs to be scored in the upcoming over, using support vector regressor and multivariate linear regression. For example, to predict the runs to be scored in 26th over of an innings, we take the data of first 25 overs and train our model.

2.3 Models Used

Linear Regression. In multivariate linear regression, we effectuate a correlation among various input explanatory independent variables and a dependent output explained variable. These models can be differentiated using the type of correlation among explained and explanatory variables, which are being considered and the number of explanatory variables being used. In this paper, we have used a multivariate regression with four independent variables.

Support Vector Regression. SVR is one of the regression methods in which we try to obtain the best possible line which is a linear hyperplane having the maximum number of points. For minimizing the error, hyperplanes are individualized which gives maximum margin. SVR can be distinguished by different kernels, sparse solution, and different number of support vectors. SVR uses a symmetric loss function which penalizes both higher and lower mistakes equally.

3 Results

We are comparing the results obtained on a few individual matches below.

Figure 1 shows the predicted score, actual score and projected score for the match between England and India. Predicted score, represented by red color is what our model has predicted. Actual score is what England had scored against India and is represented by blue. Projected score is the score predicted by other models and is given by violet color. Figure 2 is a graphical representation of predicted scores of the match between India and New Zealand.

When our model was used for predicting runs for the entire tournament, the following results were obtained.

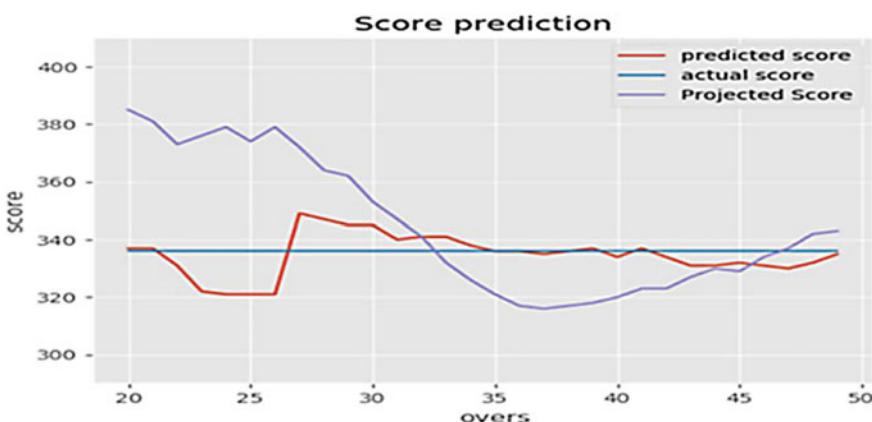


Fig. 1 England vs India

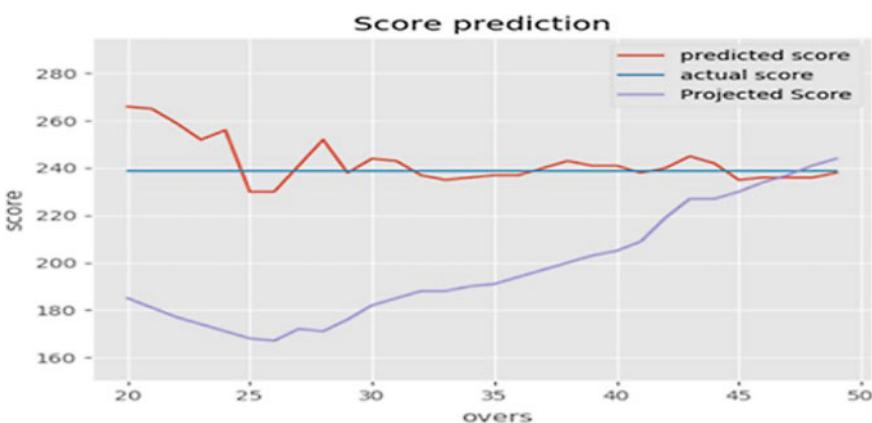


Fig. 2 India vs New Zealand

Table 1 represents the accuracy obtained. Accuracy is calculated by finding the percentage of error in run prediction to the total score.

Table 2 represents the error in run prediction. This is calculated by finding the average error in run prediction per over.

The average column represents the average value obtained from all matches across the entire tournament. The highest and least columns represent the highest and the least values obtained considering all matches in the tournament.

Figures 3, 4, and 5 represent the graphical form of error in run prediction per over. We observe that the average error keeps fluctuating between overs 20 and 29. The reason for the fluctuation is because of the limited training values. As the match progresses, we get a lot more training dataset and our accuracy will improve. This is evident when we look at the average error between the overs 30 and 40. However, there is a rise in average error post the 45th over. This is accounted by the fact that the batting teams try scoring maximum runs. Doing so, they end up losing wickets. Once a new batsman arrives, it is a little difficult to predict how the batsman performs. Even with these problems, our model has given a better result than compared to other models.

From Table 3, we come to know that our accuracy is over 2% better when compared

Table 1 Accuracy obtained

Method	Average	Highest	Least
Support vector regression	95.088	98.37	91.21
Linear regression	94.853	98.4	91.32

Table 2 Error in run prediction

Method	Average	Highest	Least
Support vector regression	1.825	1.03	2.1
Linear regression	1.958	1.133	2.23

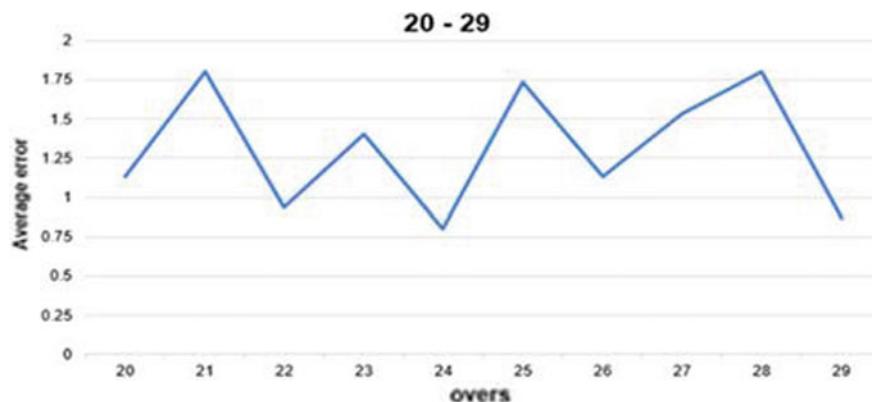


Fig. 3

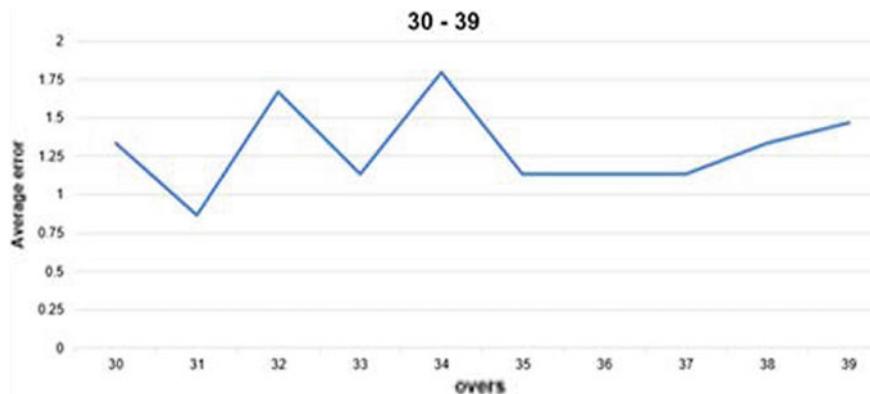


Fig. 4 .

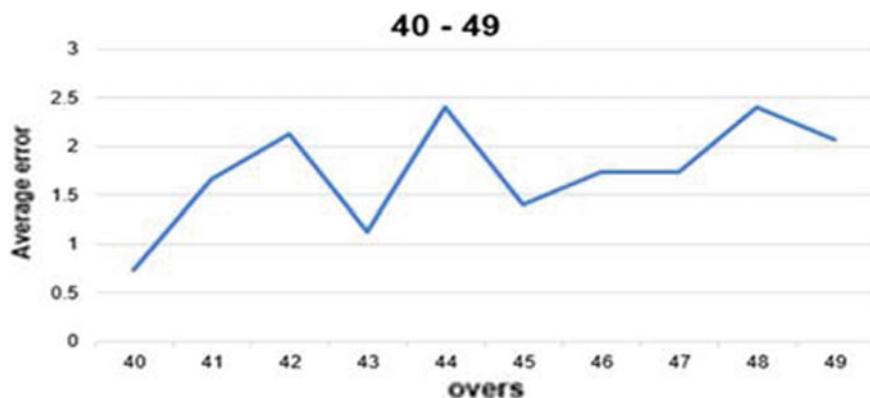


Fig. 5 .

Table 3 Comparison

Model	Accuracy	Average error
Our model	95.088	1.82
Other models	92.88	3.84

to other models. Moreover, other models give more than twice the error given by our model.

With these results, we can confidently claim that our model has predicted a team's total with a higher accuracy than compared to other models.

Acknowledgements The authors thank Dr Suryaprasad J, the Vice-Chancellor of PES University, and the management of PES University Electronic City Campus, Bangalore, for their constant support and encouragement to complete our research.

References

1. Sankaranarayanan VV, Sattar J, Lakshmanan LV Auto-play: a data mining approach to ODI cricket simulation and prediction
2. Cricket World Cup statistics (2019). https://en.wikipedia.org/wiki/2019_Cricket_World_Cup_statistics. Accessed 8 Nov 2019
3. India vs Australia, Match 14—Live Cricket Score, Commentary. <https://www.cricbuzz.com/live-cricket-scorecard/20250/ind-vs-aus-match-14-icc-cricket-world-cup-2019>. Accessed 8 Nov 2019
4. India vs Afghanistan, Match 28—Live Cricket Score, Commentary. <https://www.cricbuzz.com/live-cricket-scorecard/20264/ind-vs-afg-match-28-icc-cricket-world-cup-2019>. Accessed 8 Nov 2019
5. Bhandari I, Colet E, Parker J (1997) Advanced scout: data mining and knowledge discovery in NBA data. *Data Min Knowl Disc* 1(1):121–125
6. Luckner S, Schröder J, Slamka C (2008) On the forecast accuracy of sports prediction markets. In: Negotiation, auctions, and market engineering, international seminar, Dagstuhl Castle, vol 2, pp 227–234
7. Gartheeban G, Guttag J (2013) A data-driven method for in-game decision making in MLB: when to pull a starting pitcher. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, KDD’13, pp 973–979. ACM, New York, NY
8. Duckworth FC, Lewis AJ (1998) A fair method for resetting the target in interrupted one-day cricket matches. *J Oper Res Soc* 49(3):220–227
9. Lewis AJ (2005) Towards fairer measures of player performance in one-day cricket. *J Oper Res Soc* 56(7):804–815
10. Lemmer HH (2008) An analysis of players' performances in the first cricket twenty20 world cup series. *South Afr J Res Sport Phys Edu Recreat* 30(2):71–77
11. Allsopp PE, Clarke SR (2004) Rating teams and analysing outcomes in one-day and test cricket. *J R Stat Society Ser (Stat Soc)*, 167(4):657–667
12. Beaudoin D (2003) The best batsmen and bowlers in one-day cricket. Ph.D. thesis, Simon Fraser University

Anchor-Based Effective Node Localization Algorithm for Wireless Sensor Networks



Basavaraj M. Angadi and Mahabaleshwar S. Kakkasageri

Abstract Wireless sensor network (WSN) is usually deployed in harsh environments for collecting and delivering the data to the remotely located base station. Due to the recent achievements in WSN, tiny and expensive sensors are used with capability of sensing large information and to propagate over longer distances. In range-based localization algorithms, anchor node whose location is known plays an important role. In order to resolve the challenges of traditional localization algorithms, an anchor-based node localization algorithm is proposed for WSN. The distance between anchor or known node and unknown node is measured by optimizing anchors and creating database of optimized anchors. Using trilateration method unknown nodes are located and are used as new anchors which will reduce localization algorithm's dependency on anchors. The objective of the proposed work is to maximize the life time of the network by minimizing the energy consumption. From the simulation results, we show that the proposed algorithm increases lifetime of wireless sensor network.

1 Introduction

WSN can be used for different area of applications [1], namely health, home, industry and military. Sensor network consists of sensors densely deployed in large number. To detect the specific events, sensing task is performed by each sensor. Responsibility of the sink node is to collect the sensed data from the sensors and transmit the collected data to the task manager. Whenever the task manager wants to perform some other operation, the new task will be circulated through the sensor network. In sensor

B. M. Angadi (✉) · M. S. Kakkasageri

Electronics and Communication Engineering Department, Basaveshwar Engineering College (Autonomous), Bagalkot, Karnataka 587102, India
e-mail: bmaec@becbgk.edu

M. S. Kakkasageri

e-mail: mskec@becbgk.edu

network, communication is established based on the wireless ad hoc networking technology presented in [2]. When the sensor nodes are unable to establish communication directly with the sink, then data must be forwarded by some intermediate sensors. In WSN, there are various issues (e.g., deployment, energy, localization, coverage, etc.). One of the important issues is localization, because location information is needed for deployment, target tracking, coverage, routing, rescues and location service [3]. For example, several location-aware routing protocols presented in [4–8] are based on the geographic location. In the WSN, data can be transmitted more effectively when the sensor nodes position is accurately located. Several schemes proposed in [9–14] deals with the localization and broadly classified into range free and range-based schemes. By knowing distance between nodes or angles, location is estimated in range-based schemes.

This paper proposes an algorithm to transfer data using energy-efficient localization algorithm. Energy consumption increases when source node communicates directly with destination node. To avoid direct communication, multihop technique is employed which reduces the energy consumption. Rest of the paper is organized as follows. Section 2 presents the proposed work. In Sect. 3, simulation and result analysis are discussed. Section 4 concludes our paper.

2 Proposed Scheme

In our proposed work, sensor nodes are grouped in clusters. Here, we have considered three clusters and cluster head (CH) is assigned to each cluster. CH is chosen based on the highest residual energy and lowest ID. Sensors sense the event of interest and transmit sensed data to cluster head through various intermediate sensors through the process called multi-hopping. The environment comprises two sink nodes, mobile anchor and static anchor. When the anchor node is static as shown in Fig. 1, the CH will directly send the collected data to the base station.

When the anchor node is mobile, it continuously changes its position on the circumference of circular sensing field as shown in Fig. 2. The speed of mobile anchor can be varied according to the convenience. CH in each cluster collects data from all the sensor nodes and sent to base station through mobile anchor. When mobile anchor comes nearer to base station, it will receive the data which will be further monitored and processed.

The operational sequence of the proposed scheme is as follows: (1) Calculation of free space path loss or received signal strength indicator (RSSI) in dB; (2) distance calculation between unknown node and anchor node using RSSI; (3) position estimation using trilateration method.

- Let $A(x, y)$ be the known or anchor node and $S(x, y)$ be the unknown node whose position is not known. Using Friis transmission formula [15], free space path loss or RSS in dB is calculated using Eq. (1):

Fig. 1 Data transfer to BS without mobile anchor node

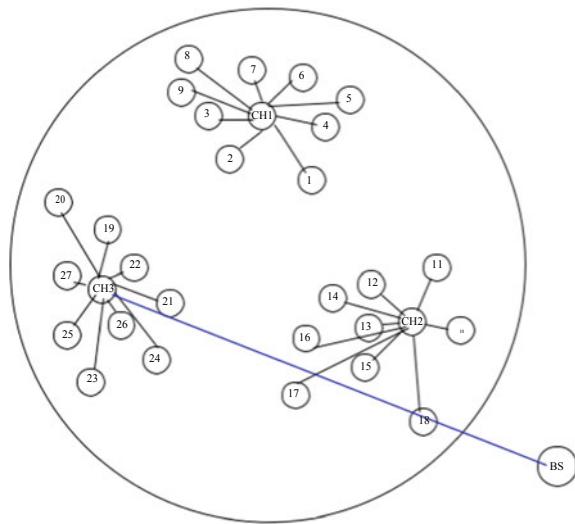
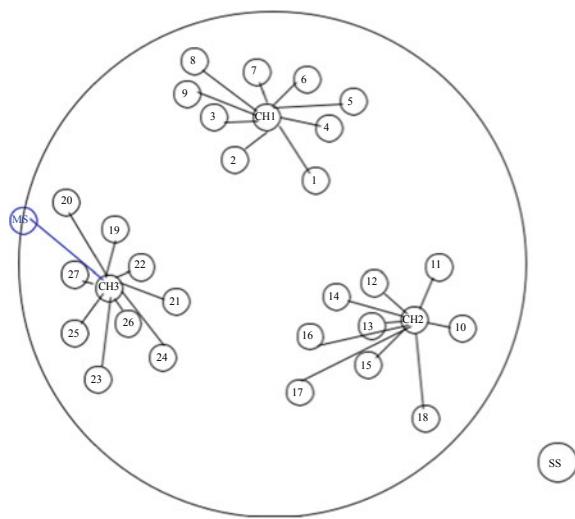


Fig. 2 Data transfer to BS with mobile anchor node



$$PL_{FS} = -27.55 + 20 \log(f) + 20 \log(d) \quad (1)$$

where PL_{FS} is free space path loss in dB, f is frequency in MHz, and d is distance in meters.

- Strength of the received signal at anchor node from the unknown node is known. Strength of the signal or path loss can be converted to distance in meters using the following Eq. (2):

$$d = 10 \frac{[\text{PL} - \text{PL}d_0 - 20 \log(f)]}{20} \quad (2)$$

where PL is path loss in dB, d is distance between anchor node $A(x, y)$ and unknown node $S(x, y)$ in meters and $\text{PL}d_0$ is the reference path loss also called as close-in reference distance. For microcell typical value of d_0 is considered as 1–10 m and for a large cell up to 1 km [16, 17].

- Trilateration method is used for position estimation of unknown node. In trilateration method, location is estimated by determining the intersection of three circles. In this case, the location is found by solving a linear system as follows:

$$(x - x)_1^2 + (y - y)_1^2 = d_1^2 \quad (3)$$

$$(x - x)_2^2 + (y - y)_2^2 = d_2^2 \quad (4)$$

$$(x - x)_3^2 + (y - y)_3^2 = d_3^2 \quad (5)$$

where d_1 , d_2 and d_3 are distances between unknown node $S(x, y)$ and anchor nodes $A(x_1, y_1)$, $B(x_2, y_2)$ and $C(x_3, y_3)$, respectively. These distances are measured using Eq. (2).

By solving Eqs. (3), (4) and (5), we get the x and y coordinates of the unknown node. The position information of unknown node is transmitted to base station with the help of mobile anchor node which is moving around the sensing field in a circular trajectory.

3 Simulation

The proposed work is simulated using Dev C. In simulation, we first create the network environment and nodes are randomly deployed. N number of nodes are divided into 3 integer parts and are made into 3 clusters, the remaining nodes are deployed in 3rd cluster. Randomly energy is assigned to each node and sensor nodes in the clusters are made to fall in circular sensing field. The cluster head is selected and using mobile anchor as the intermediate node the data is sent to base station or end user. Periodically, the cluster head is updated. The threshold distance is set by knowing the shortest distance from cluster head and nearest sensor node. To gather the data, mobile anchor moves along the perimeter of the circular sensing field. This section presents simulation inputs, performance parameters and result analysis. Some of the parameters considered for simulation are as follows: Number of nodes,

$N = 10\text{--}80$, simulation area, $A = 100 \times 100$ m, packet size, $P_s = 128$ KB, number of mobile nodes, $MS = 1$, sink mobility, $M = 2$ m/s. To test the performance of the proposed work, energy consumption, end-to-end delay and packet delivery ratio performance parameters are considered.

3.1 Result Analysis

Traditionally, energy consumed by network increases as increase in number of nodes. Figure 3 shows total energy consumed by both mobile anchor and static anchor. The energy consumed by network, when we use mobile anchor is less compared to static anchor. So the decrease in energy consumption, increases network lifetime. End-to-end delay to transfer data from source node to destination node for static anchor and mobile anchor is shown Fig. 4. Comparing both curves it is cleared that, the delay using static anchor is more compared to mobile anchor. The packet delivery ratio from source to destination node using static anchor and mobile anchor is shown in Fig. 5. Comparing both curves we understand that the total number of packets

Fig. 3 Energy consumption versus number of nodes

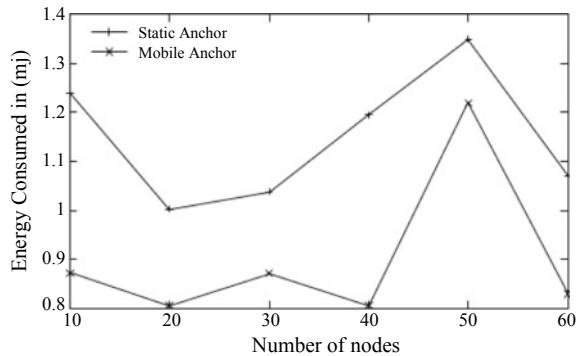


Fig. 4 End-to-end delay versus number of nodes

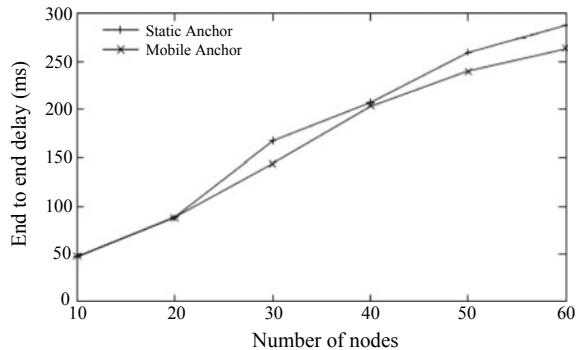
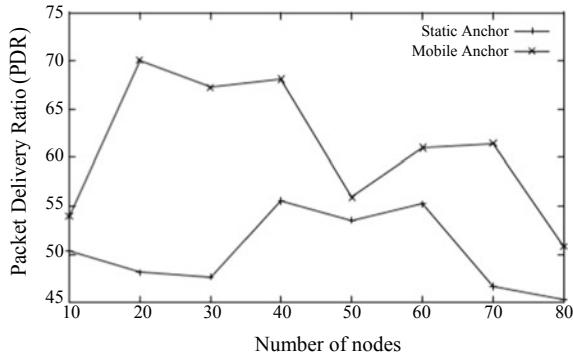


Fig. 5 Packet delivery ratio versus number of nodes



received using mobile anchor is more compared to static anchor.

4 Conclusion

Networking plays an important role in various fields such as agriculture, defense, health, environment and communication. Among the available resources, energy is a critical parameter and it has to be used effectively. Our proposed work, i.e., efficient anchor-based node localization algorithm for WSN depicts how energy consumption can be reduced by using multi-hop technique and by the use of multiple anchors (mobile anchor and static sink). We have analyzed the performance parameters for different number of nodes. Result shows that the proposed algorithm uses available energy effectively and increase the network lifetime. We have used only three clusters and two sinks. In the future, by increasing the number of clusters and sink nodes, network performance can be improved.

References

1. Kuhn JP (2004) Location-based services in mobile communication infrastructures. *AEU Int J Electron Commun* 58(3):159–164
2. Roxin , Gaber J, Wack M, Nait-Sidi-Moh A (2007) Survey of Wireless Geolocation Techniques. In: 2007 IEEE globecom workshops. Washington, DC, pp 1–9
3. Dragos N, Badrinath BR (2003) Ad hoc positioning system (APS) using AOA. In: IEEE INFOCOM 22nd annual joint conference of the IEEE computer and communications societies, vol 3, pp 1734–1743
4. Navas JC, Imielinski T (1997) GeoCast-geographic addressing and routing. In: Proceedings of the 3rd annual ACM/IEEE international conference on mobile computing and networking (MobiCom'97), ACM, New York, pp 66–76
5. Ko Y-B, Vaidya NH (2000) Location-aided routing (LAR) in mobile ad hoc networks. *J Wirel Netw* 6(4):307–321

6. Basagni S, Chlamtac I, Syrotiuk VR, Woodward BA (1998) A distance routing effect algorithm for mobility (DREAM). In: Proceedings of the 4th annual ACM/IEEE international conference on mobile computing and networking (MobiCom'98). New York, USA, pp 76–84
7. Kori GS, Chanal PM, Kakkasageri MS, Shirbur AA (2017) Energy aware multipath routing scheme for wireless sensor networks. In: Proceedings of the 7th IEEE international advance computing conference. Hyderabad
8. Karp B, Kung HT (2000) GPSR: greedy perimeter stateless routing for wireless networks. In: Proceedings of the 6th annual international conference on mobile computing and networking (MobiCom'00), ACM, pp 243–254
9. Alrajeh NA, Bashir M, Shams B (2013) Localization techniques in wireless sensor networks. *Int J Distrib Sens Netw* 9(6)
10. Singh SP, Sharma SC (2015) Range free localization techniques in wireless sensor networks: a review. *Procedia Comput Sci* 57:7–16
11. Gui L, Val T, Wei A, Dalce R (2015) Improvement of range-free localization technology by a novel DV-hop protocol in wireless sensor networks. *Ad Hoc Netw* 24:55–73
12. Tomic S, Beko M, Dinis R, Montezuma P (2017) Distributed algorithm for target localization in wireless sensor networks using RSS and AoA measurements. *Pervasive Mob Comput* 37:63–77
13. Chelouah L, Semchedine F, Bouallouche-Medjkoune L (2018) Localization protocols for mobile wireless sensor networks: a survey. *Comput Electr Eng* 71:733–751
14. Sharma G, Kumar A (2018) Modified energy-efficient range-free localization using teaching-learning-based optimization for wireless sensor networks. *IETE J Res* 64(1):124–138
15. Rappaport TS (2009) Wireless communications: principles and practice. 2nd edn, Prentice Hall PTR
16. Kurt S, Tavli B (2017) Path-loss modeling for wireless sensor networks: a review of models and comparative evaluations. *IEEE Antennas Propag Mag* 59(1):18–37
17. Klaina H, Vazquez Alejos A, Aghzout O, Falcone F. Narrowband characterization of near-ground radio channel for wireless sensors networks at 5G-IoT bands. *Sensors (Basel)* 18(1)

A Survey on Security on Medical Data and Images in Healthcare Systems



Swarnali Sadhukhan, Mihir Sing, Koushik Majumder, Santanu Chatterjee, and Subhanjan Sarkar

Abstract The development in the field of networking technologies and telecommunications has increased the popularity of telemedicine usage. In medical diagnosis, many processes have been proposed by researchers for securing patients' records and medical images that are sent from one location to another. This paper represents a technical survey on various cryptographic and watermarking processes that are applied on different medical images for secure transmission. Deterioration in the quality of the medical data or images at the time of transmission may endanger the treatment of patients, therefore, lossless and reversible methods need to be emphasized. In this study, we have analyzed some of the most relevant existing works in this area. On the basis of this analysis, we have tried to identify the open issues in the field of secure medical image transmission in order to provide secure transmission and ensure enhanced quality of treatment of the patients.

Keywords Steganography · Watermarking · Lossless · Reversible · Compression · Cryptography · Quantum image encryption · Chaotic system

1 Introduction

In medical diagnosis, many processes have been proposed by several researchers for securing patients' records and medical images that are sent from one location to another. This paper is focused on modern techniques, past works, and presents some approaches for securing medical images. The security of medical images is based on the following—(i) Confidentiality, (ii) Integrity, and (iii) Authentication.

- I. **Confidentiality:** The transmitted message must go to the intended receiver who can only recognize the content of the message.

S. Sadhukhan (✉) · M. Sing · K. Majumder · S. Chatterjee · S. Sarkar
Maulana Abul Kalam Azad University of Technology, Kolkata, West Bengal, India
e-mail: swarnali1994sadhukhan@gmail.com

- II. **Integrity:** This property ensures that the data must be exactly the same as they were sent. During the transmission, the data must not be changed by any other person.
- III. **Authentication:** It implies that the receiver has to be sure of the identity of the sender and that the message has not been sent by any other.

When patients' data and medical images are transmitted through public networks, ensuring the security of the transmitted data is a very critical issue and it needs to be protected. Cryptography is used for secure transmission and watermarking is used to maintain confidentiality and integrity. This paper is organized as follows. Section 2, presents a summary of various cryptographic and watermarking techniques for the secure transmission of medical images. This section is summarized in Table 1, which presents a comparison of various cryptographic and watermarking techniques. In Sect. 3, we have identified the open issues in the field of secure medical image transmission. Section 4 concludes the paper.

2 Literature Review

In medical diagnosis, when transferring the medical images and related data over a communication network, security is a major concern. Many researchers have presented different watermarking and cryptography techniques for the security of medical images. Thakur et al. [1] elucidated a robust watermarking method that mainly uses the transform domain technique. Here first, second-level DWT is applied on cover image followed by DCT & SVD. The watermark image also goes through DCT & SVD. After that, the transformed watermark image is embedded with the transformed cover image. Chaos-based encryption is put on a watermarked image for more secrecy. Reverse processes are used to get back the original image. The experimental results of this multilevel watermarking approach indicate that this process is completely robust and safe from different types of assault. Moreover, the extracted watermark and cover images are not much distorted.

Advantages:

- By hiding the secret information in a cover image, the confidentiality of the patient information is increased.
- Robustness is improved. NC values are also improved.
- The bandwidth is saved for telehealth applications.

Limitation:

- The performance of the proposed method is not determined on video and multiple watermarking.

Parah et al. [2] explained an IoT-driven healthcare system where the fragile watermark is used to embed the EPR (Electronic Patient Record). This type of watermark can detect whether the EPR is tampered or not during the transit. At the receiver side,

first, the receiver checks whether the fragile watermark image is the same as the one embedded in the sender side. If the extracted watermark image is not the same as the original one then the receiver cannot extract the EPR and sends a retransmission request to the sender. In this process, two address vectors are used in embedding the EPR.

Table 1 Summary of comparative study

Ref No	Methods	Advantages	Limitations
[1]	<ul style="list-style-type: none"> • DWT • DCT • SVD • Chaos-based encryption 	<ul style="list-style-type: none"> • Robust and secure • Hybrid Encryption process • improves confidentiality of patients data 	<ul style="list-style-type: none"> • Single watermark image is used • PSNR and MSE values can be improved • Not applied on video
[2]	<ul style="list-style-type: none"> • Generating stego RGB image using data embedder • Key is used for final encryption 	<ul style="list-style-type: none"> • High payload • High Imperceptibility • Detects tampering • Can be applied in real-time medical information interchange in IoT environment 	<ul style="list-style-type: none"> • Fragile to geometric attacks
[3]	<ul style="list-style-type: none"> • RSA • AES • DWT-2L 	<ul style="list-style-type: none"> • High PSNR value • High imperceptibility and capacity 	<ul style="list-style-type: none"> • Minimal deterioration
[4]	<ul style="list-style-type: none"> • Edge detection algorithm • Swapped Huffman tree 	<ul style="list-style-type: none"> • Provides security and confidentiality of patients data • The imperceptibility property is maintained • Lossless encryption 	<ul style="list-style-type: none"> • Slight variations in the peaks of the histogram
[5]	<ul style="list-style-type: none"> • SVD • Arnold Transform 	<ul style="list-style-type: none"> • Resist VQ, addition of text, copy-paste attacks • Improves tamper localization accuracy and PSNR of self-recovered image 	<ul style="list-style-type: none"> • Efficiency is not evaluated on non-fragile tampered images
[6]	<ul style="list-style-type: none"> • RG algorithm. • SHA-256 • Elliptical curve cryptography algorithm • Arithmetic coding algorithm 	<ul style="list-style-type: none"> • Better PSNR value and embedding capacity (bits) • Computational complexity is less 	<ul style="list-style-type: none"> • Used single watermark image • System robustness can be improved
[7]	<ul style="list-style-type: none"> • Calculate the non-zero AC coefficient of each sub-image • Calculate cost value matrix C by cost function 	<ul style="list-style-type: none"> • Better anti-steganalysis performance • Better results than the previous J-UNIWARD process and JPEG image steganalysis 	<ul style="list-style-type: none"> • Time complexity is high to obtain higher security performance

(continued)

Table 1 (continued)

Ref No	Methods	Advantages	Limitations
[8]	<ul style="list-style-type: none"> • DNA encoding by 1D logistic system • DNA addition and DNA subtraction by MSB manipulation • PWLCM System scrambling 	<ul style="list-style-type: none"> • No statistical relation between neighboring pixels • High immunity to defend statistical attacks • Performance is increased by DNA and the chaotic map • PSNR is infinity 	<ul style="list-style-type: none"> • Complexity is very high
[9]	<ul style="list-style-type: none"> • NEQR (Novel Enhanced Quantum Representation) • Quantum controlled NOT image • Key is generated by logistic sign map and generate quantum Key by NEQR • XOR operation is used for final encryption 	<ul style="list-style-type: none"> • PSNR is infinity • Robust against co-relation based attacks • Secure against entropy attacks • The watermark image is exposed only by the correct secret key 	<ul style="list-style-type: none"> • High complexity for simulating on classical computer
[10]	<ul style="list-style-type: none"> • Steganography • Controlled “NOT” gate • Arnold’s cat map 	<ul style="list-style-type: none"> • High PSNR values and capacity of embedding and excellent visibility • Career image is not required for extraction of the secret image 	<ul style="list-style-type: none"> • High complexity for simulating on classical computer

- It can be applied in real real-time medical data exchange in IoT environment.
- Payload is high. It can detect tamper caused by some attacks.

Limitation:

- Embedded data can't survive on geometric attacks.

In the work of Elhoseny et al. [3], RSA algorithm is applied to the odd part of the medical data, and AES algorithm is used on the even part. On the host image, second-level DWT is applied. Next, that image is embedded with the Cipher Data and finally stego image is generated. The reverse process is used to extract the data and image.

- Higher PSNR value is achieved.
- It also provides high capacity and imperceptibility.

Limitation:

- Minimal deterioration is found in the received stego image.

Muhammad Arslam Usman [4] applied the steganography technique for achieving better medical data security. They have elucidated a data embedding process by using the edge location of the cover image. Lossless compression is applied to the watermark images for ensuring high capacity. Here, the encrypted image is compressed by swapped Huffman tree encoding (SHT) algorithm. The experimental result of the histogram of host image and cover image shows that the hidden data is highly imperceptible.

- It ensures confidentiality and secrecy of the patient's data.
- Hidden data is imperceptible.

Limitation:

- Histogram shows a slight variation in the peak in comparison to the host image.

Shehab et al. [5] explained an SVD-based fragile watermarking process that uses grouped block method to increase security. They also developed the mechanism to identify the attacked portion of the medical images. The Vector Quantization attack is prevented with the use of two specially designated authentication bits—block authentication and self-recovery bits. To recover the tampered region from the neighboring block, Arnold Transform has been used. This method increases the NCC and PSNR of the reconstructed host image.

Advantages:

- Vector Quantization, copy-paste, text addition, and content removal attacks are prevented and this method can efficiently locate the attacked blocks.
- Improves tamper localization accuracy and the PSNR of the self-recovered image.

Limitation:

- Efficiency is not evaluated in the case of non-fragile tampered images.

Aparna et al. [6] explained a compression and cryptography algorithm that is used for the medical image watermarking technique. At first, the ROI part is separated from watermark image by Region Growing Algorithm and then (SHA)-256 algorithm is applied on ROI part. Next elliptical curve cryptography (ECC) is applied to encrypt the EHR document. By using the Arithmetic coding algorithm (AC), the image and EHR document are concatenated and compressed. Finally, the complete bitstream is embedded into the original medical image. Experiment results indicate that this method improves the quality of the watermark image and enhances embedding performance.

Advantages:

- Computational complexity of proposed method is less.

Limitation:

- This paper used only a single watermark image.

Liao et al. [7] elucidated a JPEG steganography scheme for medical images based on preserving inter-block dependencies. First, the JPEG image is separated into four non-overlapping sub-images consisting of 8×8 blocks such that adjacent DCT blocks belong to two different sub-images. After that, non-zero AC coefficients of sub-images are calculated and cost values are calculated from DCT coefficients using cost functions which are used to update the stego image. This process will be continued until all the sub-images have been embedded. This proposed model can effectively cluster the inter-block embedding changes and give better anti-steganalysis performance.

Advantages:

- It gives a better result than the previous J-UNIWARD method and modern JPEG image steganalysis.

Limitation:

- Considering complexity analysis, this process takes more time for giving higher security.

Praveen Kumar et al. [8] proposed a secure medical image encryption process using a cognitive approach. Here, first, the medical image is encoded by DNA encoding using 1D Logistic System. Then, DNA addition and DNA subtraction are applied on the encoded image and the result is concatenated. After that 1D logistic system is applied to that concatenated image. Next, by Piecewise Linear Chaotic Map (PWLCM) system scrambling is done on that image and the final encrypted image is generated. For sensing the unused frequency band, the spectrum sensing method is used. The experimental result shows that the pixels are equally distributed in the gray levels. So the attacker cannot calculate the relationship between the adjacent pixels. Moreover, the histogram deviation values reflect good results which ensure the robustness of this scheme.

Advantages:

- There is not any relationship between the adjacent pixels ensuring that the algorithm gives high immunity against statistical attacks.
- The infinity value of PSNR proves that the reconstruction of the secret image is hassle-free.

Limitation:

- The proposed method has a very high complexity.

Ahmed et al. [9] explained a robust quantum encryption process. In this process, a healthcare staff is sending a medical image into the cloud and from the receiver side, another healthcare staff is receiving that medical image from cloud. For this process, at first, the original medical image is changed into quantum controlled NOT image by Novel Enhanced Quantum Representation (NEQR) method. Next, the gray code is used to scramble the quantum image. The key is generated by a logistic sign

map and then the key is transformed into a quantum image representation by Novel Enhanced Quantum Representation (NEQR) method. Further, the resultant image is finally encrypted using quantum XOR operation based on the logistic-sine map-controlled key generator. The experimental results show that the zero-correlation requirement is fulfilled and has high resistance against correlation-based attacks.

- Based on histogram analysis, this method is robust against correlation-based attacks.
- This approach is secure against entropy attacks and can't survive to any small changes in the value of the pixel of the images.
- Only the correct secret key can expose secret information. Any small changes in secret key will not give proper results.

Limitation:

- Performance analysis using simulations on the classical computer has high complexity.

Ahmed et al. [10] elucidated an efficient quantum medical information hiding process. A steganography technique is used to hide the quantum watermark image into quantum cover image. This quantum watermark image is encrypted using a controlled NOT operation. After that, the encoded watermark image is embedded into the quantum cover image using the two most and least significant qubits. Before embedding, Arnold's cat map is applied to the resultant image. In this paper, only key and watermarks are needed to separate the watermarked image and cover image is not required. The proposed encryption process provides clear visibility and high embedding capacity.

- Naked eye cannot differentiate the cover and stego image.
- This method presents high PSNR values, high embedding capacity, and excellent visibility.
- The carrier image is not required for extraction of the watermark image.

Limitation:

- Performance analysis using simulations on the classical computer has high complexity.

3 Comparative Analysis

Medical images are extremely sensitive. A tiny change may lead to the wrong diagnosis. Multi-layer security is explained in [1] where DWT, DCT, and SVD are used at the time of watermark embedding and a chaotic encryption process is used to encrypt the watermark image. In [2], a fragile watermark is used to embed Electronic Patient's Record (EPR) in color image for an IoT-driven health care system. A secure and hybridized medical data transmission process is proposed in [3]. It

is an IoT-based system where the data is encrypted using RSA and AES algorithm and DWT is used for watermarking. In [4], a swapped Huffman tree, which is a lossless process, is used in a steganographic approach to secure and encrypt medical data and patient's data. Determining image authenticity and self-recovery of attacked images has been done using an SVD-based approach in [5]. Compression technique along with the cryptography algorithm for watermarking has been proposed in [6] for use in e-healthcare systems. A JPEG steganographic scheme is proposed in [7] where images are encrypted based on preserving inter-block dependencies. In [8], Praveen Kumar et al. Adopt a Cognitive Radio (CR) technology and image encryption techniques to securely and efficiently transmit medical images. [9] proposed a chaos-based quantum encryption framework for securing healthcare images. In [10], El-Latif et al. hides quantum secret image in a quantum cover image using the quantum steganography approach. The following table summarizes and compares the above studies.

4 Open Issues of Medical Image Security

After reviewing the current research in this area, we feel that the following are the open issues where further investigation may be done:

- (1) Most encryption algorithms do not consider the minimal distortion or slight peak variation of histograms in medical images. As the patients' treatment is dependent on the accurate medical image, this distortion cannot be ignored.
- (2) In the cloud platforms, execution time is a very important issue. Users expect minimum latency from the applications. So the algorithms having high time complexity may increase the latency and response time for the users. This will cause deterioration in the Quality of Experience (QoE) for the users.
- (3) Nowadays, research in quantum information hiding is also very popular. But for performance analysis, a classical computer is required, where simulating quantum algorithms is time-consuming and costly.
- (4) Most of the proposed schemes reviewed do not consider the time complexity. We feel that reducing the time complexity of these schemes is important and might lead to real-time secure medical image transmission.

5 Conclusion

With the increasing popularity of cloud computing, there has been a huge demand for telemedicine. In countries like India where there is a lack of medical facilities in rural areas, telemedicine can be a great boon for the common people. But with the huge potential of the telemedicine comes the risk of it. As deterioration in the quality of the medical data or images at the time of transmission may endanger the treatment of patients, therefore, we need to devise techniques to secure the medical

images during transmission in order to ensure the accuracy of medical data. This paper reviews the current research in the field of secure medical image and medical data transmission. On the basis of the review, an attempt has been made to identify the open issues in this area. These open issues provide us the future directions in which research should be conducted for achieving secure transmission and thereby ensuring enhanced quality of treatment to the patients.

References

1. Thakur S, Singh AK, Ghrera SP, Elhosny M (2018) Multi-layer security of medical data through watermarking and chaotic encryption for tele-health applications. *Multimed Tools Appl* 1–14
2. Parah SA, Sheikh JA, Ahad F, Bhat GM (2018) High capacity and secure electronic patient record (EPR) embedding in color images for IoT driven healthcare systems. In: Internet of things and big data analytics toward next-generation intelligence, 409–437. Springer, Cham
3. Elhosny M, Ramírez-González G, Abu-Elnasr OM, Shawkat SA, Arunkumar N, Farouk A (2018) Secure medical data transmission model for IoT-based healthcare systems. *IEEE Access* 6:20596–20608
4. Usman MA, (2018) Using image steganography for providing enhanced medical data security. In: Consumer communications & networking conference (CCNC), 2018 15th IEEE annual, pp 1–4. IEEE
5. Shehab A, Elhosny M, Muhammad K, Sangaiah AK, Yang P, Huang H, Hou G (2018) Secure and robust fragile watermarking scheme for medical images. *IEEE Access* 6:10269–10278
6. Aparna P, Kishore PVV (2018) An efficient medical image watermarking technique in E-healthcare application using hybridization of compression and cryptography algorithm. *J Intell Syst* 27(1):115–133
7. Liao X, Yin J, Guo S, Li X, Sangaiah AK (2018) Medical JPEG image steganography based on preserving inter-block dependencies. *Comput Electr Eng* 67:320–329
8. Praveenkumar P, Devi NK, Ravichandran D, Avila J, Thenmozhi K, Rayappan JBB, Amirtharajan R (2018) Transreceiving of encrypted medical image—a cognitive approach. *Multimed Tools Appl* 77(7):8393–8418
9. El-Latif AAA, Abd-El-Atty B, Talha M (2018) Robust encryption of quantum medical images. *IEEE Access* 6:1073–1081
10. El-Latif AAA, Abd-El-Atty B, Hossain MS, Rahman MA, Alamri A, Gupta BB (2018) Efficient quantum information hiding for remote medical image sharing. *IEEE Access* 6:21075–21083

Pothole and Speed Bump Classification Using a Five-Layer Simple Convolutional Neural Network



Anju Thomas, P. M. Harikrishnan, J. S. Nisha, Varun P. Gopi, and P. Palanisamy

Abstract The presence of abnormalities like a pothole, street gutter, and speed bump on the road surface creates a lot of problems like wear and tear of vehicles, fuel consumption, and may lead to accidents also. There are many literature carried out in this area related to sensor-based techniques, computer vision-based techniques, and mobile applications. Speed bumps are also different types of marked bumps, unmarked bump, and rumble strips. In this paper, an image-based system for the classification of a speed bump and pothole using a five-layer simple convolutional neural network is proposed. The proposed network achieved a classification accuracy of 97.7%.

1 Introduction

Nowadays, road accidents are increased due to over speed, careless driving, and road's bad conditions. Ministry of Road Transport & Highways, Government of India made a report on road death due to pothole are a total of 22,656 in the year of 2018 while in 2017 the death rate was 20,457, these figures indicate that why road maintains is important in a developing country like India [1]. It is difficult to identify

A. Thomas (✉) · P. M. Harikrishnan · J. S. Nisha · V. P. Gopi · P. Palanisamy
National Institute of Technology Tiruchirappalli, Tiruchirappalli, Tamilnadu, India
e-mail: anjukandathil.thomas@gmail.com

P. M. Harikrishnan
e-mail: haripm033@gmail.com

J. S. Nisha
e-mail: nishajs2007@gmail.com

V. P. Gopi
e-mail: varun@nitt.edu

P. Palanisamy
e-mail: palan@nitt.edu

the depth of pothole when it is filled with water, and if anyone who does not have any idea about this may be trapped in the pothole. Speed bumps are mainly used for the speed limit in the area of school roads, parking lot, garages, hospitals, etc. According to the IRC099 protocol of the Indian Road Congress, speed bump should be painted to ensure that solar cat eyes are illuminated and visible to assist the driver in the presence of bump during night time, but it is not followed everywhere.

An unmarked speed bump, during nighttime, is very difficult to be identified, and this kind of a speed bump may lead to an accident. Hence, identification of pothole and speed bump is an important task for saving a life. There are different methods used for the detection of a pothole and speed bumps like signal vibration-based (using the accelerator and mobile applications) and image-based (computer vision, stereo imaging, and laser) [2]. In [3], they proposed an automatic detection of the pothole and bump using ultrasonic sensors. A particular threshold value is chosen in this method, and the detected value is greater than the threshold is considered as a pothole. In [4], they used three methods for the classification of a pothole, speed bump, and gutter using a mobile application. The three methods for classification are long short-term memory network, convolutional neural network (CNN), and deep learning model. The proposed method used image-based classification of a speed bump and pothole using a simple five-layer convolutional neural network (CNN).

2 Related Work

2.1 Pothole Detection from Images

There are different works carried out in the detection of the pothole and bump separately. In [5], the authors recommended two different pothole detection algorithm using machine learning techniques, that are the least squares support vector machine (LS-SVM) and artificial neural network (ANN). This work attained classification accuracy as 86% and 89% using ANN and LS-SVM, respectively. Tedeschi and Benedetto [6] suggested an automatic pavement distress recognition (ADPR) system that is capable of performing in real time by identifying road distress and potholes. This approach used a combination of different OpenCV library and obtained greater than 70% performance measures for the local binary pattern (LBP) classification. Koch and Brilakis [7] implemented a model that deals primarily with the separation of non-defect and defect regions in the image using a threshold based on histogram structure. The shape of a pothole is calculated based on a region's geometric characteristics. Then, the authors eventually took the shape of a pothole as being about elliptical and obtained 86% accuracy with 82% precision and 86% recall.

Pothole detection using an unsupervised vision-based method is suggested in [8]. The authors extracted the pothole area from the RGB image followed by the image segmentation process. Then, the search of the pothole is done only in the extracted areas and achieved accuracy as 82%. In [2], they used CNN-based ResNet model over

the thermal images of road pothole under various weather conditions and achieved an accuracy of 97.08%. Ryu et al. [9] discussed a method to detect potholes both for concrete or asphalt road surfaces using videos. The work mainly has three steps of operation like image segmentation, extraction of candidate region, and making a decision. Due to shadows of objects present in real-world road videos, the device cannot detect potholes in darker images. In [10], they proposed a two-stage deep learning method for the identification of pothole from stereo vision with different climatic conditions and they felt it difficult to find the irregular shape of a pothole.

2.2 *Speed Bump Detection from Images*

In [11], the segmentation of speed bumps is accomplished by separating the area of bump and the area of the usual road. Even though some noise is present in the segmentation stage, it is removed by using Gaussian mixture model (GMM). Then, it is followed by the morphological operations (dilation and erosion) and also use the Regionprop model to find out the adequate position of the speed bump. This work achieved 94.7% accuracy in the daytime and 70.8% accuracy in the night. Devapriya et al. [12] proposed a method for bump identification, and they used the combination of Gaussian filtering and median filtering approach to eliminate noise in the image. Later image subtraction is done by subtracting median filtered image from Gaussian filtered image, then converting the resulting image into a binary image and using the connected component approach to analyze the region. A method for finding speed bump in video is proposed in [13]. In this method, initially the video is converted into images and each image undergone preprocessing steps to enhance the image for further processing. The second step is the morphological operation (dilation and erosion). Then, the result of morphological operation produces a horizontal and vertical projection spread. The plot of the projection is helped to identify the presence of the speed bump in the image.

One of the recent works [14] used deep learning techniques for the classification of unmarked and marked speed bump classification. This work obtained an accuracy of 97.44% for a marked speed bump and 93.83% for an unmarked speed bump. In “automatic bump detection and 3D view generation from a single road image,” the input image is processed to find out the vanishing point [15]. Then, the speed bump is detected using morphological operation. In the third step, the distance from the point of view to the bump is determined, and the vehicle speed may be decreased based on this value. Finally, all data in the 3D domain will be projected.

3 Methodology

A simple five-layer CNN architecture of the proposed work is given in Fig. 1. Due to different sizes of images in the dataset, each image is resized into 112×112 and given to the network for classification.

The CNN architecture for feature extraction is demonstrated in Fig. 1, and the learnable parameters for each layer are listed in Table 1. In the designed network, five convolutional layers and three fully connected layers are used. From the last fully connected layer, two relevant features are obtained and used for classification.

It is essential to resize the image appropriately before feeding them into CNN. In the proposed work, the input images are resized to a dimension of $112 \times 112 \times 3$ pixels equivalent to the breadth, height, and the three color channels (RGB) indicating the depth of the input image. The function of the CNN is to reduce the input images to a form that is simpler to process without losing critical features to obtain a good prediction. The function of each layer in the CNN is narrated above.

- **Convolutional layer:** It calculates the output of each neuron as a dot product with the corresponding weights of a small portion of the image. Along the length and breadth, this process is repeated. These layers use the parameter sharing system to regulate the number of parameters.
- **Rectified linear unit (ReLU):** It is the simplest nonlinear activation function and it is used here. This layer removes all negative activations with 0 by introducing system nonlinearity and applying the $f(n) = \max(0, n)$ function. There, the neuron input is n .
- **Batch normalization:** It enables each network layer to learn somewhat more independently of other layers. It helps to avoid over-fitting as it has a slight effect on regularization. Batch normalization normalizes the output of a previous activation layer by eliminating the batch mean and dividing it by the standard batch variance to improve a neural network's strength.

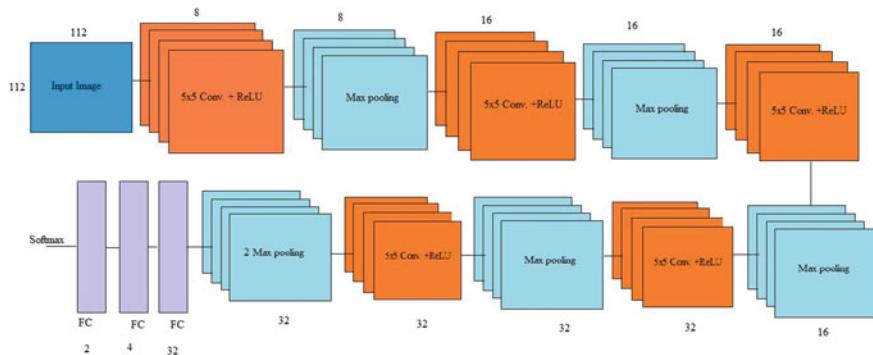


Fig. 1 Proposed network architecture

Table 1 Proposed network parameters

Name of the layer	Size	Weight	Bias	LP
Input image	$112 \times 112 \times 3$	–	–	–
Conv 1	$108 \times 108 \times 8$	600	8	608
BN 1	$108 \times 108 \times 8$	8		8
Max 1	$54 \times 54 \times 8$	–	–	–
Conv 2	$50 \times 50 \times 16$	3200	16	3216
BN 2	$50 \times 50 \times 16$	16		16
Max 2	$25 \times 25 \times 16$	–	–	–
Conv 3	$21 \times 21 \times 16$	6400	16	6416
BN 3	$21 \times 21 \times 16$	16		16
Max 3	$10 \times 10 \times 16$	–	–	–
Conv 4	$6 \times 6 \times 32$	12,800	32	12,832
BN 4	$6 \times 6 \times 32$	32		32
Max 4	$3 \times 3 \times 32$	–	–	–
Conv 5	$3 \times 3 \times 32$	1024	32	1056
BN 5	$3 \times 3 \times 32$	32		32
Max 5	$1 \times 1 \times 32$	–	–	–
Max 6	$1 \times 1 \times 32$	–	–	–
FC 1	32×1	1024	32	1056
FC 2	4×1	128	4	132
FC 3	2×1	8	2	10
Total				25,330

- **Pooling layer:** It reduces the number of inputs to the next layer of feature extraction, thus allowing us to have many more different feature maps. Max pooling is a method of discretization based on samples. The goal is to down-sample an input representation that reduces its dimensionality and enables assumptions about features contained in binned subregions to be made. Max pooling is performed by applying a max filter to (generally) the initial representation's non-overlapping subregions.
- **Fully connected layer:** The neurons, as seen in regular neural networks, have complete links to all prior layer activations. Therefore, their activations can be calculated with a matrix multiplication accompanied by a bias offset.

3.1 Dataset

The input dataset of the pothole as well as speed bump is accessible in the public domain. Here, we have used Kaggle dataset for the pothole, which is having 618

pothole images with different climatic conditions (different climatic conditions cause pothole to become water-filled, dry and wet) [16]. The speed bump dataset is taken from Mendeley data [17], and it contains marked as well as unmarked speed bumps of a total of 543 images with different sizes. The total number of images in two sets is different, and we need to make sure that each set having the same number of images otherwise the system will get biased to a particular set and make errors in classification.

4 Result and Discussion

The list of the learnable parameter in each stage of the CNN network is described in Table 1. The total parameters used in this network are only 25,330. This is less compared to the existing works (one of the existing works used ResNet-50 architecture, and ResNet-50 architecture has almost 25 million learnable parameters. We used only very less amount of parameters, and hence, the computation time and complexity of our network are very less compared to the existing methods). Data distribution and the confusion matrix of the testing are given in Tables 2 and 3. During testing, speed bumps are correctly detected and two pothole images are wrongly detected as a speed bump and achieved validation and testing accuracy of 98.1% and 97.7%, respectively.

4.1 Comparison with Existing Work

Since the classification of pothole and speed bump using vibration is done in so many previous works, our field of interest of classification is using images. Hence, for the comparison of existing work, there is not much work carried out in this area

Table 2 Data distribution

	Training	Validation	Testing
Pothole	450	50	43
Speed bump	513	57	43

Table 3 Confusion matrix of the proposed method with weighted classification layer

Predicted class	Targeted class	
	Pothole	Speed bump
Pothole	41	2
Speed bump	0	43

Table 4 Comparison with existing methods

	References	Methods	Accuracy (%)
Classification of pothole using images	[5]	ANN	86
	[5]	LS-SVM	89
	[7]	Approximate the shape of pothole as elliptical	86
	[8]	Unsupervised method	82
	[9]	Segmentation, candidate extraction, and decision	73.5
	[2]	CNN using ResNet	97.08
Classification of speed bump using images	[11]	GMM (at daytime)	94.7
	[11]	GMM (at nighttime)	70.8
	[13]	Gaussian and median filtering (marked bumps)	90
	[13]	Gaussian and median filtering (unmarked bumps)	30
	[14]	Deep learning technique (marked bumps)	97.44
	[14]	Deep learning technique (unmarked bumps)	93.83
Classification of pothole and speed bump using images	Proposed method	CNN	97.7

of pothole and speed bump classification using images. Also, the final results of the previous works of pothole and speed bump can be used for comparison. However, it should be remembered that for the purpose of classification, the previous works used their own datasets. Therefore, the only common point we can compare is that they all use images to classify. Table 4 compares the proposed classification technique with existing works.

5 Conclusion

In this work, a simple five-layer CNN network for the classification of pothole and speed bump is proposed. The number of neurons used in the network is very less compared to the other CNN networks used in previous works. It will make sure that

the system is less complex and the computational time is less due to less number of learnable parameters. The proposed CNN model gives an accuracy of 97.7% with less number of learnable parameters and makes it superior when compared with existing methods.

Funding

This work was supported by the Vandi Technologies PTE LTD Singapore [Grant No. VANDI/PS01/NITT1821 dated 10-09-2018].

References

1. Ians (2019) The news minutes, 2015 Indians lost their lives due potholes 2018
2. Bhatia Y, Rai R, Gupta V, Aggarwal N, Akula A et al (2019) Convolutional neural networks based potholes detection using thermal imaging. *J King Saud Univ Comput Inf Sci*
3. Madli R, Hebbar S, Pattar P, Golla V (2015) Automatic detection and notification of potholes and humps on roads to aid drivers. *IEEE Sens J* 15(8):4313–4318
4. Varona B, Monteserin A, Teyseyre A (2019) A deep learning approach to automatic road surface monitoring and pothole detection. *Pers Ubiquit Comput* 1–16
5. Hoang N-D (2018) An artificial intelligence method for asphalt pavement pothole detection using least squares support vector machine and neural network with steerable filter-based feature extraction. *Adv Civil Eng* 2018
6. Tedeschi A, Benedetto F (2017) A real-time automatic pavement crack and pothole recognition system for mobile android-based devices. *Adv Eng Inform* 32:11–25
7. Koch C, Brilakis I (2011) Pothole detection in asphalt pavement images. *Adv Eng Inform* 25(3):507–515
8. Akagic A, Buza E, Omanovic S (2017) Pothole detection: an efficient vision based method using RGB color space image segmentation. In: 2017 40th international convention on information and communication technology, electronics and microelectronics (MIPRO). IEEE, pp 1104–1109
9. Ryu S-K, Kim T, Kim Y-R (2015) Image-based pothole detection system for its service and road management system. *Math Probl Eng* 2015
10. Dhiman A, Klette R (2019) Pothole detection using computer vision and learning. *IEEE Trans Intell Transp Syst*
11. Srimongkon S, Chirachart W (2017) Detection of speed bumps using Gaussian mixture model. In: 2017 14th international conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON). IEEE, pp 628–631
12. Devapriya W, Babu CNK, Srihari T (2016) Real time speed bump detection using Gaussian filtering and connected component approach. In: 2016 world conference on futuristic trends in research and innovation for social welfare (startup conclave). IEEE, pp 1–5
13. Devapriya W, Babu CNK, Srihari T (2015) Advance driver assistance system (ADAS)-speed bump detection. In: 2015 IEEE international conference on computational intelligence and computing research (ICCIC). IEEE, pp 1–6
14. Varma V, Adarsh S, Ramachandran K, Nair BB (2018) Real time detection of speed hump/bump and distance estimation with deep learning using GPU and ZED stereo camera. *Procedia Comput Sci* 143:988–997

15. Murali S et al (2014) Automatic hump detection and 3d view generation from a single road image. In: 2014 international conference on advances in computing, communications and informatics (ICACCI). IEEE, pp 2232–2238
16. Patel S (2019) Pothole image data-set, web scrapped road images for pothole detection
17. Varma SKP (2018) Speed hump/bump dataset

Smart Ecosystem to Facilitate the Elderly in Ambient Assisted Living



Ashish Patel and Jigarkumar Shah

Abstract Ambient Assisted Living (AAL) facilities offer personalized care to inhabitants using their profile and surrounding environments. The services provided by AAL listed as health, indoor activities, daily routines and many others. The current focus of the research is to analyze the user's behaviour to offer different services to them. The AAL system mostly uses various sensors to capture the user's information. The smart environment uses environmental sensors, object sensors, body sensors and visual sensors as the primary devices. The collected information generally used to detect the activity of daily living (ADL). The research community requires to address added parameters to extend the services offered. The work presented here explores some new strategies to use the firings of environmental sensors. The environmental readings are useful to find the time of detection of any disease using the particular season, indoor air quality, the intensity of sun-rays and finally, the probability calculation of various illnesses during the year. This paper describes the potential role of AAL systems in elderly care by focusing on specific issues using a scenario-based approach.

1 Introduction

Ambient assisted living environments (AALs) sometimes called smart home systems is a solution to make the elderly capable of independent living. The primary objectives of the AAL system are [1]:

- Living in a preferred environment with independence Constant health monitoring
- Enhancing security and privacy, avoiding social isolation Promoting smart systems for better living

A. Patel (✉) · J. Shah

School of Technology, Pandit Deendayal Petroleum University, Gandhinagar, India

e-mail: ashish.patel@svmit.ac.in

J. Shah

e-mail: jigarkumar.shah@sot.pdpu.ac.in

© Springer Nature Singapore Pte Ltd. 2021

501

V. K. Gunjan and J. M. Zurada (eds.), *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, Advances in Intelligent Systems and Computing 1245, https://doi.org/10.1007/978-981-15-7234-0_46

Sarah Stevenson [2], in her article, explained various ways to improve the quality of older people. She addressed that health is a primary concern for an aged person living independently. Any ideal AAL solution must recognize disorders and other chronic conditions. The smart system must be capable of keeping seniors active in society while maintaining their autonomy. Another issue the smart environment must consider is mobility. The smart system must play a vital role in supporting of doing daily activities. The system must also encourage people to remain alive in society. Shah and Patel [3] identified major domains and sub-domains of the AAL system with their current status and future challenges.

Wireless sensor networks and body area networks are being used frequently for building smart home systems and continuously monitoring the residents in these environments [4–8]. These networks provide information to adequately supervise health, daily activities, critical situation alert and other vital parameters [9–11]. Another critical parameter is the location of the user using technologies like RFID (Radio Frequency Identification) and GPS (Global Positioning System) [12].

Intelligent algorithms play an essential role to integrate the above techniques to deliver effective services. Ambient assisted living solution is the result of a combined solution to many such ideas which improves the ability of independent individuals (Fig. 1). State-of-the-art technologies accomplish some of these requirements and experimentations are in progress to obtain new penetrations into these challenges and to meet the limits. Patel and Shah [13] mentioned some critical challenges like the requirement of the generalized structure, the ability to recognize multiple residents, real-time solutions and many more.

Several concepts were presented to address the issues of a smart home. With these many strategies available, it is demanded to transform these proposals into applications that help the inhabitants at large. Smart health framework, which covers the initial phase of sensor readings to the final stage of expected smart assistance, is essential to convert the proposal into a practical approach. This work explores some new approaches to use the firings of environmental sensors to address the parameters like seasonal diseases, air quality, the intensity of light and the probability of various diseases. Additionally, a framework is presented to support the proposal. The rest of the paper is organized as follows. Section 2 explores recent advances. Section 3 discusses potential issues for the smart ecosystem by presenting various scenarios. Section 4 explores validation for the proposed smart ecosystem. Finally, Sect. 5 concludes the work.

2 Related Work

In his article, Ex-Apple CEO, John Sculley [14] discussed two trends related to human health. He said, ‘The first is that the number of available hospital beds will continue to drop, having already decreased from 1,213,327 to 931,203 beds, or 23%, over the past 30 years. This occurred while the population grew by over 30%. The second trend is the rise of “super-users” in our health care system. In 2016, 5% of the

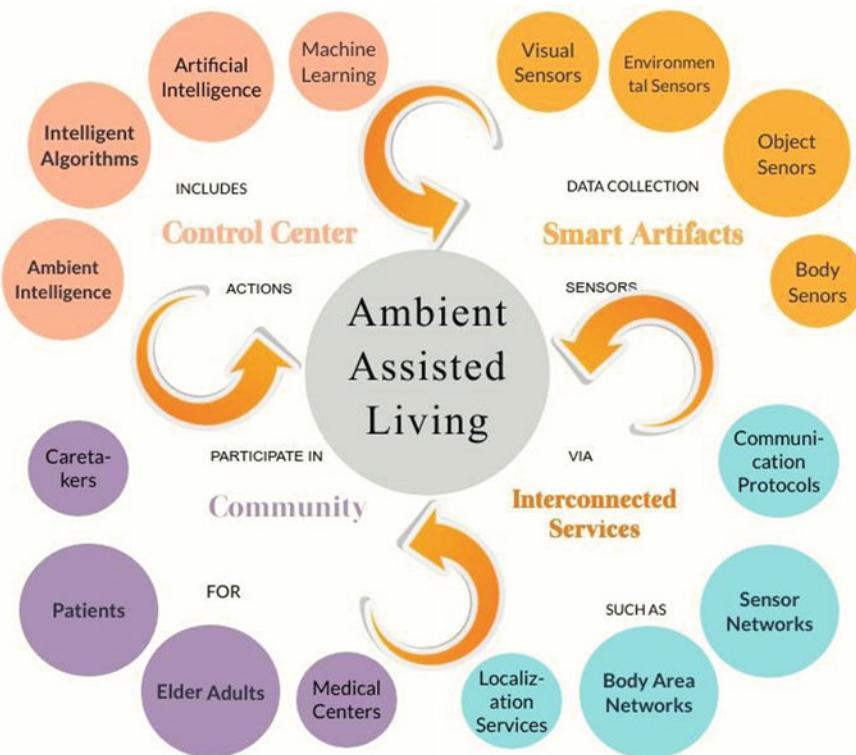


Fig. 1 Ambient-assisted living environment: components and services

population was responsible for half of all dollars spent on health care in the USA. These patients are stressing the sustainability of the health care system with their high cost of care'. The situation is alarming and needs to be addressed with the use of advanced technological solutions. Assisted living facilities are capable of handling the job with a broader context. Nehmer et al. [15] classified living assistance in three categories: comfort services, autonomy enhancement services and emergency treatment services. The emergency treatment is a mandatory and most important feature that must exist in any ideal AAL solution. Human health-related issues come under this category and should be efficiently handled. Human health influenced by many parameters like environmental effects on the body, air quality, the presence of sun-rays and many more. The ability of the AAL system to predict the possible health issues, during a specific time of the year by examining the history of that person, also plays a critical role in its success story.

Richard and Frank [16] showed that the environment has an independent effect on human health and health-related behaviours. They experimented based on greenery present near the living environment. The result concluded that environments that promote good health might be crucial in the fight to reduce health inequalities. Meng

et al. [17] proposed an ontology-based approach to manage health-related issues in geospatial sensor web. They exhibited a model of the Semantic Sensor Observation Service (S-SOS) to address health issues using air sensor data. The environmental sensing system used in their work handled the issue of air quality; it prevents handling the seasonal change in indoor locations or the capability to deal with sunrays. Marques and Pitarma [18] developed a smart-watch-based application to check the indoor air quality. They incorporated luminosity, temperature, CO₂, humidity and PM10 sensors to alert a user for the contemporary situation via mobile application. Their application, however, not utilized the collected information to analyze seasonal effects. Hay et al. [19] predicted malaria season using the satellite sensor data. Using regression analysis, they asserted the probability of malaria in a particular season. The methodology they used was principally utilizing satellite sensor data; the environmental, body and object sensors can enhance the effectiveness of the AAL approach.

3 Potential Issues for Smart Ecosystem

3.1 *Environmental Effects on Disease During Particular Season*

Scenario 1: Mr. X is 70 years old male living independently in an apartment of the smart colony prepared for the seniors. The township consists of the number of smart homes and an intelligent medical center where everything furnished with state-of-the-art technologies. Mr. X health is stable, and he is enjoying his stay. Suddenly he caught by the flue and his health declined. The best possible reason found he was more exposed to thunder during the last few days. The season sensors data exhibits the right. By learning the health problems of Mr. X, every season in which there is a freezing atmosphere, the windows are automatically sealed. The heaters set as per the necessity and recommendations for diet and drink issued to the resident. The behaviour is used as input to the control center of the ambient assisted living environment. From the next season cycle, Mr. X found no difficulty to deal with the flue.

This scenario provides knowledge of the expectations of an individual from AAL systems. There are various probable diseases due to seasonal change. In summer, there is a high probability of malaria, dengue, diarrhea, food poisoning and flu. Also, waterborne diseases like typhoid and jaundice, chickenpox, heatstroke and sunburn observed during this season. In winter, people face cold, cough, flu, bronchitis, dry and itchy skin. In monsoon, the possible causes are malaria, diarrhea, typhoid, dengue, chikungunya and cholera. Sometimes there are possibilities of Hepatitis A, stomach infections, viral diseases such as viral fever and conjunctivitis.

A good number of research proposals are available in the sphere of environmental effects on the diseases during a particular season, the AAL domain, however,

contributes very little to the inhabitants in the specific area which this work emphasizes [20–23]. Season sensors are devices that are used to get some signal depending on the current season. By integrating the sensors to find the impact of season on the human body and by learning the behavioural change, the smart environments can be made more valuable.

3.2 Effect of Air Quality on Human Health

Scenario 2: Mr. X is living independently in the environment used in scenario 1. Suddenly he found unconscious by security cameras in his room. The level of carbon dioxide in the atmosphere was the reason behind this incident. The smart system alerts the user to change the room if he can. At the same time, the doors and windows opened, airflow increased through the exhaust fans. Sensing system starts the air purification system. Soon the atmosphere is healthy, and the situation is under control. The system learns from the experience and maintains the future possibilities.

The study of Schulte [24] revealed that exposure to carbon dioxide in a range of 7–10% may result in unconsciousness. Other symptoms are headache, increased heart rate, rapid breathing, visual and hearing dysfunction and many more [25]. The minimum oxygen concentration in the air required for human breathing is 19.5%. The Occupational Safety and Health Administration, OSHA, determined the optimal range of oxygen in the air for humans runs between 19.5 and 23.5%. Oxygen and carbon dioxide sensors are used to measures the proportion of O_2 and CO_2 . The industrial applications mainly use this type of sensors; their integration in the AAL system is an added advantage.

3.3 Effect of Intensity of Light on Human Health

Scenario 3: Mr. X maintains his health by evening walking in the surrounding area. Recently he used to spend more time in an open area, after a few days he notices access aging skin. His eyesight is also drastically reduced. The cause of this effect on Mr. X was the sun-rays. The smart home system intelligently senses the intensity of light and adjust it according to the user's requirement. Additionally, it also maintains the requirement of sunlight indoor as well as the outdoor presence of the user.

The intensity of light has several benefits like sunlight triggers the release of the breezy chemicals, which boosts the mood and relieve stress. Sunlight also helps the body to maintain vitamins. Sunlight is useful against cancer as well as exposure to light during the day and darkness at night can help you maintain health. There are several drawbacks as well, as excessive exposure to the sun may increase the risks of various skin cancers, eye diseases and stimulated skin aging. Intelligent use of sun-rays in a smart environment significantly affects the human body [26]. A photometer is a device that estimates electromagnetic radiation in the range from ultraviolet to

infrared. The intelligent use of sun-rays in a smart environment significantly affects the human body.

3.4 Probable Month for Various Disease

Scenario 4: Mr. X suffers from body-ache in every ‘Y’ month. The atmospheric effect of some months is not suitable for living independently. Also the test for cancer performed in November, for flue in December and vitamins in April. The probability of a particular disease is high in a specific time of the year. The smart system regulates the indoor atmosphere, specifically for that month. Additionally, as a preventive step, it suggests the tests require for the specific season by learning from the historical records of the user.

There are numerous diseases one can observe in a specific period of the year. For example, the cancer detection rate is higher in November, December and January [27]. In the monsoon, the rate of malaria patients is very high. There are various methods, technologies and frameworks available to predict the disease effectively. The AAL domain needs to integrate potential opportunities.

4 Validation for the Proposed Smart Ecosystem

To support our proposal, we have presented a framework (Fig. 2) to utilize the role of environmental sensors in assisted living facilities. The structure uses body sensor data to recognize simple human activities like walking, standing, lying and sitting.

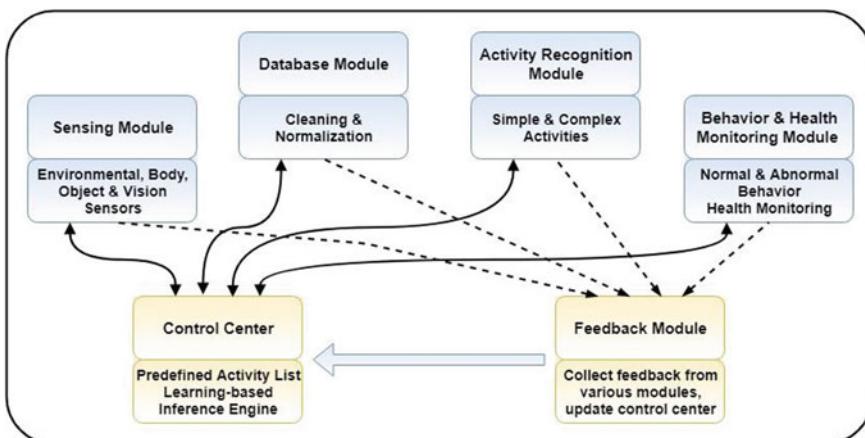


Fig. 2 Conceptual framework: behaviour and health monitoring

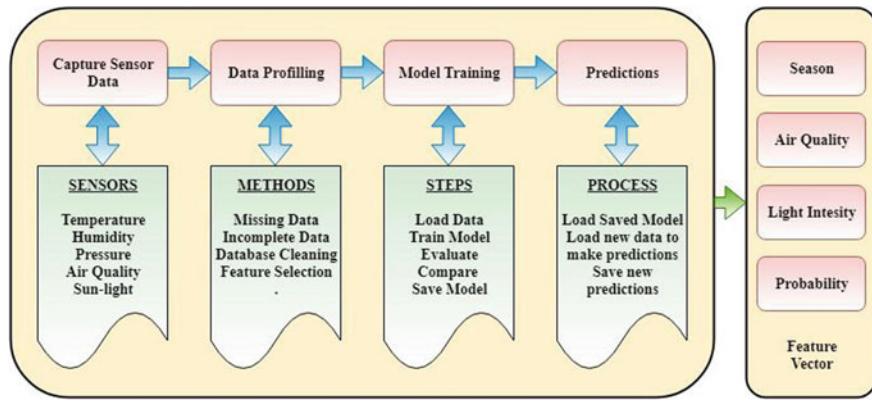


Fig. 3 Use of environmental sensors to predict the health of inhabitant

The addition of the object sensor provides the ability to understand more complex activities like watching TV, preparing breakfast, using washrooms and many more. The camera sensor is useful to identify the person to define his/her behaviour. The environmental sensors enable the AAL system to deal with the advanced services we have mentioned in Fig. 3. Sensor firings from various environmental sensors are collected and data profiling is performed to make the dataset ready for machine learning algorithms. The models are required to train and employ to predict the health of the elderly by providing the testing dataset.

Most of the available solutions use body and object sensors to model human behaviour [28]. Distinct categories to promote AAL services primarily focus the issues like activity recognition, behaviour monitoring, location, communication and data analysis. In our previous work, we explored the ambient assisted living domain with in-depth analysis and real-time human behaviour monitoring [13, 29]. Below is the list of critical challenges in smart environments recognized in the articles.

Architecture and framework

Interleaved and concurrent activities

Anomaly detection and ambiguity of interpretation of the sensor data Single occupant versus multiple occupants

Pervasiveness and acceptability of the system Unknown activity data

Quality of data used in intelligent algorithms

Risk management, user's requirements and data processing

Various research articles in the AAL domain covered the areas mentioned earlier [30–34]; we found very less contribution to the topics discussed in the presented work (Sect. 3). Other research contributions gave many novel proposals; still, there is a scope of widening the ideal AAL system [35–39]. Work mentioned here explored various AAL frameworks to identify their ability to deal with the proposal. We can perceive that more efforts expected in assisted living environments which can accommodate the possible issues of human health. Figure 4 throws the light on the

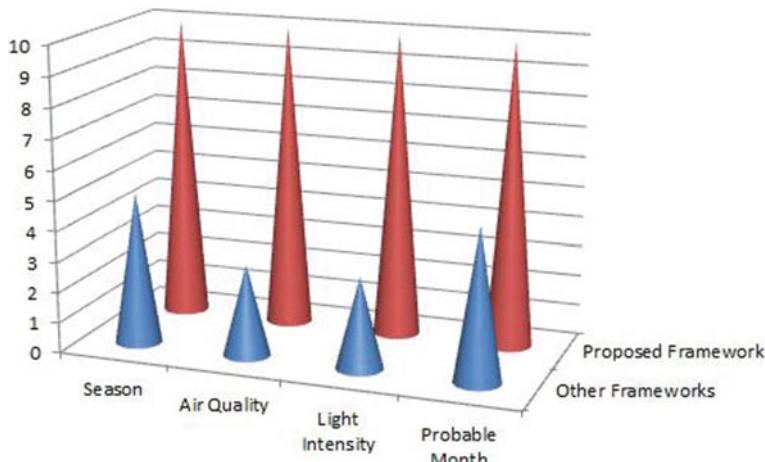


Fig. 4 Ability of the frameworks to deal with the environmental issues

work accomplished in the field of implied future issues on a scale of 1–10. The graph indicates that minimal effort contributed to environmental issues compared to others like the movement of the body, various smart objects and cameras to observe the behaviour of the occupant in an assisted living facility.

5 Conclusion

Many developed countries insist their people work up to a higher age to decrease the burden of pension, other perks and medical resources. The less-developed countries have to face similar challenges in a near feature. More advanced technological parameters need to be considered to help the elderly for an improved life to fulfill the promises of the ideal AAL system. We have considered seasonal diseases, indoor air quality, the effect of sunlight and probable time of various diseases. Most of the issues discussed here have applications in various fields of science; ambient assisted living technology hardly addresses these problems. Assisted living technologies provide an effective solution to face future challenges; provided the technology offers an ideal solution to the inhabitants of the smart environment. The research community must address the issues discussed in this article to make the AAL system future proof. As a future scope, we want to extend the convictions by designing a working model that can be used to offer extended services and compare the outcomes with the available solutions to measure its efficiency.

References

1. Calvaresi D, Cesarini D, Sernani P, Marinoni M, Dragoni AF, Sturm A (2017) Exploring the ambient assisted living domain: a systematic review. *J Ambient Intell Humaniz Comput* 8(2):239–257
2. Stevenson S (2014) Connecting families to senior care. A Place for Mom. Accessed 20 June 019. <https://www.aplaceformom.com/blog/10-29-14-ambient-assisted-living/>
3. Shah JH, Patel AD (2018) Ambient assisted living system: the scope of research and development. In: 2018 international conference on EECCMC
4. Di Martino C, D'Avino G, Testa A (2012) ICAAS: an interoperable and configurable architecture for accessing sensor networks. In: Technological innovations in adaptive and dependable systems: advancing models and concepts. IGI Global, pp 93–108
5. Gaddam A, Mukhopadhyay S, Gupta GS, Guesgen H (2008) Wireless sensors networks based monitoring: review, challenges and implementation issues. In: 2008 3rd international conference on sensing technology. IEEE, pp 533–538
6. Li HB, Kohno R (2007) Introduction of SG-BAN in IEEE 802.15 with related discussion. In: 2007 IEEE international conference on ultra-wideband. IEEE, pp 134–139
7. Patel A, Jhaveri R, Dangarwala K (2013) Wireless sensor network-theoretical findings and applications. *Int J Comput Appl* 63(10)
8. Testa A, Coronato A, Cinque M, Augusto JC (2012) Static verification of wireless sensor networks with formal methods. In: 2012 eighth international conference on signal image technology and internet based systems. IEEE, pp 587–594
9. Chen M, Gonzalez S, Vasilakos A, Cao H, Leung VC (2011) Body area networks: a survey. *Mob Netw Appl* 16(2):171–193
10. Marques G (2019) Ambient assisted living and internet of things. In: Harnessing the internet of everything (IoE) for accelerated innovation opportunities. IGI Global, pp 100–115
11. Testa A, Cinque M, Coronato A, De Pietro G, Augusto JC (2015) Heuristic strategies for assessing wireless sensor network resiliency: an event-based formal approach. *J Heuristics* 21(2):145–175
12. Munir MW, Perälä S, Mäkelä K (2012) Utilization and impacts of GPS tracking in healthcare: a research study for elderly care. *Int J Comput Appl* 45(11):35–37
13. Patel A, Shah J (2019) Sensor-based activity recognition in the context of ambient assisted living systems: a review. *J Ambient Intell Smart Environ* 11(4):301–322
14. John S (2019) Why sensors are the future of health care tech. Fortune. Accessed 26 January 2020. <https://fortune.com/2019/07/17/apple-health-watch-app/>
15. Nehmer J, Becker M, Karshmer A, Lamm R (2006) Living assistance systems: an ambient intelligence approach. In: Proceedings of the 28th international conference on software engineering. ACM, pp 43–50
16. Mitchell R, Popham F (2008) Effect of exposure to natural environment on health inequalities: an observational population study. *Lancet* 372(9650):1655–1660
17. Meng X, Wang F, Xie Y, Song G, Ma S, Hu S, Bai J, Yang Y (2018) An ontology-driven approach for integrating intelligence to manage human and ecological health risks in the geospatial sensor web. *Sensors* 18(11):3619
18. Marques G, Pitarma R (2018) Smartwatch-based application for enhanced healthy lifestyle in in-door environments. In: International conference on computational intelligence in information system. Springer, pp 168–177
19. Hay SI, Snow RW, Rogers DJ (1998) Predicting malaria seasons in Kenya using multi-temporal meteorological satellite sensor data. *Trans R Soc Trop Med Hyg* 92(1):12–20
20. Castillejo P, Martinez JF, Rodriguez-Molina J, Cuerva A (2013) Integration of wearable devices in a wireless sensor network for an e-health application. *IEEE Wirel Commun* 20(4):38–49
21. Jara AJ, Zamora-Izquierdo MA, Skarmeta AF (2013) Interconnection framework for m-health and remote monitoring based on the internet of things. *IEEE J Sel Areas Commun* 31(9):47–65
22. Koop CE et al (2008) Future delivery of health care: cybercare. *IEEE Eng Med Biol Mag* 27(6):29–38

23. Yang G, Xie L, Mantysalo M, Zhou X, Pang Z, Da Xu L, Kao-Walter S, Chen Q, Zheng LR (2014) A health-IoT platform based on the integration of intelligent packaging, unobtrusive bio-sensor, and intelligent medicine box. *IEEE Trans Ind Inform* 10(4):2180–2191
24. Schulte J (1964) Sealed environments in relation to health and disease. *Arch Environ Health* 8:438–452
25. OSHA (1989) Industrial exposure and control technologies for OSHA regulated hazardous substances. Occupational Safety and Health Administration, U.S. Department of Labor I, II, Washington, DC
26. LiveWell (2019) 13 ways the sun affects your body: the good & the bad. unitypoint.org. Accessed 20 January 2020. <https://www.unitypoint.org/livewell>
27. WHO (2019) Global health and aging. World Health Organization. Accessed 10 December 2019. <https://www.who.int/ageing/publications/globalhealth.pdf>
28. Patel A, Shah J (2019) Performance analysis of supervised machine learning algorithms to recognize human activity in ambient assisted living environment. In: INDICON 2019 (forthcoming). IEEE
29. Patel A, Shah J (2020) Real-time human behaviour monitoring using hybrid ambient assisted living framework. *J Reliable Intell Environ.* <https://doi.org/10.1007/s40860-020-00100-7>
30. Garces L, Ampatzoglou A, Avgeriou P, Nakagawa EY (2017) Quality attributes and quality models for ambient assisted living software systems: a systematic mapping. *Inform Softw Technol* 82:121–138
31. Garcia ACB, Vivacqua AS, Sanchez-Pi N, Marti L, Molina JM (2017) Crowd-based am-bient assisted living to monitor the elderly's health outdoors. *IEEE Softw* 34(6):53–57
32. Ghayyat H, Mukhopadhyay S, Shenjie B, Chouhan A, Chen W (2018) Smart home based ambient assisted living: recognition of anomaly in the activity of daily living for an elderly living alone. In: 2018 IEEE international instrumentation and measurement technology conference (I2MTC). IEEE, pp 1–5
33. Queiros A, Dias A, Silva AG, Rocha NP (2017) Ambient assisted living and health-related outcomes—a systematic literature review. *Informatics* 4:19
34. Van Grootven B, van Achterberg T (2019) The european union's ambient and assisted living joint programme: an evaluation of its impact on population health and well-being. *Health Inform J* 25(1):27–40
35. Ferreira A, Teles S, Vieira-Marques P (2019) Sotraace for smart security in ambient assisted living. *J Ambient Intell Smart Environ* 11(4):323–334
36. Offermann-van Heek J, Schomakers EM, Ziefle M (2019) Bare necessities? how the need for care modulates the acceptance of ambient assisted living technologies. *Int J Med Inform* 127:147–156
37. McConalogue E, Davis P, Connolly R (2019) Health technology assessment: the role of total cost of ownership. *Bus Syst Res J Int J Soc Adv Bus Inform Technol (BIT)* 10(1):180–187
38. Pace P, Aloisio G, Caliciuri G, Gravina R, Savaglio C, Fortino G, Ibanez-Sanchez G, Fides-Valero A, Bayo-Monton J, Uberti M et al (2019) Inter-health: an interoperable IoT solution for active and assisted living healthcare services. In: 2019 IEEE 5th world forum on internet of things (WF-IoT). IEEE, pp 81–86
39. Viana J, Ramalho A, Valente J, Freitas A (2019) Ambient assisted living—a bibliometric analysis. In: World conference on information systems and technologies. Springer, pp 218–228

Data-Driven Stillbirth Prediction and Analysis of Risk Factors in Pregnancy



Aravind Unnikrishnan, K. Chandrasekaran, and Anupam Shukla

Abstract One of the main issues in developing countries is the lack of policies for ensuring good public health conditions in rural areas. Maternal and child health care is one such area that has not improved in developing countries. Although child health has improved noticeably over the years, infant or under-5-mortality has not become any better. There remain major knowledge gaps in our understanding of how factors such as prenatal care, antenatal care, social and economic backgrounds, living conditions and lifestyle of pregnant women and their family members affect the pregnancy outcomes. Understanding such factors that affect the poor pregnancy outcome helps in formulating plans to prevent such issues and to treat them effectively. Determining health policies will be easier from a deeper analysis of such factors involved. This paper discusses some of the key machine learning techniques to predict the pregnancy outcome as a stillbirth or not and analyze some of the factors that majorly cause stillbirth.

Keywords Maternal health · Stillbirth · Neonatal care · Antenatal care · Prenatal care · Machine learning · Classification

A. Unnikrishnan (✉) · A. Shukla

Department of Computer Science and Engineering, Indian Institute of Information Technology, Pune, India

e-mail: aravindmathradan@gmail.com

A. Shukla

e-mail: anupamshukla@iiitm.ac.in

K. Chandrasekaran

Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal, Karnataka, India

e-mail: kchnitk@ieee.org

1 Introduction

With the advancement in data science and machine learning techniques, extracting information from and identifying patterns in large datasets have become very effortless. A large amount of useful information and patterns that are otherwise not so intuitive for the human brain can be identified and extracted from datasets. In this information age, data is available for almost any subject in the world. The medical field, for instance, collects a lot of medical data from every device, a patient admitted in a hospital, is assisted with, and stores them as Electronic Health Records (EHR). Health surveys conducted by public or private organizations targeting specific regions or groups of people are similar sources of data. Public Health can be improved substantially from a deeper understanding of the lifestyle of common people and its effects on their health from these survey datasets. One such area that needs improvement in the public health sector of developing countries is maternal and new-born health care. Even though the overall child health care domain has improved considerably in developing countries, the rate of infant and new-born mortality has not been found to decrease. Moreover, the highest rates of neonatal and prenatal deaths seem to occur in developing countries due to their poor socioeconomic conditions [1]. According to a World Health Organization blog of 16 February 2018 every day, almost 830 women die of preventable maternal and infant health problems [2]. It was reported that 99% of all maternal deaths occur in developing countries. Another WHO blog of 28 September 2018 states in 2017 approximately 2.5 million children died within 1 month of delivery in which around 7000 new-born deaths occur every day with about 1 million infant deaths on the first day and almost 1 million dying by the next 6 days [3]. This paper focuses on the use of machine learning and data analysis techniques to predict the chances of poor pregnancy outcomes and to build an understanding of the factors contributing to such outcomes in developing countries. The challenge in the current situation lies in the lack of knowledge in our understanding of how the supply of nutrition, prenatal care and antenatal care delivery, mental support and environmental and social factors contribute to the risk of poor maternal and child health results [4]. If we could figure out the main factors and how much they are weighing then we could bridge this knowledge gap and thus the results from the analysis will be useful in planning public health interventions in the form of health policies, constant medical care, supervision, etc. and decide what kind of intervention is required for what group of mothers at what time in their pregnancy period. This study uses “Annual Health Survey: Women Schedule (Sects. 1 and 2)” dataset published by the Department of Health and Family Welfare, Government of India in Open Government Data Platform India portal under Govt. Open Data License—India [5]. The women schedule comprises two sections, Sects. 1 and 2.

The remainder of this paper is arranged as follows: Sect. 2, begins with an overview of the dataset used and gives a detailed description of the same; Sect. 3, proposes the model in detail including steps for preprocessing and testing; Sect. 4, gives the results

of the model and gives a detailed analysis of the results including graphs for comparison of the results; Sect. 5 discusses some of the limitations of this work. Finally, the conclusions and future scopes are drawn in Sect. 6.

2 Dataset Description

The dataset was created from a survey conducted in nine ($8 + 1$) states in India that are classified as Empowered Action Group (EAG) states, who are lagging behind demographic transitions and showing worst statistics in infant and maternal mortality rates [5].

These nine states house almost 48% of the total Indian population and 59% of births while also accounting for 70% of infant mortality, almost 75% of deaths of under-5-aged and 62% of maternal mortality in the country and are under focused supervision for helping them improve their health and lifestyle conditions. The baseline survey was conducted in 2010–11, the first update was received in the 2011–12 and second update in 2012–13.

Section 1 has 197 variables and Sect. 2 350, some of which are the same as in Sect. 1. We have mainly used Sect. 1 of the dataset and hence will mention only the main attributes of this section.

Section 1 contains variables about the end result of pregnancy(s) (born alive/stillbirth/aborted); history and information about the birth; information about medical care received during delivery; details of antenatal, natal and post-natal health care received; immunization is given to the child; breastfeeding practices and nutrition supplements' details; health problems that affected child (Pneumonia, Diarrhea and fever), if any; registration of births, etc [5].

2.1 Properties of Dataset

The variety of questions asked in the survey of EAG states causes the data set to have unique characteristics. Some of the main challenging characteristics that the dataset possess are the imbalanced distribution of samples, high dimensionality and missing values. For optimizing our model for the best stillbirth prediction results, these challenges have to be solved.

- (I) High dimensionality: As seen in the description of the women's pregnancy schedule, the dataset comprises of more than 350 features, most of which are unimportant. This staggeringly high dimensionality of the data causes the model to perform poorly and come up with any patterns from the data. Moreover, most of the variables in this dataset will have linear relations between one another which may lead to redundancy of data. Hence to avoid this redundancy

- many variables will have to be removed using feature reduction techniques and also combine the related features.
- (II) Highly imbalanced sample distribution: The other main challenge that needs attention is the high imbalance of dataset. The number of samples in one prediction class (more than 95% of survival cases) of the dataset is too high compared to the number of samples in the other prediction class (less than 5% of mortality cases). This characteristic may lead to severe bias in the dataset among both labels of prediction. This imbalance is visible in almost all of the variables of the dataset. Therefore, it is important to handle the problem of the imbalanced dataset and the steps used will be discussed in detail later.
- (III) Missing data and Human errors: Due to possibilities of occurrence of human errors the data may have many incorrect variables and variables with missing values. This is a problem that has to be dealt with.

3 Machine Learning Model for Risk Factors in Pregnancy

The procedures for building the model are as follows. The challenges discussed in the previous section are something to be solved carefully. This comes under the preprocessing part which is the first step in the modelling. Preprocessing involves cleaning of data, removing missing values, etc. The next step is building the classification model that best fits the data and produces respectable prediction accuracy. In this, we have listed out few classification models that give the maximum accuracy and have given the comparisons.

3.1 *Preprocessing*

As the survey data is noisy and skewed, preprocessing is a must before feeding the data to the model. This is the most time-consuming and most important step in the whole work. Various steps are involved in the complete preprocessing of the dataset.

Combining Datasets from the Different States

The survey as mentioned above was conducted separately in the nine EAG states. For increasing the size of the dataset to be fed to the model, we have combined the datasets from all the states into one by removing some inconsistent attributes and values.

Variable Selection

The original dataset contains 350 attributes most of which are irrelevant for this work. Irrelevant attributes can cause bias in prediction. Hence important variables that contribute the most to the result are to be selected. Various methods are available for the same. Since we were working on the prediction of risk of stillbirth, attributes

that contain information about post-natal care, immunization, etc. are irrelevant and were removed.

Another method is removing variables from the domain knowledge. For this work, a lot of domain study was done to identify the main causes for poor pregnancy results and a lot of variables like information about household appliances, information about educational qualification, family planning methods used etc., were removed.

Next, the attributes that had missing values in small numbers were also dropped making sure it does not contribute to the output. Some of the variables that are dependent on each other and produced better information together, were combined into new variables through appropriate methods. For example, the dataset contains information about different tests performed by mothers during pregnancy period viz., abdominal tests, blood tests, urine tests, etc. These were combined into one variable ‘tests’ to get a better insight. Table 1 shows variables that are combined into single variable categories.

Stratification of Variables Most of the instances of certain numerical attributes will be clustered around a specific range of values while the instances (samples) beyond that range will be very less in number. It is important to have a sufficient number of instances in the dataset for each range of values. Clustering the attributes into strata of ranges will help to reduce the bias caused by such imbalanced attributes. This was done in the dataset for certain attributes such as age, IFA, etc.

Assigning Costs and Normalization

Due to the same reason of a high imbalance in unique values of certain attributes, to avoid bias in learning, it is a good practice to assign some positive and negative costs to the samples of such attributes. This should be in such a way that it makes a difference even after the normalization of those attributes. We have not used any particular rule to decide the costs but have come up with different ways of assigning costs to different attributes through experiments and by determining feature importance scores of each attribute.

The numerical attributes were scaled to values from 0 to 1 for normalizing.

Data Imputation

After deciding upon the standardized features to be used, the next step is to deal with the missing values. The methods for imputation of missing values were also decided through repeated experiments on different attributes. The main methods used to fill were mean, previous-values, next-values, and zero. Some rows were also dropped to remove missing values.

3.2 *Building the Model*

Preprocessing of data reduced the number of necessary variables to 19 variables including the output label ‘kind_of_birth’. Training set selected 70% of the data randomly and the remaining was used for testing and validating the model.

Table 1 Variables grouped to obtain better results

Individual variables	Combined variable
is_abdominal_tested is_bp_tested is_weight_tested is_urine_tested is_blood_hb_tested is_blood_for_othr_tested is_ultra_sound_tested is_breast_examination is_blood_for_group_tested	Tests
swelling_of_hand_feet_face paleness_giddiness_weakness visual_disturbance excessive_fatigue convulsion_not_from_fever	general_problems
weak_or_no_movement_foetus abnormal_position_foetus	foetus_problems
excessive_bleeding vaginal_discharge excessive_vomiting jaundice hypertension_high_bp other_preg_problems	pregnancy_problems
premature_labour excessive_bleeding_during_birth prolonged_labour obstructed_labour breech_presentation convulsion_high_bp others_prob_during_delivery	delivery_problems
consumption_of_ifax consumption_of_ifa_syrup	Ifa
weight_of_baby_grams weight_of_baby_kg	baby_weight

To solve the high imbalance in data (a large percentage of data consist of positive cases), the dataset was under-sampled to an almost equal ratio of positive and negative cases. This helps to prevent the bias for positive cases.

Finally, the dataset used for this research comprises of 10,146 samples for training out of which 4806 belongs to the dead class and 5340 to survival class. 4376 samples were used for testing and validation with 2073 samples belonging to the negative class and 2303 to survival class. A slightly higher number of samples in the survival case of the training set is used as an impurity for the models to learn better and it was experimented and found to serve the purpose. However, the samples in test cases are in a 9:10 ratio of negative cases to positive cases to reflect the slight bias in the

original dataset. The training set is 70% of the original dataset and the test set 30% of the original.

Model Selection

The diversity of models available for classification tasks makes it difficult to determine which one performs better and under what conditions. We have tried to perform the task on as many classification models as possible and have come up with five models that fit the data and predict best. They are Logistic Regression model, Gradient Boosting Classifier, Random Forest Classifier, AdaBoost Classifier and Artificial Neural Network (ANN). The data was trained and tested repeatedly to learn the optimal parameter tunings for each model. We have also provided a comparison of each model in terms of accuracy, confusion matrix and other measures in the results section. Algorithms used are briefed below.

(1) Boosting algorithms

Boosting is a machine learning technique to improve the accuracy of a model, by weighting different classification rules based on a self-rated confidence score that accounts for the usefulness of that rule and combines them to give the best accuracy [6]. Boosting algorithms are used to improve the accuracy of a model which is otherwise producing poor accuracy or taking too much time. AdaBoost is the most basic boosting algorithm of all and it is the foundation for other boosting algorithms. AdaBoost is mainly a classifier and it usually works on a decision tree with one level.

Gradient Boosting (GB) is another boosting algorithm used in this research. It is basically a combination of gradient descent and boosting methods.

$$\text{Gradient Boosting} = \text{Gradient Descent} + \text{Boosting(AdaBoost)}$$

Extreme Gradient Boosting (or XGBoost) is a high performing implementation of Gradient Boosting.

(2) Random Forest

Random Forest is an ensemble learning method that combines a large number of decision trees by using the bagging method of classification trees [7]. It decides the model by analyzing the classification output of each decision tree for the highest number of same class predictions. The individual trees are uncorrelated. The higher the number of such trees, the higher is the probability of obtaining a better prediction accuracy. A forest of uncorrelated trees is made using feature randomness.

Feature Importance

Feature Importance score is Feature Selection technique used to determine the importance of different attributes and to understand how the outcome is affected by each attribute. Determining Feature Importance score will be an integral part of this research work as it suggests the main factors causing poor pregnancy outcomes

in women and gives an insight on the extent to which each factor contributes to the same. The higher the score, the more important the variables are.

Performance Measures

Accuracy score of a model over testing does not always give the actual information about the predictions, especially for classification models. Therefore, the metrics used in this work are precision, recall and F_1 scores of confusion matrices and Area Under (AUC) the Receiver-Operating Curves (ROC) values.

Precision, Recall and F_1 scores can be computed from the confusion matrix as follows.

$$\text{Precision Score} = \frac{\text{True Positive (TP)}}{\text{TP} + \text{False Positive (FP)}} \quad (1)$$

$$\text{Recall Score} = \frac{\text{TP}}{\text{TP} + \text{False Negative (FN)}} \quad (2)$$

$$F_1 \text{ Score} = \frac{2}{\left(\frac{1}{\text{precision}}\right) + \left(\frac{1}{\text{recall}}\right)} \quad (3)$$

AUC-ROC curve is a performance measure for classification models. ROC is a probability curve and AUC represents its degree or ability to distinguish the classes. Better the AUC score, better the model is in predicting the actual surviving cases as surviving and actual dead cases as dead. The ROC curve is plotted as True Positive Rate (TPR) on y -axis versus False Positive Rate (FPR) on the x -axis.

The formulas for FPR and TPR are given below.

$$\text{True Positive Rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (4)$$

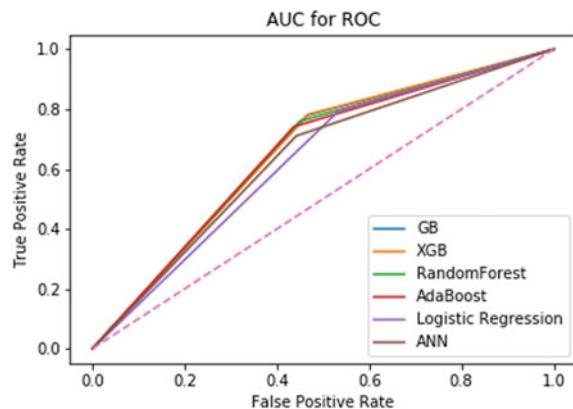
$$\text{False Positive Rate (FPR)} = 1 - \frac{\text{False Positive (FP)}}{\text{TN} + \text{FP}} \quad (5)$$

4 Results and Analysis

A variety of classifier models were implemented and tested to identify the best fitting model. All the experiments were performed on the Python-based kernel with additional libraries for each classifier model. The comparison of full prediction results including accuracy of the model, precision-recall scores, F_1 scores and AUC (ROC) scores and graphs of all the models are given below. Refer to Table 2 for scores comparison.

Table 2 Comparison of models

Model	Accuracy	Precision	Recall	F_1	AUC
Gradient boosting	0.684	0.674	0.771	0.719	0.679
XGBoosting	0.682	0.671	0.775	0.719	0.677
Random forest	0.687	0.683	0.756	0.718	0.683
AdaBoost	0.677	0.677	0.738	0.706	0.674
Logistic regression	0.660	0.642	0.796	0.711	0.652
ANN	0.644	0.622	0.823	0.709	0.635

Fig. 1 ROC plots

4.1 AUROC Plots

AUC of ROC tells us about the capability of the model in distinguishing between the classes. Figure 1 shows the comparison of AUROC plots of all the models.

The highest AUC score is 0.683 by Random Forest model which means that 68.3% of the time, the model will be able to distinguish between survival and death cases.

4.2 Neural Network Performance

The neural network model was implemented with six layers using ReLu activation function except for the last layer which uses sigmoid activation. The batch size of the model was fixed 16 and the model was trained for 100 epochs.

Figure 2 shows the training and validation accuracies of the neural network model vs epochs.

Figure 3 shows the training and validation loss of neural network model versus epochs.

Fig. 2 Training and validation accuracies of neural network

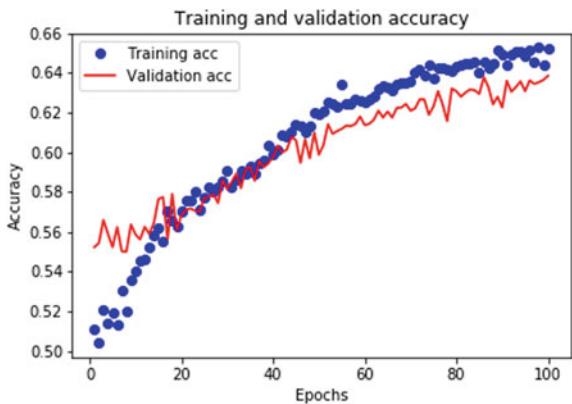
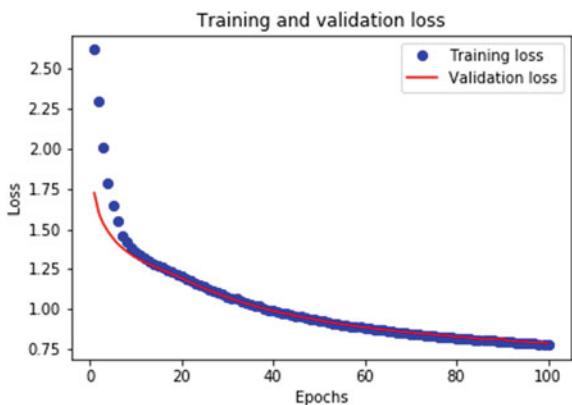


Fig. 3 Training and validation losses



4.3 Feature Importance Results

Feature Importance scores for the variables of this dataset was generated from the Random Forest Classifier model. Since many of the variables were combined into one, the importance of score of such features will be a generalized one of all such features. The importance of variables is visualized in Fig. 4. From the figure, it is clear that the type of birth, i.e. single or multiple has a role to play in deciding the risk of pregnancy. Similarly, general variables (Those that are mainly social or environmental rather than personal) such as state, rural, age, gender, etc., seem to be affecting the results. This might be due to the bias in the number of samples across these features like more number of samples from a particular state. The variable ‘diagnosed_for’ consists of the diseases that pregnant women have suffered in the past and it seems to affect the results in a high fraction. Antenatal Care (ANC) is another variable that has a huge impact on the outcome of the pregnancies. ANCs are

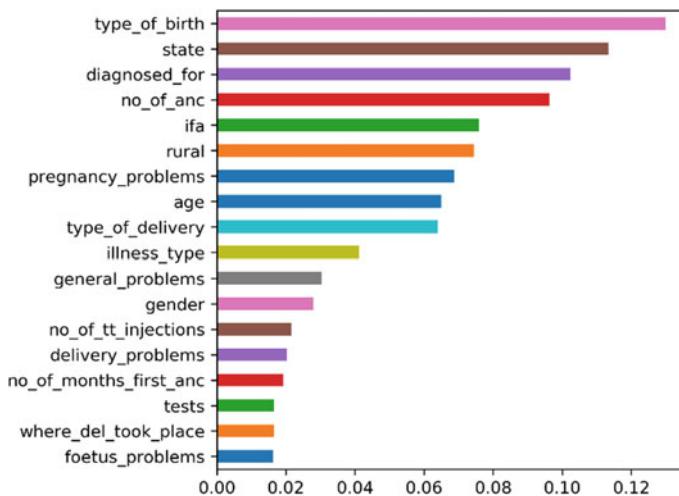


Fig. 4 Feature importance scores

preventive health care treatments and check-ups given to pregnant women throughout their pregnancy period to promote a healthy lifestyle to benefit both the mother and child.

Iron and Folic Acid (IFA) is such a treatment method given to pregnant women during ANC period to prevent and treat anaemia. The graph shows that this variable is also important for a good pregnancy outcome.

The variables ‘illness_type’, ‘pregnancy_problems’, ‘general_problems’, ‘foetus_problems’, ‘delivery_problems’ has grouped all the diseases and problems faced by the mother during pregnancy period and delivery as shown in Table 1 and they all seem to contribute to the outcome reasonably well except ‘foetus_problems’ which didn’t give a good score compared to others despite being an important factor. Some of the variables that have the least influence in outcome are, the tests performed during pregnancy, the number of months of pregnancy at the time of first ANC, the place of delivery, etc.

The most important thing to be taken care of during pregnancy is the lifestyle and living standards. Antenatal Care should be taken seriously and done as prescribed. Prevention of pregnancy-related problems is the important and immediate treatment if suffering from any such issues, is necessary.

5 Limitations

This research work has a lot of limitations. The dataset is comparatively new and does not have any major works done on it previously. Some of the main challenges faced are listed below.

1. Irregular distribution of values of samples in each feature. This is a major problem because the denser values of the feature will create a bias in learning the final outcome.
2. Despite being a very large dataset, the number of samples available for training and testing is negligibly less after the preprocessing step has removed most of the samples. This reduces the accuracy considerably.
3. High imbalance in the dataset was causing heavy bias for positive cases and this had to be tackled by sacrificing a huge amount of samples.
4. The purpose of this research work is to mainly predict the number of negative cases well and identify the factors causing it, both of which were successful to an extent. However, the model has not displayed any good performance in predicting the number of positive cases.

6 Conclusion and Future Work

The increase in NGOs, self-help groups, health schemes from government, etc., is helping improve the maternal and child health scenario in developing countries. Along with this, the number of health surveys conducted across the country is considerably increasing and the data from such surveys are a great source of information if analyzed properly. Insights gained from such analysis can be useful to study the cause of poor health situations and to decide upon the measures to solve the problems causing such situations. However, there are a lot of challenges coming along with the great opportunities from these datasets. This paper presents methods to predict risks involved in pregnancy and a tool to predict stillbirths to prevent it from a relatively untouched field of health care, which is maternal and child health. This paper may help to build more useful tools for predicting stillbirth and health care recommendation systems. As a future work dataset can be pre-processed well and other Machine Learning Models can be used accordingly.

References

1. Bhutta ZA, Darmstadt GL, Hasan BS, Haws RA (2005) Community-based interventions for improving perinatal and neonatal health outcomes in developing countries: a review of the evidence. *Pediatrics* 115(Supplement 2):519–617
2. Maternal Mortality. Available [Online]: <https://www.who.int/news-room/fact-sheets/detail/maternal-mortality>
3. Newborns: reducing mortality. Available [Online]: <https://www.who.int/news-room/fact-sheets/detail/newborns-reducing-mortality>
4. Reddy UM (2007) Prediction and prevention of recurrent stillbirth. *Obstet Gynecol* 110:1151
5. Annual Health Survey (2017) Woman schedule, open government data platform India. Ministry of Health and Family Welfare. <https://data.gov.in/catalog/annual-health-survey-woman-schedule>. Published under national data sharing and accessibility policy (NDSAP). Available [Online]: <https://demodata.nic.in/sites/default/files/NDSAP.pdf>

6. Schapire RE, Singer Y (1998) Improved boosting algorithms using confidence-rated predictions.
In: Proceedings of the eleventh annual conference on computational learning theory
7. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32

A Comparative Study of Heuristics and Meta-heuristic Algorithms with Issues in WSN



Prince Rajpoot, Kumkum Dubey, Nisha Pal, Ritika Yaduvanshi, Shivendu Mishra, and Neetu Verma

Abstract Sensors deployed in an area are collectively make a network that is known as WSN. Use in remote areas make it non-rechargeable and improvise maximum utilization of battery for longer lifetime of WSN. There are numerous issues in the development of a WSN that affect total lifetime. Clustering is an approach followed by many researchers for enhancing the lifetime of WSN. Clustering can be done by using heuristics method or meta-heuristic methods. Heuristic algorithms are applicable only for sensor applications, whereas meta-heuristic algorithms are used in WSN and also accustomed for various other fields. In this paper, we have discussed various issues in WSN and also various heuristic and meta-heuristic algorithms along with their comparisons with each other.

Keywords Wireless sensor network (WSN) · Multi-objective optimization (MOO) · Coverage · Connectivity · Network lifetime

1 Introduction

Sensors deployed in an area are collectively make a network that is known as WSN. Sensors are lower in cost and have a limited amount of energy. Sensor nodes sense the field's environmental condition in which they are deployed. Depending upon the nature of sensors, WSN has enormous applications in every area. WSN is used in agriculture, military, medical, space and other areas. In clustering approach, cluster of these sensor nodes is made. A cluster head (CH) is chosen that collect the info of all the neighbors that belong to same cluster. After performing some computation and calculation over this received data from all member nodes, CH delivers this processed packets to sink. It is assumed that the BS has infinite amount of energy since we can provide massive amount of energy by changing batteries externally.

P. Rajpoot · K. Dubey · N. Pal · R. Yaduvanshi · S. Mishra (✉) · N. Verma
Rajkiya Engineering College, Ambedkar Nagar, Akbarpur, India
e-mail: shivendu0584@gmail.com

This collected information is used for taking further decisions. The selection of CH is a big deal that plays an important role for enhancing the lifetime.

The selection of CHs is done using some parameters like residual energy of nodes, delay, distance to CH, lifetime of CH, average distance to nodes, etc. Heuristic and meta-heuristic algorithm are two types of algorithms. LEACH, LEACH-C, Hybrid LEACH are some of the popular heuristic algorithms that verify that the selection of CHs leaves a great impact for enhancing the lifetime of WSN. Whereas, genetic, artificial bee colony (ABC), particle swarm optimization (PSO), etc., are various clustering approaches come under meta-heuristic in WSN using multi-objective optimization. Meta-heuristic algorithms contain the nature-inspired approaches [1] that are divided as evolutionary algorithms and swarm-based algorithms. Evolutionary algorithms initialize the population in random manner and optimize the solution by generating offspring from the populations. Differential evolution (DE) [2], genetic algorithm (GA), etc., comes under this category. Swarm-based algorithms also start with random population and make the solution optimize by using previous or best solution. PSO [3], ACO [4], ABC [5], etc comes under this category.

In this paper, we have discussed many issues in WSN along with the algorithm that can be used to resolved those problems. Some of the comparison among popular and frequently algorithm are also done at the last section of this paper.

2 Issues in WSN

Some of the main issues are as follows:

- (a) **Fault Tolerance:** We need to make our system fault tolerant in such a way that failure of some node might not affect the whole network working. Protocols to various layers should also be able to adjust their functionality so that failure of some nodes might not cause failure of whole network. Computation of probability of no failure at t time can be done as:

$$R_n(\text{time}) = e^{-\lambda_n(\text{time})} \quad (1)$$

λ_n —failure rate of n th node.

- (b) **Delay:** It is also known as network latency. If expansion of a network is done by adding nodes then it may possible that the delay due to waiting time for path is reduced because of many new paths from new nodes. These nodes will work as the intermediate nodes so it enhances the queuing delay. This also lead to increment in buffering due to high channel contention. Computation of delay can be done as [6]:

$$\text{Delay} = (D_{\text{queuing}} + D_{\text{propagation}} + D_{\text{transmission}}) * \text{Packets(source_to_BS)} \quad (2)$$

- (c) **Consumption of Energy:** The consumption of energy can be given as [7]:

$$E_{\text{consume}} = \sum_{p=1}^{\text{Hop}} (t_p^{\text{access}} + t_p^{\text{process}}) * E_{\text{operate}}^p + E_j^t * t^{\text{msg}} \quad (3)$$

- (d) **Lifetime of Network:** Lifetime of WSN is dependent to particular application. If complete data is important then death of single node will affect the whole process, this lead to the network crash with dead of first node and can be given as:

$$\text{Life}_{\text{network}} = \min(\text{Life}_{\text{sensor_j}}) \quad (4)$$

- (e) **Connectivity of Network:** If a network is connected, then it means whole sensor nodes are directly or indirectly connected with all other nodes. Connectivity confirms the data communication among the sensors. Connectivity fitness can be computed as [8]:

$$\text{Connect_fitness} = \sum_{k=1}^{x*y} 1 - e^{-(R_{\text{Communicate}_k} - R_{\text{Sensing}_k})} \quad (5)$$

- (f) **Coverage:** Coverage shows the how much portion of an area can be sensed by the sensor. It takes all the points one by one of an area and check whether there is at least one sensor or not that could cover that point. If all the points are covered by at least one sensor, then it means that whole area is covered by the WSN. If the size of area is $x * y$, then the coverage can be given as [9]:

$$\text{Cover_portion} = \frac{\sum_{x'=0}^x \sum_{y'=0}^y S(x', y')}{x * y} \quad (6)$$

where $S(x', y')$ —supervision status of point at (x', y') .

- (g) **Number of Sensors Deployed:** More sensors mean more cost, so we need optimum number of nodes to deploy in an area but maximum area should be covered by the nodes. So we need to restrict the nodes number.

3 Comparative Study

We have compared some heuristic and meta-heuristic algorithms used for lifetime enhancement of WSN as shown in Table 1.

Table 1 Comparative study of various algorithms used in WSN based on various factors

Algo	Type	Type of decision for CHs	Topology	CH dispensation	Cluster generation	Goal	Power vanishing
LEACH [10]	Heuristic	Arbitrary	Direct transfer	Weak as well as arbitrary	Nearest CH	Power saving	No balancing
Energy LEACH [11]	Heuristic	Arbitrary but after chosen finalized based on energy	Single Hop	Weak and arbitrary	Remaining energy plays role after selection	Chose high residual energy CHs	No balancing
LEACH-C [10]	Heuristic	Arbitrary	Direct transfer	Weak and arbitrary	Remaining energy greater than average	Power saving	No balancing
EECS [12]	Heuristic	Arbitrary with election	Single hop and location unaware	Strong using localized communication	Three attributes based cost	Power preservation	Balance up to a level
HEED [13]	Heuristic	Probability with residual energy	Single hop	Well-distributed cluster	Closest CH	Power preservation	No Balancing
DWEHC [14]	Heuristic	Arbitrary	Multi hop	Strong since clusters size are balance and also not overlap	Closest CH	Power preservation	Not available
FLOC [15]	Heuristic	Arbitrary	Single hop	Strong since clusters size are balance and also not overlap	Closest CH	Self-healing and preservation	Not available
GA [16]	Meta-heuristic	Residual memory and residual energy-based fitness function	Single hop	Unequal but unequal cluster	Closest CH	Energy preservation	Somewhat balanced
PSO [3]	Meta-heuristic	Multi-objective	Single hop	Unbalanced cluster	Closest CH	Low distance with high energy	Unbalance

4 Applications of Multi-population Methods

We have described various applications of multi-population methods in two categories: first is according to optimization classes that are explained in Table 2 and second is according to area of applications that are described in Table 3.

Table 2 Applications according to optimizations classes

Problems	Ref.	Important analysis
Optimization using multiple models	[17]	According to [17], there are many enhancements in the PSO algorithm for the multi-model optimization but this proposed algorithm gives the best result as compared to these other algorithms
Active optimization	[18]	According to [18], they have proposed an algorithm that is better than other algorithms for solution of dynamically optimized problem
MOO	[19]	In paper [19], it is described that this evolutionary PSO is better as compared to other enhancements in PSO for the solution of the MOO problems
Large-scale optimization	[20]	In [20], it is described that a parallel adaptive ABC leads in performance as compared to other algorithms that are being used for the solution of this type of optimization problems
Connective optimization	[21]	According to the author in [21], the algorithm used is better than the other algorithm used for the solution of these types of problem in optimization
Strained optimization	[22]	In [22], it is said that this hybrid PSO gives better result as compared to the other algorithm used with other enhancements in PSO for solving these types of problems in optimization
Clamorous optimization	[23]	A less number of algorithms are discussed for these types of optimization problems

5 Conclusion

Here, we have presented fundamental of WSN with various applications continuing with various issues in WSN. We have also provided a comparative analysis among various heuristic and meta-heuristic approaches used in WSN for various purposes. We have also proposed various applications of multi-population-based approaches.

Table 3 Applications according to application areas

Problems	Ref.	Important analysis
Way scheduling	[24]	In [24], it is described that this algorithm gives better result than the other algorithms that are being used for solving these way searching problems
Analytical information	[25]	According to the author in [25], some genes-based programs are used that gives better performance as compared to other methods which are being used for solving these types of problems
Criterion assessment and authority	[26]	According to paper [26], fruit fly-based optimizing methods is performing better than the other algorithms that are being used nowadays and this algorithm is also efficient as compared to other methods
Electronic dilemma	[27]	In [27], a hybrid MSO algorithm have been used that gives us better result than the other enhanced PSOs used for solving these problems
Mathematic equalization dilemma	[28]	According to paper [28], a multi-layered optimization algorithm have been used whose performance is better. This algorithm is used for solving these types of problems

Acknowledgements This paper is financially supported by TEQIP III of REC Ambedkar Nagar.

References

- Coello CAC, Toscano G, Salazar M (2004) Handling multiple objectives with particle swarm optimization. *IEEE Trans Evol Comput* 8(3):256–279
- Das S, Suganthan PN (2011) Differential evolution: a survey of the state-of-the-art. *IEEE Trans Evol Comput* 15(1):4–31
- Clerc M (2006) Particle swarm optimization. ISTE Publishing, London
- Dorigo M, Gambardella L (1997) Ant colony system: a cooperative learning approach to the traveling salesman problem. *IEEE Trans Evol Comput* 1(3):53–66
- Karaboga D, Basturk B (2007) A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J Glob Optim* 39(3):459–471
- Ammari HM, Das SK (2005) Trade-off between energy savings and source-to-sink delay in data dissemination for wireless sensor networks. In: Proceedings of 8th ACM international symposium on modeling, analysis and simulation of wireless and mobile systems (MSWiM'05), Montreal, Canada, Oct 2005, pp 126–133

7. Rajagopalan R, Mohan CK, Varshney P, Mehrotra K (2005) Multiobjective mobile agent routing in wireless sensor networks. In: Proceedings of IEEE congress on evolutionary computation (CEC'05), Edinburgh, UK, Sept 2005, pp 1730–1737
8. Syarif A, Benyahia I, Abouaissa A, Idoumghar L, Sari RF, Lorenz P (2014) Evolutionary multi-objective based approach for wireless sensor network deployment. In: Proceedings of IEEE international conference on communications (ICC), Sydney, Australia, Jun 2014, pp 1831–1836
9. Chen J, He S, Sun Y, Thulasiraman P, Shen XS (2009) Optimal flow control for utility-lifetime tradeoff in wireless sensor networks. *Comput Netw* 53(18):3031–3041
10. Heinzelman WB, Chandrakasan AP, Balakrishnan H (2002) An application-specific protocol architecture for wireless microsensor networks. *IEEE Trans Wirel Commun* 1(4):660–70
11. Xiangning F, Yulin S (2007) Improvement on LEACH protocol of wireless sensor network. In: 2007 international conference on sensor technologies and applications (SENSORCOMM 2007). IEEE, pp 260–264
12. Ye M, Li C, Chen G, Wu J (2005) EECS: an energy efficient clustering scheme in wireless sensor networks. In: Performance, computing, and communications conference, IPCCC 2005. 24th IEEE international, pp 535–540
13. Younis O, Fahmy S (2004) HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans Mob Comput* 3(4):366–79
14. Ding P, Holliday J, Celik A (2005) Distributed energy-efficient hierarchical clustering for wireless sensor networks. In: International conference on distributed computing in sensor systems. Springer, Berlin, Heidelberg, pp 322–339
15. Demirbas M, Arora A, Mittal V (2004) FLOC: a fast local clustering service for wireless sensor networks. In: Workshop on dependability issues in wireless ad hoc networks and sensor networks (DIWANS/DSN), pp 1–6
16. Khan I, Sahoo J, Han S, Glitho R, Crespi N (2016) A genetic algorithm-based solution for efficient in-network sensor data annotation in virtualized wireless sensor networks. In: 2016 13th IEEE annual consumer communications & networking conference (CCNC). IEEE, pp 321–322
17. Niu B, Huang H, Tan L, Duan Q (2017) Symbiosis-based alternative learning multi-swarm particle swarm optimization. *IEEE/ACM Trans Comput Biol Bioinform* 14(1):4–14
18. Li C, Nguyen T, Yang M, Mavrovouniotis M, Yang S (2016) An adaptive multipopulation framework for locating and tracking multiple optima. *IEEE Trans Evol Comput* 20(4):590–605
19. Castro O, Santana R, Pozo A (2016) C-Multi: a competent multi-swarm approach for many-objective problems. *Neurocomputing* 180(SI):68–78
20. Zhou J, Yao X (2017) Multi-population parallel self-adaptive differential artificial bee colony algorithm with application in large-scale service composition for cloud manufacturing. *Appl Soft Comput* 56:379–397
21. Michalak K (2016) Sim-EDA: a multipopulation estimation of distribution algorithm based on problem similarity. In: Proceeding of European conference on evolutionary computation in combinatorial optimization, pp 235–250
22. Srivastava L, Singh H (2015) Hybrid multi-swarm particle swarm optimization based multi-objective reactive power dispatch. *IET Gener Transm Distrib* 9(8):727–739
23. Li J, Li M, Yang X (2009) Cluster based multi-populations genetic algorithm in noisy environment. In: Proceeding of Chinese conference on pattern recognition, Nanjing, China, Nov 2009, pp 161–165
24. Turky A, Sabar N, Song A (2016) A multi-population memetic algorithm for dynamic shortest path routing in mobile ad-hoc networks. In: Proceeding of IEEE congress on evolutionary computation (CEC), Vancouver, Canada, Jul 2016, pp 4119–4126
25. Maua G, Grbac T (2017) Co-evolutionary multi-population genetic programming for classification in software defect prediction: an empirical case study. *Appl Soft Comput* 55:331–351
26. Li S, Lu Z (2015) Multi-swarm fruit fly optimization algorithm for structural damage identification. *Struct Eng Mech* 56(3):409–422

27. Nawaz A, Mustafa E, Saleem N (2017) Solving convex and non-convex static and dynamic economic dispatch problems using hybrid particle multi-swarm optimization. *Tehnicki Vjesnik-Technical Gazette* 24(4):1095–1102
28. Yeh J, Lin J (2017) Learning ranking functions for information retrieval using layered multi-population genetic programming. *Malays J Comput Sci* 30(1):27–47

A Systematic Review on an Embedded Web Server Architecture



Sumanth Kumar Panday, R. V. V. Krishna, and Durgesh Nandan

Abstract An embedded web server is an integration of an embedded system and web server. An embedded system is related to a microprocessor which comes under IoT (Internet of things) and VLSI (Very Large-Scale Integration). As technology is still improving rapidly and it is possible to fabricate the number of the component in a small area by that size of a microprocessor is reducing. And parallelly speed of the Internet also covering and hope after launching of 5G network, Internet speed will be improved, and with high-speed internet, the management of device and network troubleshoot will improve, we can come out from the buffering problem and it will provide remote access to the webserver. This day EWS application is in trends due to its performance and effective time management and fixing of problems automatically without informing the lord of the device. In this research article authors study the systematic growth of EWS, its feasibility, suitability for various applications and deeply study about positive and negative including future scopes.

Keywords Embedded webserver · IoT · FPGA · Raspberry-Pi · Arduino board

1 Introduction

This Justification of embedded web server is simply the integration of more than one device which runs over command of the web server which is provided by the client through the server, where web server uses TCP/IP protocol for linking devices

S. K. Panday · R. V. V. Krishna

Department of Electronic and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, East Godavari, Andhra Pradesh, India

e-mail: sumanth.kumar.panday06@gmail.com

R. V. V. Krishna

e-mail: rvvkrishnaece@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India
e-mail: durgeshnandano51@gmail.com

with the help common address number which is also called an IP address and finally gives remote access from web server which will manage device effectively in time. Cascading of different devices and interfacing each other with the help of system software and then after uses for different-different application in general motive of embedded web server such that user can free in monitoring without facing any complexity web-based monitoring system is more advantageous for user like more security, reliability, flexible device user-friendly generate of stability and nowadays we are free from worries or problem of designing of hardware after introducing to Arduino and raspberry-pi, already they consist cascading pin connect with multiple devices. Various sections of paper organizations are as follows. Section 2 covers detailed about the systematic work done in this field. Section 3 covers the overview of essential part of EWS architecture like FPGA and core processor. Section 4 focuses on what is the exact methodology of EWS. Section 5 deals about to utility of EWS. Section 6 covers about findings and at last Sect. 7 covers the conclusion and future scopes.

2 Literature Review

In this section, we try to show the systematic growth of EWS architecture. In 2001, Ju presents interface mechanisms for use embedded management applications and the embedded web server. He presents HTTP/1.1 protocol also called POS-EWS which works on four mechanisms which are SSI-type, java SNMP type, SSI SNMP type and CGI-type [1]. In 2003 Saba Mynaganam gives an idea of monitoring the parameter of a connected device through online using PLC with an embedded web server [2]. In 2004 again design concept is presented by Tao Lin so that consumers can access their device remotely. This device used for home application and industrial monitoring devices, the main motto of all this to provide an effective algorithm to the interface of traditional equipment to the webserver [3]. In 2005, AT90S8515 single chip which uses RISC technology and RTL-8019 network interface controller hardware platform comes in the picture which provides a flexible remote access device [4]. In 2006 little advancement in network occurred which optimized for home automation, education laboratories, industrial and instrumentation [5]. In 2007 power quality (PQ) design is introduced by Kaohsiung where Samsung S3C4510 is used as an embedded system that works on the Linux operating system [6]. At same year methodology on telemonitoring ECG was represented by Yanzheng LI using dynamic web server is using for dynamic web page, where software is developed on ARM9 [7]. In 2008 web server is designed through the ARM processor and analyzed the software and hardware configuration and also explains the selection of embedded operating systems [8]. In the same year, Atmega128 was introduced with a web server, which comes with many advantages like small size, low cost, low power consumption and flexible designing [9]. In 2010 EWS achieve by building ARM920T-S3C2410S chip as its core and Linux as its operating system and through CGI gateway dynamic

web page is successfully realized [10]. In 2011 SQLite architecture was introduced which helps for managing past performance data which store in the database uses SQLite.

3 Technical Review

Every EWS architecture has two essentials components: (a) FPGA and (b) Soft core processor.

(a) FPGA

FPGA is defined as a field-programmable gate array, and it belongs to the integrated circuit. FPGA is a variable gate array which can be reconfigured or redesign after manufacturing through FPGA its very relevant to make a complex and large circuit without confusion and errors, the main important characteristics of FPGA is, it gives high speed, high efficiency with re-usable.

(b) Soft core processor

Softcore processor is fabricated by FPGA, ASIC, and CPLD which is reprogrammable, and the speed of softcore processor is less than 200 MHz but disadvantage of the softcore processor was that it runs with less speed which overcomes with hardcore processor which is fabricated with silicon and its speed is 100's of MHz up to 1 GHz of speed. Many architecture techniques are used to manufacture the processor, some are mention below as Based on ARM, AVR, MicroBlaze, MS-51, MIPS, RISC-V, SPARC, X86, and many more others and where MicroBlaze architecture of the 32-bit processor is most popular.

4 Methodology

The strategy of building an embedded web server is depicted in Fig. 1.

1. Software required

Web server means massive data-center, which can store millions of websites, pictures and many more, the strategy of a webserver is to fetch, store and again deliver the data, and this is possible only with the help of HTTP protocol.

In first step, in the strategy of design for EWS, the building of frontend, we can use HTML5, which interfaces database using PHP, but today MEAN STACK is in trend where angular.js is used for frontend designing and MongoDB is using a database (mongo is a database and MongoDB is a server) which is connected by the help of node.js and with express which is called as middleware.

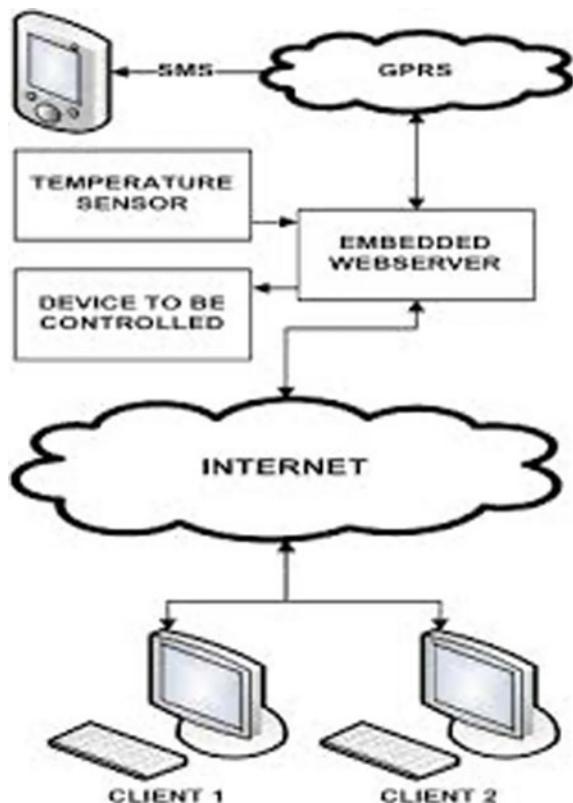


Fig. 1 Strategy of building an embedded web server

2. Hardware required

Raspberry-pi

It is a minicomputer which is a design by Eben Upton in the United Kingdom at the Cambridge University, at that time Eben Upton design this in the application as improving the programming skill of their student. In market Rasberry-pi comes with the fully functional and operating system used is ‘Raspbian’, Rasberry-pi consist of two packs to connect with another device, one is used for input whereas other for output. Figure 2 is a Rasberry-pi.

Arduino Board

At very first an Italian person Massimo Benzi design Arduinio, It is simply a micro-controller board, it does not have any operating system. Arduinio has 20 pins (analog and digital pin) for input and output, the storage capacity of the Arduinio board is 32 kbs. Figure 3 shows the Arduinio board.

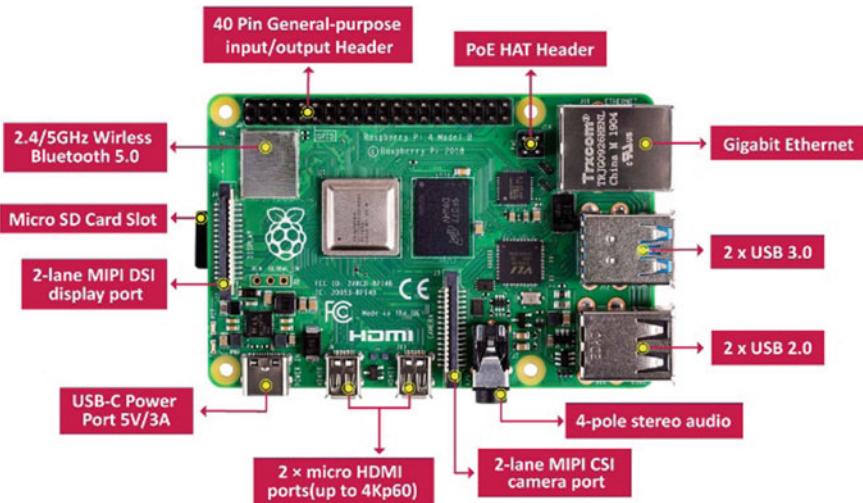


Fig. 2 Essential part of Rasberry-pi board

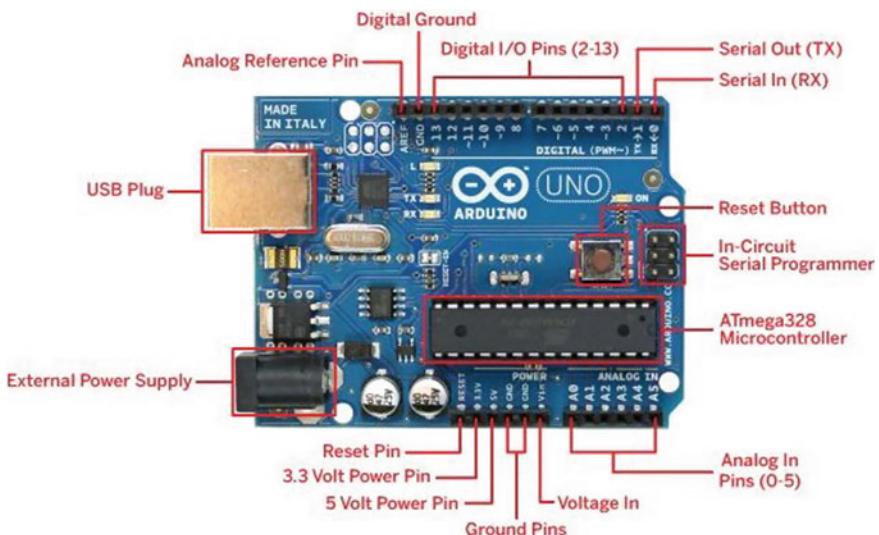


Fig. 3 Essential part of Arduino board

For EWS mainly used open-source electronic platform device is Arduino board which consist of ARM cortex-M3 processor (32-bits), which is connected with local device like controlling sensor, temperature sensor, etc. and pass the data to ESP8266

which is WiFi microchip which is fully under-control of the microprocessor, it works as middleware who is sharing the current updates with web server.

Working

Client can control board component or the peripherals, how much client will instruct that only it is going to perform for example AC, tube light, the fan is controlling if you want to ON AC and tube light then press ON button in front designed page then AC and tube light will start working and updates comes in a dynamic web page.

5 Applications

EWS architecture has been popular in various fields. Here the author tries to explain the utility of EWS in some popular domain.

1. Healthcare

We can use it in hospitals for monitoring the patient body's every parameter like pulse rate, blood pressure, etc. In each movement embedded device sends a parameter to a web server which can be seen by a doctor from anywhere, if any emergency required it can also send a notification to the doctor.

2. Transport and travels

For tracking the current status of a motor vehicle, we can use an embedded web server.

3. Agriculture

EWS can also use in agriculture parameter monitoring the soil purity, salt percentage added there, humidity.

4. Education

Nowadays, distance-learning programming is famous, it can also possible by EWS, for example of some educations applications are live classes, attendance systems, either institute can monitor by sitting in one place either inside, or form distance.

5. Automotive industry

EWS can also give big roles in doing automation, for example, machine running without raw material.

6. Domestic use

EWS also uses for various home parameters like water tank automation, kitchen monitoring, fan or AC can do automatically by sensing temperature, etc.

Table 1 EWS performance evaluation

Performance element	CPU usage	Maximum user connectivity	Code size	Run time memory
Value	15%	64 users	35 kb	6 kb

Table 2 Web server and PC server performance [11]

Item	PC server	Webserver
Size	Large	Small
Power	100 of watt	A few watts
Power supply	High volts	Battery
Consumption cost	1000 of RMD	100 of RMD

6 Results

Embedded web server performance evaluation is done by the help of parameters like CPU usage, maximum user connectivity, code size, run time memory and many more. Table 1 shows the performance analysis of EWS. The CPU usage is 15% for running an embedded web server and rest are for other work, and it supports maximum user connectivity of 64 users more than that it not supports, all its code consists size of 35 kb, and uses 6 kb run-time memory allocation.

Web server and PC server performance have been analyzed based on size, power, power supply source and consumption cost. Comparative analysis has been discussed about Web server and PC server performance in Table 2.

Here we have given an analyzed result which divided into different fields and these are architecture scalability, sensor connectivity, internet connectivity, and microprocessor and microcontroller. Here architecture scalability is measured by the Iota server, web server and TINI, which gets value for the same parameter differently, sensor connectivity data analyzed for 12C, ESBus, 1-wire. Comparative analysis has been discussed about web server enabling architecture in Table 3 [7].

7 Conclusion and Future Scope

The main motto of implementation of the embedded web server is creating a virtual world, through EWS we can create a virtual environment in many fields, like virtual laboratories, monitoring the property from a distance, etc. Embedded web server is also called as embedded web technology. Nowadays it gradually extended in the personal computer market due to high efficiency, speed, and reliability. Embedded web server is the central building block of microprocessor and microcontroller are integrated with memory are used for observation and controllability, the advantage of EWS is web browser does not applet permanently, it's user interface software

Table 3 Comparison of web server enabling architecture [7]

Parameter	Architecture		
	IoT a Server	Web sensor	TINI
<i>Architecture scalability</i>			
Minimum number of intelligent node types	1	2	1
Scalability by number of sensors	Yes	limited	Yes
<i>Sensor connectivity</i>			
Type	12C	ES-bus	1-wire
Network concept	Multiple masters, multiple slaves	Single master, multiple slaves	Single master, multiple slaves
Number of signal lines	2	4	1
Max nodes	>100	9	30
The data rate, bps	100 k nom 3.4 M max	10 k	16.3 k nom 142 max
<i>Internet connectivity</i>			
Type	GSM	Ethernet	Ethernet
Data, rate, bps	<85.6 K	10 M	10 M
Conversation algorithm encoding	JavaScript, VB-Script, Java Applet	Embedded application	Java servlet
User interface	Text graphical	Text	Text graphical
Presentation layer	ESP	ASCII	Dynamic HTML
<i>Microprocessor/Microprocessor</i>			
Min CPU count	1	>=2	1
CPU/ALU, bit	8/16	8/8	8/32
ROM size	32 kb	8 kb	512 kb
RAM size	2 kb	368 bytes	512 bytes

upgrade and another provides output to the remote user in any format. According to the previous data problem is analyzed and use its intelligence and work on it, think what is next, for example at moderate temperature automatically switch on the fan and after certain higher temperature AC will switch ON if any troubleshoot is happening then fix by its own. Embedded web server takes us to complete automation, which reduces human effort and increases employability and human being become interested in working in their specialist field.

References

1. Choi MJ, Ju HT, Cha HJ, Kim SH, Hong JWK (2000) Efficient embedded web server for web-based network element management. In: IEEE symposium record on network operations and management symposium, pp 187–200
2. Ju HT, Choi MJ, Hong JWK (2001) EWS-based management application interface and integration mechanisms for web-based element management. *J Netw Syst Manag* 9(1):31–50
3. Ju HT, Choi MJ, Han S, Oh Y, Yoon JH, Lee H, Hong JW (2002) An embedded web server architecture for XML-based network management. In: IEEE symposium record on network operations and management symposium, pp 5–18
4. Mylvaganam S, Waerstad H, Cortvriendt L (2003) From sensor to web using PLC with embedded web server for remote monitoring of processes. *Proc IEEE Sens* 2(2):966–969
5. Lin T, Zhao H, Wang J, Han G, Wang J (2004) An embedded web server for equipments. In: Proceedings of the international symposium on parallel architectures, algorithms and networks, I-SPAN, pp 345–350
6. Fu C, Zhu Z, Gao X, Wang P (2005) The design and implementation of a general reduced TCP/IP protocol stack for embedded web server
7. Klimchynski I (2006) Extensible embedded web server architecture for internet-based data acquisition and control. *IEEE Sens J* 6(3):804–811
8. Chang WF, Wu YC, Chiu CW (2006) Development of a web-based remote load supervision and control system. *Int J Electr Power Energ Syst* 28(6):401–407
9. Li Y, Wu S, Li J, Bai Y (2007) The ECG tele-monitor based on embedded web server. In: 2007 1st international conference on bioinformatics and biomedical engineering, ICBBE, pp 752–755
10. Chen TH, Huang JX (2007) Design and realization of CGI in embedded dynamic web technology. In: Proceedings—2007 IFIP international conference on network and parallel computing workshops, NPC 2007, p 774–777
11. Yang J, Xie Y, Chen T (2009) Research on web server application on multi-core embedded system. In: Proceedings—2009 international conference on embedded software and systems, ICRESS 2009, pp 412–416

Water Sharing Marketplace Using IoT



Dendukuri Ravi Kiran, Aki Rohith, Kothur Dinesh Reddy,
and G Pradeep Reddy

Abstract The objective of the proposed system is to eradicate the conventional water management system and create a market place where any house can buy and sell water in their neighbourhood. Water sources like tankers and public sector water have impurities, come at irregular time intervals and tankers surge their prices based on the demand and season. A marketplace within communities would create competition; regulations can help in eradicating this informal sector and create a fair market with water of better quality and price. Machine learning algorithms are used to give the buyer and sellers the predicted value of how much water each household uses on a daily or monthly basis to make it easier for the user for purchasing and selling of water.

Keywords Water sharing · IoT · Machine learning

1 Introduction

As per the recent World Bank Report, on average an Indian household consumes 80 Liters of water every day. In India, 85% of its water supply is dependent on groundwater and remaining on rivers and other water bodies. Groundwater is the water that flows beneath the Earth's surface filling the porous spaces in the soil and

D. R. Kiran (✉) · A. Rohith · K. D. Reddy · G. P. Reddy

Department of ECE, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad,
India

e-mail: davidhume3690@gmail.com

A. Rohith

e-mail: aki.rohithgupta007@gmail.com

K. D. Reddy

e-mail: kothurdineshreddy@gmail.com

G. P. Reddy

e-mail: pradeepreddygriet@gmail.com

rocks. Groundwater is mainly accumulated due to rains [1]. This water is extracted by drilling a well into the aquifer, a water pump is then used to escalate the water into the households. India is now the world's largest user of groundwater, withdrawing about $250 \text{ km}^3/\text{year}$. If the current rate of groundwater depletion persists, India can solely have 22% of the current daily per capita water offered in 2050, probably forcing the country to import its water. Mainly in urban areas, water bodies like lakes and rivers are also used as water sources. These water bodies depend on rain and atmospheric temperature [2], due to global warming and rapid climatic changes these water bodies are drying up. This was seen in Chennai in the year 2019 when the city's reservoirs and lakes were parched, and its wells had dried out due to two years of scanty rain which led 9 million individuals starving for water.

Alternate methods that could solve this issue are rainwater harvesting and sewage water treatment. Rain water harvesting is a technique that is used to collect and store rainwater in subsurface of aquifers; it is the best way of creating an artificial water storage facility without disturbing nature. Sewage water treatment is the method of removing contaminants from municipal wastewater, containing the main ménage and industrial waste material [3]. The main reason these methods are not adopted in mass is that these methods have capital and storage limitations. These methods can be implemented in bulk if they had a way of generating income and recover their investment but as the water selling industry is an informal sector in India, an average Indian household cannot afford to install its own rainwater harvesting or sewage water treatment unit.

To overcome such a complication an IoT based water management system is proposed, where the system consists of a group of houses with interconnected water pipe linings, where each house's water supply pipe is connected to a mainline which ultimately connects to the main water tank. The water tank acts as a medium where every water transaction takes place. The pipelines contain the solenoid valves, flow sensors and water pump motors which ensures the transfer of water from one location to another. A mobile application is used as a medium where one individual can find all the buyers and sellers in his/her community. When a transaction of water takes place between two individuals, the water is transmitted from sender to receiver. The entire system is controlled using an Arduino and a Raspberry Pi.

2 Proposed Method

Development of IoT based water management system is designed, where the individual household can request water from another house in the same community. In this method, the houses with excess water can sell theirs to the houses that are in need. The main tank is present which acts as a medium of water flow on every transaction; this avoids unnecessary congestion and allows smoother transition.

Water level in the individual's storage tank and main tank is monitored using an ultrasonic sensor [4]. Based on the requested amount of water by the buyer, water is sent from the seller comparing the amount of water present in his storage and main

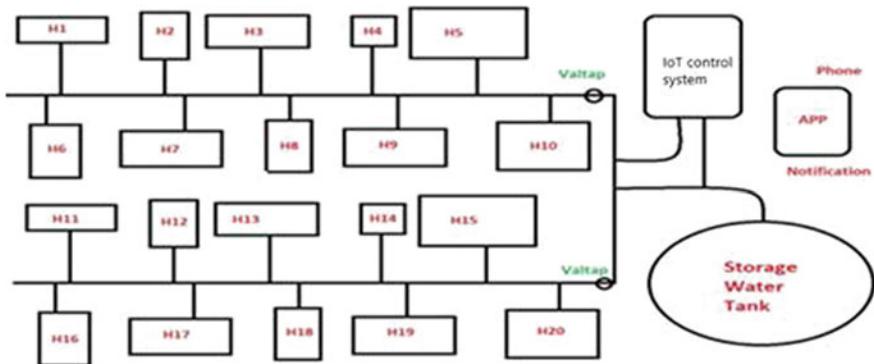


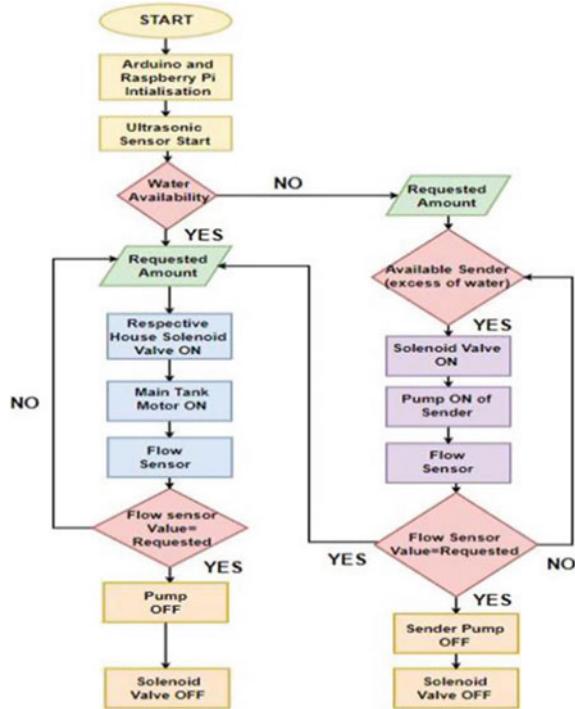
Fig. 1 Outline of the proposed method

tank. A mobile application contains a dashboard of all the information on potential buyers and sellers.

Figure 1 represents the schematic view of the implemented community and water management system. H_1 to H_{20} represents the houses in the community, a water storage tank is used as a medium. The main water tank is used to reduce any disproportionate water transfer or congestion in the community. Pipeline of every house consists of a flow sensor and a solenoid valve. IoT control room is a virtual hub where all the requests and response actions are recorded, implemented and monitored [5].

Figure 2 describes the flow of water in the water management system [6]. If any householder in the community requests water (this can be done through the app) then request information is received at the controller. The community's main water tank acts as a water bank where potential sellers can store their water in case of storage limitations or autonomous transactions, the tank has an ultrasonic sensor which is used to sense the level of water present in it. If the requested amount of water (by the receiver) is present in the main tank, then the respective solenoid valve of the receiver's house gets on and the motor or pump of the main water tank pumps the water from the main tank to the respective house. The flow sensor present in the receiver's house will calculate the amount of water flowed to the house. If the requested amount becomes equal to the water flowed, then the pump of the main tank gets off and the solenoid valve of the receiver's house gets off. Finally, the money will be credited to the seller who stored his water in the main tank. If the requested amount of water is not present in the main water tank, then the controller requests water from the entire houses present in the community in the form of message or notification through an application [7]. If any household has an excess amount of water, then their water is sent to the main tank. This is done as follows; first the solenoid valve of the sender's house and the main tank gets on. The pump of the respective house pumps water from the house to the tank. The flow sensor calculates the amount of water flowing. If the requested amount reaches, then the motor of the respective house gets off and the solenoid valve also gets off. After the amount is received at the main tank it is sent to the house that requested water the main tank

Fig. 2 Flowchart describing the implementation



and the money is credited to the sender. The process takes place as the transactions mentioned above.

The water management system consists of various stages which include sensors, actuators, communication units, and design of database.

2.1 Sensor Details

2.1.1 Ultrasonic Sensor

Ultrasonic Sensors measure the distance to the target by measuring the time between the emission and reception. When mounted vertically on the water tanks, the sensor provides data on the amount of water level present in the tank, which can be used to find the daily water usage of an individual house or in a household. Figure 3, the sensor provides data on the amount of water used by a selected household in a day by obtaining the difference between the initial reading of the sensor at the day and the final reading at night.

Fig. 3 Water usage in the selected house



2.1.2 Water Flow Sensor

The Flow sensor works on the hall-effect which outputs the corresponding pulse signal. The sensor provides data on how much water a seller sends, and a buyer receives [8].

2.2 Actuator Details

2.2.1 Water Pump Motor

A water pump is a mechanical device that changes the energy flow from electrical to mechanical. The actuator is used to pump the water from the sender to the receiver [9].

2.2.2 Solenoid Valve

Solenoid valve is a control unit which, when electrically energized or de-energized, either shut off or allow fluid flow. The actuator is used to guide the water flow in the right direction to the receiver.

2.3 Mobile Application

Users can find the analytics of daily water consumption in their household based on their water tank data. Figure 4, the application is developed in MIT app inventor

Fig. 4 Mobile application

platform and using Firebase as the cloud storage platform. The user when desires to buy water can request the app which then sends notifications and messages to all the potential sellers in the community. The first responding seller is the recipient of the transaction and sends his water to the buyer based on the requested amount. The application also displays the predicted usage of water per house based on an algorithm. The application records the monetary transactions and banking statements using UPI portal.

2.4 UPI Portal

The Unified payment interface is created using easy UPI payment API. This library arranges installed UPI applications on the user's mobile and when the consumer wishes to pay with the application, it connects with the application via deep-linking and completes the payment procedure and returns to the application with the transaction details [10].

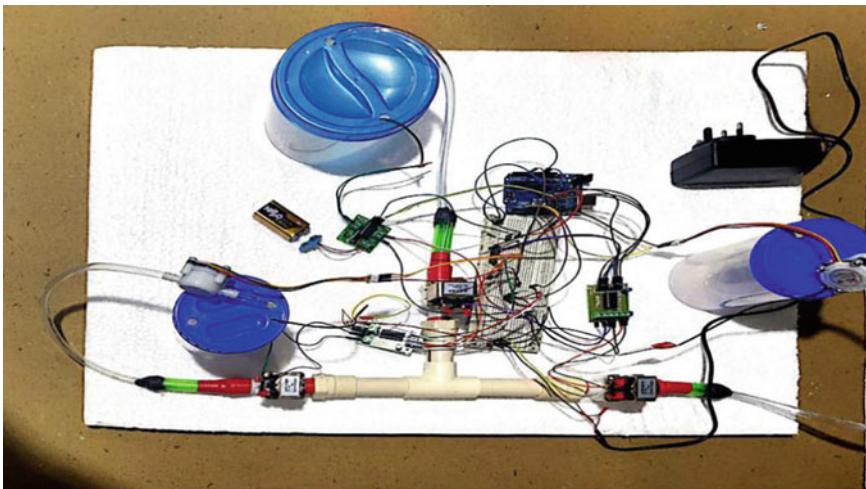


Fig. 5 Prototype of the implemented system

2.5 Machine Learning Algorithm

Linear regression is used to estimate real values based on continuous variables. Ultrasonic sensor mounted on the water tanks provides variable data with water consumption [11] of a user on a daily basis. The day at which the consumption was recorded, the number of people in the house at the respective reading and the season of the year are also considered as independent variables. These variables are used to find a dependent variable which is the predicted water usage for the user. In the proposed method 20 houses respective variable data was taken in the period of one year for training and the predicted expression was generated.

In Fig. 5, prototype is provided with the connection of all the sensors and actuators mentioned above to the respective storage tanks of the houses. The small tumblers in the figure describe the storage water tanks of the houses and the big tumbler describes the main storage tank of the system.

3 Result

The system is thereby comprised of a water managing marketplace that assists users with their daily water usage and predicts their future expenditure and allows them to sell and buy water within their community. Research on a community with 20 houses was done and a report is generated with the cost incurred for purchasing water when they used the platform to buy excess or privately generated water using rainwater harvesting or sewage water treatment within the community versus buying it traditionally from water tankers.

The chart for plotting differences of various water sources and the proposed system in terms of price is shown in Fig. 6.

Figure 6 is the comparison of cost readings of water consumed in 20 houses within the proposed system taken every month for a period of a year and compared it with another 20 houses where tankers and private players are their main water source. The graph shows the decrease in the price of water when the proposed system is used. Figure 7, the data is being made available to the user in the form of notifications and

Fig. 6 Cost comparison of selected scenario

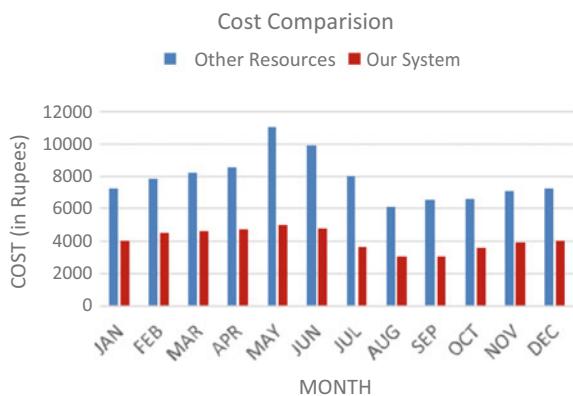
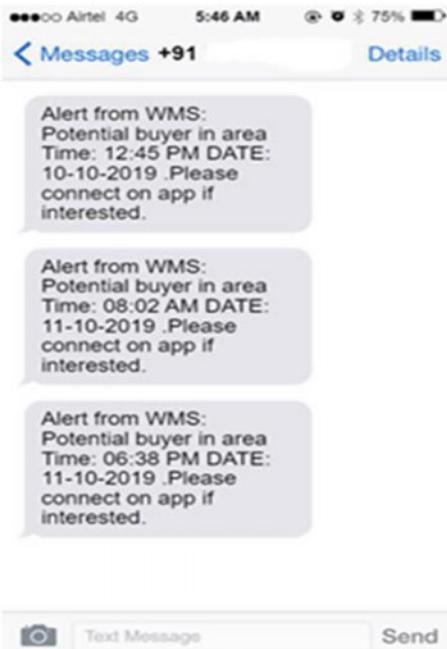


Fig. 7 Notifications received by the user



application information using IoT Network.

4 Conclusion

A precautionary system avoids or eliminates the issue of overpricing water by water tankers during water scarcity times which is being a major problem in many big cities. It also acts as a platform for doing business with water at reasonable prices. The proposed system also encourages people for implementing rainwater harvesting and sewage water treatment as they can sell the water that they refined or retreated. This results in an increase in levels of ground water which is very much needed in the present world.

References

1. Kamyotra JS, Bhardwaj RM (2011) Municipal wastewater management in India. IDFC report, Chapter 20
2. Olagunju A, Thondhlana G, Chilima JS, Sène-Harper A, Compaoré WRN, Ohiozebau E (2019) Water governance research in Africa: progress, challenges and an agenda for research and action. Water Int
3. Bhatt J, Patoliya J (2016) IoT based water quality monitoring system. Int J Ind Electron Electr Eng
4. Mishra G (2017) Ultrasonic ranging and detecting using Arduino and processing. Int J Sci Res Comput Sci Eng Inform Technol 2(3). ISSN 2456-3307
5. Robles RA, Martin D, Navarro M, Calero R, Iglesias S, Manuel L (2015) An IoT based reference architecture for smart water management processes. J Wirel Mob Netw Ubiquitous Comput Dependable Appl
6. Ebere EV, Oladipo FO (2013) Microcontroller based automated water level control system. Int J Innov Res Comput Commun Eng 1(6):1390–1396
7. Gowthamy J, Reddy CR (2018) Smart water monitoring system using IoT. Int Res J Eng Technol 5(10)
8. Amin RSS, Annaswamy A, Moura S, Bulusu C (2015) Smart cities and control. IEEE Control Syst Mag
9. Ormsbee LE, Lansey KE (1994) Optimal control of water supply pumping systems. J Water Resour Plann Manag 120(2)
10. Reza SMK, Tariq SAMd, Reza SMM (2010) Microcontroller based automated water level sensing and controlling: design and implementation issue. In: Proceedings of the world congress on engineering and computer science
11. Tomas O (2013) Evaluating machine learning for predicting next-day hot water production of a heat pump. In: 4th international conference on power engineering, energy and electrical drivers

Augmenting the Existing CBPM Maintenance Philosophy Currently in Use in the Marine Sector with Intelligent Predictive—CBPM Philosophy



Minakshi Gautam, Vaishnavi S. Ramu, Sachin Sinha, Pranay Kumar Reddy, Monica Kondur, and S. Suresh Kumar

Abstract The maintenance philosophy of equipment/system used in maritime environment which is highly corrosive follows CBPM (Condition-Based Predictive Maintenance) philosophy, where the maintenance of any equipment/system is based on the existing running condition of the same. Based on the running condition of the equipment/system, they are subjected to maintenance so as to have enhanced exploitation of the equipment/system. Among the various maintenance philosophy of equipment/system in our ecosystem, predictive maintenance has become a widely used term and the same is finding deep roots in maritime applications. The concept of predictive maintenance is being constantly updated by engineers and researchers based on monitoring historical data, modelling, simulation, and failure probabilities to predict fault and system deterioration over their useful life. Leveraging on various existing data science techniques and thereby sensor adaptation(s), the current paper proposes real-time predictive analytical techniques using R/PYTHON/JULIA and its associated libraries, such that the source of defect is localized thereby resulting in enhanced and better exploitation of the equipment/system with minimal downtime. The authors et al. are of the opinion that the real-time (existing) CBPM techniques that are currently being followed in maritime environment are time-consuming and require enhanced level of human monitoring and intervention for enhanced exploitation of the same. The authors et al. have also designed a customized self-healing RS (Recommender System) that liquidates majority of problems onboard a marine vessel utilizing advanced ML concepts and the same also recommending to the watchkeeper/EOW/Bridge in audio/video mode of possible upcoming defects on any maritime equipment/system.

Keywords Maritime · CBPM · Analytics · Predictive

M. Gautam · V. S. Ramu · S. Sinha (✉) · P. K. Reddy · M. Kondur · S. S. Kumar
CSE, Presidency University, Bangalore, India
e-mail: sachinsinha5695@gmail.com

1 Introduction

Marine operations worldwide require that the equipment/systems be exploited on a 24*7*365 basis so as to cater to the prolonged deployment of the marine vessel. Thus, in order to keep a marine vessel to float and to move, it is required that regular and periodic maintenance of the equipment/system be undertaken periodically. Majority of marine vessels (both merchant navy and warships) have aging equipment/systems and thus the maintenance costs associated with these turn out to be higher. The high maintenance costs require the maintainer to define clearly the maintenance objectives, with the element of indigenization added to the same and the same also replying as far as possible on human intervention [1].

There are three kinds of maintenance philosophies currently in vogue in the marine sector—namely Breakdown Maintenance (BM), Planned Preventive Maintenance (PPM) and Condition-Based Predictive Maintenance (CBPM). Due to the various merits and de-merits of these maintenance philosophies, almost the entire marine sector resorts to the CBPM philosophy which is like a well-oiled machinery, i.e. the same has been time tested and proved.

In the maritime sector, failure of any equipment/system is attributed to the improper maintenance being undertaken and the same can result in any kind of a catastrophe which may also lead to loss of human life. The maintenance decisions of any marine vessel are taken by the central administrators (Operation team in case of Merchant vessels and the Planning department of the Ship Repair Yard in case of Warships). Such maintenance decisions are taken by experienced planners according to the maintenance manual provided by the OEM (Original Equipment Manufacturer), the reported breakdown history or failure data, and the operating experience and wisdom of the maintenance staff and technicians. Under any situation of uncertainty (which is a common scenario in marine sector), it becomes very difficult to plan the maintenance activities in advance, as the same depends on the operational cycle of the marine vessel and also other important operational commitments of the marine vessel. All the above referred herein lead to the maintenance organization (Ship Repair Yard) to undertake the maintenance of the marine vessel in a “Fire Fighting” mode.

2 Condition-Based Predictive Maintenance

Condition-Based Predictive Maintenance (CBPM) is a maintenance philosophy used to reduce the uncertainty of maintenance activities and is carried out according to the need indicated by the running condition of the equipment/system. The key feature of CBPM is that it can detect and quantify possible failures of equipment/system before it actually occurring. The ‘key features’ implicate those running parameters that have a direct or indirect bearing on the equipment health.

The common problems faced by any marine equipment/system is ageing, corrosion and deterioration. These causes are due to the operation of the equipment/system in a very hostile corrosive environment and the same leading to early ageing and deterioration of the equipment/system, all the way right from the foundation of the equipment through the base plate to the superstructure of the equipment/system. There are a variety of methods/devices that are used to measure the equipment health and thus the EHM (Engine Health Monitoring) techniques play a very important role in the effective exploitation of the equipment. The trend of deterioration of any equipment/system is measured periodically in a marine environment and necessary checks and balances for the same are also done, viz ICCP (Impressed Current Cathodic Protection), Sacrificial Anodes etc [2]. The final maintenance decisions depend on the actual measured abnormalities, the last running condition and availability of spares with the depot.

Keeping in mind the hostile operating environment of any marine equipment/system, a proper maintenance plan is drawn out by the Command and Control authority with the maintainers (Ship Repair Yard). Both the above also keep in mind the availability of spares which is forecast based ion the maintenance activity and an extension of RH (Running Hours) is given to the marine vessel to exploit the equipment/system beyond the stipulated RH before the equipment/system enters into repair phase.

Elaborate study was undertaken by the authors et al. reveal that the existing practices of maintenance of marine vessels are planned refits of marine vessels which invariably slip due to various factors such as availability of dock space, spares, etc.

3 The Need for Augmenting the Existing CBPM Maintenance Philosophy

Ship Owners have in-depth knowledge in the usage of decision support system (DSS) which is a computerized information system which contains domain-specific knowledge and analytical decision models to assist the decision-maker by presenting information and the interpretation of various alternatives.

Although the existing DSS is literally a foolproof system in place, the authors et al. with their profound experience are of the view that the same could be replaced with the state of art Decision System using PowerBI deploying Bokeh Plots. Recommender System (RS) augmented with the same is also proposed as the same will serve to recommend the maintainer/decision-maker on the possible options of maintenance of the equipment/system. This shall provide the decision-maker in taking intelligent and SMART decisions. Intelligence in the said tool of PowerBI shall also be accelerated using the components of Bayesian theory, fuzzy networks, etc.

After reviewing the current maintenance management practices, we find out that condition-based fault diagnosis and the prediction of the equipment deterioration trend are vital in maintenance management approaches as the same has a direct

bearing on the life of individuals' present onboard a marine vessel. The authors et al. opine that a self-healing technique on intelligent condition-based fault diagnosis and Cloud-Based Recommender System needs to be integrated with the current EHM techniques for better Engine Health and enhanced operational efficiency of marine vessels as the same is the need of the hour.

4 Intelligent Condition-Based Predictive Maintenance System with Self-healing Features Using Predictive Analytics

The Intelligent Condition-Based Predictive Maintenance System with Self-healing Features augmented using Predictive Analytics shall integrate the following concepts:

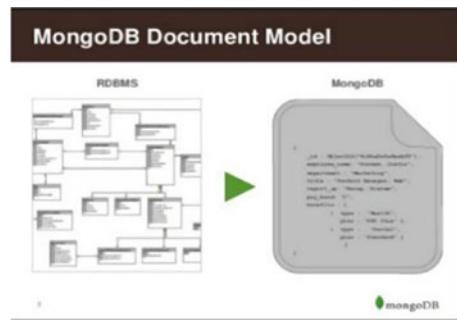
- (a) Unmanned real-time equipment monitoring
- (b) Self-healing based on equipment health
- (c) Predictive analytics solution for monitoring and predicting equipment health
- (d) Cloud-based Recommender System (RS)

A. *Unmanned real-time equipment monitoring*

Real-time equipment monitoring is defined as the collection and interpretation of the relevant equipment parameters for the purpose of the identification of the state of equipment changes from normal conditions and trends of the health of the equipment. Various Equipment parameters such as S&V (Shock and Vibrations) readings, temperatures, pressure remain stable (normal) as long as the equipment is in good health. These parameters indicate abnormality when the running health of the equipment degrades. Information on the potential problem can be easily obtained from monitoring devices fitted with the equipment/system that forms a part and parcel of the equipment/system. Efforts are underway by our Indian Navy to install IPMS (Integrated Platform Management System) onboard all Ships, wherein all marine vessels shall be networked using OFC backbone. The same will serve the purpose of local as well as remote monitoring of health of the equipment/system.

Information captured by the above leads the maintainer to undertake maintenance of the marine vessel. Simulation done on unmanned exploitation of the marine equipment/system has yielded 95% results. It is thus proposed by the authors et al. to introduce unmanned real-time monitoring of equipment/systems. The real-time monitoring shall be done by sensors and gauges fitted on the equipment/system. Such data captured by the sensors/probes/gauges will in turn be communicated to the cloud wherein this data will be stored in a NoSQL database, viz MongoDB. Data from this cloud database will be retrieved by the user/maintainer/decision-maker via JULIA and real-time plots of the same will be generated using BOKEH which will serve as a focal point to the decision-maker/maintainer to take sound decision regarding maintenance of the equipment/system (Fig. 1).

Fig. 1 MongoDB document model



B. *Self-healing based on equipment running health*

Since this issue is in its infancy and under extensive trials, the authors have considered various cases prevalent in marine sector viz, Low L.O Level in a sump, High exhaust temperature of DA, Low Load running of Engines and such other cases to prove the concept of self-healing systems. In cases as mentioned herein, self-healing concepts such as refilling to the desired level of L.O sump, forced jacket cooling of an exhaust manifold, etc. have proved to be successful. The efficiency of the same has been tested using simulations and found to be satisfactory.

C. *Predictive analytics solution for monitoring and predicting equipment health*

Data Science solutions have proved to be unmatched in predictive analytical solutions wherein predictive analytical solutions to 100% accuracy have been found to be achieved within industry. Leveraging on the same, it is proposed to predict the equipment health considering all inputs such as L.O, F.O temperature/pressure and such and various other parameters that will aid in predicting the equipment health. The various steps that are followed is ad-seriatim:

- (i) Data Collection
- (ii) Statistical modelling of data
- (iii) Optimization using Machine Learning algorithms
- (iv) Training and testing the model
- (v) Data Visualization.

Sample studies have been undertaken by the team on providing predictive analytics solutions for monitoring and predicting equipment health and the same has been found to be highly successful in the limited studies carried out.

D. *Cloud-based Recommender System (RS)*

Maintenance of marine equipment/systems just by CBPM techniques is felt is not a complete solution to address the complete and correct maintenance of the equipment/system. A cloud-based RS designed by the team is proposed for the same by the authors et al., which shall reside in the private cloud of the shipowner and input

of data will be via the public cloud [3]. Thus the hybrid cloud approach will serve the complete purpose, both even for merchant navy ship owners as well for defense applications.

A snippet of code for the same is as under.

```
corr_contact = pd.DataFrame(similar_to_contact, columns=['Correlation'])
corr_contact.dropna(inplace=True)
corr_contact.head()#corr_AFO = pd.DataFrame(similar_to_air_force_one, columns=['correlation'])
corr_AFO.dropna(inplace=True)
corr_AFO.head()
```

5 Smart Maintenance

This concept of SMART maintenance proposed by the authors et al. is on undertaking maintenance ‘on-demand’. Herein the entire data, viz the detailed maintenance routines on any equipment/system and all its associated sub components will be an IoT-based solution wherein the power of Internet will be leveraged to solve the maintenance issues of marine vessels.

The authors et al. are of the opinion that unplanned maintenance can be disastrous in the low-margin shipping business, particularly for small fleet owners whose vessels haul cargo or nudge large transport ships through small spaces. This shall also be beneficial for tight schedules of marine business, wherein if a tugboat is under unplanned maintenance which really is not needed, then the vessel owner loses business which has a disastrous effect on the company’s exchequer.

This concept of IoT enabled maintenance proposed herein steps here wherein the same shall have sensors embedded or attached to onboard equipment/system. The bespoke software designed by the authors et al. using PYTHON to augment this IoT device, shall listen for potential malfunctions, such as erratic operations, abnormal L.O, F.O pressure/temperatures, etc. This software acquires, cleans and forms a data frame of the data at rate of 50,000 times per second in majority of instances. This software identifies patterns of failure before they become potential failures thereby disrupting the operation of the equipment. For example, the L.O temperature despite going beyond safe operating levels may not trigger an alarm if the L.O pressure is within a safe operating regime. Our software detects this anomaly and alerts the watchkeeper and maintainer and warns him of a potential disruption in the operation of the equipment.

Analyzing the real-time marine data of a sea-going vessel, the authors et al. are of the opinion that this could result in huge savings. As a test case, a faulty (as reported by the ship maintainer and owner) single-stage reciprocating fuel pump was taken for the purpose of analysis, which otherwise would have resulted in the marine vessel being under repairs for at least 2 months. IoT enabled analysis on this pump whilst in operation would have resulted in saving of time and money for the shipowner.

6 Probing Deeper

There is a still higher payoff when data of such kind is analyzed and combined to uncover failure trends that would otherwise result in docking of the marine vessel. OLAP cubes are preferred herein that shall map data from different sensors on ship in a way that the user can create analyzers and visualizations to benefit in higher downtime of the marine vessel. The authors et al. propose that analyzing data using the right tools and techniques as mentioned in Para 4.3 herein would result in saving time and money to the shipowners (Merchant Navy and Defense).

We are aware that small savings in terms of smaller efficiency improvements can add up to big rupee savings across many vessels. Three issues that fleet owners are generally concerned about were taken up for the purpose of analysis: fuel efficiency, unscheduled downtime and environmental compliance [4, 5].

For example, some cargo boats carry up to five onboard generators to power things like refrigerators and compressors. Each of those generators may output different amounts of power so that they can be mixed and matched according to demand. You'd think operating fewer generators would use less fuel, but it turned out that isn't always the case.

Multidimensional analytics was done by the authors et al. on the three areas as mentioned herein. The case of a deterioration of ship hull is considered herein. Ships speed and efficiency are reduced by external hull growths such as barnacles and seaweeds. This results in excess drag in the ship and thus poorer fuel economy. The authors et al. observed that the marine vessels (Merchant Navy and defense) do cleaning of hull of ship by jet blasting and manual scrapping whenever it is subjected to scheduled maintenance that involves dry-docking of the ship. However, it was observed that no calculation of real-time deterioration of ship hull was being done by the ship owners. Overcoming this problem using analytical solutions resulted in savings of Rs 70 Million INR.

The real-time plot of actual versus the regression line using BOKEH of a marine vessel based out of Kochi is as shown in Fig. 2 [6].

7 Conclusions

In this paper the authors et al. have proposed ways by which the existing CBPM maintenance philosophy can be augmented with the Intelligent Predictive CBPM Maintenance Philosophy. Post extensive field study, our results show that the proposed Intelligent Predictive CBPM Maintenance Philosophy shall have positive results to all the sea-going vessels in the long run.

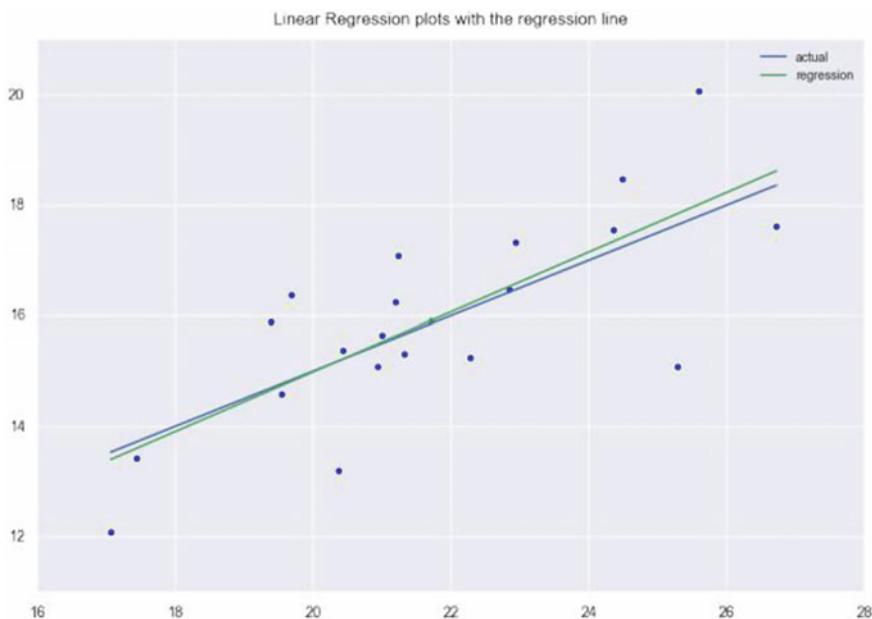


Fig. 2 Real-time plot of actual versus regression line of a marine vessel based out of Kochi

References

1. Kolumban G, Kennedy MP, Kis G, Jako Z (1998) FM-DCSK: a novel method for chaotic communications. In: Proceedings of the 1998 IEEE International Symposium on Circuits and Systems, ISCAS'98, vol 4, pp 477–480
2. Sushchik M, Rulkov N, Larson L, Tsimring L, Abarbanel H, Yao K, et al (2000) Chaotic pulse position modulation: a robust method of communicating with chaos. *IEEE Commun Lett* 4(4):128–30
3. Kennedy MP, Kolumban G, Kis G, Jako Z (2000) Performance evaluation of FM-DCSK modulation in multipath environments. *IEEE Trans Circ Syst I Fund Theor Appl* 47(12):1702–1711
4. Rulkov NF, Sushchik MM, Tsimring LS, Volkovskii AR (2001) Digital communication using chaotic pulse position modulation. *IEEE Trans Circuit Syst I Fund Theor Appl* 48(12):1436–1444
5. Proakis JG (1989) Digital communications. McGraw-Hill, New York
6. Pahlavan K, Levesque AH (1995) Wireless information networks. Wiley, New York

An Investigation on Rolling Element Bearing Fault and Real-Time Spectrum Analysis by Using Short-Time Fourier Transform



M. Siva Santhoshi, K. Sharath Babu, Sanjeev Kumar, and Durgesh Nandan

Abstract Feature extraction has more importance in fault diagnosis and also to identify the important changes of rotary machines. Rolling elements are an important part of a rotary machine. The working condition of the rotary machine is based on the performance of rolling elements. Rolling element produces the fault vibration signals which are non-stationary so time-frequency distribution (TFD) is used. And time-frequency distribution is depending on Short Time Fourier Transform (STFT). This paper combines the concept of TFD and STFT. This paper also presents the different approaches of the Short-Time Fourier Transform. Another thing discussed in this paper is the real-time spectrum analysis of discrete short-time Fourier Transform. This paper is a simple analysis of the rolling element fault diagnosis problem of a rolling element with the use of TFD and STFT.

Keywords STFT · TFD · DFT · FFT · Rolling element bearing · Fault vibrations · RTSA · RBW · POI

1 Introduction

Rolling element bearing is used for low friction and it is also called a roller bearing. Roller bearings are in a different number of shapes and sizes. It is a crucial part of a rotary machine. By using their two bearing rings it can carry their rolling elements by placing them between their two rings [1, 2]. This work is related to the rotary

M. S. Santhoshi

Department of Electronics and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India

e-mail: mutyalasivasanthoshi99@gmail.com

K. Sharath Babu · S. Kumar · D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd, CL Educate Ltd., New Delhi, India
e-mail: durgeshnandano51@gmail.com

S. Kumar

e-mail: sanjeev.kumar@accendere.co.in

machine working condition, so it is very important to identify the faults that occurred, to avoid the breakdowns of the rotary machine. The main and widely used method to avoid faults is vibration-based monitoring [3]. Here fault vibration signals are non-stationary, so here the traditional diagnosis techniques are applied on these waveforms in frequency or time domain, and then to identify the working condition of rolling-element make the criterion function. Because of rollers, balls and loads, there is a lot of impact on vibration signals, so it is very difficult to evaluate the working condition of the rolling element in frequency or time domain. So, the time-frequency distribution (TFD) can be used, it will denote the components of the vibration signal by using the combination of both time and frequency domain in 2D [4]. Because of the energy distribution capacity of TFD, it is preferred. Here TFD uses the STFT [1, 4–6].

Short-time Fourier transform is a linear transform that offers a back and forth between spectral resolution and temporal resolution. Here spectral resolution is the electromagnetic spectrum region width and it can sort out the spectral features and bands into their different components and temporal resolution is nothing but how much time required to make a revisit and obtain data for the same location [5, 6]. STFT used in spectrum analysis. In digital signal processing, the method used for spectral analysis is the discrete Fourier transform (DFT). DFT is developed by using a fast Fourier transform (FFT), DFT is calculated by using FFT. FFT can be used for the finite-length signal collection of stationary sinusoidal components [5, 7, 8].

For a 100% probability of intercept (POI) in a highly dense environment Real-time spectrum analysis (RTSA) is used due to its advantage of overlapping Fast Fourier Transform (FFT) and high speed of memory [9].

2 Literature Review

Fault vibration signals which are obtained from rolling elements are non-stationary because for this reason TFD is used to characterize the local information of these fault Vibration signals [1, 2]. The periodic impulses of roller bearings represent the reasons for faults occurring in components. These impulses are noisy and weak so it is complex to detect them [1, 10, 11]. Here FFT based methods are not used because by using FFT, it is not possible to get the existing information of fault vibration signals [2–4]. So we go for other methods and in DFT (discrete Fourier transform) we go for FFT because it is calculated by using FFT [7]. Time-frequency analysis is the type of method which is mostly used for the signals which are lasting for a short time that is known as transient signals (any type of sudden change in signals is known as transient) [3, 12]. By making the average of those signal events in frequency domain vibration analysis methods like the power spectrum, the transient events do not appear clearly in spectrum lines [3]. But in time domain methods which are also applicable to transient signals, it has a chance of losing the different machine components frequency information [3, 4, 13]. The real-time feedback and EEG (electroencephalogram) use the STFT and coherence analysis. By using these methods EEG'S real-time frequencies are obtained [14, 15]. The wavelet Transform

and Wigner-viler distribution is applied to biomedical signals to detect both spectral and temporal characters of the biomedical signals [4, 13, 14]. To get real-time in broadband digital receivers, mono bit implementation of FFT has been done. And to increase the detection capacity of non-stationary signals, time-frequency analysis (TFA) uses the STFT and also implemented on the FPGA platform [16–18]. STFT is related to the Fourier transform. It is used to identify the phase and frequency of a signal which changes over time. The method for operating STFT has divided the longer time signals into shorter of the same length and then apply Fourier transform on the individual segment and for real-time spectrum analysis [4, 7, 9, 19]. The disadvantages of STFT are overcome by new architectures like feed-forward STFT. The heart rate variability signals which are extracted from ECG and STFT are used to estimate the 24 h power spectral density [13]. The time-dependent spectrum was calculated by using STFT. Feed forward and feedback architecture have been calculated by using key design characteristics like bit width, radix, and scaling. Mono bit FFT is developed, to get faster real-time performance in broadband digital receivers. In STFT feed-forward methods are more suitable for real-time constraints. STFT is of two types are Continuous-time STFT and Discrete-time STFT [1]. In the first type, the signal which is transformed is getting multiplied with a window function in a short period. Here window function is nothing but a SQL function where intervals are taken from one or more rows in the results set of select statements.

$$STFT\{x(t)\}(\tau, \omega) \equiv X(\tau, \omega) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-j\omega t dt} \quad (1)$$

Here (τ) is known as window function which is centered to zero, $x(t)$ is a signal which is going to be transmitted. $X(\tau)$, is a Fourier transform of $x(t)$ ($t - \tau$). Here w is the window function and ω is the frequency. Complex function (it is the function of a complex number), it represents the signal phase and magnitude over time and frequency.

In discrete-time STFT the data which is transformed should be in the form of chunks and frames.

$$STFT\{x[n]\}(m, \omega) \equiv X(m, \omega) = \sum_{n=-\infty}^{\infty} x[n]w(n - m)e^{-j\omega n} \quad (2)$$

Here m is discrete and ω is continuous. The maximum frequency span and real-time bandwidth offer zero-gap overlapping FFT processing. It is an important parameter of RTSA which can be used to get more information about the spectrum based on signal content type. If real bandwidth is very large then there may be a chance of a large amount of noise occurring and damage the energy of the pulsed signal. If the real bandwidth is too narrow then it is not possible to represent the signal energy accurately. So, for 100% POI a signal must have enough duration to confirm its presence in an entire FFT [8].

3 Methodology

In this paper, we go through the bearing elements of the rotary machine. Vibration signals of bearings which are non-stationary are complex because of the rich content of information [1, 2]. According to the state of bearings, vibration signals will change simultaneously. To identify and represent the changes, one analysis is required which makes identification much easier and more reliable, that analysis is known as time-frequency analysis (TFA), which is using mostly in signal processing [6, 12].

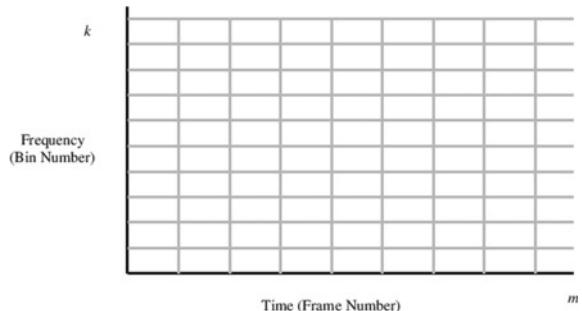
To evaluate the changes in both frequency and time domain, time-frequency analysis is used. Here in Fig. 1 shows the vertical line indicates the time resolution and horizontal lines indicate the frequency resolution. To multiply the time series in which the non-stationary signals are nearly considered as locally stationary and then transformed to the time-frequency method, this approach uses the window function [1, 2]. TFA is using a Short Time Fourier Transform [4]. As we have already seen in the literature part STFT can be also written as

$$S(t, f) = \int x(t + \tau)w(\tau)e^{-j2\pi f\tau}d\tau \quad (3)$$

Usually, $S(t, f)$ is not considered because the amplitude spectrum is convenient and sufficient to deal with. In an industry environment, there are more chances for fault developing, for example, they are like single-sided impulse component and double-sided impulse component. These faults have occurred because in industrial environment vibration signals are usually affected by different parts. So that's why STFT is used and it has good capacity and easy principles. STFT is calculated in two ways one is using feed-forward Fourier transform and another one is by using an iterative fashion obtaining each frequency values independently [5–7]. For discrete signal $x(n)$ the STFT can be written as

$$X(n, p) = \sum_{m=n}^{n+(N-1)} x(m)e^{-j\frac{2\pi p}{N}m} \quad (4)$$

Fig. 1 Frequency versus time frame graph



where, $p = 0, 1 \dots N - 1, n = \text{Time}$. For certain time n , STFT has a close similarity to the FFT of the sequence from sample $x[n]$ to $x[n + (N - 10)]$. To calculate STFT we have to use N , FFT processors by keeping them in parallel (here for each FFT processor has $2 \log_2 N$ adders and $\log_2 N$ multiplier with a total memory size of N) so by this information we know that the STFT has N number of FFT, so the total no of adders that STFT have $2N \log_2 N$, number of multipliers are $N \log_2 N$ with memory (N^2).

Another approach is obtaining values at each time instant for each frequency

$$\text{STFT}x[n, k] = e^{j\frac{2\pi}{N}k} [\text{STFT}x[n - 1, k] + x[n - 1 + N] - x[n - 1]] \quad (5)$$

For this, there are N adders and N multipliers and memory size of $2N$ [5].

A low duty cycle may be able to capture by the RTSA. But it is complex to observe in the real-time regular display. By strengthening the signal at the receiver to background noise and increasing the interference by dispersing the power of modulated signal over a wideband width and this technique is known as spread spectrum modulation which is used by GPS, Bluetooth. Some RTSA has a feature that they can record the output of the processed spectrum for future and deeper analysis. Frequency mark trigger is only possible with RTSA [7–9].

4 Result

As we discussed in the above STFT can be calculated by two approaches they are FFT and another one is iterative and feed-forward STFT (Table 1).

When a comparison is made between the FFT and feed-forward STFT, the feed-forward STFT can reduce the multipliers, memory size and adders. And feed-forward STFT does not have accumulative error because of its advantage of having a sufficient number of multipliers and adders when compared to iterative STFT.

Table 1 Comparison between types of approaches of STFT

STFT implementation	FFT	Iterative	Feedforward STFT
Accumulative error	No	Yes	No
Multipliers	$N \log_2 N$	N	$N - 1$
Adders	$2N \log_2 N$	N	$2N - 2$
Memory	N^2	$2N$	$(N/2) \log_2 N$

5 Conclusion

In this paper we discussed faults occurred in rolling elements and how those faults affect the working condition of the rotary machine and we also go through the method to avoid the fault vibrations. Fault vibrations occurring in rolling elements are non-stationary, so time-frequency distribution is used because of its flexibility of energy distribution capacity. But time-frequency distribution depends on short-time Fourier Transform. Short-time Fourier transform is also used in spectral analysis. STFT is divided into continuous-time STFT and discrete-time STFT which was discussed in literature review. STFT can be calculated by using two approaches which were mentioned in methodology. Discrete Fourier Transform is developed by using FFT which is one type of approach to calculate the STFT. RTSA has an advantage of overlapping FFT and high speed of memory for 100% (POI) in a dense environment. RTSA is used in military and defense applications such as radar and so on.

References

1. Gao Huizhong, Liang Lin, Chen Xiaoguang, Guanghua Xu (2015) Feature extraction and recognition for rolling element bearing fault utilizing short-time Fourier transform and non-negative matrix factorization. *Chin J Mech Eng* 28(1):96–105
2. Cocconcelli M, Zimroz R, Rubini R, Bartelmus W (2012) STFT based approach for ball bearing fault detection in a varying speed motor. In: Condition monitoring of machinery in non-stationary operations. Springer, Berlin, Heidelberg, pp. 41–50
3. Boufenar M, Rechak S, Rezig M (2012) Time-frequency analysis techniques review and their application on roller bearings prognostics. In: Condition monitoring of machinery in non-stationary operations. Springer, Berlin, Heidelberg, pp. 239–246
4. Ivanović VN, Stojanović R, Jovanovski S, Stanković L (2006) An architecture for real-time design of the system for multidimensional signal analysis. In: 2006 14th European signal processing conference. IEEE, pp. 1–5
5. Garrido Mario (2016) The feedforward short-time fourier transform. *IEEE Trans Circ Syst II Express Briefs* 63(9):868–872
6. Liu KJR (1993) Novel parallel architectures for short-time Fourier transform. *IEEE Trans Circ Syst II Analog Digit Sig Process* 40(12):786–790
7. Zhang S, Yu D, Sheng S (2006) A discrete STFT processor for real-time spectrum analysis. In: IEEE Asia pacific conference on circuits and systems APCCAS 2006–2006. IEEE, pp. 1943–1946
8. Gupta S, Abielmona S, Caloz C (2009) Microwave analog real-time spectrum analyzer (RTSA) based on the spectral–spatial decomposition property of leaky-wave structures. *IEEE Trans Microw Theory Tech* 57(12):2989–2999
9. Gupta V, Mittal M (2019) QRS complex detection Using STFT, chaos analysis, and PCA in standard and real-time ECG databases. *J Inst Eng India Series B* 100(5):489–497
10. Lei Y, He Z, Zi Y, Qiao H (2007) Fault diagnosis of rotating machinery based on multiple ANFIS combination with GAs. *Mech Syst Sig Process* 21(5):2280–2294
11. Li W, Shi T, Liao G, Yang S (2003) Feature extraction and classification of gear faults using principal component analysis. *J Qual Maintenance Eng*
12. Pathiran, AR, Erikiananda K, Getachew T, Gziabher HG (2019) Performance and predict the ball bearing faults using wavelet packet decomposition and ANFIS. *Int J Eng Sci Technol* 11(2):33–47

13. Malarvili MB, Mesbah M, Boashash B (2007) Time-frequency analysis of heart rate variability for neonatal seizure detection. *EURASIP J Adv Sig Process* (1):050396
14. Weidong Z, Yingyuan L (2001) EEG real-time feedback based on STFT and coherence analysis. In: 2001 conference proceedings of the 23rd annual international conference of the IEEE engineering in medicine and biology society, vol. 2. IEEE, pp. 1869–1871
15. Thornburg H, Leistikow RJ, Berger J (2007) Melody extraction and musical onset detection via probabilistic models of framewise STFT peak data. *IEEE Trans Audio Speech Lang Process* 15(4):1257–1272
16. Sanchez MA, Garrido M, Lopez-Vallejo M, Grajal J, Lopez-Barrio C (2005) Digital channelised receivers on FPGAs platforms. In: IEEE international radar conference, 2005. IEEE, pp. 816–821
17. Sanchez MA, Garrido M, Lopez-Vallejo M, Grajal J (2008) Implementing FFT-based digital channelized receivers on FPGA platforms. *IEEE Trans Aerosp Electron Syst* 44(4):1567–1585
18. Amornwongpeeti S, Ono N, Ekpanyapong M (2014) Design of FPGA-based rapid prototype spectral subtraction for hands-free speech applications. In: Signal and information processing association annual summit and conference (APSIPA), 2014, Asia-Pacific. IEEE, pp. 1–6
19. Cabal-Yepez E, Garcia-Ramirez AG, Romero-Troncoso RJ, Garcia-Perez A, Osornio-Rios RA (2012) Reconfigurable monitoring system for time-frequency analysis on industrial equipment through STFT and DWT. *IEEE Trans Industr Inf* 9(2):760–771

A Review of 4-2 Compressors: Based on Accuracy and Performance Analysis



P. Venkata Ganesh, E. Jagadeeswara Rao, and Durgesh Nandan

Abstract A very urgent need for high-speed adders needed in signal processing applications (SPA). Due to parallel addition circuits play a major role in SPA but it appears the errors after adding the input bits. So, few researchers focus to reduce the error in recent years, and also few researchers focus to improve the speed and reduce the area, also they develop the Compressors (COMP) to compare with the conventional methods. COMP is also more useful in image processing application like edge detector method and it also plays a major role in approximate multipliers (MULP's). In this paper, we mainly focus to give the review of how to improve the performance of COMP. In the last few years which were developed by different researchers work in this area. Finally, this paper gives some recent COMP performance analysis in terms of Delay, Power consumption (PC), and power delay product (PDP).

Keywords COMP · PC · PDP

1 Introduction

In this paper, we mainly discuss 4-2 compressors (COMP) and about its applications. Mainly we use the COMP used for power saving techniques. Examples of these CMOS applications are image processing, real-time images, real-time speech recognition graphic accelerators, etc. Multipliers are responsible for performing the

P. Venkata Ganesh · E. Jagadeeswara Rao

Department of Electronics and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India

e-mail: venkataganesh789@gmail.com

E. Jagadeeswara Rao

e-mail: emandi.jagadeesh@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India

e-mail: durgeshnandano51@gmail.com

Arithmetic (ARTH) Operations (OPER). It is also responsible for DSP and micro-processors applications which improve the speed. Filtering and convolutions are the main two important DSP algorithms. The overall discussion about the mathematical expression of 4-2 COMP. In Sect. 3, we discuss table form. In Sect. 4, we discuss the result, and finally, we discuss conclusion and reference in Sect. 5. Circuit depends on the speed of the MUX and power consumption (PC). In COMP the main significations are power (PW), area, and delay. The speed of the MUX how fast the processor will run and high speed with low PC and area occupied by the architecture. A multiplication process mainly depends on three stages there are (1) to generate the Partial product (PP), (2) to reduce the PP, (3) producing the final product. The following stage is reducing the performance (PER) of the MUX with respect to PWD and speed. The main focus of the paper is to decrease the PC, delay, and area of the 4-2 COMS by using Full Adders (FA)

In our paper, we are discussing on a less PC and highly Energy Efficient 4-2 COMP is proposed. In Sect. 1, we discuss the behavior of 4-2 COMP and design of 4-2 COMPS. In Sect. 2, we will be discuss about systematic development of literature.

2 Systematic Literature

In this paper the driving output is between the fan-out from 1 to 10. A new 4-2 CMOS COMP has been designed, which has a power efficiency of 20–400%. The applied voltage is about 1–3 v [24]. In 2000, D. Radhakrishnan explain about a novel CMOS 4-2 COMP using pass logic is represented. A XOR–XNOR gates are used to design the circuit. In this circuit design, we use 28 TRANS [14]. In 2001, Kaurna Prasad. In this paper, we can know about the LPW and higher-order COMP such as 4-2 and 5-2 COMP unit 11.67% improvement in speed, 11.67% improvement in speed, and 44.37% in the PDP has to be improved [25].

In 2003, Pedram Mokrian. In this paper, we deeply discuss 4:2 COMP, a new novel layout scheme is presented for optimal placement. It reduces the number of 4-2 COMP. The average reduction is 2.74% of the complete cell, the average decrease in 4-2 COMP reduction is 8.171% for the total number of cells [26]. In 2003, Jungmin Gu and Chip-Hong Chung. In this paper, we see about a new LPW of 4-2 COMP. This is capable of functioning an ultra-low voltage. The architecture of 4-2 COMP is analyzed. The architecture of 4-2 COMP is analyzed [6]. In 2004, Chip-Hong Chang. In this paper, they have represented several architectures (ARCH) and designing of LPW for 4-2 and 5-2 COMP. A 5-2 COMP ARCH of 4Δ delays is introduced [12].

In 2005, Xien Ye. This paper described about the LPW tree mux based on adiabatic logic. The simulation results show that the CPAL tree MUX achieves considerable energy savings [29]. In 2005, Yangbo Wu. This paper presents two adiabatic 4-2 COMP with complementary pass-trans logic (CPAL). They have proposed two new CPAL COMP using a four-phase AC PW supply. The PC of proposed CPAL based on full adder is very low [17]. In 2005, G. Michael Howard. In this paper, we study

several LPW and high speed 4:2 COMP circuit designs of several various digital logic styles. HSPICE simulation results are given in terms of power delay (PD), PWP [21].

In 2006, Himanshu Thapliyal. In this paper, 4-2 COMP design and Modified Montgomery multiplications ARCH are represented [5]. In 2008, Sreehari Veeramachaneni. In this, a new ARCH and design the maximum speed, LPW 32, 4-2 and 5-2 COMP can operate at ULV (ultra-low voltages) are proposed. Simulations results have been varied by changing the input applied voltages. The voltages are varied from 0.9 to 3.3 v. The ARCH of different types of COMP are analyzed by CMOS and CMOS+ with the implementation of some logic gates (MUX and XOR) [28].

In 2009, Peng Chang Majid Ahmadi. This paper represents an HS LPW of 4:2 COMP circuit design based on some logical design. Domino logic (DL) design is the basic 4-2 COMP design. In this paper, they have described different designs of 4:2 COMP at the circuit level and logic level. The 4:2 COMP, which can be designed in split DL, which has the best results in terms of area, delay, PDP, PW [9]. In 2010, Shirinpourashraf. In this, we proposed logic of 4:2 COMP which uses the Pass transistor and DL structures. The ckt has 60% imp in PWD, and 49% imp in speed, then compared to the Split domino (SD) [23].

In 2010, P. Aliparast. In this paper a new design is introduced for designing a VHS CMOS 4-2 COMP this is the main imp in fast Digital ARTH OPER. It has the highest PC. It consists of 43 TRANS. Active area is $13 \mu\text{m} \times 59 \mu\text{m}$. So this is an ideal sub-circuit to implement fast digital arithmetic units [20]. In 2010, Zhou Meng. In this, we proposed a HS ACS ARCH base on 4:2 compression arrays. We designed it on radix-4 HS ACS unit. It consists of 32-bit operand. The result data shows the worst-case delay which is better than 50% at $0.25 \mu\text{m}$ CMOS [1].

In 2012, Jorge Tonfat. This paper presents 2(Adders) AD COMP ARCH addressing HS and LPW, to decrease the PC of the circuit [15]. In 2012, Amir Fathi. In this paper, they have discussed the designing of a fast 4-2 COMP. [4].

In 2013, Ardalan Najafi. According to this paper, a new 4-2 COMP ARCH is proposed. This architecture uses Carry Generator Module (CGEN). ARCH is simulated in 180 nm and 130 nm technologies. Reduces PDP—17.01%, parameter by—39.04%. The new COMP is proposed using CGEN blocks. It also reduces delay by 6.11 and 18.15% in 180 and 130 nm technologies, respectively [18]. In 2013, Abdoreza Pishvai. In this paper, they discussed about the merits to propose a new 4-2 COMP design, and verify its PER, and the OP is compared with it is previous designs. Finally, the results show a better improvement than compared to the best of the reference design, based on the terms PDP, PW, delay. The PW—13%, delay—17%, PDP—30%. In 54×54 -bit binary MUX We use more than 1300 4-2 COMP [3].

In 2014, Sanjeev Kumar. In this paper, a LPW HS 4-2 COMP circuit is proposed to fast digital ARTHI IC. They have proposed an XOR–XNOR logic design which LPC of 180.89 pW and the applied voltage is 1.3. The PC—718.72. PDP—(315.01×10^{-22}) j at 1.8 V [19]. In 2014, Mehdi Ghasemzadeh. The researchers have attributed this article to a 4-2 COMP which is based on a new COMP structure with some

external special features [13]. In 2014, Sanjeev Kumar. In this paper, LPW HS 4-2 COMP circuit is proposed for FD ARTHI IC. The proposed logic gates (XOR-XNOR) design. It shows the PC of 180.89 pW when the applied voltage is 1.8 V. XOR provides max opd of 3.1702 ns. XNOR shows delay of 1.9342 ns at 1.8 V. The PC of 718.72 pW with max opd of 43.83 ps [2].

In 2016, p saha. This paper provides design and arch of 4-2 and 5-2 COMP has be noted. The propagation delay is 41–61 ns of COMP [27]. In 2016, Dinesh Kumar. In this paper, we identified the modified 4-2 COMP by improve the MUX design. The modified circuit diagram shows 32.25% less PC and 13.04% of reduction in PDP with traditional design the applied volt is 1.8 V [30]. In 2016, Omid Akbari. In this paper, they have proposed four 4:2 COMP, which can have both exact modes and approximates operating modes. The delay—46%, PC—68% in approximate mode These COMP consists of 32-bit structure [30]. In 2016, Amirali Amirsoleimani. In this paper, mersister devices are used to implement a new 4-2 COMP. Finally, we have to compare both the memristor and COMP based [22]. In 2016, A. Chandrakala1. In this paper, an LPW HS 4:2 COMP circuit is proposed for FD ARITH IC. OP values have been calculated by verilog HDL [10].

In 2017, Asif Rashid. In this paper, FPGA implemented by using 4:2 COMP circuit. The proposed one shows the implementation and improvement in performance over the older approach [16]. In 2017, Minho Ha and sunggu lee. In this paper, APP MULP is a common OPER used in APP computing methods for high performance and LPW compute. An APP for 8-32 bit MUX which required smaller areas than compared to the present ckt design. It has been reduced from 23.2 to 24.4% in area and 22.4–24.5% in LPW and 11.2–17.0% delay than an exact MUX [11]. In 2017, Raphael Dornelles. In this paper, they have designed a three 4-2 adder COMP. The delay—32.45, area—7.4%, PW—22.41% [8].

In 2017, Manish Kumar1. In this paper, they have used only 35 TRANS to design a 4-2. The ARCH of this COMP consists of TRANS 8. The delay, PC, and PDP are varying with the applied voltage from 1 to 3 v [7].

3 Methodology

A 4-2 COMP is a combinatory device that COMP 4PP into 2PP. The basic 4-2 COMP 5 inputs namely there are given as N1, N2, N3, N4, Cin. While adding the total 5 inputs we generate 3 output in FD. The outputs are mainly given as (1) S, (2) C1, (3) CO is the output coming from the least preceding significant. C1 is the output coming from the input of the preceding state of the COMP. Initially, all the inputs are assigned with I equally, and carry is assigned with 1 bit more than the inputs I + 1. By the truth table, we mainly get the three basic mathematical equations for the S, CO and C1 which are shown in Eqs. (1), (2), (3) and (4). Finally, a basic 4-2 COMP has been designed with the FA and mathematical equations are given. The basic block diagram of Comp as shown in below.

$$N1 + N2 + N3 + N4 + Cin = s + 2 * (C0 + C1) \quad (1)$$

$$S1 = N1(\text{XOR}) + N2(\text{XOR}) + N3(\text{XOR}) + N4(\text{XOR})CIN \quad (2)$$

$$C0 = (N1\text{XOR } N2) * N3 + \text{COM}(N1\text{XOR } N2) * N3 \quad (3)$$

$$\begin{aligned} C1 &= (N1\text{XOR } N2\text{XOR } N3\text{XOR } N4) * Cin \\ &\quad + \text{COMP}(N1\text{XOR } N2\text{XOR } N3\text{XOR } N4) * N4 \end{aligned} \quad (4)$$

By using the cadence virtuose tool the simulations are done in 181 nm CMOS TECH. The results on various parameters of the proposed COMP are shown below. By varying the supply voltages. Finally, the proposed results show the max op delay, average PC, and PDP. The following equations give the OP of 4-2 COMP that has been shown in Table 1. The common implementation of a 4-2 COMP is accomplished by using 2FA are shown in Figs. 1 and 2. This design of 4-2 COMP is not accuracy because it has 17 incorrect OP of total 32 actual combinations which is error rate 4-2 COMP. The maximum error rate is 53% and the related truth table shown in Table 1.

4 Results and Discussion

In the section mainly consider the recent 4-2 Comp different performance parameter consider to proposed different authors and see in Table 2.

The basic applications are image processing, signal processing, edge-detector, real-time image processing ext. High-speed signal processing is used in wireless communications. LPW devices are used for biometric signal processing. A 4-2 COMP has been designed with several Trans and logical gates. Among them, we compared the PDP, PW, delay, no. of Trans are shown below.

In Fig. 3: The maximum output delay occurs at Comp 3 which is of (11.67 ns). Some researchers have worked to reduce the delay time of the Comp. In [13] the delay has been reduced which is of (0.24 ns).

In Fig. 4: we discuss about number of transistors are used to construct the 4-2 comp. If we use more transistors the circuit design will be complex and cost also increases. In [3] we can see that less number of transistors have been used but a delay is more to overcome the delay we increase number of transistors.

In Fig. 5: Area depends on the number of transistors that have been used. The area occupied is less when we reduce number of transistors. In [3] the area has occupied by 19% (Fig. 6).

Table 1 General truth table of 4-2 comp

Fig. 1 Block diagram of comp

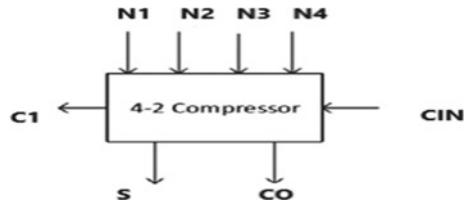


Fig. 2 Full adder based on 4-2 compressors

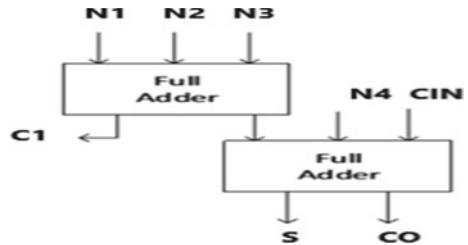


Table 2 Hardware performance comparison of various 4-2 Comp

Comp	Power	Delay (ns)	No. of trans	Area	Pdp
Comp [3]	3.3	11.67	8	452	126
Comp [5]	0.6	6.77	80	374	123
Comp [6]	0.6	6.29	80	372	105
Comp [9]	2.5	0.27	30	564	196
Comp [12]	1.8	0.47	44	402	62
Comp [13]	1.8	0.24	17	550	18.25
Comp [14]	1.8	2.12	43	767	39.9
Comp [16]	1.8	2.62	10	520	61.04
Comp [17]	1.2	0.342	36	1375	17.01
Comp [18]	1.8	0.934	12	798	17.01
Comp [23]	1.8	0.363	24	750	7.855

5 Conclusion

In this paper, mention the different 4-2 Comp with different techniques developed by different researchers till now. Also, provide the basic construction of 4-2 Comp and mentioned performance analysis of different 4-2 Comp. The maximum delay is 11.67% [3] it has reduced to 2.4% [13]. The less number of transistors are 8 [3] the maximum number of transistors used are 80 [5] due to more number of transistors area has increased. The maximum area is 1375 [17] due to more area the cost has increased so that some researcher has been tried to reduce the area it has reduced to 372 [6]

Fig. 3 Delay comparison of different 4-2 comp

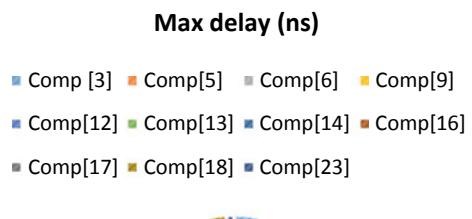


Fig. 4 No. of transistors comparison of different 4-2 comp

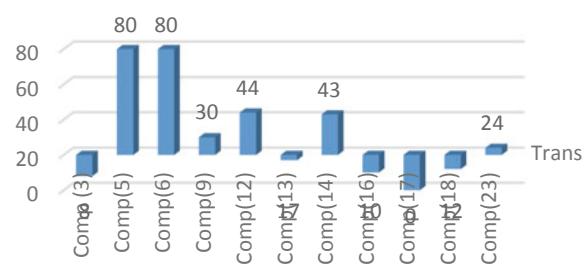


Fig. 5 Area comparison of different 4-2 comp

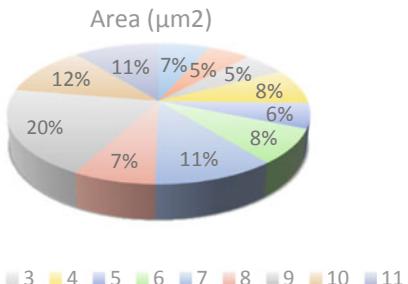
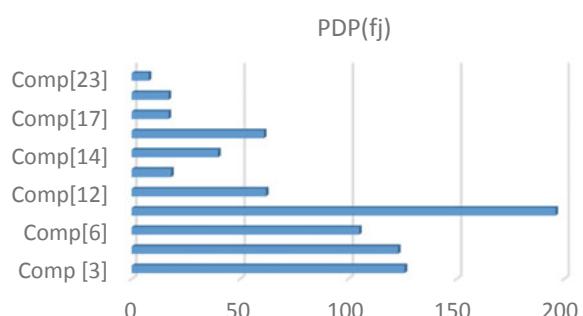


Fig. 6 PDP comparison of different 4-2 comp



A new 4-2 comp has been designed in this paper with increasing efficiency, LPW, high speed. The delay, pdp, pc, area has been calculated by varying the applied Voltages.

References

1. Akbari O, Kamal M, Afzali-Kusha A, Pedram M (2017) Dual-quality 4:2 compressors for utilizing in dynamic accuracy configurable multipliers. *IEEE Trans Very Large Scale Integr Syst* 25:1352–1361. <https://doi.org/10.1109/TVLSI.2016.2643003>
2. Aliparast P, Koozekanani ZD, Khiavi AM, Karimian G, Bahar HB, Aliparast P (2010) A new very high speed CMOS 4-2 compressor for fast digital arithmetic circuits. *Proc 17th Int Conf Mix Des Integr Circuits Syst Mix* 191–194
3. Amirsoleimani A, Ahmadi M, Teimoory M, Ahmadi A (2016) Memristor-based 4:2 compressor cells design. *Proc IEEE Int Symp Circuits Syst* 1242–1245. <https://doi.org/10.1109/ISCAS.2016.7527472>
4. Chandrakala A, Sreeramulu A, Srinivas L (2016) Design and implementation of 4-2 compressor design with new Xor-Xnor. 3:175–178. <https://doi.org/10.17148/IARJSET.2016.3735>
5. Chang C, Member S, Gu J, Member S (2004) Ultra low-voltage low-power CMOS 4-2 and 5-2. 51:1985–1997
6. Chang P, Ahmadi M (2009) A high speed low power 4:2 compressor cell design. 2009 Int Symp Signals Circuits Syst ISSCS 2009. 1–4 <https://doi.org/10.1109/ISSCS.2009.5206178>
7. Dornelles R, Paim G, Silveira B, Fonseca M, Costa E, Bampi S (2017) A power-efficient 4-2 adder compressor topology 281–284
8. Fathi A, Azizian S, Hadidi K, Khoei A, Chegeni A (2012) CMOS implementation of a fast 4-2 compressor for parallel accumulations. *ISCAS 2012—2012 IEEE Int Symp Circuits Syst* 1476–1479. <https://doi.org/10.1109/ISCAS.2012.6271526>
9. Ghasemzadeh M, Akbari A, Hadidi K, Khoei A (2014) A novel fast glitchless 4-2 compressor with a new structure. *Proc 21st Int Conf Mix Des Integr Circuits Syst Mix* 127–130. <https://doi.org/10.1109/MIXDES.2014.6872170>
10. Gu J, Chang CH (2003) Ultra low voltage, low power 4-2 compressor for high speed multiplications. *Proc IEEE Int Symp Circuits Syst* 5:321–324
11. Ha M, Lee S (2018) Multipliers with approximate 4-2 compressors and error recovery modules. *IEEE Embed Syst Lett* 10:6–9. <https://doi.org/10.1109/LES.2017.2746084>
12. Howard GM, Mokrian P, Ahmadi M, Miller WC (2005) Power and delay analysis of 4:2 compressor cells. *Proc IEEE Int Symp Circuits Syst* 3559–3562. <https://doi.org/10.1109/ISCAS.2005.1465398>
13. Kumar D, Kumar M (2016) Modified 4-2 compressor using improved multiplexer for low power applications. 2016 Int Conf Adv Comput Commun Informatics, ICACCI 2016. 236–242. <https://doi.org/10.1109/ICACCI.2016.7732053>
14. Kumar S, Kumar M (2014) 4-2 compressor design with new XOR-XNOR module. *Int Conf Adv Comput Commun Technol ACCT* 106–111. <https://doi.org/10.1109/ACCT.2014.36>
15. Margala M, Durdle NG (2008) Low-power low-voltage 4-2 compressors for VLSI applications. *Proc IEEE Alessandro Volta Meml Work Low-Power Des* 84–90. <https://doi.org/10.1109/lpd.1999.750407>
16. Meng Z, Minglun G (2010) A high-speed ACS design based on 4:2 compression array. *ICSPS 2010 Proc 2010 2nd Int Conf Signal Process Syst* 2:V2-547-V2-550 (2010). <https://doi.org/10.1109/ICSPS.2010.5555671>
17. Mokrian P, Howard GM, Jullien G, Ahmadi M (2003) On the use of 4:2 compressors for partial product reduction. *Can Conf Electr Comput Eng* 1:121–124
18. Najafi A, Mazloom-nezhad B, Najafi A (2013) Low-power and high-speed 4-2 compressor. 2013 36th Int Conv Inf Commun Technol Electron Microelectron 66–69

19. Parhi KK Low-power 4-2 and 5-2 compressors 129–133
20. Pishvaie A, Jaberipur G, Jahanian A (2013) Redesigned CMOS (4; 2) compressor for fast binary multipliers. *Can J Electr Comput Eng* 36:111–115. <https://doi.org/10.1109/CJECE.2013.6704692>
21. Pourashraf S, Sayedi M (2010) A novel 4:2 compressor for high speed and low power applications. *Proc 2010 18th Iran Conf Electr Eng ICEE 2010* 471–475. <https://doi.org/10.1109/IRANIANCEE.2010.5507022>
22. Radhakrishnan D, Preethy AP (2000) Low power CMOS pass logic 4-2 compressor for high-speed multiplication. *Midwest Symp Circuits Syst* 3:1296–1298. <https://doi.org/10.1109/MWSCAS.2000.951453>
23. Rashid A, Mir AG (2017) Achieving performance speed-up in FPGA based 4:2 compressor using fast carry-chains. *2017 4th Int Conf Signal Process Integr Net, SPIN 2017* 5–9. <https://doi.org/10.1109/SPIN.2017.8049905>
24. Ravi A, Sznajder L (2017) Design of an energy efficient 4-2 compressor design of an energy efficient 4-2 compressor. <https://doi.org/10.1088/1757-899X/225/1/012136>
25. Saha P, Samanta P, Kumar D (2017) 4:2 and 5:2 decimal compressors. *Proc Int Conf Intell Syst Model Simul ISMS* 424–429. <https://doi.org/10.1109/ISMS.2016.87>
26. Thapliyal H, Ramasahayam A, Kotha VR, Gottimukkula K, Srinivas MB (2006) Modified montgomery modular multiplication using 4:2 compressor and CSA adder. *Proc Third IEEE Int Work Electron Des Test Appl DELTA 2006* 414–417. <https://doi.org/10.1109/DELTA.2006.70>
27. Tonfat J, Reis R (2012) Low power 3-2 and 4-2 adder compressors implemented using ASTRAN. *2012 IEEE 3rd Lat Am Symp Circuits Syst LASCAS 2012 Conf Proc* 1–4. <https://doi.org/10.1109/LASCAS.2012.6180303>
28. Veeramachaneni S, Kirthi Krishna M, Avinash L, Puppala SR, Srinivas MB (2007) Novel architectures for high-speed and low-power 3-2, 4-2 and 5-2 compressors. *Proc IEEE Int Conf VLSI Des* 1:324–329. <https://doi.org/10.1109/VLSID.2007.116>
29. Wu Y, Zhang W, Hu J (2005) Adiabatic 4-2 compressors for low-power multiplier. *Midwest Symp Circuits Syst 2005*:1473–1476. <https://doi.org/10.1109/MWSCAS.2005.1594391>
30. Ye X, Hu J, Tao W (2005) Complementary pass-transistor, pp 270–273

Emotion Recognition Using Chatbot System



Shraddha Pophale, Hetal Gandhi, and Anil Kumar Gupta

Abstract It is observed that the number of suicide attempts among students is increasing day by day. The reasons behind these attempts can be tough competitions, comparisons with others in college, or at home or partial nature of teachers towards students. In this context, the computer chatbot system may be built to recognize the emotions from the text conversations done by students. Such a system will also help teachers to understand student's psychological behavior to monitor the information of students and take prior precautions. The goal of this paper is to provide an effective conversational interface for the same. We are using the hybrid model to improve the performance of recognizing human emotions. The system employs the structural characteristics of chatting using NLP techniques.

Keywords Chatbot · Natural language processing · Machine learning · Emotion classification · Emotion recognition

1 Introduction

Many interactive dialog systems are now available for instant communication and interaction. However, identifying the emotions of the users who are interacting is a difficult task. There is a need for such a system so that preventive measures can be taken in advance for persons having emotions like sadness, anger, depression, etc. Human-machine dialogues are trying to close the human-human communication.

S. Pophale (✉) · H. Gandhi
Walchand College of Engineering, Sangli, India
e-mail: shraddha.popuale@walchandsangli.ac.in

H. Gandhi
e-mail: hetal.gandhi@walchandsangli.ac.in

A. K. Gupta
Centre for Development of Advanced Computing, Pune, India
e-mail: anilg@cdac.in

One important factor is that we expect machines to understand our emotions and purpose and respond with the ability to understand and share the feelings.

Now almost every college or university we enter, the problem is, they take training from psychological consultation centers. However, college students are afraid to express private experiences so that they are having problems meeting to consult psychological consultants [1].

2 Related Work

Till date, very few research papers are available for emotion detection using Natural Language Processing techniques. Yoon SR [2] presented the hybrid system in which two models were used, (1) keyword-based model and (2) machine learning model. The keyword-based model used the keywords to indicate emotions and the proposed system used the knowledge-based artificial neural network (KBANN) for extracting features and domain knowledge. The model was built for eight categories of emotions, i.e., hope, love, sadness, anger, happiness, neutral, fear. More than 1000 sentences were used by this system not only for training but also for testing. The accuracy with emotional keywords in sentences was 90% for every emotion and without emotional keywords, accuracy was 45–65%. Hence, the sentences with emotional keywords were having high accuracy than the knowledge-based artificial neural network.

In [3], Christos Troussas presented the language learning for sentiment analysis by applying the Naive Bayes classifier. This system used Facebook status to apply the Naive Bayes classifier for learning the language and sentiment analysis. The proposed system was designed for classifying opinion at sentence level into negative, positive, or neutral sentiments. This model accepts inputs from status updates. The system built three models—Naive Bayes, Perceptron, and Rocchio to predict the associated sentiment. For Facebook status analysis, they collected more than 7000 status updates from 91 users. The performance of models was compared using recall, precision, and F-score. Rocchio classifier gave a better F-score for the best performance. Naive Bayes proved to be the best of all three performance measures.

More SR [4] proposed the typical representation of character and action with some degree of realism. They suggested emotion detection indicating the cause of entered text on social networking websites. Visual images were generated according to the text's emotion. This paper deals with emotional sentences that do not contain any emotional words as well. The dataset contains blogs and twitter data. Naive Bayes classifier was used for training the data provided as input data. The input data set contains one thousand sentences of every emotion type such as happiness, anger, sadness, and fear. The classifier model was built and different evaluation measures were evaluated.

The Lea Canales [5] proposed two main approaches, lexical approach, and machine-learning approach. First, the lexical approach depends on available lexicons or ontologies. Linguistic features are used under machine-learning approach.

Their evaluation includes testing by utilizing the model, they construct the information contained in the type of EmotiNet. They can identify the emotion communicated in new examples about the classes in ISEAR, i.e., International Survey of Emotional Antecedents and Reactions. EmotiNet model results in the chain of activities and their corresponding emotional impact utilizing an anthropological illustration.

According to Chopade CR [6], the human-computer connection is extremely incredible and most recent area of research in light of the fact that the human world is getting more and more digitized. This needs the computerized frameworks to emulate human behavior accurately. The emotion is one part of human behavior which assumes a significant job in human-computer connection, the system interfaces need to perceive the emotions of the user to show intelligent behavior. Humans express the emotion in the form of text, our facial expression, and speech. Consistently, a huge measure of printed text data is assembled into the web, for example, web journals, social media and pictures, and so on. This involves a difficult style as it is framed with both plain text and short informing language. Chetan centered on an outline of emotion discovery from text and described the emotion detection techniques. These strategies are partitioned into the following four primary classifications: keyword-based, Lexical Affinity strategy, learning-based, and hybrid-based methodology. Restrictions of these emotion recognition strategies were presented and the text normalization was addressed taking care of methods for both plain text and short informing language (Fig. 1).

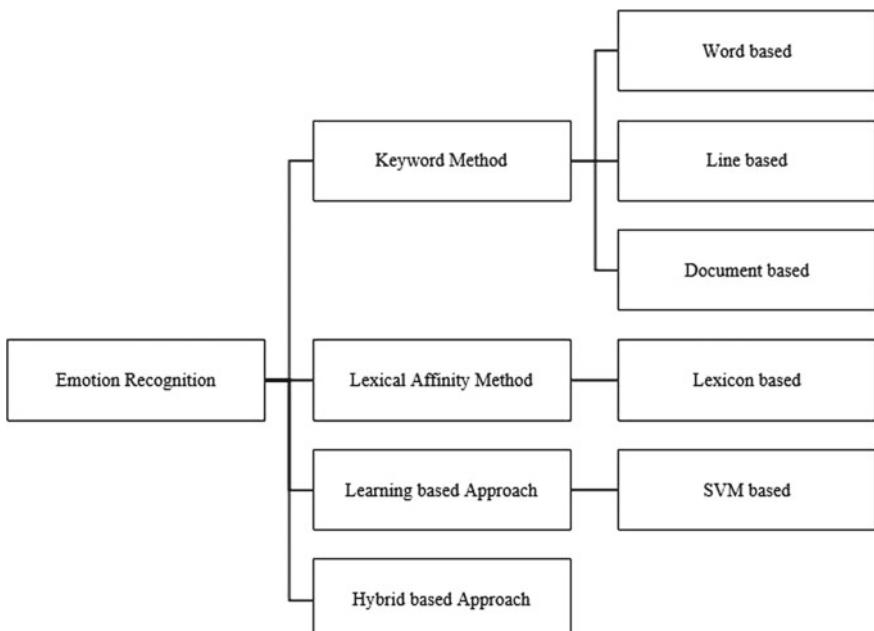


Fig. 1 Text based emotion recognition techniques

Bruna O [7], manages textual emotion classification which picked up consideration as of late. Emotion classification is utilized in user understanding, mentoring applications, national security, and product evaluation. It attempts to recognize the emotional substance in the text message and dependent on various methodologies to set up what sort of emotional substance is available, assuming any. Textual emotion classification is the hardest to deal with since for the most part depends on linguistic resources and it acquaints numerous difficulties with the task of text to emotion represented by an appropriate model. An essential piece of every emotion detector is an emotion model. The focal point of this paper is to present emotion models utilized for classification. Unconditional and dimensional models of emotion are clarified and some further developed methodologies are referenced.

Binali H [8], proposed a novel strategy dependent on the idea of AI for Emotion Detection utilizing different algorithms of Support Vector Machine, i.e., SVM and significant emotions portrayed were connected to the Word-Net for upgraded accuracy. The word-based approach with NLTK (Natural Language Toolkit) package was used as they were the best to use for human language data. The learning-based approach divided the training and testing dataset in the ratio 3:1.

Shivhare [9], Emotion detection in text records is a substance-based game plan issue including thoughts from the areas of NLP and ML. Keyword Spotting System, the watchword design coordinating issue can be depicted as the issue of discovering events of keyword from a given set as substrings in a given string. Lexical Affinity approach is an expansion of the keyword spotting method. It distributes a probabilistic affinity for a specific emotion to discretionary words separated from getting emotional keywords. The emotion detector algorithm figures out the weight for specific emotion by including loads doled out at each degree of the pecking order and ascertains the same for its counter emotion, at that point looks at the two scores and more prominent one is taken as the detected emotion.

Sreeja studied emotion models [10] and contributed to better approaches to improve the correspondence between sensitive humans and systems which are apathetic. Numerous scientists and Psychologists have given responses to questions, for example, how would we have emotions and what makes us have these emotions. They have proposed various hypotheses to clarify why humans have emotions and recommend computational models to classify how to arrange their emotions. A categorical model has the boundaries of an identification task in attempting to distinguish the exact emotional states perceived by masses. For example, studies cannot help choosing one of the basic emotions (e.g., Anger, disgust, fear, joy, sorrow, and surprise) even though they feel neutral and want to pick out that category. Their limitations regarding emotion classes, and emotion recognition using the categorical model, especially for poems, we introduce a novel categorical model based on “Navarasa” descriptions in “Natyashastra” to add value to emotion which a rarity in emotion detection.

The Jain [11], an emotion is a specific inclination that describes a perspective, for case in point, love, joy, fear, anger, etc. An incredible body of work exists in the field of emotion extraction. The work done around there remembers recognizing abstract portions in text, discovering sentiment direction and, in few cases, deciding

fine-grained differentiations in sentiment, for example, emotion and appraisal types. Fuzzy Logic is very true for a sentiment analysis procedure in which the system must be able to recognize the sentiment expressed by a customer in a review based on the statements about various features of the product or service.

Hajar [12] through the extension of Wireless use, there is a creating necessity for slicing advanced features that plan to Phone users an intelligent connection. The system used an unsupervised machine learning algorithm that performs emotion arrangement, on account of a data corpus, worked from YouTube remarks. The clarification for such a decision is the comparability between YouTube remarks and text composing style. To assemble a book segment into a particular emotion class, process its closeness to each target emotion, using the Pointwise common Information measure. The role of the corpus is to perform information words stemming from the remarks utilizing techniques from the “NLP Component”. Stemming gives a similar structure for words from a similar family, for instance: “am”, “are” and “is” become “be”, plural things become solitary, present constant types of action words become infinitive structures, etc. We run numerous tests over sentences communicating a specific emotion. The procedure yields an overall exactness of 92.75%, which reflects the feasibility of the system.

Binali [8] suggested that emotion is an imperative piece of human life and notwithstanding different things, significantly sway decision making. The typescript is part of sentences utilizing a sentence splitter. These sentences structure the essential unit for sentence characterization. From that point, we utilize a POS tagger to comment on the information with syntactic data recognizing all adjectives, verbs, adverbs, and nouns. It shows how these models have been used by discussing computational approaches to manage emotion recognition. The author proposes a hybrid based architecture for emotion detection. The Support Vector Machine algorithm is used for approving the proposed architecture and achieves a forecast accuracy of 96.43% on web blog data.

In this system, Dini [13], present an analysis to identify emotions in tweets. The Machine Learning Approach used a multiclass linear classifier associated with a Quasi-Newton minimizer, under the Stanford NLP implementation. The main lexical resource used in SentiMiner is a gazetteer of emotions (1577 lemmas) automatically extracted from the WordNet Affect database. The assessment displays that a machine learning classifier performs best on emotion detection, while a representative methodology is better for distinguishing relevant (i.e., emotional) tweets.

The expression of the emotions in man and animals by charles darwin [14], presented despite various endeavors by more than a few researchers, ordering emotional measurements identified with prosperity and state of mind stays a troublesome errand, with moderately low accuracies, going from 55 to 80%. Instances contain by means of smartphone statistics to model social communications, to study the affiliation between mood and sleep, happiness, and disposition, and to detect tension, to expect depressive indicators. Others have likewise endeavored prediction of acceptable grained symptoms on an uninterrupted scale using smartphone data and wearable devices (Table 1).

Table 1 Comparative summary of literature survey for emotion recognition

References	Data set	Technique	Additional Features	Accuracy (%)
Bruna [7]	Ekman	WordNet-Affect		73
Ramalingam [8]		WordNet	Keywords based approaches	64
Shivhare [9]	Blogs	Keyword spotting technique		56
Sreeja [10]		Dimensional emotion model	Navarasa based	
Jain [11]	ISEAR	Probabilistic classifier—Naïve Bayes algorithm	TFIDF	80
Yasmina [12]	YouTube comments	Corpus building	Statistical classification method	91–95
Binali [8]	Ekman	Keyword based	OMCS knowledge, WordNet, emotional weight	43
Dini [13]	Tweets	Emotion tweet corpus for classification	Dictionary substitution approach	74
Kumar [15]	Blog	Fuzzy logic	Entropy classifier and a knowledge-based tool	78

3 Methodology

At the beginning time of work, the area developed are recognized. The sentiment viewpoints are recognized as the positive and negative descriptive words and different weight is assigned to these descriptors dependent on the criticality vectors. After assigning these weights, the following work is to process a single message. This procedure includes the sentence adjustment and recognizable proof of area angles and sentiments. In light of the aspect and adjective relationship, the weight of a specific sentence is recognized. At the last stage, an order approach will be applied to identify the characteristics of specific sentiment (Fig. 2) [16].

3.1 Input Text

In this input consider as chats done with the system.

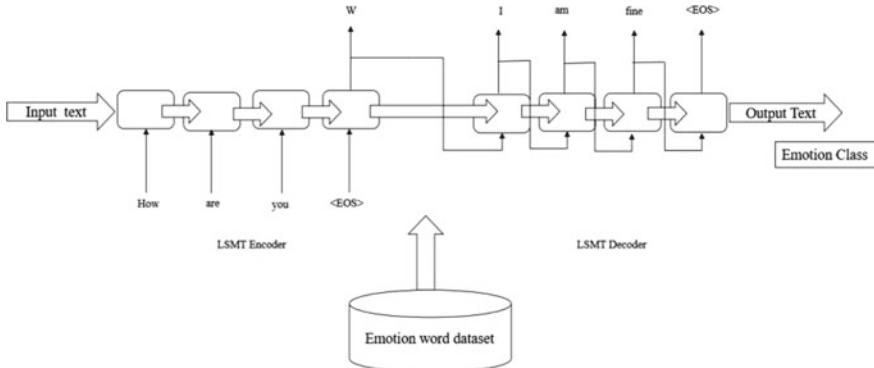


Fig. 2 LSTM encoders and decoders for chatbot system and the output text processed for emotion class prediction

3.2 Emotion Decoder

There are three ways for detection of emotion: Learning-based approach, hybrid Based method, and keyword Based method. Semantic and syntactic data such as n -grams and these approaches are utilized for the detection of the emotions. Learning-based methodology utilizes prepared classifiers for identification of emotion. It uses keywords as features classify input textual data into emotion classes. Obscured limits between emotion classes and lacks context analysis generated these are the major disadvantages of use of learning-based approach. To explain the classification task support vector machine means that SVM is utilized in this approach. It has achieved great execution in a few fields identified with text mining. Keyword-based methodology includes recognizing words to scan in textual information. It is straight forward and simple to implement. This methodology depends on specific emotional keywords and requirements preprocessing for better exact results. Hybrid based approach in emotion detection is a blend of keywords and learning-based approaches. This methodology gives higher accuracy results via training many classifiers and including information-rich linguistic data from word dictionaries [17].

3.3 Seq2seq Model

The Sequence to Sequence model means that seq2seq involves of twofold Recurrent Neural Networks—an encoder as well as a decoder. The encoder delivers the input sequence, word as a result of word and gives out a context that is a function of the final hidden state of encoder, which would ideally capture the essence that is a semantic precipitate of the input sequence. Based on this context, the decoder produces the output sequence, one word at a time while observing at the context and the previous

word through each time step. This is present an unreasonable generalization, but it gives you an idea of what occurs in Sequence to Sequence.

3.4 Emotion Detection Algorithm

Emotion of text data recognizes with support of emotion detection algorithm. This algorithm procedure calculates the weights of specific emotion by giving weight for every level of hierarchy structure and finds similar for emotion counter, after that see the difference of both value scores and which the maximum that one detected emotion is.

3.5 Output Text

Output from the Chatbot given in the form of emotion like happiness, sadness, fear, etc. using the seq2seq model. The seq2seq model overcomes the drawback of bag of words such as fixed-sized input, doesn't take order into account, and fixed-sized output. Seq2seq model is also used for encoding and decoding the text.

4 Conclusion

Generating responses from a sequence of chatbot conversations and identifying the emotions of a user is the need of an hour. With this dissertation work, we will try to address this problem using Natural Language Processing techniques and the seq2seq model.

5 Future Scope

The implementation of Artificial Intelligence (AI) chatbots will increase the features respective witness, mainly in the consumer-based facilities or services. The developing trends indicate that chatbots will be matching providing similar services and human behavior. Emotion will recognize with speech, text, facial, etc. build with chatbot system and notify class also helpful for prevention.

References

1. Sahni D, Aggarwal G (2015) Recognizing emotions and sentiments in text: a survey. *Int J Adv Res Comput Sci Softw Eng* 5(5)
2. Yoon SR, Lee YS, Yang SH, Ahn KH, Lee JH, Lee JH, Ryu SG (2010) Generation of donor natural killer cells from CD34+ progenitor cells and subsequent infusion after HLA-mismatched allogeneic hematopoietic cell transplantation: a feasibility study. *Bone Marrow Transplant* 45(6):1038–1046
3. Troussas C, Virvou M, Espinosa KJ, Llaguno K, Caro J (2013) Sentiment analysis of Facebook statuses using Naive Bayes classifier for language learning. In: IISA 2013, IEEE, pp 1–6
4. More N, Jadhav D (2018) A survey on emotions generation using text mining for social networking websites, pp 1554–1560
5. Engineering S (2014) W001-2014 Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC 2014), no. Jisic
6. Chopade CR (2015) Text based emotion recognition: a survey. *Int J Sci Res* 4(6):409–414
7. Bruna O, Avetisyan H, Holub J (2016 November) Emotion models for textual emotion classification. In: Journal of physics: conference series, vol 772, issue 1. IOP Publishing, p 012063
8. Binali H, Wu C, Potdar V (2010 April) Computational approaches for emotion detection in text. In: 4th IEEE international conference on digital ecosystems and technologies. IEEE, pp 172–177
9. Shivhare SN, Khethawat S (2012) Emotion detection from text. arXiv preprint [arXiv:1205.4944](https://arxiv.org/abs/1205.4944)
10. Sreeja PS, Mahalakshmi GS (2017) Emotion models: a review. *Int J Control Theor Appl* 10(8):651–657
11. Jain U, Sandhu A (2015) A review on the emotion detection from text using machine learning techniques. *Int J Curr Eng Technol* 5(4):2645–2650
12. Hajar M (2016) Using YouTube comments for text-based emotion recognition. *Procedia Comput Sci* 83:292–299
13. Dini L, Bittar A (2016 May) Emotion analysis on twitter: the hidden challenge. In: Proceedings of the tenth international conference on language resources and evaluation (LREC’16), pp 3953–3958
14. Book: The expression of the emotions in man and animals by charles darwin
15. Jain VK, Kumar S, Fernandes SL (2017) Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *J Comput Sci* 21:316–326
16. Kanger N, Bathla G (2017) Recognizing emotion in text using neural network and fuzzy logic. *Indian J Sci Technol* 10(12)
17. Ghandeharioun A, McDuff D, Czerwinski M, Rowan K (2019 September) EMMA: an emotion-aware wellbeing chatbot. In: 2019 8th international conference on affective computing and intelligent interaction (ACII). IEEE, pp 1–7

A Hybrid and Improved Isolation Forest Algorithm for Anomaly Detection



G. Madhukar Rao and Dharavath Ramesh

Abstract Anomalies defined as patterns or data points that do not conform to a well-defined notion of normal behavior. Anomaly detection is a significant research problem that caters to the interest of a large number of research scientists. It is a significant step in every useful data mining framework. Several techniques involving one or more of the following fields, namely statistical analysis, machine learning, soft computing, deep learning, and information theory, which have used for making better anomaly detection systems. Anomaly detection finds its applications in various fields such as detecting malicious behavior in online social media networks, detecting fraud in credit card transactions, fault detection systems. This paper presents a hybrid anomaly detection algorithm that outperforms the existing Isolation forest algorithm. A basic introduction of the existing algorithms given and then a comparative study performed between the existing algorithms and our hybrid algorithm.

Keywords Anomaly detection · Decision tree · Ensemble learning · Isolation forest · Supervised machine learning

1 Introduction

Outliers of anomalies are extreme values that deviate from other observations on data; they may indicate fluctuation in the estimation, test errors, or curiosity. There are two kinds of outliers: univariate and multivariate. Univariate outliers refer to looking for outliers in a single feature distribution, and multivariate outliers refer to looking for outliers in an n-feature distribution. Modern-day anomaly detection techniques designed, keeping in mind the multivariate outliers. The common cause

G. Madhukar Rao · D. Ramesh ()

Department of Computer Science and Engineering, Indian Institute of Technology (ISM),
Dhanbad, Jharkhand, India

e-mail: drramesh@iitism.ac.in

G. Madhukar Rao

e-mail: madhukar.iitism@gmail.com

of outliers includes data entry errors, measurement errors, data processing errors, intentional errors. Many anomaly detection techniques that build a profile of regular instances first tend to provide only a binary categorization, i.e., whether an instance is an anomaly or not. A general utility expected from a suitable anomaly detecting mechanism is to be able to measure the degree to which an instance is an anomaly. Isolation Forest is the first model-based learning algorithm that is fundamentally different from other model-based algorithms. In a way that it explicitly isolates anomalies while other previous models tend to construct models which first create a profile of non-anomalous instances, termed standard instances, and then classify the instances into being non-anomalous (standard) or anomalous [1]. The base estimator of Isolation Forest is an Isolation Tree, which is a binary decision tree, which uses randomization to select a splitting attribute at each node and then picks a random value between the base and highest value of the splitting attribute as the split value. The base estimator has replaced with the C4.5 decision tree [2], which uses Shannon's entropy-based information gain to decide on the splitting attribute [3]. As per the observations, it outperforms the Isolation Tree-based Isolation Forest ensemble. The anomaly score evaluation of the Isolation Forest algorithm has used, thus making the algorithm a hybrid of Isolation Forest and C4.5 decision trees (also called J48 decision trees). The standard Isolation Forest algorithm, as well as the hybrid Isolation Forest Algorithm, has been proposed to assign anomaly scores to all the instances in the dataset such that the more the anomaly score, the higher is the chance of being an anomaly. An anomaly detection problem can be of the following variants [4].

1. Find all the data points $x \in D$ with an anomaly score higher than some pre-decided or calculated threshold t for a given dataset D .
2. Find all the data points $x \in D$ having top-n anomaly scores for a given dataset D .
3. The dataset D consisting of normal (but unlabeled) data points, and a test point x , compute the anomaly score of x concerning D .

The variants 1 and 2 dealt with using a supervised learning algorithm if then dataset is labeled [5]. Variant 3 has unlabeled data and thus require an unsupervised learning algorithm [6, 7]. The algorithm that has proposed is a supervised learning algorithm. Isolation forest is an effective method for fraud detection. The basic principle of isolation forest is that outliers are few and are far from the rest of the observations. Frauds are outliers too. Isolation forest explicitly prunes the underlying isolation tree once the anomalies identified. It observed that C4.5 decision trees perform better than isolation trees while using the underlying expected value mathematics of the isolation forests in the ensemble built using C4.5 decision trees. In this paper, a hybrid algorithm that uses isolation forests and C4.5 decision trees to assign anomaly scores efficiently described. The binary classification by setting a certain threshold done, after which the F1 score metric is to measure anomalies as a measure of the proposed model's performance concerning other models.

The content of the paper as follows: Sect. 2 describes the related works. The proposed methodology described in Sect. 3. Section 4 describes the results of the proposed algorithm, which shows better improvement in performance metrics concerning anomalies. It also shows the comparisons with other classifiers at last, conclusion, and future work, which describes in Sect. 5.

2 Related Work

Anomaly detection has been a topic of research for several decades. Many researchers were working on problems like intrusion detection, fault detection, transaction fraud detection. Tend to employ various techniques, as it observed on reviewing much literature on anomaly detection. Support vector machines (SVMs) have been there for a long time as a very widely used tool for classification. Anomaly detection tends to be a by-product of all model-based classifiers that profile normal instances first. C-SVMs [8] first used for classification of instances. But, C-SVMs tend to perform poorly when the data set is imbalanced, hence making them not suitable for anomaly detection. But, kernel models such as SVMs can be quite sensitive to overfitting the model selection criterion. One-class SVMs also used for anomaly detection [9], but they tend to be suited better for unlabeled data only. An anomaly detection system using an entropy-based technique has done [10], which uses Shannon's entropy for classification of instances into anomalies or not. The C4.5 decision trees used as a part of several outlier detection systems [11]. The C4.5 is the most widely used decision tree out of all the decision tree variants for classification problems. C4.5 decision trees have been observed to handle missing data better than the other popular variant, ID3 decision trees [12]. Decision trees often tend to over fit, and many individual models fail because of the same issue. Ensemble learning used to minimize the overfitting issue [13]. Ensemble learning has seen to improve performance in many cases and for many problems [14–16]. Isolation Forests were the first ensemble-based supervised learning model that was dedicated to anomaly detection and gave anomaly scores to each test instance, and not just a binary classification. Hence, it is different from other models in the sense that anomaly detection is the prime usage here and, unlike other classifiers, is not a by-product of classification. The application of Isolation forests for anomaly detection demonstrated by Liu et al. [17], and it shows that Isolation based anomaly detection outperforms ORCA [18], one-class SVM LOF [19], and random forests (RF) [20]. In this paper, C4.5 trees are taken as the base model and build an ensemble upon it. It observed that this hybrid algorithm produces a better F1 score than Isolation Forests having Isolation Trees as the base model.

3 Proposed Work

3.1 Preliminaries

3.1.1 Isolated Random Forest

Isolation Forest is an ensemble-based supervised learning algorithm for anomaly detection. The underlying estimator of an Isolation Forest ensemble is Isolation Tree. It is a binary tree such that each node has either 2 children or no child. In these trees, partitions at every node are created by first randomly choosing a feature out of the complete feature set and then choosing a random split worth between the minimum and most worthy of the chosen feature. Instances with the value of the selected attribute less than the split value from the left child of the node and the instances with the value of the selected attribute higher than the split value form the right child of the node. It goes until all instances of a node belong to the same class, or the max depth of the tree reached. On a basic level, outliers are less successive than regular perceptions and are not quite the same as them as far as qualities. It is a direct result of utilizing such irregular partitioning, and they ought to be recognized nearer to the root of the tree, with fewer parts vital. Similarly, as with other outlier discovery strategies, for decision-making anomaly score is required. On account of Isolation Forest, characterized as:

$$s(x, n) = \frac{E(h(x))}{c(n)} \quad (1)$$

The path length can be represented in $h(x)$ of perception x , $c(n)$ is the normal path length of un-effective pursuit in a Binary Search Tree having n nodes. $E(h(x))$ is the average of $h(x)$ in the entire ensemble of Isolation Trees (Fig. 1).

3.1.2 C4.5 Decision Tree

C4.5 is similar to ID3 as it utilizes the idea of information entropy. Information entropy is the normal rate at which a stochastic source of data produces information. There are a few numerical variations of entropy. The one utilized in C4.5 decision tree is Shannon's entropy characterized as:

$$H = -\sum_{k=1}^K p_k \log_2 p_k \quad (2)$$

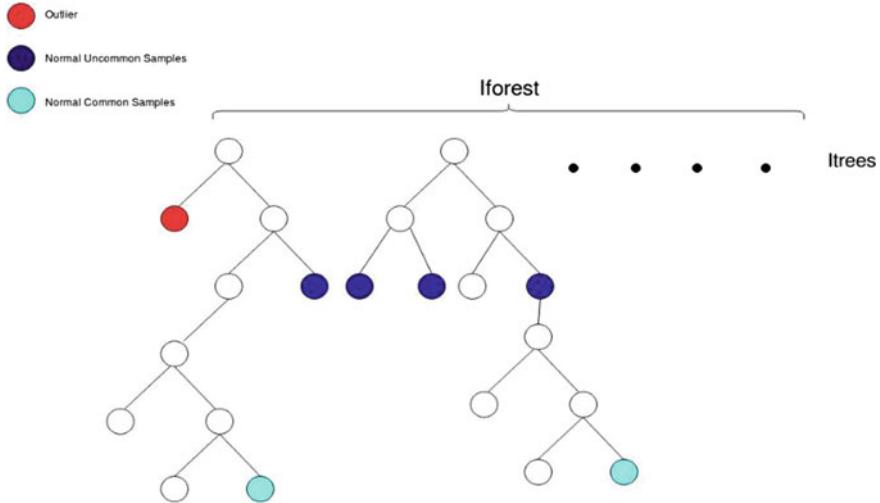


Fig. 1 Isolation forest. Two trees of the ensemble shown for understanding

3.1.3 Proposed Hybrid Algorithm (Hybrid i-Forest)

In this approach, the improvement shown upon the standard Isolation Forest, which is an ensemble of Isolation Trees, by replacing the base estimator (base model) of the ensemble with the most widely used variant of decision trees, namely C4.5 decision trees. The following algorithm represents the outline of the overall algorithm. It used to build the ensemble from C4.5 trees.

Algorithm 1: Hybrid i-Forest ($X t \ limit$)

Input: X —input data, t —number of trees, subsampling size, $limit$ —the maximum height of each C4.5 decision tree in the ensemble (can be predefined by the user, though there is a default value depending upon the sample size, as explained in the algorithm)

Output: a set of t iTrees

1. Initialize Forest
2. If the $limit$ is not specified, set height limit $l = \text{ceiling}(\log_2)$
3. **for** $i = 1$ to t **do**
4. $X_0 \leftarrow \text{sample}(X, \dots)$
5. Hybrid-Forest \leftarrow Hybrid-Forest \cup C4.5-Decision-Tree X_0 (, None, 0, l)
6. **end for**
7. **return** Hybrid-Forest

To evaluate any test instance x on a particular C4.5 decision tree, Algorithm 3 can be used, which is the same as the algorithm used to evaluate any test instance

on a particular Isolation Tree. To evaluate any test instance \mathbf{x} on the entire ensemble, Algorithm 3 runs on all the C4.5 decision trees in the ensemble. Then Eq. (1) can be used to generate the overall anomaly score of the instance \mathbf{x} . The utility functions used in Algorithm 5 given below. The following algorithm used to calculate the entropy of a given dataset based on the class labels.

Algorithm 2: Calculate entropy (X)

Input: X —the dataset whose entropy is calculated

Output: Shannon's entropy of the dataset

1. Ones = Number of records in X belonging to anomaly class
2. Zeroes = (Total rows in X)—Ones
3. Entropy = 0
4. $P(\text{Ones}) = \text{Ones}/(\text{Ones} + \text{Zeroes})$
5. $P(\text{Zeroes}) = \text{Zeroes}/(\text{Ones} + \text{Zeroes})$
6. if $P(\text{Ones})! = 0$, then do,
7. Entropy = Entropy + $P(\text{Ones}) * \log_2(P(\text{Ones}))$
8. end if
9. if $P(\text{Zeroes})! = 0$, then do,
10. Entropy = Entropy + $P(\text{Zeroes}) * \log_2(P(\text{Zeroes}))$
11. end if
12. return Entropy

The following algorithm used to calculate the information gain obtained on splitting the dataset present on a particular node of the C4.5 decision tree. The dataset is split on a particular attribute's particular split value, as explained in Algorithm 6.

Algorithm 3: Calculate_information_gain ($X, attribute_index, split_value, entropy$)

Input: X —the dataset which split at a particular Node,

$attribute_index$ —the index of the splitting attribute in the attribute list
 $split_value$ —the split value of the splitting attribute $entropy$ —the entropy of the dataset X

Output: Information gain on the splitting X at the splitting value of the splitting attribute

1. L = Total records in X
2. New_entropy = 0
3. Make datasets for the left and right node using the $attribute_index$ and

$split_value$. Records X with value at the splitting attribute higher than the value present in X_{left} and the records with the value at the splitting attribute lesser than or equal to the $split_value$ present in X_{right} .

4. $\text{New_entropy} = \text{New_entropy} + \text{calculate_entropy}(X_{left})$
5. $\text{New_entropy} = \text{New_entropy} + \text{calculate_entropy}(X_{right})$
6. Return entropy—New_entropy

The following algorithm used to assign a class label to a leaf node of the C4.5 decision tree

Algorithm 4: Classify_leaf (X)

Input: X —the dataset present at the leaf node, for which class label decided

Output: Label of the Node

1. Ones = Number of records in X , belonging to anomaly class
2. Zeroes = (Total rows in X)—Ones
3. If Ones \geq Zeroes, then return 1
4. Else return 0
5. End If

4 Results and Experimental Analysis

The imbalanced dataset contains the positive class (frauds) account for 0.172% of all transactions. It has the class label 1 in the dataset. The dataset was collected from (<http://mlg.ulb.ac.be>) of Université Libre de Bruxelles (ULB). The fraudulent transactions represent anomalies/outliers and thus form the anomalous class in the labeled dataset (Fig. 2).

Fig. 2 Imbalance of the dataset. This dataset presents 492 fraud and 284807 normal transactions

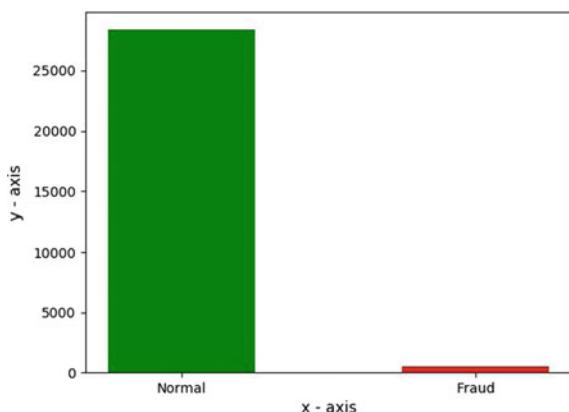
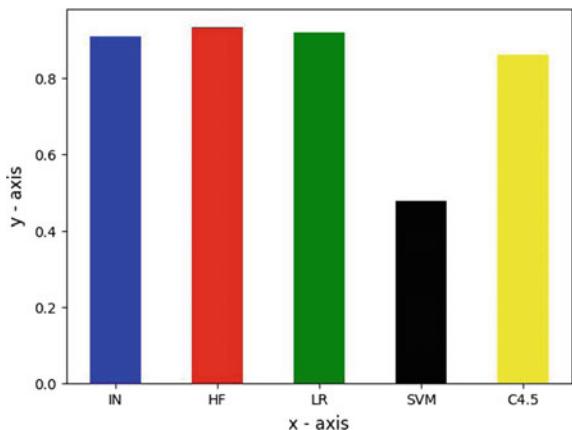


Fig. 3 Comparison of precision score w.r.t anomalous class



The data, after all the data preprocessing, was tested on five models—logistic regression, C4.5 decision trees, support vector machines (SVMs), Isolation Forests with Isolation Trees as the Base Estimator, and our ensemble, having C4.5 decision trees as the base estimator for Isolation Forests. The proposed method outperforms all other methods concerning the precision score as well as recall score, both w.r.t. anomalous class. The following three figures show the comparison of Precision, Recall, and F1 score w.r.t anomalous class. As it is evident from the histograms, the proposed algorithm performs slightly better than the Isolation Forests with Isolation trees.

Figure 3 denotes that SVMs perform the worst with a precision score of 0.478. A single C4.5 decision tree achieves a precision of 0.864, and a logistic regression classifier, after hyperparameter tuning achieves 0.921 precision score. Naive isolation forests achieve 0.911 precision score, and our algorithm Hybrid i-Forest (isolation forest with C4.5 decision tree as base estimator achieves 0.935 precision score, which is the highest among all.

Figure 4 depicts that recall scores of all the algorithms are very close, while our algorithm slightly improves the recall score. The recall score of 0.894 achieved using a hyperparameter tuned logistic regression classifier. A single C4.5 decision tree achieves a recall score of 0.893, and SVM achieves a 0.932 recall score. Naive isolation forests achieve 0.936 recall score, and our algorithm Hybrid i-Forest (isolation forest with C4.5 decision tree as base estimator achieves 0.949 recall score, which is the highest among all.

Figure 5 shows that SVM performs the worst with an F1 score of 0.632. A single C4.5 decision tree achieves an F1 score of 0.879, and logistic regression classifier, after hyperparameter tuning achieves 0.907 F1 scores. Naive isolation forests achieve 0.923 F1 scores, and our algorithm Hybrid i-Forest (isolation forest with C4.5 decision tree as base estimator achieves 0.942 F1 score, which is the highest among all.

Fig. 4 Comparison of recall score w.r.t anomalous class

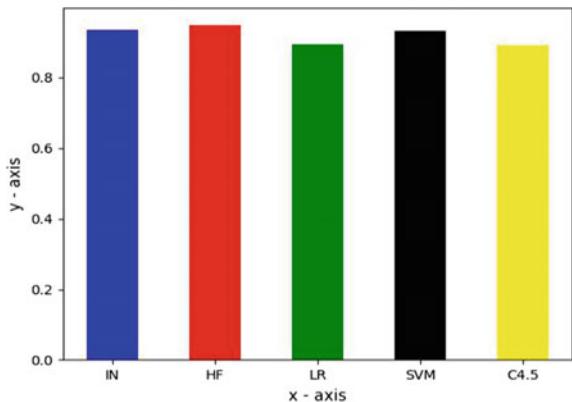
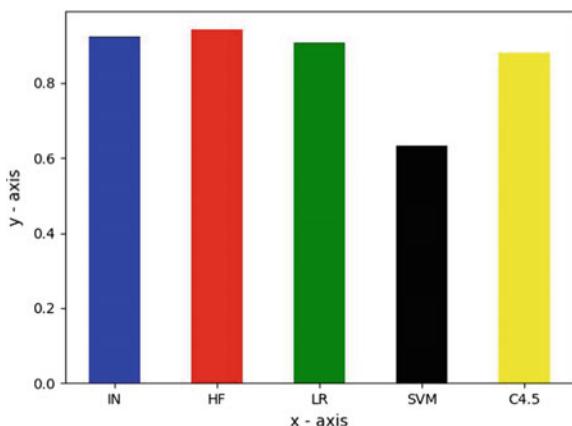


Fig. 5 Comparison of F1 score w.r.t anomalous class



5 Conclusions and Future Work

The proposed approach performs better than the Isolation Forest with isolation trees, which is generally the most used approach for outlier detection. The Isolation Forest works on unlabeled datasets, but this approach brings isolation forest's mathematics to labeled datasets as replacing isolation trees with C4.5 decision trees brings supervised learning into the picture. Several methods to calculate entropy exists, which tried in the future instead of Shannon's entropy for C4.5 decision trees. Anomaly detection using Tsallis entropy [21], (which can be generalized from Shannon's theorem [22] to combine the effects of Gini Impurity and Shannon's entropy in constructing decision trees may further improve the results.

Acknowledgements The authors thank the Department of Computer Science and Engineering, Indian Institute of Technology (ISM) Dhanbad, Government of India, for providing their research support.

References

1. Liu FT, Ting KM, Zhou Z-H (2008) Isolation forest. In: Eighth IEEE international conference on data mining. <https://doi.org/10.1109/icdm.2008.17>
2. Quinlan JR (1996) Improved use of continuous attributes in C4.5. *J Artif Intell Res* 4
3. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J*
4. Tan P-N et al (2019) Introduction to data mining. Global edition. Pearson Education Limited
5. Madhukar Rao G, Ramesh D (2016) Supervised learning techniques for big data: a survey, vol 9, IJCTA. International Science Press, pp 3811–3891
6. Mehrotra KG, Mohan CK, Huang HM (2017) Clustering-based anomaly detection approaches. In: Anomaly detection principles and algorithms. Springer, Cham
7. Siddiqui S, Khan MS, Ferens K (2017) Multiscale Hebbian neural network for cyber threat detection. In: International joint conference on neural networks (IJCNN). IEEE
8. Oku K et al (2006) Context-aware SVM for context-dependent information recommendation. In: Proceedings of the 7th international conference on mobile data management. IEEE Computer Society
9. Cawley GC, Talbot NLC (2010) On over-fitting in model selection and subsequent selection bias in performance evaluation. *J Mach Learn Res*
10. Berezinski P, Jasius B, Szpyrka M (2015) An entropy-based network anomaly detection method. *Entropy* 17(4)
11. Jiang SY, Yu W (2009) A combination classification algorithm based on outlier detection and C4.5. In: International conference on advanced data mining and applications. Springer, Berlin
12. Hssina B et al (2014) A comparative study of decision tree ID3 and C4.5. *Int J Adv Comput Sci Appl* 4(2)
13. Zhang C, Ma Y et al (2012) Ensemble machine learning: methods and applications. Springer Science & Business Media
14. Kim MJ, Kang DK (2010) Ensemble with neural networks for bankruptcy prediction. *Expert Syst Appl* 37(4):3373–3379
15. Hindman M (2015) Building better models: prediction, replication, and machine learning in the social sciences. In: The ANNALS of the American Academy of Political and Social Science, vol 659.1
16. Erdal HI, Karakurt O (2013) Advancing monthly stream flow prediction accuracy of CART models using ensemble learning paradigms. *J Hydrol* 477
17. Liu FT, Ting KM, Zhou Z-H (2012) Isolation-based anomaly detection. In: ACM transactions on knowledge discovery from data (TKDD), vol 6.1
18. Bay SD, Schwabacher M (2003) Mining distance-based outliers in near linear time
19. Breunig MM et al (2000) LOF: identifying density-based local outliers. In: ACM sigmod record, vol 29, no 2. ACM
20. Madhukar Rao G, Ramesh D, Kumar A (2020) RRF-BD: ranger random forest algorithm for big data classification. In: Computational intelligence in data mining, AISC, vol 990. Springer, Berlin
21. Ziviani A et al (2007) Network anomaly detection using nonextensive entropy. *IEEE Commun Lett* 11(12):1034–1036
22. Dos Santos RV (1997) Generalization of Shannon's theorem for Tsallis entropy. *J Math Phys* 38(8):4104–4107

Ranger Random Forest-Based Efficient Ensemble Learning Approach for Detecting Malicious URLs



G. Madhukar Rao and Dharavath Ramesh

Abstract The massive quantity of data generated from a variety of resources is hard to manage the process and analyze. Big data enables machine learning algorithms to find meaningful patterns and make accurate predictions. Big data analytics can be used in organizations to process and analyze the vast amount of data to find the insights of data, such as risks, threats, and incidents. These incidents create more security issues. A wide range of non-constructional activities happening in WWW requires the detection of malicious URLs for internet security. Malicious URLs or Websites are hosting spontaneous content and involve many users to become victims of scams. This paper presents ensemble learning-based, faster, and memory-efficient random forest algorithm for detecting malicious URLs. The proposed method proves scale best with the number of instances and variables.

Keywords Big data · Cybersecurity · Machine learning · Malicious URLs · RRF

1 Introduction

An enormous amount of data generated at an exceptional rate from various sources (e.g., Financial, Health, Government, Internet of things, social networks) leads to bigdata [1]. The accumulation of large and complex datasets is hard to process and manage with the traditional database techniques. The datasets are accessible in various formats such as standardized, semi-organized, and unstructured data in petabytes and even more [2]. Big data analytics used for many applications such as social networks, Cyber Security, and Health care. Machine Learning is a tool used to make predictions of data based on history [3]. Following five supervised

G. Madhukar Rao · D. Ramesh

Department of Computer Science and Engineering, Indian Institute of Technology (ISM),
Dhanbad, Jharkhand, India

e-mail: drramesh@iitism.ac.in

G. Madhukar Rao

e-mail: madhukar.iitism@gmail.com

learning models such as, (Random Forest, Artificial Neural Network and Decision Tree, Support Vector Machine, K-Nearest Neighbor) are used to compare the proposed variant, that described below. KNN is an instance-based learning technique that supports classification and regression [4]. K-nearest training samples used for testing the instances. In a classification problem, the class values of the selected neighbors treated as output. One of the powerful supervised learning techniques is a support vector machine (SVM), which used for both the classification and regression problems [5]. The SVM algorithm determines the set of support vectors (training examples) by marking these training examples to one of two classes, as a non-probability twofold linear classifier. C4.5 is a popular machine learning algorithm derived from the divide-and-conquer technique, and it used in several applications. Build the decision trees from training data; it requires the concept of information entropy and attribute selection based on the information gain. C4.5 is a simple, efficient technique and handles the missing values in the data [6]. Multiple models are combined to boost the performance of classification problems known as ensemble learning [7]. Ho et al. [8] proposed a method called random decision forests to overcome the limitations of the Decision Tree. The significance of this method is to build multiple decision trees from the given feature space using randomly selected subspaces. The random forest model is fast to construct and even faster to forecast the future of data. It integrates sampling, subspace approach, and ensemble approaches with many decision trees [9].

Random Forest uses the random selection of features and bagging sampling approach to collect the multiple decision trees with a controlled variation. However, the generated forests used for the further experimental method for the detection of variable interactions [10]. A neural network represented as a direct graph where nodes (neurons) are connected to map a set of input data into a set of related outputs. Three types of layers used in neural networks such as input nodes, hidden layers, and output nodes, which are known as Multilayer Perceptron. Each neuron is associated with a non-linear activation function (Softmax, hyperbolic, logistic) to fire the positive output [11]. Big data creates countless opportunities for cybersecurity analytics. Big data analytic techniques are beneficial for detecting different types of cyber-attacks, which include spamming, botnets, malware, denial of service phishing, and website threats. This paper presents a faster and memory-efficient Ranger Random Forest model for detecting malicious URLs. The remaining sections are as follows. The related works described in Sect. 2. The proposed methodology is described in Sect. 3. Description of Experimental setup is given in Sect. 4. Section 5 illustrates the comparative analysis followed by the conclusion section as Sect. 6.

2 Related Work

Kruppa et al. [12], classified web pages using their URLs to fetch the page content without any delay. In this model, classification features extracted from the segmented URL tokens. The authors have used high-quality segmentation and feature extraction to improve classifier performance. Ma et al. [13] present a method to perceive

malicious URLs from the lexical and host-based features. Their approach is capable of identifying the essential features from the given URL data to improve the classification performance. Liaw and Wiener [14] proposed a cost-effective malicious URL detection system, which uses two costs sensitive online active learning (CSOAL) algorithms. To achieve better performance, they evade the class imbalance problem in the detection of URLs and optimize the performance using two cost-effective algorithms. Naoum et al. [15], presents an ANN model with backpropagation for the development of the intrusion detection system. This paper utilizes the NSL-KDD dataset, and it has shown the enhanced performance with a detection rate of about 94.7%. Frank et al. [16] addressed the binary classification problem of malicious URLs and evaluated the performance with several well-known machine learning classifiers. They have used a public dataset containing 2.4 million URLs and 3.2 million features for evaluation. Random Forest (RF) and Artificial Neural Network (ANN) classifiers have shown the highest accuracy than the other classifiers. The implementation of Random Forest has its strengths and weaknesses. Zhao and Hoi [17] proposed a random forest using R, which is rich in programming and widely used, but it has not optimized for high dimensional datasets. Genuer et al. [18] projected a Random Forest algorithm for big data research, in which the authors have used five variants of RF to evaluate the performance using a big dataset containing millions of instances. They achieve superior results in terms of out-of-bag error, efficiency, and run time. A Ranger (**R**ANdom forest **G**Ene**R**ator) is a technique developed by Bandari and Suthaharan [19] to build the platform-independent and modular framework for considerable data analysis. The ranger package is available in R, and this achieves better results than the traditional RF.

3 Proposed Work

This section presents our work for the detection of malicious URLs. This approach uses a new variant of a random forest algorithm named Ranger Random Forest (RRF) for detecting malicious URLs. RRF used for faster implementation of the random forest without affecting the accuracy of the original RF classifier. RRF performs well for a high dimensionality dataset, where RF cannot apply. The working of RRF [20] as follows.

Ranger Random Forest(RRF) algorithm:

Stage1: For a given training data D, build a random forest.

Stage2: Counting the no of trees in test data used for calculating the weights of all training examples.

Stage3: Build a weighted random forest for the training statistics for every test instance, the usage of the weights calculated in Stage2. Estimate the result of the test example as usual.

Kruppa et al. [21] describe the process of how class probabilities and multiclass classification problems evolved. The Ranger Random Forest utilizes all the features

of Random Forest, and some of the extra features were added to it, which makes it perform well than the Random Forest classifier. Ranger can handle the different types of input data by determining the bottlenecks and optimizing the algorithms. There are two choices given to the users to operate in two modes, either in runtime or memory-efficient mode. The first approach sorts the values of features before and gets them by utilizing their index. The second approach says that unprocessed values are fetched and then arrange them at the time of splitting. The first approach used for run time mode and the second approach is used for small nodes in-memory mode. In RF and RRF, the vital factor is splitting the node, where *mtry* variables require to be deciding the splitting contestants. Ranger Random Forest algorithm uses two different approaches for node splitting. Drawing the *mtry* contestant was another bottleneck for every node with *bigmtry* values with many features. Efficient memory utilization achieved by avoiding the copies of original data along with a simple data structure. In classification trees, the Gini index used to measure the node impurity, and in regression trees, variance response measured. The node with less impurity would be the criteria for splitting the nodes for classification and regression of Random Forest. For Survival Random Forest trees, the log-rank test is the criterion for splitting. Out-of-bag data can be used as prediction error in Random Forest. The mean decrease accuracy calculated over the out-of-bag across all the predictions to find the variable importance. The mean decrease in accuracy calculated as

$$\text{Mean Decrease Accuracy} = \frac{\text{mean}}{\text{standard deviation}} \quad (1)$$

The ranger package implemented with two essential functions, namely, *ranger* () and *predict* (). Ranger used for building the trees and predict function used for predicting the responses for datasets. Ranger optimized for high dimensional data by utilizing less computation time with fewer memory requirements.

4 Experimental Setup

This section presents the experimental instances validated through the proposed ensemble methodology. The proposed variant and other models have implemented in R programming and use standard libraries such as ranger, random forest, nnet, cart, e1071, and class support to build the machine learning models for detection, which are capable of handling more massive datasets. The results for Random Forest and Ranger Random Forest uses 500 trees, KNN with $K = 3$, SVM uses radial basis function, and in Multilayer Perceptron (ANN), two hidden layers obtained. Experiments were carried out on a system with an Intel Core i7 processor, windows 10 OS, and 16 GB RAM. For experimental purposes, we use the dataset related to the Canadian Institute of Cyber Security [22]. In this, we evaluate four malicious URL datasets, namely; benign, phishing, malware, and spam URLs. Table 1 displays the

Table 1 Description of a dataset

Dataset	No. of instances	No. of features
Phishing	10,000	7
Malware	11,566	7
Spam	12,000	7
Benign	20,223	7

description of the utilized datasets for evaluating the proposed model. We have used training (70% of the sample data) and testing (30% of sample data) sets to evaluate the models.

5 Result Analysis

For a valid comparison, the following six models (RRF, RF, ANN, DT, SVM, and KNN) evaluated. Run time, Memory Consumption, and accuracy metrics used for evaluating the six classifiers. The statistical analysis of all the models is described with tables and figures, as shown below.

Tables 2 and 3 show the run time and memory consumption of the proposed method using four different URL datasets. The classification methods sorted according to the run time and memory consumption across different methods. The first observation is that all the methods achieve high accuracy scores with minimal difference. The second observation is that the proposed method records less computation time and less memory utilization to achieve high accuracy than the other methods. Figures 1 and 2 depict the visualization of different performance measures per classification method and phishing URL dataset.

The comparison based on runtime and memory consumption using the malware URL dataset depicted in Figs. 3 and 4.

Table 2 Comparison of evaluated models using phishing and malware datasets

Approach used	Phishing			Malware		
	Run time elapsed	Memory	Accuracy	Run time elapsed	Memory	Accuracy
RRF (prop.)	0.19	2.62	99.35	0.55	2.20	98.96
AN	0.30	10.60	99.28	0.72	12.50	98.70
RF	0.52	62.02	99.20	1.59	269.25	98.55
DT	1.20	95.06	98.80	2.90	350.28	98.02
SVM	2.50	120.36	98.02	4.20	421.06	97.82
KNN	3.80	180.40	97.79	6.05	513.02	97.01

Table 3 Comparison of evaluated models using spam and benign datasets

Approach used	Spam			Benign		
	Run time elapsed	Memory	Accuracy	Run time elapsed	Memory	Accuracy
RRF (prop.)	0.06	3.17	98.87	0.69	4.14	98.40
AN	0.25	15.35	98.65	1.40	19.50	98.25
RF	0.95	193.22	98.50	3.47	385.22	98.03
DT	2.05	255.6	98.00	5.25	450.55	97.68
SVM	3.95	312.24	97.82	6.50	550.35	97.02
KNN	5.22	400.16	97.13	8.40	650.65	96.98

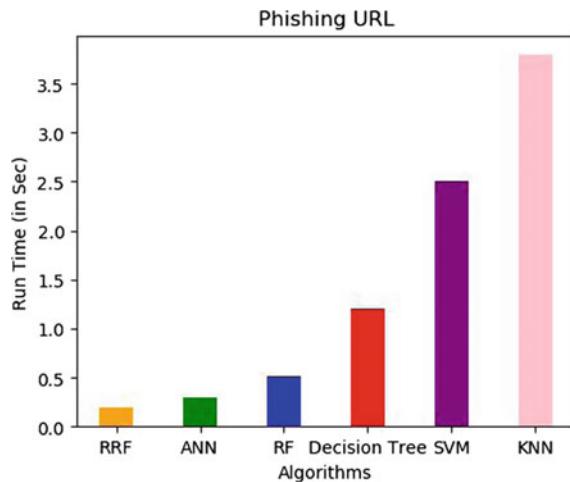
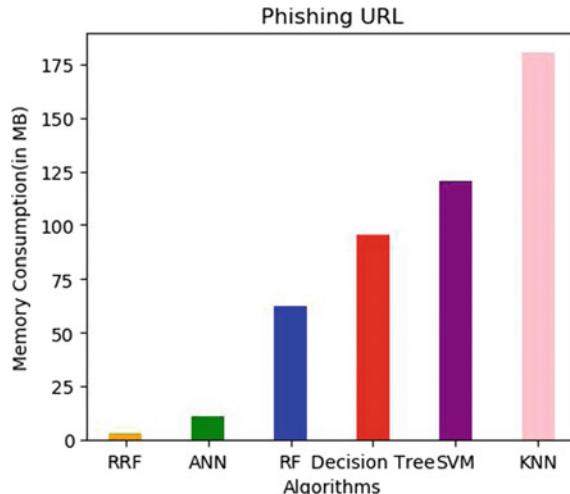
Fig. 1 Runtime comparison for phishing URL dataset**Fig. 2** Memory utilization comparison for phishing URL dataset

Fig. 3 Runtime comparison for malware URL dataset

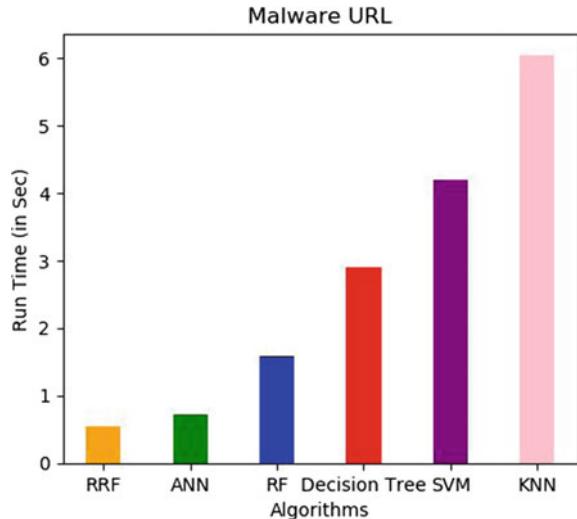
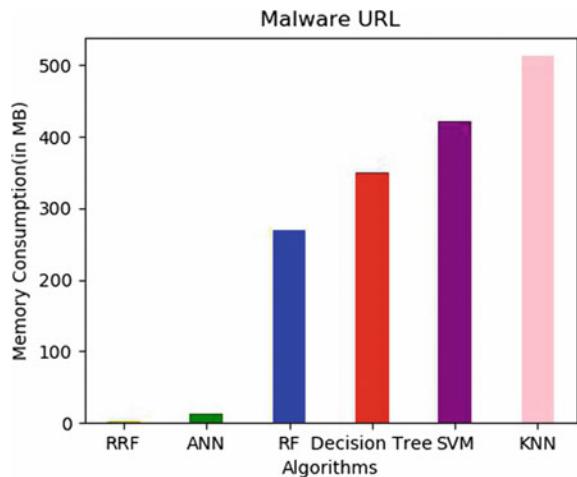


Fig. 4 Memory consumption comparison for malware URL dataset



The comparison based on the factors mentioned above using the spam URL dataset depicted in Figs. 5 and 6.

Ranger outperforms the other methods for the benign URL dataset. It explores that ranger used for high volume and high dimensionality datasets. Figures 7 and 8 depict the visualization of the performance of all methods. From this, it observed that the proposed method performs well across the different datasets and achieves efficiency.

Fig. 5 Runtime comparison for spam URL dataset

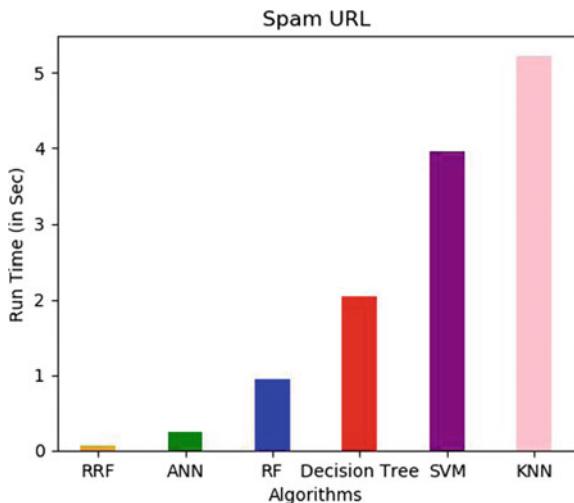
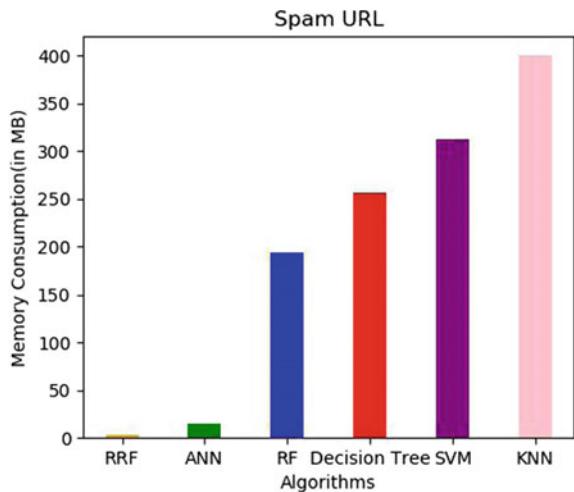


Fig. 6 Memory consumption comparison for spam URL dataset



6 Conclusion

In this paper, we have explored cybersecurity problems in terms of big data processing and how big data machine learning used to analyze cybersecurity attacks. Traditional tools and techniques are not much suitable to analyze the big data streams. Big data creates countless opportunities for cybersecurity analytics.

This paper shows the machine learning approaches for cybersecurity analytics to overcome challenges such as data volume and scalability. The proposed method minimizes the computation time and uses less memory to detect malicious URLs

Fig. 7 Runtime comparison for benign URL dataset

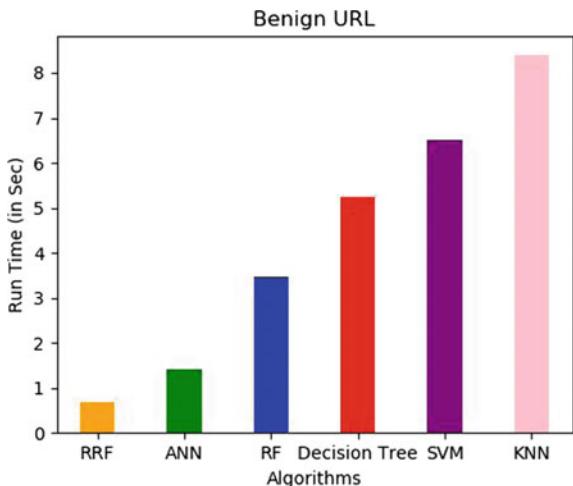
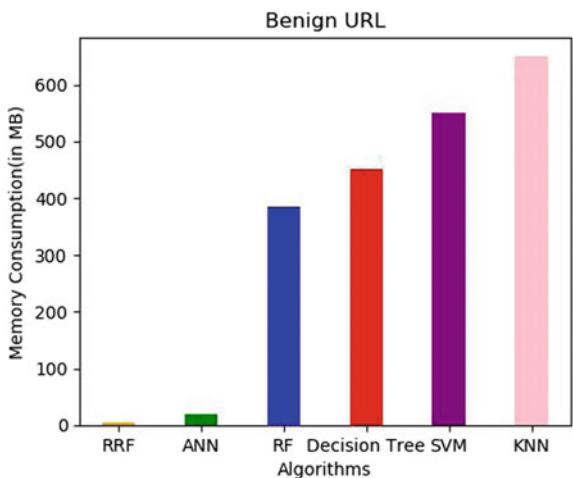


Fig. 8 Memory consumption comparison for benign URL dataset



efficiently. In recent years big data machine learning enabled to derive the intelligence and meaningful information from real-time data, which strongly needed for cybersecurity analytics.

Acknowledgements This paper is an extended version of the paper published in Computational Intelligence in Data Mining, Advances in Intelligent Systems and Computing 990, Springer. The authors thank the Department of Computer Science and Engineering, Indian Institute of Technology (ISM) Dhanbad, Government of India, for providing their research support.

References

1. Hashem IAT, Yaqoob I, Anuar NB, Mokhtar S, Gani A, Khan SU (2015) The rise of “big data” on cloud computing: review and open research issues. *J Inf Syst* 47:98–115
2. Madhukar Rao G, Ramesh D (2016) Supervised learning techniques for big data: a survey. *IJCTA, Int Sci Press* 9:3811–3891
3. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning. Springer, New York
4. Bandari S, Suthaharan S (2012) Intruder detection in public space uses suspicious behavior phenomena and wireless sensor networks. In: Proceedings of the 1st ACM international workshop on sensor-enhanced safety and security in public spaces at ACM MOBIHOC, pp 3–8
5. Hearst MA, Dumais ST, Osman E, Platt J, Scholkopf B (1998) Support vector machines. In: Intelligent systems and their applications, vol 13(4), IEEE, pp 18–28
6. Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kauffman Publishers
7. Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Mach Learn* 51(2):181–207
8. Ho TK (1995) Random decision forests. In: Document analysis and recognition, proceedings of the third international conference, Montreal, Quebec, Canada, vol 1. IEEE, New York City, NY, pp 278–282
9. Breiman L (2001) Random forests. *Mach Learn* 45(1):5–32, (2001)
10. Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. *Syst Sci Control Eng Open Access* J 2(1):602–609
11. Rosenblatt F (1961) Principles of neurodynamics. Perceptrons and the theory of brain mechanisms. DTIC document, Technical report
12. Kan MY, Thi HON (2005) Fast webpage classification using URL features. In: Proceedings of the 14th ACM international conference on information and knowledge management. ACM, pp 325–326
13. Ma J, Saul LK, Savage S, Voelker GM (2009) Beyond blacklists: learning to detect malicious websites from suspicious URLs. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 1245–1254
14. Zhao P, Hoi SC (2013) Cost-sensitive online active learning with application to malicious URL detection. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 919–927
15. Naoum RS, Abid NA, Al-Sultani ZN (2012) An enhanced resilient backpropagation artificial neural network for intrusion detection system. *Int J Comput Sci Net Secur* 12(3):11–16
16. Vanhoenshoven F, Nápoles G, Falcon R, Vanhoof K, Köppen M (2016) Detecting malicious URLs using machine learning techniques. In: 2016 IEEE symposium series on computational intelligence (SSCI). IEEE, pp. 1–8
17. Liaw A, Wiener M (2002) Classification and regression by random forest. *R News* 2(3):18
18. Genuer R, Poggi JM, Tuleau-Malot C, Villa-Vialaneix N (2017) Random forests for big data. *Big Data Res* 9:28–46
19. Wright MN, Ziegler A (2015) ranger: a fast implementation of random forests for high dimensional data in C++ and R. arXiv preprint [arXiv:1508.04409](https://arxiv.org/abs/1508.04409)
20. Madhukar G, Ramesh D, Kumar A (2020) RRF-BD: ranger random forest algorithm for big data classification. In: Computational intelligence in data mining. Springer, Singapore, pp 15–25
21. Kruppa J, Liu Y, Biau G, Kohler M, König IR, Malley JD, Ziegler A (2014) Probability estimation with machine learning methods for dichotomous and multicategory outcome: theory. *Biometric J* 56(4):534–563
22. <https://www.unb.ca/cic/datasets/url-2016.html>

Smart Camera for Traffic Control by Sing Machine Learning



Priya Tiwari, Santosh Jagtap, and Dattatray Bade

Abstract In today's growing population, vehicular congestion is exponentially leading to long traffic jams on roads. In India, traffic conditions of roads are not similar to those of foreign countries; therefore, detection, counting, and classification of vehicles in real-time have become difficult. In the world's growing population, traffic is one of the biggest problems in megacities. Hence, there is a need to understand traffic density conditions on the roads in real time for the betterment of the city's traffic congestion. The need for optimization in traffic control and simulation is increasing day by day due to different reasons for traffic congestions viz uncontrolled demand, considerable delay of red light, insufficiency of roads capacity, etc. The dissertation work presents the method of traffic optimization by using a machine learning approach. The data utilization is optimum. The algorithm will run on the system on chip (SoC) to achieve high speed and throughput without any additional overhead related to deploying the algorithms in the cloud. It can also help people for traveling safely with the reduction of waiting time and fuel consumption. Hence, the data which has been collected for real-time traffic management can also be used for making new plans and analyzing the situation.

Keywords Machine learning (ML) · Artificial intelligence (AI) · Convolutional neural network (CNN) · Switching algorithm · Traffic congestion · Intelligent traffic light controller (ITLC)

1 Introduction

India is one of the developing countries where rural to urban population is shifting rapidly, and this change is reflected on lives of urban population. In India, population has been increased from 27.8% (286 million) in the year 2001 to 31.2% (377 million)

P. Tiwari (✉) · S. Jagtap · D. Bade

Electronics and Telecommunication Engineering, Vidyalankar Institute of Technology, Mumbai, India

e-mail: tiwaripriya0101@gmail.com

in 2011, and it is predicted that it can grow to 40% by 2030 and more than 50% by 2050 [1]. Hence, the impact of increased population in cities directly affects the infrastructure management and various other services provided by government, and among them traffic congestion is major issue to dealt with.

For most cities, it is not possible to build new roadways, rail systems, or ports. However, capacity can be increased by embedding sensors and location technology into the transportation infrastructure and using cloud-based real-time analytics to reduce congestion and transport times. The main objective is to develop such system, so that whole transportation system can be managed intelligently across all modes of transport [2].

According to their needs and traffic situation, every city can take good advantage of sensor network and other technical means to change the traditional transport system and establish the smart traffic management system, including adaptive traffic signal (automatic controlling traffic lights with changing the flow of vehicles) control system, traffic controlling of urban area, and so on.

The smart traffic management system can be achieved when it will be combination of various parameters, for example, urban planning, operations management, construction details, and many more supporting smart urban development. Using urban IOT along with traffic monitoring, air quality and noise monitoring is possible [3].

Nowadays, many cities have setup of cameras for traffic monitoring system, such cameras have been deployed in many cities that consume less power, and they provide large amount of information required. It has been realized that sensors are the devices used for sensing purpose, and GPS is also installed on modern vehicles to monitor the traffic density on roads, having a combination of acoustic sensors and air quality along with them [4].

Many attempts were made to transform a traditional traffic light into smarter one, but these attempts are either sensor based or manual. The information collected by these devices is very advantageous for governmental body of the city and citizens to maintain the traffic discipline.

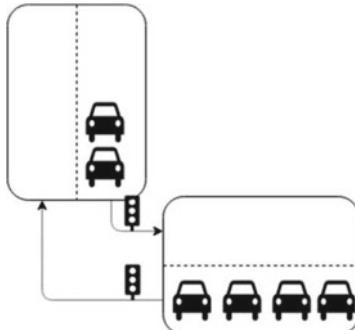
Introducing microcontroller for monitoring traffic lights helps in applying algorithmic approach to traffic light to make it smarter in real time. Image processing algorithm such as Haar-cascade and subtraction of background method is used for dealing with different conditions. These algorithms need a training period and could be made more accurate. This approach will allow the traffic timer to change in real time.

In order to give a real-time view of the upcoming traffic light, region of influence concept was created, which allows user to fetch data and statistics from the upcoming traffic light directly and to decide better route. By one click, user can receive all the statistics and data along with real-time images from the traffic light in layman language traffic light is directly communicating user [5].

2 Methodology

2.1 Smart Traffic Light

Consider the following scenario:



Let us consider that after a traffic light turns green from red, it takes approximately X_a seconds for the car to cross the lane and reach the other side of the road. Here, the time required is dependent on the position of the car.

In this scenario, let us consider two cases:

Case 1: Static Traffic Light

Let us assume that the traffic light is going to remain ON (green) for 10 s and OFF (red) for 10 s. Now, let us say that the traffic light A just turned ON, thereby cars from lane A are free to move. Let us assume 3 s for the movement of a single car to the other side of the road (ignoring the extra positional time considerations). So, for a total of two cars, it will take 6 s. In this case, the signal is still ON for an extra 4 s. After this, the traffic light B turned ON, thereby cars from lane B are free to move. Since the signal is ON for 10 s, only three cars from this lane will be able to pass within that time frame. The signal from lane A turns ON again for 10 s while the one car at lane B must wait. Finally, when the signal at lane B turns ON, the last car will be able to cross the road. The total time required for all the cars to cross the several roads will be $3 + 3 + (10 + 3) + (10 + 3) + (10 + 10 + 3) = 68$ s.

Case 2: Smart Traffic Lights

Let us assume that our smart traffic light adjusts the ON time based on the traffic intensity. So, for lane A, detecting low traffic, it chose duration of 8 s, and for lane B, detecting high traffic, it chose duration of 12 s. Now, let us say that the traffic light A just turned ON, thereby cars from lane A are free to move. We have assumed 3 s for individual car movement across the road. So, for a total of two cars, it will take 6 s. In this case, the signal is still ON for an extra 2 s. After this, the traffic light B turned ON, thereby cars from lane B are free to move. Since the signal is ON for 12 s, all the cars will be able to move across the road within this time frame. Therefore, the total time required for all the car to cross the several roads will be $3 + 3 + (8 + 3) + (8 + 3) + (8 + 3) = 50$ s.

The above scenario can easily be more optimized, but let us a little worse scenario. Let us say that for lane A, detecting low traffic, it chose a duration of 9 s and for lane B, detecting high traffic it chose duration of 11 s. In this case, the change in the duration from the static scenario is slim. Now, let us say that the traffic light A just turned ON, thereby cars from lane A are free to move. We have assumed 3 s for individual car movement across the road. So, for a total of two cars, it will take 6 s. In this case, the signal is still ON for an extra 3 s. After this, the traffic light B turned ON, thereby cars from lane B are free to move. Since the signal is ON for 11 s, only three cars from this lane will be able to pass within that time frame. The signal from lane A turns ON again for 10 s while the one car at lane B must wait. Finally, when the signal at lane B turns ON, the last car will be able to cross the road. Therefore, the total time required for all the cars to cross the respective roads will be $3 + 3 + (9 + 3) + (9 + 3) + (9 + 3) + (9 + 9 + 3) = 63$ s.

Thus, it can be easily seen that due to the dynamic nature of the smart traffic lights, there is going to be an optimization of the signal duration leading to more smoother traffic and saving of the commuter time.

3 Real-Time Implementation

In the proposed method, real-time scenario is considered. In this implementation, six hours of videos have been collected from the traffic control department of Thane district. The videos collected from different-different roads of Thane showing all possibilities of traffic density. Depending on road conditions, the density of traffic is divided into three situations that are low, medium, and high. Applying machine learning algorithm, traffic density can be estimated. The machine learning algorithm works on running video and simultaneously shows the numerical representation of traffic congestion, which is represented as follows:

- Low—0
- Medium—1
- High—2

In the below images, a situation of the road is considered when a number of vehicles are more to check the density of roads. Hence, machine learning algorithms are applied to the images for density estimation with various steps. Initially, a standard image is captured, as shown in Fig. 1. In Fig. 2, image processing starts, and on a standard image, gray scaling is done. After the grayscale region of interest is taken, and the image is cropped, as shown in Fig. 3. Then, in Fig. 4, the cropped image of road enlarges, and Figs. 5 and 6 show actual density estimation by separating the black and white pixels by thresholding.

Hence, the given data is processed for separating the black and white pixels to estimate the density of vehicles on the road. Below, data shows that a number of black pixel is more from the images processed above; therefore, traffic density is high that is above 3.

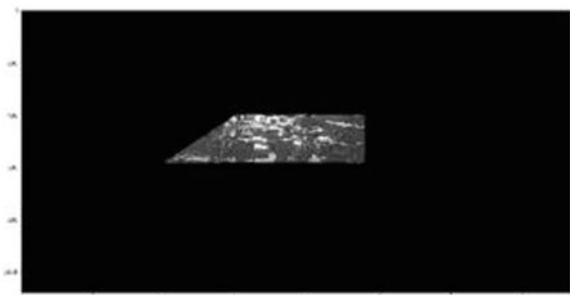
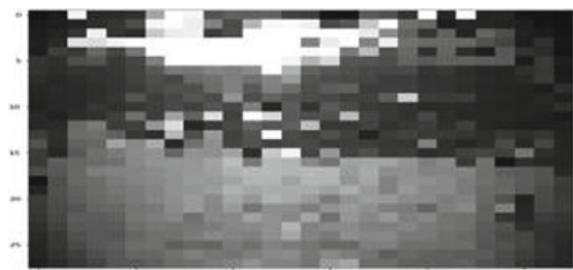
Fig. 1**Fig. 2****Fig. 3****Fig. 4**

Fig. 5**Fig. 6**

White Pixels Count: 28949
Black Pixels Count: -14249
Traffic Density: 3.031651343953962

4 Results

Comparison between training and testing of accuracy; (Fig. 7; Table 1).

By the end of the machine learning training task, a training accuracy of 97.5% and testing accuracy of 98.0% are observed, which clearly shows that our model performed very well on the dataset.

Comparison between training and testing of loss (Fig. 8; Table 2).

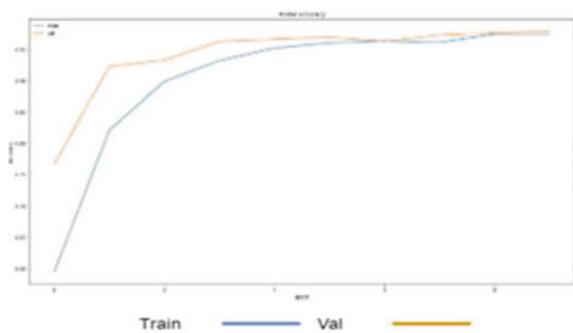
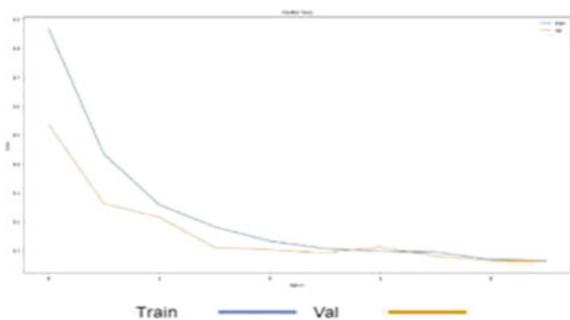
Fig. 7

Table 1 Training and testing data of accuracy

Training accuracy	Testing (validation) accuracy
0.5941667	0.7666666507720947
0.8208333	0.9233333468437195
0.89916664	0.9333333373069763
0.9325	0.9633333086967468
0.9525	0.9666666388511658
0.9608333	0.9700000286102295
0.96416664	0.9633333086967468
0.96166664	0.9733333587646484
0.975	0.9766666889190674
0.975	0.9800000190734863

Fig. 8**Table 2** Training and testing data of loss

Training loss	Testing (validation) loss
0.8692614555358886	0.5367411835988363
0.4346064607302348	0.26262855251630146
0.2584100373586019	0.21539190967877705
0.18211153556903203	0.11062057733535767
0.13252320845921833	0.10290105948845546
0.10641482119758924	0.09115208889047305
0.097242996742328	0.1123387015859286
0.09391615514953931	0.08037679026524226
0.06898216631263494	0.06533852199713389
0.06454030486444633	0.06337897154192129

By the end of the machine learning training task, a training loss of 0.0645 and testing loss of 0.0633 is observed, which is very practical and healthy to work on a model in real-life scenarios.

The observed graphs are pretty smooth, so that we can decline the chances of any overfitting.

5 Conclusion

In the proposed method, the machine learning algorithm is used to tackle all the existing problems of traffic congestions on roads. The machine learning algorithm is performed on real-time videos captured by CCTV footage of roads at the junctions. The proposed solution has an added advantage since it will be using a machine learning algorithm which gets better every time by learning through experience or multiple data processing and are able to make decisions by considering various parameters. It gives very good accuracy and also minimizes the loss of data, and hence, the optimum results are obtained. The main focus of the algorithm developed is for the switching of traffic lights according to the vehicles present on the road, the type of vehicle on the road, and hence, reduction of the traffic congestion is possible by automatic real-time data processing done by machine providing the safe drive to people and reducing consumption of fuel.

6 Future Scope

In the modern era of developing countries, vehicular congestion has become a serious issue to work on, and hence, the automatic system is much needed to control and improve the traffic conditions. Hence, the proposed work makes the traffic management simpler. It will also provide significant data that can be used for future planning of roads for analysis.

Supplementary benefits also include city-wide environment monitoring with real-time mapping, which can be digested and utilized by various governing bodies. Further, this method can be used to convert all traditional traffic lights to smarter ones to get the maximum benefit of machines by reducing human efforts.

References

1. Gupta K, Hall RP (2017) The Indian perspective of smart cities. Prague
2. Harmon RR, Castro-Leon EG, Bhide S (2015) Smart cities and the internet of things. Portland State University, Portland, USA
3. Su K, Li J, Fu H (2011) Smart city and the applications. School of Computer Wuhan University Wuhan, Hubei

4. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. *IEEE Internet Things J* 1(1):22–32
5. Dubey A, Akshdeep Rane S (2017) Implementation of an intelligent traffic control system and real time traffic statistics broadcasting. In: International conference on electronics ICECA. Army Institute of Technology, Pune

Systematic Review on Full-Subtractor Using Quantum-Dot Cellular Automata (QCA)



Sri Sai Surya, A. Arun Kumar Gudivada, and Durgesh Nandan

Abstract In earlier days, we used a CMOS-based subtractors; these are having lower specifications in terms of power dissipation, switching speed, and area. In this era, we are having nanotechnology which is of range 10^{-9} , to reduce the overall specifications of the circuit. For this, we are having many tools for designing of circuits instead of using bulky circuits, and we have QCA Pro tool for the designing of different layouts and architectures of a particular circuit. In this paper, different subtractors are compared, these designs are because of quantum-dot cellular automata (QCA), and also the basic concept of the full-subtractor is discussed. QCA is emerging nanotechnology in digital circuits. QCA can also be used in computational architecture. This tool is to check the power dissipation and several clock delays, and the area is verified. The drawbacks of CMOS technology like area, switching speed, complexity are overcome by QCA technology.

Keywords Full-subtractor · Half-subtractor · CMOS · QCA · Actin · ALU · DSP · BCD · DKG · RFAS · MV gate · Intricacy · Latency

S. S. Surya · A. Arun Kumar Gudivada

Department of Electronics and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India

e-mail: sadhanalasurya@gmail.com

A. Arun Kumar Gudivada

e-mail: arunkumarg@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India

e-mail: durgeshnandano51@gmail.com

1 Introduction

The basic arithmetic logic circuits, adder and subtractor, are used in digital computational works. The performance of digital systems can directly affect the operation of adder and subtractor circuits. In this direction, several QCA digital approaches have been devised to increase the performance of adder and subtractor circuits [1].

- Toffoli and Fredkin designed and implemented gates having the least delay, circuit complexity, and area using QCA. Present QCA technology will vanquish the issues like scaling and dimensions compared to the older CMOS innovations. But, in the QCA design, data demolition causes squandering of heat and uses large power to design QCA.
- Subsequently, to obtain low control structures and to control the data stream of QCA design, reversible logic is used. The QCA subtractor which is reversible logic is completed studied, and at that point, we first talk about a proposed reversible structure which is dependent on three and five-input large parted gates, and also powerful one-layer hybrid plan.
- The new full-subtractor improves the plans likely matrices adjustments, cell requirements, delay, circuit complexity, etc. The QCA-based subtractor uses polarizing of electrons, whereas CMOS uses voltage levels as binary states, i.e., logic HIGH (1) and LOW (0). The construction of chip can be possible with a minimum number of XOR gates since for any mathematical design modulo-2 addition is used most, and this function is accomplished using XOR gated structure, thereby cell area is reduced in a great extent [2].
- For the processors having high working speeds, the 8-bit reversible adder/subtractor can be efficiently utilized. This circuit applies profoundly to fast ALU processors [3]. The reversible gate in this exploration will be used in different digital signal processing applications where half-adder and half-subtractor are going to work at the same time instant.

2 Literature Review on QCA Subtractor

In 2010, S. Karthigai Lakshmi et al. have reported different subtractors designed using quantum-dot cellular automata (QCA) and explained the basic concept of QCA. They suggested that it has the potential for attractive features than a transistor-based innovation [4]. By using the QCA technique, we can configure fascinating computational models. Half-subtractor and full-subtractors are organized and regenerated using QCA designer, rather than that of attractive features which are reported by S. Karthigai Lakshmi, to produce a relatively simple and efficient implementation.

In 2012, using QCA, Nimit Gupta and Nilesh Patidar et al. proposed reversible logic plan. Instead of using normal logic gates like AND, OR, NOT, etc., they used reversible (DKG) logic gate. This reversible gate acts as a conventional adder-subtractor. The simulation result of this gate shows that there is a high-speed rate,

small in size, and low power consumption. And also these outcomes demonstrate that it is very superior to the transistor-based variant [5]. Meanwhile using in large circuits, there is a design of QCA reversible subtractor by using subtractor and XOR gate with proper wiring which is reported by Ali Shahidinejad [6].

In 2013, Bahram Dehghan represented that half-subtractor circuits are delivered by utilizing a basic recursive course of action of two-inputs, two-outputs AND–NAND as well as OR–NOR logic gates and MV gates for full-subtractor feasible with QCA innovation, and therefore, utilizing MV gates, the digital plan structure is energy-saving structure of logic circuits and extraordinary advance toward plan of complex logic circuits [7]. With adjustable cell count and area, Vikram kumar Pudi presented productive QCA plans of single-bit and multi-bit subtractors [8].

In 2014, Prameela Kumari et al. implemented subtractor in which the QCA cells are arranged vertically and their components are in a stacked format. By this arrangement, the cell requirement for the interconnection and error detection time is decreased. The structures of adders, subtractors, ripple adder's, ripple-cum subtractor use vertically stacked QCA cell's arrangement, and by using this thought of structures Prameela Kumari et al. implemented a QCA subtractor. These structures are implemented with the least number of cells and timing zones. After the simulation, the results show that there is a most decrement in the number of cells utilized, i.e., 77% for general structure and 90% for interconnection of cells [9]. Nagamani et al. reported about adder–subtractor which distributes the overflow logic of BCD subtractor using negative control lines. They also stated that adders and subtractors are important structural parts of any processor. BCD subtractor reduces quantum cost, delay, and gates count and fasten the subtractor. For the next generations, this subtractor shows a way for the approach of an increase in the field of reversible processing [10]. Afterward, rotation of cells is reduced, and the 8-bit circuit can be utilized in bigger circuits, for example, arithmetic-logical unit (ALU), and later they also designed a QCA designer software and also verified the proposed structure [11]. With single-layered, obtaining the low number of garbage outputs and without requiring any rotated cells, Elham Taherkhani et al. proposed a QCA-based novel reversible full adder–subtractor circuit. This circuit has better design plans like cell count, area, and energy dissipation by 50% when compared with present QCA-based single-/multi-layered reversible adders [12].

Later, Rajon Bardhan et al. designed and suggested a 32-bit adder/subtractor circuit using QCA 3-dot cell, this circuit improves 90% of cell count and 99% of the area than existing circuits, and it has the ability of ultralow power consumption with a higher speed capability [13]. Afterward, to use at more complex circuits and to use at a molecular level, Actin is used as QCA cells, and it is responsible for the movement of the cells in QCA [14]. Here, QCA uses polarization of electrons for the representation of binary states, whereas CMOS uses voltage levels as the binary states.

In 2016, to design arithmetic circuits, C. Labrado and H. Thapliyal proposed a single-layered QCA structure. When compared to a single-layered QCA plan, proposed structures of a full-adder, four-bit ripple carry adder, full-subtractor, and four-bit ripple borrow subtractor using QCA have higher specifications [15]. Later

Md. Abdullah-Al-Shafi suggested an XOR layout of subtractor, generally, parity generator and parity checkers having low power consumption, and similarly, the suggested XOR layout can be applied to low power consummating comparators. It is used in upcoming computing like quantum computers, nanotechnology, and low power architectures [1].

In 2017, Not only subtraction but also several operations such as addition, subtraction, multiplication, and division are performed using nanocalculator. This calculator is structured by QCA innovation utilizing QCA designer. In this nanocalculator, a solitary molecule in a quantum dot assumes the responsibility for activities. This architecture is intended to improve the framework execution level with increasingly math tasks contrast with existing CMOS structure strategy. Using QCA designer 2.0.3, Ramanand Jaiswal et al. designed a five-input majority gate. The simulation results show that it has a minimum cell area with less cell count and clock cycles. To determine energy dissipation, they used the QCA Pro tool [16]. The designed QCA full- and half-subtractors occupy less area in terms of $0.06 \mu\text{m}^2$ and $0.10 \mu\text{m}^2$ when compared to conventional subtractors. These subtractors have 38 and 83 cells while the conventional half- and full-subtractors have 93 and 122 cells, respectively.

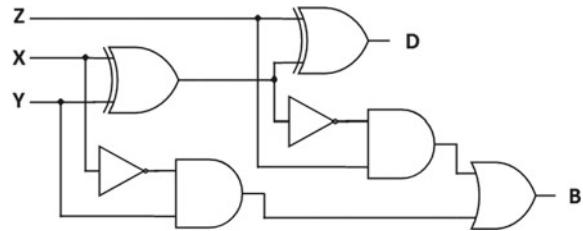
In 2018, Firdous Ahmad et al. proposed several new reversible subtractors, they are 3×3 new reversible gate and modified Feynman gate, and they structured an ideal reversible single-layer full adder–subtractor circuit. These gates are better regarding the reduced area and clock delays over existing traditional designs and verified using the QCA designer tool [17]. In this present era, atomic- or molecular-level QCA implementation is developed. Ritika Jain et al. presented a new structure of reversible adder and subtractor gate. Using this gate, addition and subtractions are carried out without using any reversible circuits which were never performed in recent years [18].

In 2019, to design a full adder–subtractor with the least number of cells, Saeid Zoka et al. designed a one-bit full-adder with the least delay in cells using QCA designer tool. This structure uses a single layer without having any fixed cells. And also, there is an improvement in the design of architecture with 44 QCA cells in full-adder and just 83 QCA cells in full adder–subtractor. It has a minimum cell area of $0.09 \mu\text{m}^2$ which seems to be better than the previous subtractors area [19].

2.1 Description of the Subtractor

The full-subtractor performs subtraction between two or more bits and gets borrow of the lower to organize in some condition. It has an output which shows that 1 has been borrowed. Here, the basic full-subtractor has Z, Y, and X as inputs and borrow and difference as the outputs. For the below subtractor, we need two XOR gates, one OR gate along with two NOT gates and two AND gates which is shown in Fig. 1. The circuit is designed using logic gates, but in later era, we used QCA-based full-subtractor and we can design any gate using quantum-dot cellular automata. It reduces number logic gates, power consumption, circuit complexity, delay, etc. This

Fig. 1 Basic logic diagram of full-subtractor

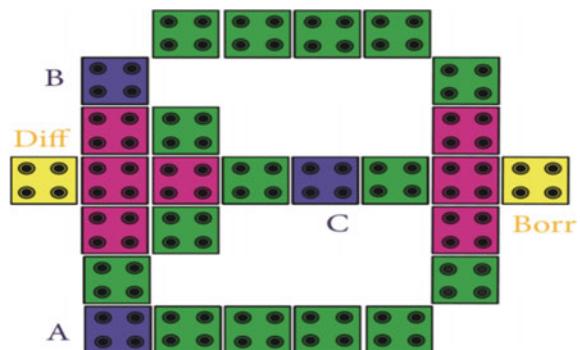


automaton is used in more digital computational systems. Here, we have different inputs, and for these input gates, it will produce different outputs. Figure 1 shows the basic logic diagram of full-subtractor using logic gates. By using Karnaugh maps, we will calculate borrow and difference equations. The logic “0” is nothing but a low supply voltage or LOW input voltage, and logic “1” is logic HIGH voltage. If the inputs esteem is $Z = 0$, $Y = 0$, and $X = 0$ at that point, the output esteems for difference (Diff) and borrow (Borr) are both 0. For input esteems $Z = 1$, $Y = 1$, and $X = 1$, the output esteems for difference (Diff) and borrow (Borr) are both 1.

2.2 Architecture of QCA-Based Full-Subtractor

QCA designer is the simulation engine for the QCA estimations. Each cell in QCA can be performed in four ways. The extension of the QCA cell is 18×18 with 5 nm thickness quantum dots. For neighboring cells, center-to-center space is 20 nm. The below outline has 27 QCA cells. There is polarization in the QCA cells, and it is nothing but the systematic condition of a cell. In cell polarization, $P = -1$ to be regarded as logic “0,” and $P = +1$ to be regarded as logic “1.” Three cells with marks A, B, and C are input cells and outputs as diff, borr. This architecture is proposed by using the XOR gate-based full-subtractor [2] (Fig. 2).

Fig. 2 Layout of QCA subtractor [2]



3 Implementation Outcomes

Table 1 shows the different comparison outputs of cell intricacy, extent, latency, and wire crossing of proposed designs. Table 1 shows the five proposed architectures outcomes, and these designs have different value ranges where extent is in terms of micrometers. Most of the designs have wire crossing. But Md. Abdullah-Al-Shafi et al. designed without cross wiring and with latency of two clock zones, the extent of $0.44 \mu\text{m}^2$ along with 27 of cell intricacy in 2018 which is better compared to all the proposed ones [2].

3.1 Complexity Study

It deals with the area utilized, latency (time interval between stimulation and response), and the quantity of XOR gates used, and so on. Table 2 demonstrates the complexity study which was already proposed [2]. Here, there is a great reduction in the cell area of $0.0087 \mu\text{m}^2$ and also decrement in the area usage of the circuit with only 29% of the total area along with covered area $0.02 \mu\text{m}^2$. Thus, it shows that this subtractor has a wide range of usage in digital computations. All the subtractors have different specifications, but only one of full-subtractor [2] specifications are studied in this paper.

Table 1 Implementation outcomes of the subtractors

QCA architecture	Cell intricacy	Extent in μm^2	Latency	Wire crossing
Presented in [18]	136	0.168	7	Yes
Presented in [6]	52	0.039	2	Yes
Presented in [3]	233	0.132	8	Yes
Presented in [19]	272	0.44	9	Yes
Presented in [2]	27	0.02	2	No

Table 2 Complexity study

Outline	XOR gates used	Cells used	Covered area (μm^2)	Cell area (μm^2)	Area usage (%)	Latency
Full-subtractor [2]	1	27	0.02	0.0087	29	2

4 Conclusion

The basic concept of subtractor is discussed, and the full-subtractor truth table is shown. Different proposed designs are compared to likely implementation outcomes, complexity study. Thus, we conclude that the full-subtractor in the QCA approach is better than that of other approaches. Transistor-based is time-consuming and complex. Architecture for the subtractor using QCA is discussed. In further, there are many designed architectures of QCA subtractor, but in my point of view with lesser cell intricacy and complexity, the discussed architecture [2] is better in all conditions.

References

1. Abdullah-Al-Shafi M, Bahar AN (2016) QCA: an effective approach to implement logic circuit in nanoscale. In: 2016 5th international conference informatics, electronics vision, ICIEV 2016, pp 620–624. <https://doi.org/10.1109/ICIEV.2016.7760076>
2. Abdullah-Al-Shafi M, Bahar AN (2018) An architecture of 2-dimensional 4-dot 2-electron QCA full adder and subtractor with energy dissipation study. *Act Passiv Electron Compon.* <https://doi.org/10.1155/2018/5062960>
3. Abedi D, Jaberipur G, Sangsefidi M (2015) Coplanar full adder in quantum-dot cellular automata via clock-zone-based crossover. *IEEE Trans Nanotechnol* 14:497–504. <https://doi.org/10.1109/TNANO.2015.2409117>
4. Ahmad F, Ahmed S, Kakkar V, Bhat GM, Bahar AN, Wani S (2018) Modular design of ultra-efficient reversible full adder-subtractor in QCA with power dissipation analysis. *Int J Theor Phys* 57:2863–2880. <https://doi.org/10.1007/s10773-018-3806-3>
5. Ahmad F, Bhat GM, Khademolhosseini H, Azimi S, Angizi S, Navi K (2016) Towards single layer quantum-dot cellular automata adders based on explicit interaction of cells. *J Comput Sci* 16:8–15. <https://doi.org/10.1016/j.jocs.2016.02.005>
6. Shahidinejad A, Selamat A (2012) Design of first adder/subtractor using quantum-dot cellular automata. *Adv Mater Res* 403–408:3392–3397. <https://doi.org/10.4028/www.scientific.net/AMR.403-408.3392>
7. Angizi S, Alkaldy E, Bagherzadeh N, Navi K (2014) Novel robust single layer wire crossing approach for Exclusive OR Sum of products logic design with Quantum-dot Cellular Automata. *J Low Power Electron* 10:259–271. <https://doi.org/10.1166/jolpe.2014.1320>
8. Bardhan R, Sultana T, Lisa NJ (2016) An efficient design of adder/subtractor circuit using quantum dot cellular automata. In: 2015 18th international conference on computer and information technology, ICCIT 2015, pp 495–500. <https://doi.org/10.1109/ICCIITechn.2015.7488121>
9. De D, Sadhu T, Chandra Das J (2016) Bioprocess modeling and simulation of half subtractor using actin based quantum cellular automata. *Mater Today Proc* 3:3276–3284. <https://doi.org/10.1016/j.matpr.2016.10.009>
10. Dehghan B (2013) Generating subtractor design by QCA gates under nanotechnology. *Int J Sci Eng Investig* 2:30–34
11. Gupta N, Patidar N, Katiyal SK, Choudhary K (2012) Design of hybrid adder-subtractor (HAS) using reversible logic gates in QCA. *Int J Comput Appl* 53:1–7. <https://doi.org/10.5120/8494-2442>
12. Taherkhani E, Moaiyeri MH, Angizi S (2017) Design of an ultra-efficient reversible full adder-subtractor in quantum-dot cellular automata. *Optik (Stuttg)* 142:557–563. <https://doi.org/10.1016/j.ijleo.2017.06.024>

13. Karthigai Iakshmi S, Athisha G, Karthikeyan M, Ganesh C (2010) Design of subtractor using nanotechnology based QCA. In: 2010 IEEE international conference on communication control and computing technologies, ICCCCT 2010, pp 384–388. <https://doi.org/10.1109/ICCCCT.2010.5670582>
14. Kianpour M, Sabbagh-Nadooshan R (2017) Novel 8-bit reversible full adder/subtractor using a QCA reversible gate. *J Comput Electron* 16:459–472. <https://doi.org/10.1007/s10825-017-0963-1>
15. Kumari NP, Gurumurthy KS (2014) QCA system design using blocks with vertically stacked active elements. In: Proceedings of the IEEE international caracas conference on devices, circuits and systems, ICDCDS, vol 1, pp 1–6. <https://doi.org/10.1109/ICDCSyst.2014.6926132>
16. Nagamani AN, Ashwin S, Agrawal VK (2014) Design of optimized reversible binary adder/subtractor and BCD adder. In: Proceedings of the 2014 international conference on contemporary computing and informatics, IC3I 2014, pp 774–779. <https://doi.org/10.1109/IC3I.2014.7019664>
17. Pudi V, Sridharan K (2013) Efficient QCA design of single-bit and multi-bit subtractors. In: Proceedings of the IEEE conference on nanotechnology, pp 1155–1158. <https://doi.org/10.1109/NANO.2013.6720958>
18. Sarma R, Jain R (2018) Quantum gate implementation of a novel reversible half adder and subtractor circuit. In: Proceedings—2nd international conference on intelligent circuits and systems, ICICS 2018, pp 77–80. <https://doi.org/10.1109/ICICS.2018.00027>
19. Zoka S, Gholami M (2019) A novel efficient full adder–subtractor in QCA nanotechnology. *Int Nano Lett* 9:51–54. <https://doi.org/10.1007/s40089-018-0256-0>
20. Ahmad PZ, Quadri SMK, Tantary SM, Wani GM, Ahmad F, Bahar AN (2017) Design of novel QCA-based half/full subtractors. *Nanometer Energy* 6:59–66. <https://doi.org/10.1680/jnaen.15.00020>
21. Jaiswal R, Sasamal TN (2018) Efficient design of full adder and subtractor using 5-input majority gate in QCA. In: 2017 10th international conference on contemporary computing, IC3 2017, pp 1–6. <https://doi.org/10.1109/IC3.2017.8284336>
22. Labrado C, Thapliyal H (2016) Design of adder and subtractor circuits in majority logic-based field-coupled QCA nanocomputing. *Electron Lett* 52:464–466. <https://doi.org/10.1049/el.2015.3834>
23. Reshi JI, Banday MT (2016) Efficient design of nano scale adder and subtractor circuits using quantum dot cellular automata. In: IET conference publication 2016. <https://doi.org/10.1049/cp.2016.1508>
24. Roohi A, DeMara RF, Khoshavi N (2015) Design and evaluation of an ultra-area-efficient fault-tolerant QCA full adder. *Microelectron J* 46:531–542. <https://doi.org/10.1016/j.mejo.2015.03.023>
25. Sangsefidi M, Karimpour M, Sarayloo M (2016) Efficient design of a coplanar adder/subtractor in quantum-dot cellular automata. In: Proceedings—EMS 2015 UKSim-AMSS 9th IEEE European modelling symposium of computer modeling and simulation, pp 456–461. <https://doi.org/10.1109/EMS.2015.74>

Intelligent Resource Identification Scheme for Wireless Sensor Networks



Gururaj S. Kori and Mahabaleshwar S. Kakkasageri

Abstract In this paper, an intelligent resource identification mechanism using game theory strategy is used to resolve the resource identification issues involving the bandwidth and energy of wireless sensor network. Resource identification is done using the game theory-based approach which mainly relies on the reputation value of nodes. Proposed intelligent resource identification mechanism uses non-cooperative game theory to calculate the utility or payoff functions. The main objective of the proposed scheme is to efficiently identify the available resource of wireless sensor network for information transmission, consuming less energy, bandwidth, and reducing time delay. Our simulation results show that the proposed scheme enhances the performance in terms of number of resources identified, resource identification delay, and end-to-end delay w.r.t. energy and reputation, etc.

1 Introduction

Sensors have found their importance in all the fields of science and engineering. Hardware, software, and algorithms are very much necessary for wireless sensor networks to sense, communicate, and to compute. Nowadays in any computing and communication applications, sensors perform multiple tasks which are cost effective, utilize low power, and are effectual with precise results. Emerging sensor networks are heterogeneous in nature, in addition with an increase in onboard memory and computational speed, with built-in global positioning system. But the resource

G. S. Kori

Electronics and Communication Engineering Department, Biluru Gurubasava Mahaswamiji Institute of Technology, Mudhol 587313, Karnataka, India
e-mail: korigurus@yahoo.com

M. S. Kakkasageri (✉)

Electronics and Communication Engineering Department, Basaveshwar Engineering College (Autonomous), Bagalkot, Karnataka 587102, India
e-mail: mskec@becbgk.edu

limitations such as unpredictable communication conditions, restricted battery life, and available radio resource restrict wireless sensor network (WSN) applications [1, 2].

Resource identification refers to the action or process of identifying or finding available resource in the network such as available number of alive nodes and battery, radio resources, and bandwidth. The work given in [3] shows that an energy-efficient dynamic energy management technique reduces power consumption by sensor node just by shutting down some components of sensors, which yields better savings and enhanced life time. Energy conservation in WSN by exploiting the interactions in cross-layer of WSN protocol stack is presented in [4]. Authors have also concentrated on routing algorithms, scheduling, data aggregation, and MAC protocol research for optimization power. Detailed concepts and applications of game theory in improving the performance of WSNs are presented in [5]. A study of game theory-based WSNs protection from selfish behavior or malicious nodes is presented in [6]. Due to scalability, low complexity, and disseminated nature of WSNs, malicious attacks can be modeled effectively using game theory. Brief overview of classifications of games, preliminaries used in games (Nash equilibrium, Pareto efficiency, pure, mixed, and fully mixed strategies), and game models are presented [7]. Recent game theory applications for WSNs are presented in [8, 9].

Rest of the paper is organized as follows. Proposed work is described in Sect. 2. Section 3 presents simulation model, simulation inputs, and result analysis. Section 4 concludes our research work.

2 Non-cooperative Game Theory-Based Resource Identification Scheme

The issue addressed in this research work is to calculate available resources in the network such as number of active nodes, their energy, and available bandwidth. To compute these resources, non-cooperative game theory concept is used. Non-cooperative game theory is one of the strategic conditions in which there is a gain or loss for a player and the game is between two timely players. Game theory is mainly considered when there is more than one resource for computation. In these conditions average concept will not hold good and the accuracy of output will not be obtained. Hence, to avoid this, game theory is conceptualized which takes the results based on utility or payoff functions which are computed with respect to both the source and receiver. The non-cooperative game theory is utilized here to calculate the utility or payoff functions. The advantage of the non-cooperative game theory is that the decisions of the players are independent of each other.

In WSN (W_n), a number of nodes considered are n_i which are randomly deployed as shown in Eq. (1).

$$W_n = \sum_{i=1}^m n_i \quad (1)$$

The GPS position (x, y) of the nodes in the network are mentioned in Eqs. (2) and (3).

$$x_i = \{x_1, x_2, x_3, \dots, x_m\} \quad (2)$$

$$y_i = \{y_1, y_2, y_3, \dots, y_m\} \quad (3)$$

Energy values (E) and the distance of the nodes (DIS) with respect to the sender are assigned using the random function as shown in Eqs. (4) and (5). The sensor nodes are mobile and hence the reputation factor (r_p) of the nodes is random. It is identified with respect to their neighbors as shown in Eq. (6).

$$E = \{E_1, E_2, E_3, \dots, E_m\} \quad (4)$$

$$DIS = \{D_1, D_2, D_3, \dots, D_m\} \quad (5)$$

$$r_p \in (0, 1) \quad (6)$$

Threshold energy and the threshold distance for the network are based on the condition as per Eq. (7).

$$E \geq e \& \& DIS \leq dis \& \& r_p \geq 0.9 \quad (7)$$

Based on the condition, resources identified (R_k) are mentioned in Eq. (8).

$$R_k = \{R_1, R_2, R_3, \dots, R_k\} \quad (8)$$

k number of resources participate in the game theory strategy using non-cooperative game theory strategy. It calculates the utility of each player independent of the opponent player. Utility is nothing but the motivation of players. For a given player, utility assigns a number which is implied with the property that a higher number produces or gives much outcome. “ k ” resources utility (where “ i ” and “ j ” are source and destination nodes, respectively) w.r.t. the sender is given by Eqs. (9), (10) and (11).

$$U_{[j]} = \{(energy[best_res_{[j]}] \times 0.001) - ((\zeta/\eta) \times r_p[sender_node])\} \quad (9)$$

$$\{\zeta, \eta \in (0, 1) \& \& \zeta + \eta = 1\} \quad (10)$$

$$U_j = \{U_1, U_2, U_3, \dots, U_k\} \quad (11)$$

ζ number of resources are obtained whose utility is greater than zero and then for the obtained ζ resources, again senders utility is calculated with respect to the available resources as mentioned in Eqs. (12) and (13).

$$U_{\text{sender}}[j] = \{((\text{energy}[\text{sender_node}] \times 0.001) \\ - \{(\eta/\zeta) \times r_p[\text{best_res_sel}_{[j]}]\}) \} \quad (12)$$

$$U_z = \{U_1, U_2, U_3, \dots, U_z\} \quad (13)$$

Among the z resources, the node with the largest value of utility is considered to be the best node for data communication.

3 Simulation

The proposed scheme has been simulated using C++ programming language as discrete event simulator. In this section, we discuss performance metrics and results analyzed.

3.1 Performance Metrics

To examine the performance effectiveness of the proposed resource scheduling scheme, some of evaluation metrics analyzed are as follows:

- *Total number of resources identified*: It is the quantity of resources identified at various reputation ranges at particular number of nodes.
- *Resource identification delay*: Total time taken by the node to identify the sufficient resources is considered as the resource identification delay. It is measured in terms of milliseconds (ms).
- *End-to-end delay*: It is defined as the time taken to identify sufficient resources, calculate the utility or payoff function, and identify the best resource for communication. It is expressed in terms milliseconds (ms).
- *Bandwidth utilized*: It is defined as the bandwidth utilized to compute the resource identification (i.e., used for exchange of control packets) to the total available bandwidth in the network. It is measured in percentage (%).

3.2 Result Analysis

As shown in Fig. 1, resources are identified by varying the node density and reputation factor. It can be seen that as the reputation is greater, a number of resources identified are short with the increasing number of nodes and hence the resources identified are very much appropriate for the calculation of utility or payoff function. Therefore, the proposed algorithm progressively performs at a good rate and the network is maintained for a long span of time. Resource identification delay variation w.r.t. the

Fig. 1 Total number of resources identified versus node density

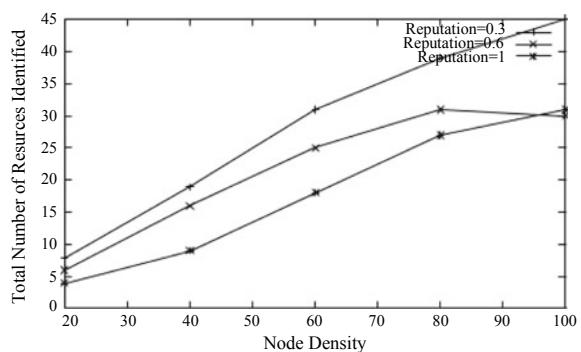


Fig. 2 Resource identification delay versus node density

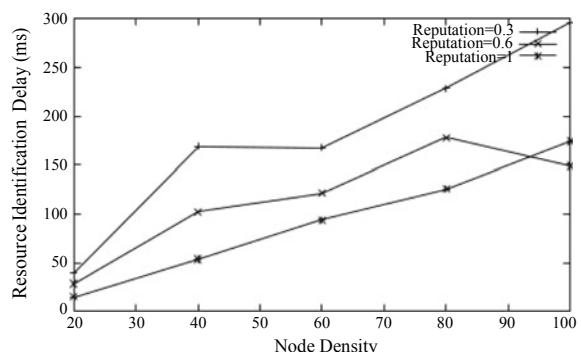
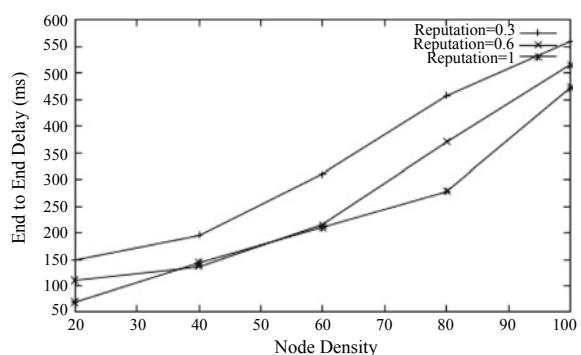
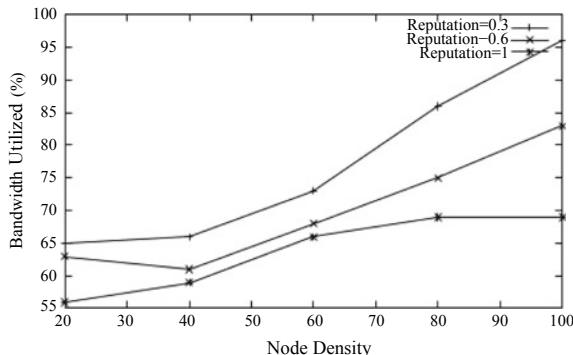


Fig. 3 End-to-end delay versus node density



variation of node density for various reputation ranges is mentioned in Fig. 2. End-to-end delay by varying the number of nodes is shown in Fig. 3. Bandwidth utilization by varying the number of nodes w.r.t. the reputation factors is depicted in Fig. 4.

Fig. 4 Bandwidth utilized versus node density



4 Conclusion

As wireless sensor network is highly dynamic, it becomes quite difficult to track and identify resources in the nodes. In this paper, we have proposed resource identification mechanism using non-cooperative game theory strategy for resource identification. The scheme is based on the reputation value of nodes. Our proposed resource identification scheme efficiently identifies the available resource of wireless sensor network for information transmission, consuming less bandwidth and reducing time delay. Our simulation results show that the proposed scheme enhances the performance in terms of number of resources identified, resource identification delay, and end-to-end delay w.r.t. various reputation factors.

References

1. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2013) Wireless sensor networks: a survey. *J Comput Netw*, pp 393–422 (Elsevier)
2. Rawat P, Singh KD, Chaouchi H, Bonnin JM (2014) Wireless sensor networks: a survey on recent developments and potential synergies. *J Supercomput* 68(1):1–48
3. Lin X-H, Kwok Y-K, Wang H (2009) Energy-efficient resource management techniques in wireless sensor networks In: Guide to wireless sensor networks computer communications and networks. Springer, pp 439–468
4. Shah K, Kumar M (2008) Resource management in wireless sensor networks using collective intelligence. In: Proceedings of the international conference intelligent sensors, sensor networks, and information processing (ISSNIP'08)
5. Shoukath Ali M, Singh RP (2017) A study on game theory approaches for wireless sensor networks. *Int J Eng Adv Technol (IJEAT)* 6(3):5–7
6. Abdalzaher MS, Seddik K, Elsabrouty M, Muta O, Furukawa H (2016) Game theory meets wireless sensor networks security requirements and threats mitigation: a survey. *Sens J* 16(07):1–27 MDPI Publications
7. Benmammar B, Krief F (2014) Game theory applications in wireless networks: a survey. In: Proceedings of the 13th international conference on software engineering, parallel and distributed systems (SEPADS'14), Poland, pp 208–215

8. Shi HY, Wang W, Kwok NM, Chen SY (2012) Game theory for wireless sensor networks: a survey. *Sens J* 12(7):9055–9097 MDPI Publications
9. Han Z, Niyato D, Saad W, Baar T, Hjungnes Are (2012) Game theory in wireless and communication networks: theory, models, and applications. Cambridge University Press, Cambridge

A Systematic Review on Various Types of Full Adders



D. Dhathri, E. Jagadeeswara Rao, and Durgesh Nandan

Abstract In recent years, Personal computers (PCS) and mobile phones (MP) are more used in the world but still now very fewer researchers focus to structure the low power (LPW) and high speed (HS) Multiplier (MUL) and accumulate unit (MACU). MACU is a main arithmetic component in PCS and MP for doing any math work. In MACU main components are MUL and adders and also adder is one main component in MUL unit. So, in this paper mention recent XOR-XNOR based adders and also mention techniques for optimizing the performance constraints. This thorough investigation incorporates the deliberate improvement, looks at the most recent design of each adder, and advocated what one is better over other revealed viper is likewise featured.

Keywords MACU · FAd · RCA · CLA CSA

1 Introduction

This paper presents an Adder. The expansion of two binary numbers is the principle and regularly utilized number-arithmetic operations. In about all digital IC layout, nowadays, the addition is the most frequently used basic operation. By utilizing this adder APPROX, we will limit area and inertness. This structure is especially helpful in computation-intensive applications that are vigorous to little mistakes in calculation. The potential applications of this APPROX Adder fall mostly in regions where there

D. Dhathri · E. Jagadeeswara Rao

Department of Electronics and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India

e-mail: sgckbabu266@gmail.com

E. Jagadeeswara Rao

e-mail: emandi.jagadeesh@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India
e-mail: durgeshnandano51@gmail.com

is no severe necessity on exactness or where super-LPC and rapid execution are a higher priority than the precision. The APPROX is most pivotal at the adder level nowadays in the field of signal processing especially digital and Image. It is also used for the neural system. Here various types of adder's design have been studied based on compactness, performance, and accuracy. Rest of the paper arranged is like that. Section to cover existing work, Sect. 3 discussed the existing adder implementation methodology, Sect. 4 covers a comparative analysis of various adder's circuit and at last Sect. 5 paper concluded.

2 Systematic Literature

In 1990, Kazuo Yano has proposed a fast MUL that is 16-bit which is enforced as a test vehicle for investigation of a new circuit technique for high-speed power (HSPW). This HSPW is completely utilized on the entire critical path to fulfilling very HSPs compare to existing MUL [1]. In 1992, Nan Zhuang has suggested two CMOS full adder (FAd) based upon the transmission theory. It is difficult to design a CMOS algebraic since the high impedance cannot be expressed in the Boolean algebra is overcome for decreasing the no. of transistors (TRANS) [2]. In 1999, Mark Vesterback a has invented a new FAd ckt of propagation delay (PD) to the carrier is split into half and the PD to the sum. The sum is expanded by 40% for the new ckt. The ckt consists of 16 MOSFETs, which has been decreased to 14 MOSFETs, and the area is around diminished [3].

In 2001, D. Radhakrishnan has proposed for XOR and adder cells by Using a 6-TRANS XOR-XNOR cell is designed which isn't influenced by the threshold voltage (TV) drop in MOS TRANS. It likewise decreased the no. of logic gates needed for executing the AC [4]. In 2002, Ahmed M. Shams and Tarek K. Darwish has proposed the Analysis of LPW-Bit CMOS FAd Cells. This study is used to increase the driving PW of the adder ckt. [5]. In the same year, Basil George and Nikhil Soni have proposed a new technique to build a total of 3 LPW 10 TRANS FAd using a logic gate (x-nor), although it has a problem of threshold loss which is presented in large ckt like MUL. They agreed to the existing adder which consumes less PW around 10–20%, than compared to the 10TRANS Static Energy Recovery FAd (SERF) [6]. In 2004, Sumeer Goel has proposed a novel XOR-XNOR circuit dependent on CPL style. The circuit has improved as far as PDP which is accessed by the FAd circuit. The best standard design of PDP is 24% when compared to existing FAd [7]. In 2005, M. Linares Aranda proposed a FAd designed with bootstrapped pass TRAN's logic. For 1bit FAd, the PD is 36% and trying to save the PD by using a new 8-bit CRA. The 8 bit CRA has a PD of 28%. [8]. In the same year, Toshiro Akino Kei Matsuura suggested a logic conspire for controlling the logic gates of the draw up and pull-down MOSFETs as current sources in the UBiCMOS ckt and the speed of the CMOS FAd driven by the U-BiCMOS + CL circuit is 1.9 events faster than a static CMOS FAd driven by and devours 20% less vitality expecting a stack capacitance of 0.3542 pF and Vdd of 1.2 V [9]. In 2006, Sumeer Goel designed an XOR–XNOR circuit depends on

CPL. In this examination, we utilize just a single static inverter as opposed to two in the routine CPL style. This XOR ckt hybrid CMOS FAd has a more prominent execution than the majority of the standard FAd cells [10]. In 2007, Chiou-Kou Tung advised a new hybrid CMOS FAd with driving capability with 1.8–35.6%. In this PC, 11.7–41.2% in time delay of Co, and 13.7–91.4% in PDP of Co [11]. In the very same year, Keivan Navi has proposed an HS adder cell called a bridge, using a new design style. This bridge layout tenders high regularity and greater density compared to conventional design. By using this bridge TRANS, the speed is improved over conventional adder [12]. In 2008 Omid Kavehei suggested FAd24T and FAd24T New designs which lead a productive FAd as far as PC, delay, and size in contrast with a current FAd structure [13]. In 2009, Saradindu Panda explained about the new advanced FAd circuit by applying transmission Gate (TG) technology. By using this FAd we can decrease the avg PW, Leakage PW, delay, and noise [14]. In 2010, Vladimir V. Shubin put forward that the one bit CMOS ripple carry FAd cell. This FAd is most reasonable for speed among different ones for N-bit adders over 3-bit for CMOS mirror configuration style and comparably PC [15]. In 2010, Aminul Islam has suggested the one-bit FAd cell, which has been effectively examined by appointing high-TV to some TRANS and low TV to others. The new existing FA possesses less avg power dissipation, PD and PDP, and wider robust across PVT variations [16]. In 2010, Manisha Patnaik studied the low leakage one-bit FAd cell with the technology of 90 nm. In this analysis latency, active PW, and ground bounce noise are diminished with a one-volt supply [17]. In 2011, Y. Berg suggested a novel HS and ultra-low-voltage (ULV) FAd circuits dependent on ULV semi floating-gate CMOS logic. The FAd adventures floating-gate TRANS both by offering HS, particularly for carry propagate, and numerous esteemed intermediate representations of the entire sum [18]. In the same year, Subodh Wairy proposed the Design Analysis of XOR (4T) passed on the LV CMOS FAd Circuit. This technique helps in decreasing the PC and the PD by keeping up the low unpredictability of logic design [19]. In 2012, Attapon Sudsakorn suggested one-bit FAd by using CMOS technology with 22 nm for the low supply voltage. The no. of TRANS using in this operation is 14 and it has less PC compared to the CCMOS circuit, TG Adder circuit, and Hybrid logic FAd circuits [20]. In 2014, Bhavani Prasad recommended that another adder design of EXOR-EXNOR. The adder cell consists of less PW, PDP, and delay [21]. In 2014, Shahbaz Khan proposed an improvement of diminished complexity Wallace MUL with decreased PC and region by utilizing Energy-Efficient CMOS FAd at the spot of traditional FAd [22]. In 2015, Yasser S. Abdalla proposed a one-bit FAd. The max no. of TRANS used in this 19 and having low PC and HS operation [23]. In 2017, Lee Shing Jie put forwarded an LPW one-bit hybrid FAd. The max TRANS used in this hybrid FAd is 13 performing great regarding delay and PC when contrasted with 1-bit FAd [24]. In 2018, Ashish Kumar Yadav suggested LPW HS one-bit FAd Ckt design at CMOS Technology with 45 nm. These existing FAd have accomplished the maximum sparing of PW 91.65% and 93.04%, max decrease in the delay of 59.37%, and 76.76% and max saving of PDP 91.64% and 96.01% [25]. In 2018, Lakshmi S has introduced a novel 1bit vitality effective hybrid Adder This adder can upgrade execution parameters like power and delay of two CMOS hybrid FAd circuits. By

utilizing this investigation, we can present progressively proficient designs as the MGDI system diminishes the general PD, area, and PDP [26]. In the same year, Zarin has proposed Comparative Analysis and Simulation of Different CMOS FAdS utilizing Cadence in 90 nm Technology. It talked about the presentation of four FAdS to be specific; traditional CMOS FAd with 28 TRANS, TG FAd with 20 TRANS, 14 TRAN's adders utilizing pass TRAN's logic & GDI based FAd containing 10 TRANS [27]. In 2019, Irfan Ahmed has proposed an investigation of prominent 1-bit FAd circuits. The investigation of FAd regarding PW, delay, PDP, region, and threshold loss [28]. Around the same time, Smita Singhal has proposed another system of PW decrease in a CMOS domino logic. This design expands less measure of PW.LPW gadgets have the component consuming low with high delay [29]. In the extremely same year, Pravin has recommended FAdS utilizing regular Hybrid and Static Logic and novel gated supply topology. The TRANS utilized in Conventional is 28 for 1-bit FAd while our current topology utilizes 16 TRANS. Static logic demonstrates that a 1-bit FAd working from 1.2 V supply at a maximum recurrence of 250 MHz consumes 47 μ W [30]. Around the same time, Aaina Nandal has presented another structured system, vitality ECRL with the sleepy keeper for LPW FAd. To diminish the power dispersal, another methodology called ECRL with the presence of a sleepy keeper [23].

3 Methodology

An adder is an advanced ckt that plays out the expansion of numbers. In various PCs and various sorts of processors, adders are used in the ALU. They are similarly used in various parts of the processor, where they are used to learn addresses, table documents, increase and decrement of operations, and practically identical tasks. Despite the way that adders can be created for some number depictions, for instance, paired coded decimal or abundance 3, the most generally perceived adders' chip away at binary numbers. In circumstances where two's complement one's complement is being used to show to negative numbers, it is inconsequential to alter an adder into an adder subtract or other checked number representations require more logic around the fundamental adder. Adders can be widely gathered into going with four classes [5]: 1. Ripple Carry Adder (RCA), 2. Carry Look-Ahead Adder (CLA), 3. Carry Select Adder (CSA) and 4. Parallel Prefix based Adder (PPA).

3.1 Ripple Carry Adder

Ripple Carry Adder [RCA]: RCA is a circuit of combinational logic. It is utilized to include two n-bit binary numbers. It requires n FAd in its circuit for including two n-bit double numbers. It is also known as an n-bit parallel adder. In RCA, the carry out produced by each FAd serves as carry-in for its adjacent most significant FAd.

Each carries bit ripples or waves into the next stage. That's why it is called "RCA". RCA works in different stages. A FAd is a combinational ckt, it contains two bits of the operand and a bit of carrying, states A, B, and C, respectively, and gives Sum (S1) and Carry bit (C0) as outputs. Every FA takes the carry in as input and produces carry-out and sum bit as output. The carry-out produced by a FAd serves as carry-in for its adjacent most significant FAd. When carry-in becomes available to the FAd, it activates the FAd. After FAd becomes activated, it comes into operation. By using RCA, we can construct 1-bit as well as 4-bit RCA.

$$C_o = (A \cdot B) + (C_i \cdot (A \oplus B)) = (A \cdot B) + (B \cdot C_i) + (C_i \cdot A) \quad (1)$$

$$S = (A \oplus B) \oplus C_i \quad (2)$$

RCA does not allow us to use all the FAds simultaneously. Every FAd needs to fundamentally hold up until the carry bit ends up accessible from its adjoining FAd. This expands the propagation time. Because of this reason, RCA turns out to be very moderate. To conquer this disadvantage, a carry looks ahead becomes possibly the most important factor (Figs. 1 and 2; Table 1).

Fig. 1 1-bit FAd

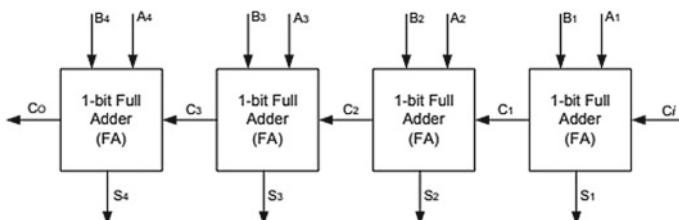
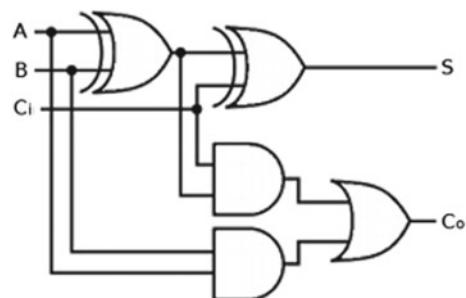


Fig. 2 4-bit ripple carry adder

Table 1 FAd truth table

Ci	B	A	C0	S1
0	0	0	0	0
0	0	1	0	1
0	1	0	0	1
0	1	1	1	0
1	0	0	0	1
1	0	1	1	0
1	1	0	1	0
1	1	1	1	1

3.2 Carry Look-Ahead Adder [CLA]

CLA is an improved variant of RCA. It generates the carry-in of each FA simultaneously without causing any delay. The time complexity of CLA adder = $\Theta(\log n)$. One such method is to infer the sum and Convey outputs by utilizing moderate terms characterized as generating (G1) and (P1). Generate term delivers a do free of the carry-in, for example regardless of what the carry-in is, the complete is constantly “1” when both of its inputs M1 and M2 are ‘1’ in this way $G1 = M1 \cdot M2$. The Propagate expression moves the input Carry as output Carry when just one of the inputs is high and thus Propagate term is characterized as $P1 = M1 \oplus M2$. Therefore, we have

$$G1(M1, M2) = M1 \cdot M2 \quad (3)$$

$$P1(M1, M2) = M1 \oplus M2 \quad (4)$$

For instance, delineate the idea of Propagate and Generate even more distinctly. In the Propagate case the carryout relies upon the carry-in, for example at the point while Carry is zero(0) and carry out is zero(0) and while carry is one(1) and carry out is one(1) and on account of generate, regardless of while the carry is carried through constantly 1 (Fig. 3).

The disadvantages of CLA adder are, it involves complex hardware, costlier, and likewise gets progressively entangled as the quantity of bits increments.

3.3 Carry Select Method

CSA is a specific method to actualize an adder, which is a logic component that registers the $(n + 1)$ bit total of two n -bit numbers. The CSA, for the most part, comprises of two RCA and a

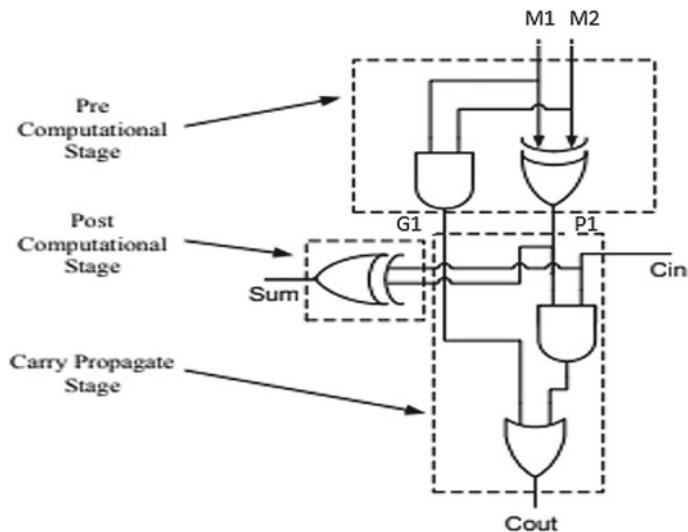


Fig. 3 1-bit FAd with carry propagate and generate

multiplexer. RCA hangs tight for the information convey (C_i) and afterward processes the sum and carries (C_o) in this way expanding its delay. To decrease the delay a CSA is presented, whatever pre-registers the sum and carry for the 2 possible cases, for, i.e., C_i = 1 and C_i = 0. The determined Sum is given to a MUL, which picks the right output build upon the C_i originating from the previous stage. The calculation of Sum decreases the delay of shipping of Carrying which is constrained to only one MUL for a particular stage.

3.4 Parallel Prefix-Based Adder (PPA)

A PPA comprises three phases, i.e., pre-computation stage, prefix network stage, and post-computation stage (Fig. 4).

Fig. 4 Prefix adder block diagram

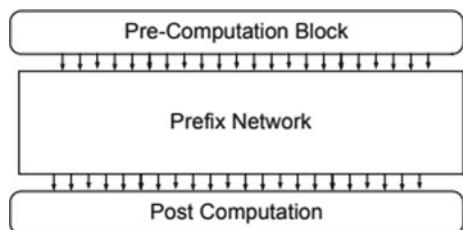
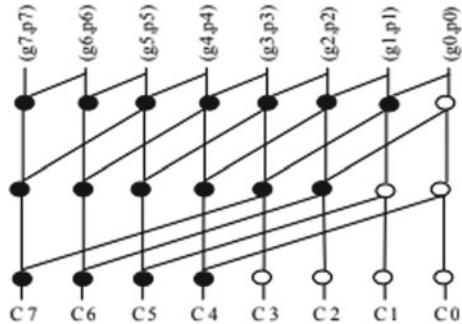


Fig. 5 Example of a Kogged-Stone prefix adder



The precomputation stage computes the carry propagate and convey generate bits for each information pair in (1) and (2) equations. Both the carry generated and carry propagated bits are the last carry which is figured out by the prefix system stage computation can be changed into a prefix issue utilizing the acquainted operator ‘ \circ ’, which associates pairs of generate and propagate bits as given

$$(g, p) \circ (g', p') = (g + p \cdot g', p \cdot p') \quad (5)$$

where g and g' represent to the general terms and $p1$ and $p1'$ represent to the propagate terms. Utilizing the operator ‘ \circ ’ back to back propagate and generate sets can be gathered to create carry as follows: (Fig. 5)

$$C_i = (g_i, p_i) \circ (g_{i-1}, p_{i-1}) \circ \dots \circ (g_1, p_1) \circ (g_0, p_0) \quad (6)$$

4 Result

Comparative study of various full adders has been shown in Table 2 in terms of power, delay, Pd, and hardware resources required for implementing various types of full adders.

Figure 6a shows the comparisons of FAd cells in terms of power (v). The maximum input power supply is given to FAd1 which is 6.53 v. Most of the researchers have worked to reduce the input power supply to FAd. In FAd7, power has been reduced to 0.0048 v.

The maximum output delay happens at FAd1 which is of (12.09 ns). Some scientists have worked to decrease the delay of the FAd. In FAd6 the delay has been decreased which is of (0.000653 ns).

The graph describes the Power dissipation of FAd cells. The maximum power dissipation is in FAd10 which is 29 μ W and then reduced to 1.1 μ W. We discuss no. of transistors are used to build the FAd. If we utilize progressively no. of TRANS

Table 2 Comparisons of various full adder

Design	Name	Power (V)	Delay (ns)	PD	No. of Trans
Robust full adder [16]	FAd1	6.53	12.09	1.1	28
LL-1BIT full adder [17]	FAd2	1	0.5	3	10
ULV full adder [18]	FAd3	0.25	0.499	9	20
LV CMOS full adder [19]	FAd4	1.8	0.042	13	16
LP CMOS full adder [20]	FAd5	1.5	4.054	18.1	12
XOR-XNOR full adder [21]	FAd6	1.3	0.000653	20.1	16
NOVEL DESIGN [23]	FAd7	0.0048	0.05	22	19
1-bit-13T hybrid adder [24]	FAd8	1.8	0.35	24	13
NEE 1-bit hybrid full adder [26]	FAd9	1.2	0.001	26	12
16T full adder [30]	FAd10	1.2	0.325	29	16

the circuit configuration will be complex and cost likewise increases. In FAd2 we can see that fewer transistors have been utilized, however, delay is more to conquer the delay we build the no. of TRANS.

5 Conclusion

In this paper, mainly give the one root map for research in HPW and LPW designs in digital system design and mentioned different types of serial adders. Mentioned different types of FAd with different techniques developed by different researchers and also mentioned performance analysis of these FAd checks in Table 1. Finally, we conclude in this different type of FAd with some different techniques developed by different researchers till now. The maximum delay is 12.09% [16] and then reduced to 0.006% [21]. The max no. of transistors is 28 [16] and then reduced to 10 [17]. Due to more no. of transistors, the area increases. As the area increases, the circuit complexity is also increased. The max PDP is 29% [30] and the minimum PDP is 11% [16]. A CMOS FAd has been existing in this paper with expanding the efficiency and diminishing the area and PDP likewise latency. This upgrade is would have liked to contribute similar enhancements in DSP frameworks and the IP areas.

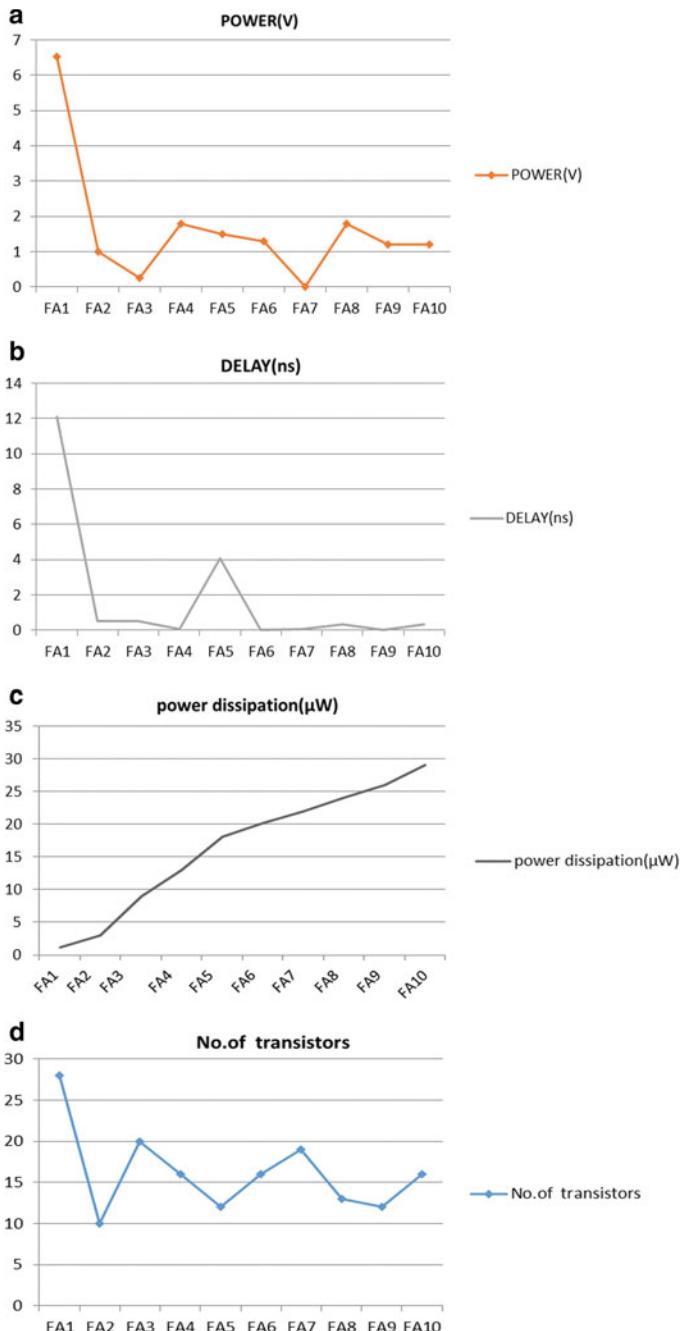


Fig. 6 **a** Comparisons of FAd cells in terms of power (V). **b** Comparisons of FAd cells in terms of delay (ns). **c** Comparisons of FAd cells in terms of PD (μ W). **d** Comparisons of FAd cells in terms of no. of transistors

References

1. CPL-multiply.pdf
2. Zhuang N, Wu H (1992) CMOS full adder 27(5):840–844
3. Vesterbacka M (1999) 14-transistor CMOS full adder with full voltage-swing nodes. In: IEEE workshop on signal processing systems. *SiPS* design and implementation, pp 713–722
4. B, no. x, pp 19–24
5. Shams AM, Darwish TK, Bayoumi MA (2002) Performance analysis of low-power 1-bit CMOS full adder cells. *IEEE Trans Very Large Scale Integr Syst* 10(1):20–29
6. Gates X, Design and analysis of low power 10- transistor full adders using novel Nikhil Soni, no. 1
7. Goel S, Gollamudi S, Kumar A, Bayoumi M (2004) On the design of low-energy hybrid CMOS 1-bit full adder cells. *Midwest Symp Circuits Syst* 2:209–212
8. Aguirre Hernández M, Linares Aranda M (2005) A low-power bootstrapped CMOS full adder. In: 2nd international conference on electrical and electronics engineering, ICEEE XI conference on electrical and electronics, CIE 2005, vol 2005, pp 243–246
9. Akino T, Matsuura K, Yasunaga A (2005) A high-speed domino CMOS full adder driven by a new unified-Bicmos inverter. In: Processing—IEEE international symposium on circuits and systems, pp 452–455
10. Goel S, Kumar A, Bayoumi MA (2006) Design of robust, energy-efficient full adders for deep-submicrometer design using hybrid-CMOS logic style. *IEEE Trans Very Large Scale Integr Syst* 14(12):1309–1321
11. Tung CK, Hung YC, Shieh SH, Huang GS (2007) A low-power high-speed hybrid CMOS full adder for embedded system, In: Processing 2007 IEEE work design and diagnostics of electronic circuits and systems, DDECS, pp 199–202
12. Navi K, Kavehei O, Rouholamini M, Sahafi A, Mehrabi S (2007) A novel CMOS full adder. In: Proceeding IEEE international conference VLSI design, pp 303–307
13. Kavehei O, Azghadi MR, Navi K, Mirbaha AP (2008) Design of robust and high-performance 1-bit CMOS full adder for nanometer design. In: Proceedings—IEEE computer society annual symposium on VLSI Trends VLSI technology and design ISVLSI 2008, pp 10–15
14. Panda S, Kumar NM, Sarkar CK (2009) Transistor count optimization of conventional CMOS full adder & optimization of power and delay of a new implementation of 18 transistor full adder by dual threshold node design with submicron channel length. In: Codec—2009—4th international conference on computers and devices for communication, pp 8–11
15. Shubin VV (2014) New high-speed CMOS full adder cell of mirror design style. In: 2010 11th annual international conference and seminar on micro/nanotechnologies and electron devices, EDM’2010—Proceedings, pp 128–131
16. Islam A, Akram MW, Pable SD, Hasan M (2010) Design and analysis of robust dual-threshold CMOS full adder circuit in 32 nm technology, In: Processing—2nd international conference on advances in recent technologies in communication and computing 2010, pp 418–420
17. Pattanaik M, Agnihotri S, Varaprasad MVDL, Arasu TA (2010) Enhanced ground bounce noise reduction in a low leakage 90 nm 1-volt CMOS full adder cell. In: Processing—2010 international symposium on electronic system design, ISED 2010, pp 175–180
18. Berg Y (2011) Ultra-low-voltage and high-speed CMOS full adder using floating-gates and multiple-valued logic. In: Processing—41st IEEE international symposium on multiple-valued logic, ISMVL 2011, pp 259–262
19. Wairy S, Singh G, Vishant, Nagaraj RK, Tiwari S (2011) Design analysis of XOR (4T) based low voltage CMOS full adder circuit. In: 2011 Nirma university international conference on engineering current trends and technology NUICONE 2011—conference proceedings, vol 2011, pp 1–7
20. Sudsakorn A, Tooprakai S, Dejhan K (2012) Low power CMOS full adder cells. In: 2012 9th international conference on electrical engineering/electronics, computer, telecommunications and information technology, ECTI-CON 2012, no. 2, pp 1–4

21. Khan S, Kakde S, Suryawanshi Y (2014) Performance analysis of reduced complexity Wallace multiplier using energy-efficient CMOS full adder. In: Processing—2013 international conference on renewable energy and sustainable energy, ICRESE 2013, pp 243–247
22. Bhavani Prasad Y, Harish Babu N, Ramana Reddy KV, Dhanabal R (2014) Comparative performance analysis of XOR-XNOR function based high-speed CMOS full adder circuits. In: ICROIT 2014—processing 2014 international conference on reliability optimization and information technology, pp 432–436
23. Nandal A, Kumar M (2019) Design and implementation of CMOS full adder circuit with ECRL and sleepy keeper technique. In: 2018 international conference on advances in computing, communication control and networking, pp 733–738
24. Jie LS, Ruslan SH (2017) A 4-bit CMOS full adder of 1-bit hybrid 13T adder with a new SUM circuit. In: Processing—14th IEEE student conference on research and development. Advances technology and the humanities, SCOReD 2016, pp 1–5
25. Yadav AK, Shrivatava BP, Dadariya AK (2018) Low power high-speed 1-bit full adder circuit design at 45 nm CMOS technology. In: International conference on recent innovations in signal processing and embedded systems, RISE 2017, vol 2018-January, pp 427–432
26. Lakshmi S, Meenu Raj C, Krishnadas D (2018) Optimization of hybrid CMOS designs using a new energy efficient 1 bit hybrid full adder. In: Processing in the 3rd international conference on communication and electronics systems ICCES 2018, ICCES, pp 905–908
27. Tabassum Z, Shahrim M, Ibnat A, Amin T (2018) Comparative analysis and simulation of different CMOS full adders using cadence in 90 nm technology. In: 2018 3rd international conference for convergence in technology I2CT 2018, pp 1–6
28. Ahmed I, Shahid MK (2019) Analysis of CMOS Full Adder Circuits. In: 2019 advances in science and engineering technology international conferences, pp 1–5
29. Singhal S, Mehra A, Tripathi U (2019) Power reduction in domino logic using clock gating in 16 nm CMOS technology. In: 2019 6th international conference on signal processing and integrated networks, SPIN 2019, pp 274–277
30. Borude PV, Agrawal SS (2019) Low power 16-T CMOS full adder design. In: Processing in the 2nd international conference on intelligent computing and control systems ICICCS 2018, ICICCS, pp 1130–1134

Design and Implementation of Smart Real-Time Billing, GSM, and GPS-Based Theft Monitoring and Accident Notification Systems



B. Jyothi Priya, Parvateesam Kunda, and Sanjeev Kumar

Abstract The rapid expansion of technology and architecture has improved our lives while increasing the traffic hazards, vehicular robberies, and road accidents, which in turn cause a tremendous death toll and property because of poor crisis facilities. As per the government demonstration, everybody ought to observe the traffic rules on the off chance that an individual disrupts any guidelines, at that point the approved individual will charge the fine (Inigo in IEEE Trans Veh Technol 38(3):112–122, [1]). As thefts increasing day today, to control this problem, the security of the system should be increased. If the car is stolen, the owner is allowed to stop ignition via GSM network communications by sending a ‘stop’ message. Then, the parked position can be easily identified by using the methods of the global positioning system. When an accident happens, the family members, friends, nearby police station, and hospitals also will get the notification, along with the location where it happens (Lotufo and Morgan in Automatic numberplate recognition (6):1–6, [2]). After that, the location of an incident can be easily identified using the GSM network and the global positioning system (GPS) technique (Collins et al. in A system for video surveillance and monitoring. Carnegie Mellon University, Carnegie, [3], Dickinson and Wan in Road traffic monitoring using the trip system, pp. 56–60, [4]). We use ARM 7 for this technology as the centerpiece of our new smart vehicle tracking system with types of equipment like RFID, GPS, and GSM. The equipment/programming co-structure of another smart vehicle tracking framework incorporates automatic billing, RFID identification, programmed accident detection, and GPS location.

B. J. Priya · P. Kunda · S. Kumar (✉)

Department of Electronics and Communication Engineering (ECE), Aditya College of Engineering and Technology (ACET), Surampalem, Andhra Pradesh, India
e-mail: sanjeev.kumar@accendere.co.in

B. J. Priya
e-mail: jyothipriyabassa@gmail.com

P. Kunda
e-mail: parvateesam.kunda@acet.ac.in

Keywords ARM7 microcontroller · GSM · GPS · RFID tag · Vibration sensor · MEMS sensor · Accelerometer · Eye blink sensor · Smoke sensor · Alcohol sensor · Temperature sensor · Motor · Buzzer

1 Introduction

The combination of increased power to compute networking technologies and advanced engineering principles makes vehicles smarter. As vehicles increase day to day, similarly, the thefts also increase day by day; so to protect a vehicle from theft, the security system should be stronger than before, and this particular design provides a keyless entry system as the key weapon to stop the vehicle from the theft. This system provides more security to the vehicles and prevents them from being stolen. If the vehicle is theft and the engine has started without the owner's approval, then the owner gets the notification that the engine has started, at that point the owner would then be able to stop the motor by sending the message STOP, so that the engine stops and the vehicle is protected. From that point onward, with the aid of GSM and GPS innovations, we can precisely locate the area of the vehicle. As vehicles increase, the number of road accidents also increases. As per the World Health Organization (WHO), every year more than 1.25 million deaths were registered across the worldwide and there were more than two hundred thousand deaths alone in India 2013. The primary reason for increasing the deaths is due to poor emergency medical services. So, there must be a solution to this problem; to take care of this issue, the framework is combined with some features like GPS tracking, ignition cut-off, pointing the location of a vehicle as a message using GSM module. If any vehicle crosses the zebra crossing or does not follow any traffic rules such as when the red light is ON if any vehicle crosses the road, consequently the individual subtleties will be sent to the nearest RTO Office. As the innovation developed, the nation has contributed an enormous amount of cash, especially on road infrastructure. As the infrastructure is increasing rapidly, the number of vehicles is also increasing. As technology increases faster, the majority of the offices deal with this issue in a customary manner, for example, manual judgment and street checking. Along these, lines of checking have a few issues like bogus checking and have substantial work to check huge vehicles, to solve this issue, we use a versatile vehicle tracking framework to supplant the customary technique.

2 Literature Survey

To monitor and control the traffic, wide-area detection system (WADS) is most reliable. We can track the vehicles up to 120 m by using charge-coupled device (CCD) cameras. To improve the tracking, we have to use two solid-state CID cameras, two frame grabbers, and a high-speed 16-bit microcomputer. This will increase the cost,

and this device is only applicable for short distances [1]. It can also monitor the traffic by using the image processing technology. In this technique, the cameras record the video and count the number of vehicles and apply the traffic signals. It not only measures the number of vehicles in the traffic but also measures the vehicle's speed, length, etc. This technique is more reliable than the normal monitoring system [3–6].

If any person breaches the traffic rules or traffic signals, then the vehicle's number plate is identified automatically. By using this, we can also identify whether the vehicle is having a license or not. Mainly it consists of three main parts. The first one is extracting the region, i.e., to which region the vehicle does belongs to by using some algorithms such as edge detection and smearing. The second one is segmentation; in this, starting letters are separated and the third one is recognition of characters by using statistical-based template matching [2].

The vehicle can be tracked by using GPS, GSM/GPRS technology. By using GPS, we can track the precise location of the vehicle and this location is sent to another device, i.e., to the mobile phone by using GSM/GPRS technology. If we want to display the vehicle on the map, we use Google APIs, and by using this, we can determine the shortest path to arrive at the destination within a short period. These modems are controlled by the microcontroller [7–10]. It can track the vehicles not only by using GPS/GSM/GPRS modems but by using some technologies like GIS, wireless pointing, and communication, we can exactly track the position of the vehicle as well as we can protect the vehicle from hijacking and thieves [11, 12].

The number of thefts increases the vehicles being stolen by them. To recover their vehicles from the thefts, we should provide a high-security system to the vehicle. To start the vehicle, we can use fingerprint, RFID, or any password system so that the security increases and thefts reduce. If any person breaks the door or glasses to start the vehicle with the help of the tilt sensor, we can notice the break and cut down the ignition. By this, we can reduce vehicle thefts [10, 13–15].

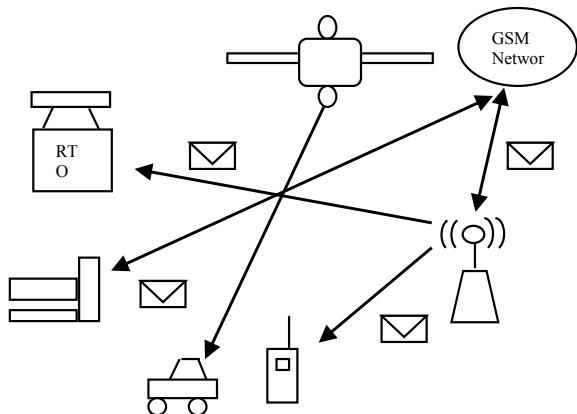
3 Function and Structure of the System

The framework is designed by ARM7, integrated software for storage, automatic billing, GSM cellular wireless technology, and GPS location technique. This is a new system that is used for tracking the vehicles, theft detection, accident detection, and automatic billing, i.e., owe the charges when they break the rules. This system consists of GSM network, mobiles, vehicles, etc., as shown in Fig. 1.

3.1 System Composition

Billing system. When the system violates any traffic rules such as zebra crossing and not following the traffic signals, the RFID reader peruses information which is

Fig. 1 Structure of the system



present inside tag and subsequently sends the information to RTO Office through GSM net.

Mishap recognition. At a point when the mishap happens, the framework recognizes the mishap by utilizing the accelerometer MEMS and vibration sensor. By using these sensors, ARM microprocessor gathers the information, afterward the notification will be sent to client personalities, emergency clinics, and the police station moreover.

Burglary detection. The proprietor can stop the vehicle remotely by sending the stop message through GSM network when the vehicle is robbed. After this, with the aid of GPS technology, the area, i.e., the location of the vehicle will be sent to the proprietor's mobile.

GPS positioning. By GPS, the framework can send the location and time of the vehicle precisely to the server room.

4 Hardware Implementation

4.1 Infrared Sensor (IR Sensor)

IR sensor is an electronic gadget that recognizes the things or items which are available in the environment. It can gauge (measure) the heat originating from the articles or things and also recognizes the movement of the item depending on the IR radiations. IR sensor takes in some sort of infrared radiation that is undetectable to our naked eye, rather than emitting it. Passive IR sensors absorb IR rays but do not emit it. All items which are available in the environment emanate some type of warm radiations in the infrared range, and these radiations are imperceptible to our unaided eye; however, these radiations are recognized by an infrared sensor. In an infrared sensor, the emitter is an IR light-emitting diode (LED) and the detector is an IR photodiode,

sensitive to IR light, which has a similar wavelength as those discharged by the IR LED. As IR light falls on the photodiode, resistors and the voltages are adjusted to the received size of the IR light.

4.2 AT89S52

The AT89S52 is an 8-bit microcontroller with a programmable device flash memory of 8 kbytes. The small-scale controller absorbs low power and is high in execution. The flash on the chip permits the memory to be reprogrammed by the developer, by the processing system. The AT89S52 is a powerful, highly flexible, and cost-effective microcontroller, because of its conjunction with the 8-bit CPU.

4.3 max232

MAX232 is an integrated circuit manufactured by Maxim Integrated Products. It serves as the voltage logic converter by converting the TTL Logic level (COM port of microcontrollers) to TIA/EIA-232-F level (laptop serial port RS232). It is used for interaction between the PC or laptop and the microcontroller. For data transfer, RS232 uses serial communication. A dual driver/receiver is MAX 232. There is a power generator that supplies a single 5 V voltage supply to RS232.

4.4 RFID

RFID stands for radio-frequency identification. It contains a tag and a reader. This uses radio waves to identify objects and individuals. The reader emits electromagnetic waves, which are absorbed by the tag. The absorbed energy from the tag is used to drive the microchip in it, and a signal containing the identifying tag number is returned to the reader to identify a human object. Therefore, it is used for security purposes.

4.5 GSM

GSM stands for global system for mobile communication, and it utilizes time-division multiple access (TDMA). It digitizes and compresses the data, and then the compressed information is sent to a channel with two different streams of client information and each one is having its own scheduled time. GSM is a circuit-switched system. It consists of eight time slots in which each slot is having 25 kHz. It drives at 900 or 1800 MHz frequency. In the USA, GSM operates at 850 and 1900 MHz.

4.6 ARM7

The ARM7 is a 32-bit microprocessor. It consumes low power utilization and provides high performance. The architecture of ARM depends on the principles of reduced instruction set computer (RISC). RISC instruction sets are much easy when compared to complex instruction set computers (CISC).

4.7 GPS

GPS stands for global positioning system. It depends on the satellite system which is used to decide an object area. The system was first used by the US military and later extended to civilians. In many commercial products, GPS recipients are used, for example, vehicles, smartphones, watches, and GIS apps. This system provides not the only place but also provides time, climate, anywhere on or near the Earth by using four or more GPS satellites. The US Government maintains the complete network. Anybody who has a GPS receiver can access the service.

4.8 L293D Driver

L293D driver is an integrated chip that is used to control different motors. It consists of two H-shaped bridge circuits that will control two DC motors in a clockwise and counter-clockwise direction. As an input, it takes low current and provides high current at the output which is used to drive various loads such as stepper motor and DC motors.

4.9 Vibrator Sensor

The vibration sensor detects vibrations in the object. We use these sensors mostly in automobiles to detect accidents.

4.10 Fingerprint Sensor

The fingerprint scanner is one form of electronic safety or security system. Fingerprint varies among individuals. Every person has a unique fingerprint that identifies and authenticates an individual to allow or refuse access to the system. Both hardware and software are employed in this technology.

4.11 Iris Sensor

Iris recognition is one type of biometric identification. This uses methods for computational pattern recognition for one or both of the irises of the eyes. Iris of an individual is unique and stable. Iris scanner scans the unique patterns of iris that are the vivid circles in the eyes of the people. Iris scanner sends the invisible infrared rays to our naked eye and scans the iris pattern, i.e., the colored circles, excluding the eyelids, the lenses, and specular reflections that normally obscure irises. The scanned iris pattern is converted into pixels, and next it is compared with the pixels stored in the database. If both are matched, it will unlock the device otherwise the object is locked.

4.12 Mems Sensor

MEMS stands for the microelectromechanical system. It is a compact machine with both electronic and mechanical components. The MEMS sensor is a small size and consists of 1–100 μm of components.

4.13 Accelerometer

It is a gadget which measures gravitational acceleration, tilt, and vibration in a device in which it is installed. They measure the capacitance between two components. Some of the accelerometers which use the piezoelectric effect measure the very small voltage changes. To measure the acceleration in multiple directions, we use multiaxis sensors or multiple linear axis sensors.

4.14 Eye Blink Sensor

The eye blink sensor recognizes the eye-wide variations that differ according to the blink of the eye. The output is high if the eye is closed otherwise the response will be low.

4.15 Smoke Sensors

Smoke sensor detects the smoke in which it indirectly indicates the fire. Nowadays, these are extremely useful for detecting fire accidents in homes, offices, schools, and industries. As their utilization increases, the cost of the sensor reduces automatically.

4.16 Alcohol Sensor

The alcohol sensor detects the level of alcohol in your breath and works just like your typical breathalyzer. This sensor is highly sensitive and has a quick response time, and the output is similar to the alcohol consumption concentration which will be in analog form. It just requires only 0–3.3 V ADC.

4.17 Temperature Sensor

The temperature sensor senses the temperature, and these sensors are used most commonly. The sensing can be carried out either directly with the heat, source or remotely utilizing radiated power. Thermocouples, resistance temperature detectors (RTDs), thermistors, infrared and semiconductor sensors are some of the temperature sensors.

4.18 Transmitter

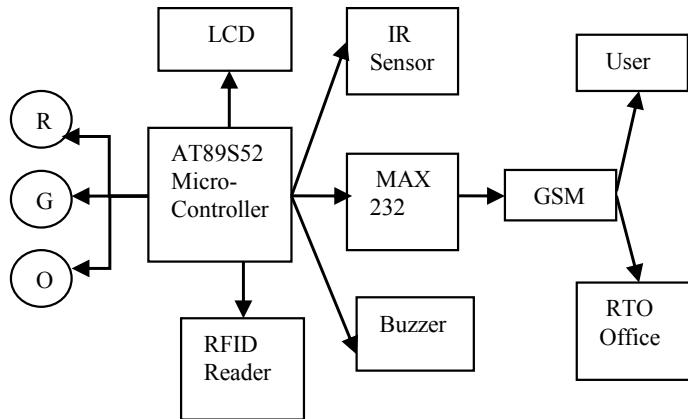
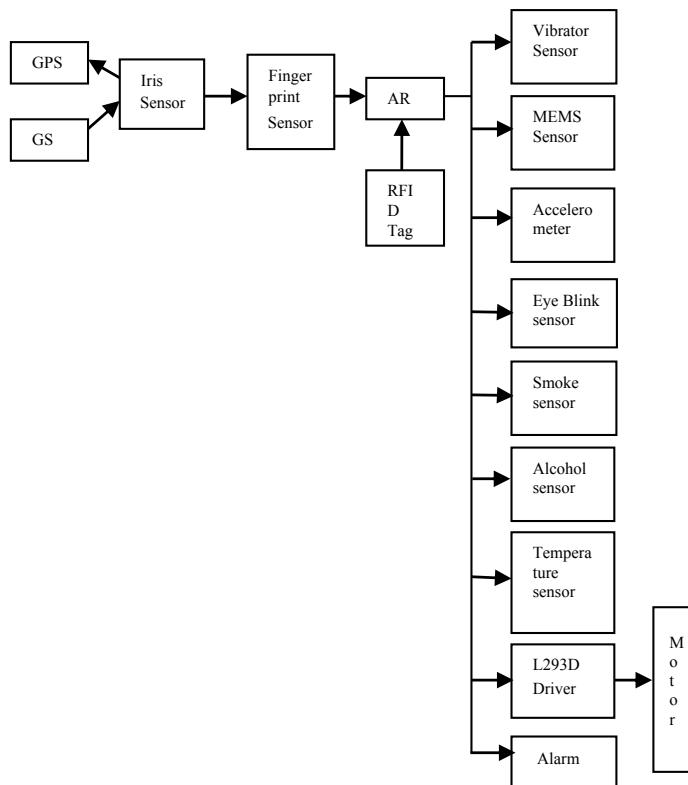
See Fig. 2.

4.19 Receiver

The framework contains two segments

- (1) Transmitter
- (2) Receiver.

Transmitter consists of microcontroller, reader, etc. The yellow light is used to exhibit that the red light will be ON, after that when the red light is ON, the reader will be started. So when any vehicle crosses the zebra crossing, then the reader peruses the data present in the tag and sends the data to ARM. So, the data received by RFID Reader is sent to the GSM network through Max 232 successive or sequential communication (Fig. 3).

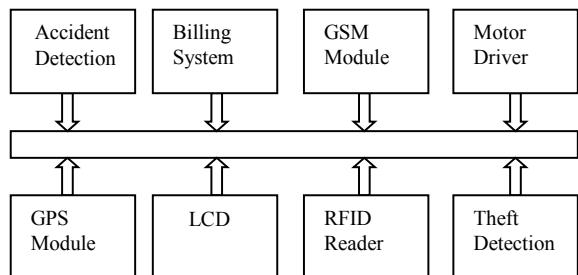
**Fig. 2** Structure of the transmitter system**Fig. 3** Structure of the receiver system

The charged amount is sent via the GSM network to the respective user. The system will be deactivated when the green light ON. Receiver consists of MEMS sensor, accelerometer, eye blink sensor, smoke sensor, alcohol sensor, temperature sensor, motor, buzzer, etc. When the mishap occurs, the vibration sensor and MEMS sensor recognize and send the sign to ARM7 and stop the engine by sending SMS via the GSM net and report the notification to the user's identity, nearby hospitals, and the police stations also. At the point when a vehicle has robbed, it is recognized by the proprietor through a message that the motor has turned over and afterward the vehicle is suddenly stopped which is remotely controlled by sending the stop message through GSM network; later the vehicle position is remotely accessed by the proprietor by using GPS technique. Not only these with the aid of remaining sensors, we can detect various functions like alcohol sensor which is used to detect the amount of alcohol consumed. If the consumed alcohol exceeds the limit, then it will send an alert message whether it is safe to drive or not. If it will not motor turn off automatically. Likewise, the temperature sensor measures the heat in the car, and smoke sensor is used to detect smoke. An eye blink sensor is used to sense the variance across the eye that will vary as per eye blink; if any one of these sensors crosses their limits, then the motor will turn off immediately by sending an alert message to the owner mobile.

5 Software Design

ADS integrated development environment is developed by the ARMS Company. ARM Developer Suite is the name for the ADS integrated development environment. And the new version of the ADS integrated development environment is ADS1.2. It is the latest version of IDE, and it supports all features of microcontroller, debug software, and JTAG simulation and also supports the language of Assembly, C, and C++. The Windows 98, Windows XP, Windows 2000, and Red Hat Linux allow you to run the software with the advantages of high compilation performance and rich application libraries, and this environment is free of cost (Fig. 4).

Fig. 4 Software modules



5.1 Software Composition

Implementation. The structure of this programming framework can be characterized into three modules: billing framework, the mishap identification, and the burglary discovery. These three modules are interlinked with each other by communicating between them via message. And these modules are executed in the principle program, and the communication module is executed in the program alone.

Design flow of Burglary Detection (Fig. 5)

Flowchart implementation of billing system (Fig. 6)

Flowchart implementation of accident detection (Fig. 7)

Fig. 5 Flowchart for theft detection

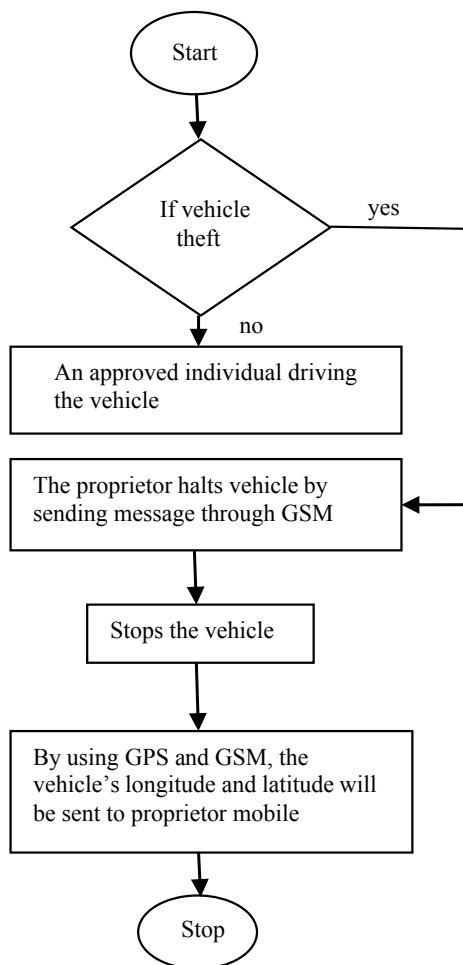


Fig. 6 Flowchart for billing system

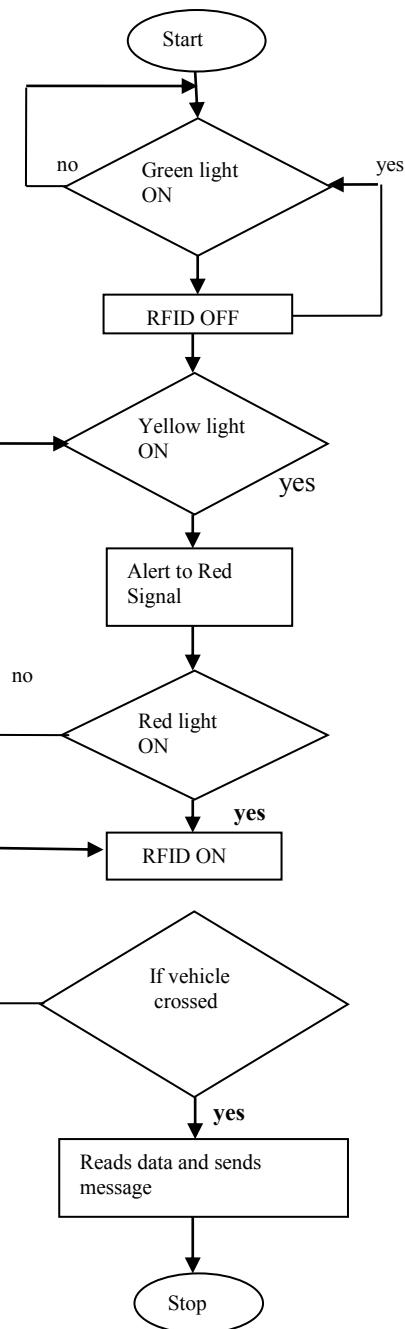
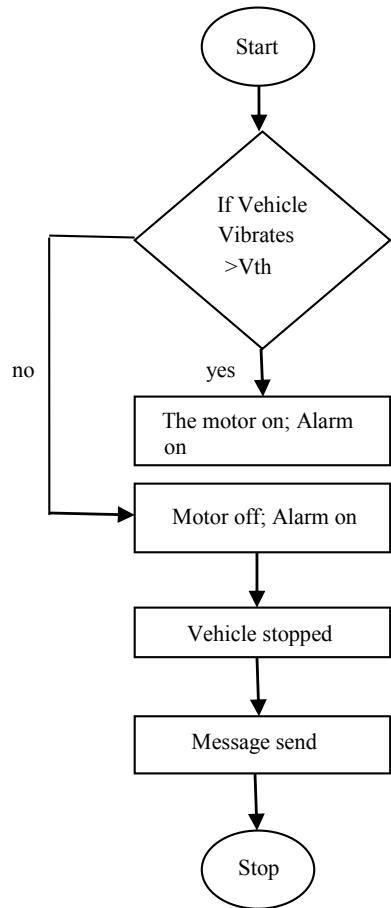


Fig. 7 Flowchart for accident detection



6 Result

With the help of this system, we can save a lot of people's life by providing certain necessary information at a certain time. Because this system consists mainly of three subsystems, when they detect accident, the message along with the longitude and latitude will be sent to the nearby police station, hospitals, and user identities, and it provides high security to the vehicle which prevents from theft; final one is automatic billing system, i.e., if anyone does not follow the rules of the traffic or signals, then the fine is charged from the cardholder and the amount will be sent to both owner and nearby RTO office.

7 Conclusion

This paper mainly focuses on vehicle tracking, increasing security in vehicle identification and accident detection system by using different modules in the system. There are three different subsystems in this system: programmed billing framework, car robbery recognition, and accident detection frameworks. These modules' functioning increases the system's total efficiency. This framework can be introduced at the passage of the checkpoints, places where it requires high security, for example, government buildings, armed force camps or at the borders, and so on. This system uses wireless communication for transmission of the data to the RTO office for automatic billing. The system is small, compact, and cost-effective.

Acknowledgements I am very grateful to express my deep sense of gratitude and respect toward my guide, Mr. K. Parvateesam for his excellent guidance right from the selection of the topic until the end of publishing the paper. He has given us tremendous support and both technical and moral front. I am thankful to Dr. T. K. Rama Krishna Rao, Principal and Mr. D. Kishore, Head of Department, ECE, Aditya College of Engineering and Tehnology (ACET), for their support during the completion and publishing the paper.

References

1. Inigo RM (1989) Application of machine vision to traffic monitoring and control. *IEEE Trans Veh Technol* 38(3):112–122
2. Lotufo RA, Morgan AD, Johnson AS (1990) Automatic numberplate recognition. In: Proceedings of image analysis for transport applications, IEE colloquium, vol 6, pp 1–6
3. Collins RT, Lipton AJ et al (2000) A system for video surveillance and monitoring. Carnegie Mellon University, Pittsburgh PA
4. Dickinson KW, Wan CL (1989) Road traffic monitoring using the trip system. In: IEEE second international conference on road traffic monitoring, vol 2, pp 56–60
5. Johnson AS, Bird BM (1990) Number; plate matching for automatic vehicle identification. In: Proceedings of electronic image and image processing in security and forensic science, IEEE Colloquium, vol 4, pp 1–8
6. McKenna K You Brandt A (1990) Moving object recognition using an adaptive background memory. Cappellini, time-varying image processing and moving object recognition. Elsevier, Amsterdam, pp 289–296
7. Grimson W, Stauffer C, Romano R et al (1998) Using adaptive tracking to classify and monitor activities in a site. In: IEEE conference ON computer vision and pattern recognition, pp 22–29
8. Stauffer C, Grimson WEL (2000) Learning patterns of activity using real-time tracking. *IEEE Trans PAMI* 22(8):747–757
9. Robson D (2007) Safe and secure in transit. Imaging and machine vision Europe, Dec. 2007—Jan. 2008. Online article: http://www.imveurope.com//features/feature.php?feature_id=50. Access 27 Nov 2009
10. Geetha M, Sangeeta B (2017) Anti-theft and tracking mechanism for vehicles using GSM and GPS
11. Tan H (2010) Design and implementation of vehicle monitoring system based on GSM/GIS/GPS
12. Hu H, Fang L (2009) Design and implementation of vehicle monitoring system based on GPS/GSM/GIS

13. Nagaraja BG, Rayappa R, Mahesh M, Patil CM, Manjunath TC (2008) Design & development of a GSM based vehicle theft control system
14. Mamun KA, Ashraf Z Anti-theft vehicle security system with preventive action
15. Ramaprasad SS, Sunil Kumar KN (2017) Intelligent traffic control system using GSM technology

Authentication of Vehicles in Vehicular Clouds: An Agent-Based Approach



Shailaja S. Mudengudi and Mahabaleshwar S. Kakkasageri

Abstract Authentication of vehicle nodes in Vehicular Cloud (VC) network is a major aspect for providing and usage of services from cloud services providers. Most of the authentication methods use encryption for authenticating the user, which depends on the secret key. But for some reason if the key is compromised, then the complete authentication framework collapses. In this paper, we present a novel agent-based authentication framework which acts as a second layer of authentication process which follows the encryption. After the communication link is established, the framework is used to monitor the behavior of the vehicle node. Authentication agent's performance is measured in terms of computational delay with number of attributes and test samples to show our proposed method takes very small delay in authenticating the user, which makes it appropriate for VC as time plays a crucial role.

1 Introduction

VANET is self-organizing wireless network with no fixed infrastructure. In order to handle VANET, we should be able to handle multiple connections between the nodes, dynamic routing protocol for each node as the network is not fixed. These can be handled efficiently by a class of inter-machine (m2m) networks [1]. VANET aims at facilitating vehicle nodes with safety and entertainment-related services. To render the services, mainly two types of communication links are used (i) vehicle-to-vehicle (V2V) communication (ii) vehicle-to-infrastructure (V2I) communication

S. S. Mudengudi (✉)

Department of Electronics and Communication Engineering,
Tontadarya College of Engineering, Gadag, Karnataka 582101, India
e-mail: psmssm@gmail.com

M. S. Kakkasageri

Electronics and Communication Engineering Department, Basaveshwar Engineering College (Autonomous), Bagalkot, Karnataka 597102, India
e-mail: mskec@becbgk.edu

[2]. Due to its wireless nature, attackers can violate authenticity, confidentiality and privacy of the vehicle node [3]. Vehicular cloud networking is an emerging technology which is a combination of vehicular cloud, infrastructure cloud and traditional back-end (IT) cloud. However, the implementation of the vehicular cloud network is now possible with the technology like Apple CarPlay, BMW Connected Drive. Vehicular cloud computing is based on mobile cloud computing which is able to provide the driver and vehicle user with resources for the ease and secured driving on effective low cost. A framework of vehicular cloud is presented in [4] collaborates the smart sensing and communication capabilities of vehicle nodes so as to present potential computing machines. As more amount of data is transferred for the secured drive of the vehicle user, the tampering of the data by the hackers is a major issue. So it is basic need to authenticate the vehicle node or the cloud service provider (CSP). The one-time authentication code is generated using selection and combination from key pool which is pre-stored at the time of registration. Authentication can be text based, bio-metric based, password based or even based on a third-party which authenticates the nodes. To address the issue of compromised secret key in this paper, we have presented KNN-based authentication method which can be installed on the any cryptography based authentication framework. If there is a notably change in the behavior of the node, it is detected immediately and classified as malicious node. Further, after detection the node may be added to black list and services can be terminated. Our work is light weight and simple as the KNN classifier used is very simple, less complex and gives fast accurate results.

The paper is organized as follows. In Sect. 2 Proposed authentication scheme which is based on KNN is presented followed by results and discussion based on the results. The work is concluded in Sect. 3.

2 Agent-Based Authentication Framework

In this section, we present the proposed agent-based authentication framework using the KNN algorithm. The vehicular cloud network (VCN) architecture is considered as shown in Fig. 1. It consist of main three units—vehicular cloud, cloud service provider (CSP) and the central authority (CA). VC is formed by resources shared by the vehicle which are having excess underutilized resources, resources from RSU and CSP's. Communication is established using vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) links. Every vehicle has to register to CA, which is a trusted third-party. During registration process, a unique identification (ID) is provided by the CA. This ID is used by the vehicle node to render all the services in the VCN. CA maintains the database of each vehicle. Database consists of details such as vehicle ID, vehicle details, cloud services opted, past history, etc. Vehicle in the network can access resources from the cloud anytime via the CA. The request from the vehicle will be processed by CA and forwarded to the cloud. The communication involves RSU of the vehicles vicinity. As vehicle nodes may be static or in motion, the RSU keeps track of the location of vehicle. Our agent is installed at vehicle nodes, RSU and the CA. As

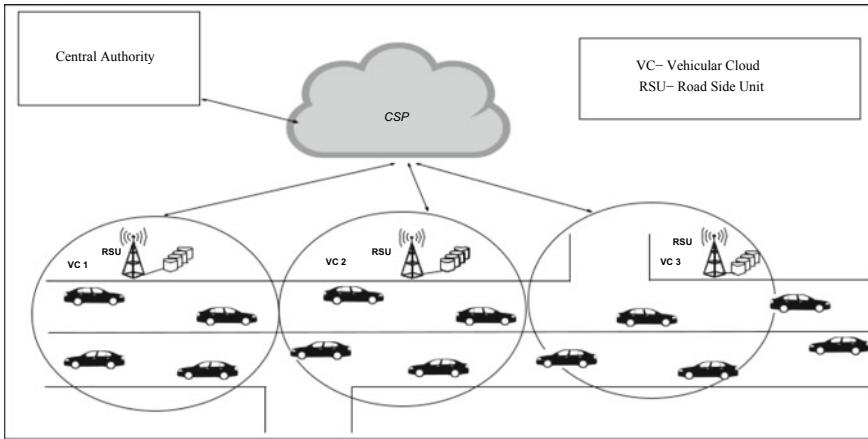


Fig. 1 Vehicular cloud network

it is lightweight, there is no substantial overhead added. The data used for processing is already present in the storage. We use KNN classifier to detect if any malicious node has acquired the access to unauthorized resources or any node being compromised due to secret key. There will be deviation in the behavior of malicious and legitimate nodes. This change will be detected by the KNN classifier, which classifies between malicious and legitimate users. The first-tier authentication involves the conventional authentication methods using encryption. Second tier involves our proposed agent. As the VCC requires fast and accurate response and minimum overhead in terms of installation, KNN is the best fit.

2.1 *K-Nearest Neighbor Classifier*

In order to improvise the presentation of choices to the users, the on line Web sites use the data mining and recommendation techniques. These techniques also makes it easy to find the right choices with in less time. Some of the data mining classifiers are the k-nearest neighbor classifier, rule-based classifier, decision tree classifier and Bayesian classifier. Among these, k-nearest neighbor classifier (KNN) is best adopted, when there is no prior knowledge about data distribution. The other advantages of KNN are scalability, reduced error, faster results, etc. An instance-based learner, KNN compare the test sample with the stored training sample and classify based on the similarities between them. The similarities are calculated using distance functions such as Euclidean distance, Manhattan distance and cosine similarity.

Let A_i be the test sample with the attributes $(A_{i1}, A_{i2}, A_{i3}, \dots, A_{in})$, where n is the number of attributes. Let m be the total samples used for training. In order to calculate the distance between the samples, there is need to normalize the large

ranges from dominating the smaller ranges. Any of the normalization methods can be used for example min–max as shown in Eq. 1.

$$\text{Norm} = \frac{V - \min Y}{\max Y - \min Y} \quad (1)$$

where $\min Y$ and $\max Y$ are the minimum and maximum values that an attribute Y can take. V can take values in the range $[0, 1]$. The Euclidean distance between the test sample A_i and A_t is calculated as shown in Eq. 2.

$$D(A_i, A_t) = \sqrt{(A_{i1} - A_{t1})^2} \quad (2)$$

In general, Eq. 2 can be reduced to Eq. 3

$$D(A_i, A_t) = \sqrt{\sum_{m=1}^n (A_{im} - A_{tm})^2} \quad (3)$$

2.2 Authenticating Agency

Proposed authentication agency based on KNN is shown in Fig. 2. Agency mainly has three entities (1) Authentication Manager Agent (AMA); (2) Authentication Information Collection Agent (AICA); and (3) Authentication Knowledge Base (AKB).

- **Authentication Manager Agent (AMA):** It is the agent responsible for classification of nodes into authenticated or malicious nodes. AMA runs in the vehicle nodes and the cloud service provider. Initially, it creates AICA and the AKB.

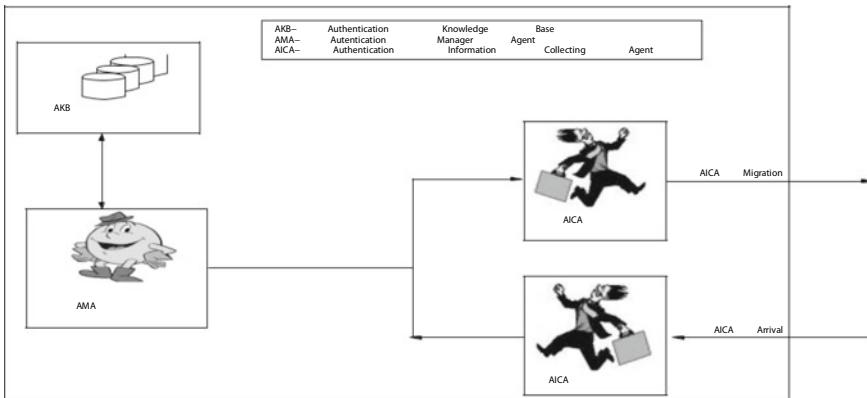


Fig. 2 Authentication agency

When the communication is established between the vehicle node and the cloud service provider (CSP), the details about the communication link are monitored by the neighbors, road side unit and the central agency. The details are updated frequently. All the activities of the AICA and AKB are co-ordinated, controlled and monitored by the AMA.

- **Authentication Information Collecting Agent (AICA):** This agent is responsible for collecting the attributes or the data related required for authentication. It is activated by the AMA, which in turn initiates the process of the AICA data collection. As the agent is mobile and can create clones, it travels around the network reaching the VC, CSP's and RSU's eventually collecting the information needed. Every entity it visits the data collected will be updated in the database AKB which is coordinated by the AMA.
- **Authentication Knowledge Base (AKB):** It comprises of information regarding vehicle IDs, CSP's IDs, available bandwidth, vehicle status (connected/disconnected to network), list of services which can be provided by the available cloud services. Data to AKB is read from or updated by AMA and AICA. It also has stored the normal attribute values for each vehicle node and the CSP's as shown in Eq. 4 where Vb is the node, 'a' are attributes and 'm' is number of attributes.

$$Vb = [a_1, a_2, \dots, a_m] \quad (4)$$

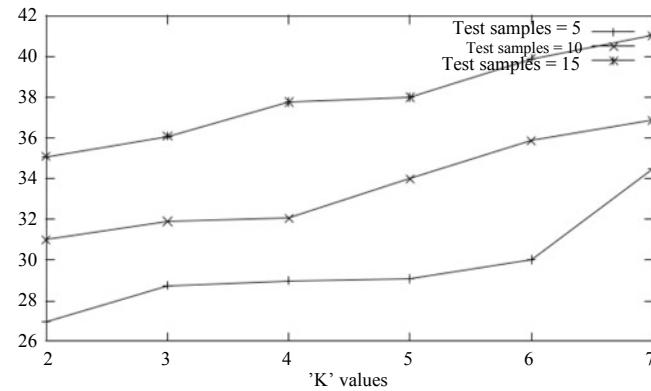
Now, the complete working of the proposed work is as follows:

1. Registration: The vehicle node as well as the cloud service provider (CSP) register with the central authority (CA) as in [5]. A key is generated K_p by the CA and sent to the registering entity, which is stored in their respective KB.
 $K_p = [ID, N, T]$, where ID—Identity of the vehicle node, N —Random number, T —Time stamp. The same key is utilized for the further communication.
2. Service request: Now, the vehicle which needs the service request the CA to provide the most trusted CSP which provides the requested service. The CA uses the method [6] to provide the vehicle node with most trusted CSP as per the requirements of the vehicle node.
3. Communication establishment: When the vehicle node chooses the CSP to utilize the service, it starts the communication using the key K_p . For every time period ' T ', the AMA triggers AICA to collect the information about the authentication related attributes related to the vehicle node (Vb) from the vehicular cloud network. $Vb_i = [A_1, A_2, \dots, A_m]$, where 'A' are attributes and 'm' are number of attributes. The AICA updates the collected information to the AKB and the AMA. The AMA then performs the KNN classification on the updated data, i.e., Vbi and the data stored in the AKB, i.e., Vb . Based on the outcome of the KNN classifier, the node is decided as either genuine node or malicious node (i.e., class-C). Next, the AMA of the node which wants to check whether Vb is genuine node or malicious node, triggers its AICA. The AICA collects the opinion of all the neighboring nodes. Now, the decision is made by checking the number of positive opinions and number of negative opinions using KNN. Based on the outcome of

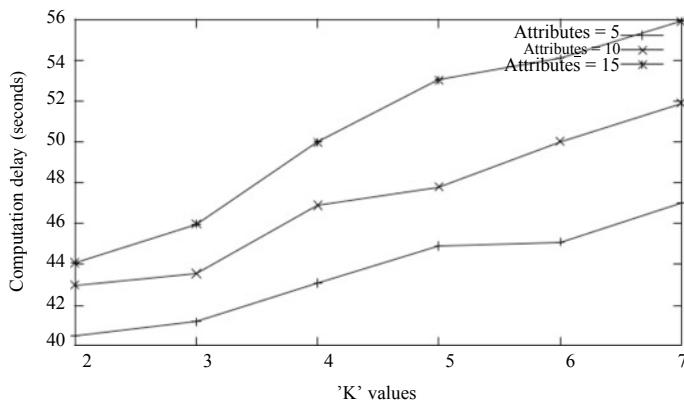
the KNN classifier, if the node is grouped as malicious, the service provided by the cloud can be truncated else the service will continue as usual.

The proposed scheme has been simulated using C++ language for various network scenarios. The simulation parameters considered to evaluate the performance effectiveness of proposed work is computational delay.

- **Computational delay with constant number of attributes:** The AMA calculates the number of nearest neighbors, i.e., 'k' and decides the class depending on class of the neighbor nodes obtained. The time required to classify with constant number of attributes but varying test samples.
- **Computational delay with constant number of test samples:** It is the time required for classification process for different number of data set samples.



(a) Computational Delay Vs. 'K' Values



(b) Computational Delay Vs. 'K' Values

Fig. 3 Computational delay for different test samples and attributes

2.3 Result Analysis

The results of the simulation are presented and discussed in this section.

For different values of ‘ K ’ versus computational delay for authenticating, a node is plotted in Fig. 3a. It is observed that as ‘ K ’ increases, the computational delay tends to increase more. The reason as ‘ K ’ increases, the number of nearest neighbors also increases. So it takes more time to make the decision for the AMA.

The computational time for different ‘ K ’ values by varying number of attributes is shown in Fig. 3b. It can be observed that as the ‘ K ’ value increases, the computational time also increases. When the number of attributes considered are less, the computation time is less. Similarly, as the number of attributes increases, the computational delay increases as the data to be processed also increases. As well the delay is affected by the time taken by the AICA to collect the information from the neighboring nodes.

3 Conclusion

In this paper, we have put forth the concept of authenticating a vehicle for VC using software agent. The software agent takes a major role in decision making. For the decision making, KNN classification algorithm is used. The KNN algorithm is the simplest and the easiest to implement. This speeds up the decision-making process in our presented concept. This system can also be used as a second layer of authentication in case of malicious node investigation. One more advantage of KNN algorithm is we can prioritize the opinions of the trusted and the nearby nodes.

References

1. Krundyshev V, Kalinin M, Zegzhda P (2018) Artificial swarm algorithm for VANET protection against routing attacks. In: 2018 IEEE industrial cyber-physical systems (ICPS), vol 5, pp 795–800
2. Deeksha, Kumar A, Bansal M (2017) A review on VANET security attacks and their countermeasure. In: 2017 4th international conference on signal processing, computing and control (ISPCC), vol 7, pp 580–585
3. Ahmad F, Kazim M, Adnane A, Awad A (2015) Vehicular cloud networks: architecture, applications and security issues. In: 2015 IEEE/ACM 8th international conference on utility and cloud computing (UCC), vol 16, pp 571–576
4. Hassanein HS, Abdelhamid S, Elgazzar K (2015) A framework for vehicular cloud computing. In: 2015 international conference on connected vehicles and expo (ICCVE), vol 8, pp 238–239
5. Thayananthan V, Albeshri A (2015) Big data security issues based on quantum cryptography and privacy with authentication for mobile data center. Procedia Comput Sci 50:149–156
6. Mudengudi SS, Kakkasageri MS (2017) Establishing trust between vehicles in vehicular clouds: an agent based approach. In: 2017 international conference on smart technologies for smart nation (SmartTechCon), vol 03, pp 529–533

ANN-Based Model to Predict Reference Evapotranspiration for Irrigation Estimation



Neha K. Nawandar, Naveen Cheggoju, and Vishal Satpute

Abstract Accurate estimation of water needs manages both crop yield and water loss occurring due to imprecise water supply to the crops. In this context, this paper proposes an artificial neural network (ANN)-based model to estimate reference evapotranspiration (ET_0) which is crucial in deciding the water needs of a crop. The Penman–Montieth (PM) method is considered as a benchmark by the Food and Agriculture Organization of the United Nations (FAO), but it lacks usage in deployments due to its heavy input requirements. The model presented in this paper targets to mimic the PM method and succeeds in obtaining the same results using minimum input variables. Corresponding results have been mentioned, which show that an infinitesimally small error of a maximum 0.4 mm/day is present in the output predicted using the proposed ANN model. It is found that the predicted ET_0 has no ill-effect on the future computations.

Keywords Artificial neural network · Reference evapotranspiration · Irrigation · FAO

1 Introduction

Agriculture plays a major role in the Indian economy as the livelihood of most of its people depends upon it and it has to cater to the ever increasing food demands. To meet the growing demands, crop yield must be improved. It is dependant upon

N. K. Nawandar (✉) · N. Cheggoju · V. Satpute

Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

e-mail: nehanawandar@gmail.com

N. Cheggoju

e-mail: naninaveen.ch@gmail.com

V. Satpute

e-mail: vrsatpute@ece.vnit.ac.in

various factors like water supplied, nutrient availability, ambient conditions, etc. Out of these, the amount of water provided to the crops is the most crucial element. As a result, water management in agriculture, i.e. precise estimation of crop water needs plays an important role.

The Food and Agriculture Organization of United Nations (FAO) [1] identifies a parameter as reference or potential evapotranspiration (ET_0) which can be used to estimate the water needs of a crop. Computed in mm/time, it is the amount of water lost due to evaporation from the soil and transpiration from the crop surface. To measure this quantity, there are radiation-based methods like Penman–Montieth (PM), temperature-based methods like Hargreaves, Blaney–Criddle, etc., where PM [2] is considered as the benchmark. It is the standard used to compute ET_0 , when all the required parameters are available from weather stations, and in case of non-availability the other temperature based methods could be used. Most of these techniques compute an approximated value of ET_0 which could reflect on the final estimated water needs. This is where machine learning algorithms come into play for mimicking the PM method using minimal inputs. In recent years, researchers have presented proposals for ET_0 estimation using algorithms like support vector machine (SVM), artificial neural networks (ANN), extreme learning machine (ELM) with the help of climatic data. Some of these works have been mentioned further.

Chauhan and Shrivastava [3] presented one-month ahead estimation of ET_0 for Mahanadi reservoir project using different ANN algorithms. The work utilized a 3-9-4 ANN which predicted the output using three different learning models and made a comparison based on the obtained results. Tabari and Talaee [4] demonstrates four models for ET_0 estimation in Iran, where data from weather station was collected to train and develop the network. Authors have also performed statistical analysis to find out best of their presented models. Trajkovic et al. [5] also prove reliability of ANN in ET_0 estimation and obtained statistical properties by using weather data of 20 years in southern Europe. Another work that compares ANN with PM method has been presented in [6], where the ANN is trained and validated and comparison against PM is made. Pandorfi et al. [7] predict ET_0 using ANN specific to sweet pepper for greenhouses using climatic data, radiation, and wind speed of 4 months.

Data gathered using multiple sensors along with the weather station data is also useful in estimating evapotranspiration. Kelley and Pardyjak [8] used low cost sensors for short training times and estimate site specific ET_0 . Similar works based on data-driven estimation have been presented in [9–13]. Apart from ANN, researchers have also used ELM- and SVM-based estimation. One such work proposed using ELM applicable to Bihar, India can be seen from [14] and another work using Jodhpur and Pali weather stations has been presented in [15].

This shows the extent to which prediction of reference evapotranspiration is important. Thus, this paper focuses on the same problem statement and presents an ANN-based model for the parameter prediction. The remaining paper has been organized as follows: Sect. 2 presents the proposed work; the results of which can be seen from Sect. 3. Section 4 concludes the paper and references have been mentioned at the end.

2 Proposed Work

This section starts with the dataset used for prediction followed by the proposed model and its pseudo-code which achieves the same.

2.1 Data Collection and Dataset

The ambient temperature minimum (t_{\min} : °C) and maximum (t_{\max} : °C) value, humidity (h : %), wind speed ($u2$: km/day) at 2 m, mean sunshine hours (p : hrs), and the radiation (r : MJ = m² day) values are the parameters using which estimation is done. Out of these parameters, the ambient temperature and humidity are acquired using sensors interfaced to a micro-controller. This gathers the data and saves it for further usage. The remaining parameters have been obtained using CLIMWAT [16]. It is a climatic database and has various weather stations for which data is available for use in combination with CROPWAT program [17].

2.2 Estimation Model

An artificial neural network (ANN) model has been used for evapotranspiration estimation. The parameters mentioned in 2.1 act as the inputs to the model and ET_0 is the predicted output. The model thus has six nodes and one node in the input and output layers, respectively. Here, the aim is to replicate the PM model using minimum parameters in order to precisely estimate ET_0 to match to its output. Algorithm 1 shows the pseudo-code for estimation and is implemented in Python.

Initially, using the sensor data and data from CLIMWAT database, input samples are created. The output using PM method is obtained using CROPWAT software which is used as the target output for training the samples. This output acts as the reference and is used to compute the error in the estimated output values. The dataset is divided in 3:1 ratio for training and testing purposes. Validation set is randomly taken from the train set and is 20% of it. The model is trained using back propagation and Adam optimizer is used as it is one of the best optimizers today. It is used to minimize the loss function which is the mean absolute error.

The model is trained for 200 epochs and losses are obtained after every epoch to check for model accuracy. Weights obtained post-training are saved and weights of the model with maximum accuracy are used for testing and further estimation of evapotranspiration. Apart from computing the error, the estimated values are used to estimate the water needs. The complete results have been presented and discussed in Sect. 3.

Algorithm 1: ET₀ estimation (t_{min}, t_{max}, h, u2, p, r, ET₀)

```

1: Inputs: tmin, tmax, h, u2, p, r
2: Output: ET0
3: for data do
4:   assign: train, test
5:   for train data do
6:     create NN model 6-x-x-1
7:     use activations for layers: relu, linear
8:     define cost function, optimizer
9:     compile and fit model
10:    save weights and model
11:  end for
12:  for test data do
13:    load model with least loss
14:    compile and predict on test data
15:    compute error
16:  end for
17: end for
18: End

```

3 Results and Discussions

This section presents samples of the sensor data, CLIMWAT database values, ET₀ using PM by CROPWAT and the ANN model results. As mentioned in Sect. 2, the presented estimation using ANN is a data driven model, that uses climatic data to predict the evapotranspiration. This predicted value needs to be validated and for this purpose, ET₀ achieved using the benchmark method (PM) is considered to be the basis for comparison. A dataset consisting of numerous samples of required inputs has been used for training and testing purpose. Table 1 gives some sample examples of the six different input parameters considered by the model.

As mentioned earlier, the values have been obtained using sensors and CLIMWAT database. The database consists of data accounting to various weather stations around the world. The weather stations in India for which database is available can be seen from Fig. 1. Using this data in the CROPWAT software provides with evapotranspiration using PM method. This computation for all the months for a sample weather station can be seen from Fig. 2.

In a similar manner, values for more input samples are computed using the software and it is set as the target output which is to be achieved by the ANN model. The inputs and the target output are then used to train the network to achieve desired accuracy (trained for 200 epochs). Post-training, the hyperparameters are obtained and the best model, i.e., model with least loss and maximum accuracy is used for testing and future predictions.

Table 2 gives comparison between the PM method and ANN estimated output for some input samples. It also mentions the error present in the predicted output. These

Table 1 Input samples used for training and testing the model

Parameter	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
t_{\max}	24.3	24.2	23	22.3	21.3	18.8	14.3	11.7	21.7	22.8	24.5	26.1	26	23.9	23.5	23.6	23.5
t_{\min}	40.7	36.1	31.1	30.7	31.1	32.7	31	29.9	31.4	31.1	31.7	32.4	32.1	29.4	28.5	28.5	28.7
h	31	55	71	71	68	48	42	40	83	83	86	87	89	100	100	100	100
$u2$	190	225	225	199	138	95	78	69	121	130	121	130	156	147	147	112	104
p	8.9	6.2	3.3	4	5.3	7.3	7.6	7.8	9.1	9.3	9.5	8.4	7.2	4.8	4.3	4.8	6
r	23.3	19.2	14.8	15.6	16.7	17.8	16.3	15.5	20	21.8	23.6	22.5	20.5	16.6	15.8	16.7	18.3

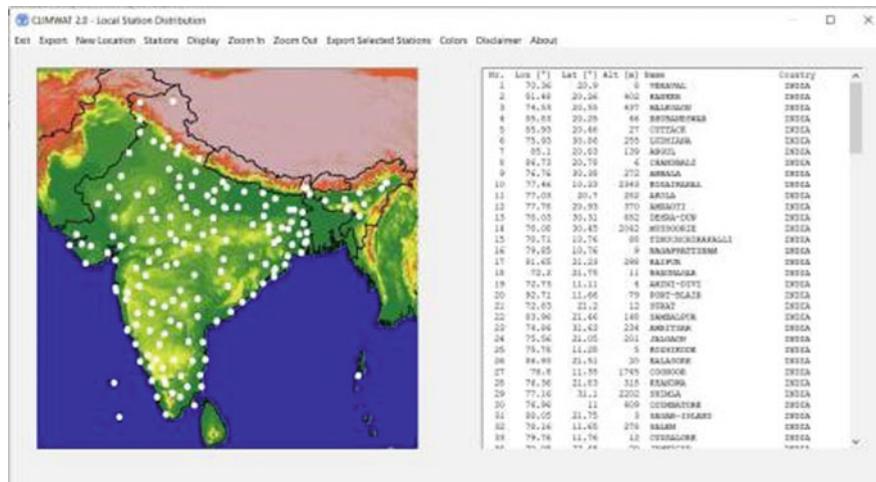


Fig. 1 CLIMWAT database with weather stations in India (marked by white bold dots) and sample list (of 33 stations)

Country		Location 12		Station		AMRAOTI		
Altitude	m.	Latitude	20.93 °N	Longitude	77.78 °E			
Month	Min Temp	Max Temp	Humidity	Wind	Sun	Rad	ET ₀	
	°C	°C	%	km/day	hours	MJ/m ² /day	mm/day	
January	15.5	28.9	38	147	7.9	16.2	4.00	
February	17.2	31.8	30	156	8.7	19.1	4.97	
March	21.2	36.2	28	164	8.8	21.3	6.14	
April	25.2	39.7	26	173	8.5	22.3	7.13	
May	27.8	42.2	27	216	8.3	22.5	8.47	
June	25.7	37.0	49	268	6.2	19.2	7.15	
July	23.4	30.3	74	294	4.1	16.1	4.54	
August	23.0	29.8	75	251	4.3	15.9	4.18	
September	22.7	30.6	72	173	5.2	16.5	4.10	
October	20.8	32.1	50	130	7.3	17.7	4.50	
November	17.4	30.1	40	130	8.0	16.7	4.12	
December	15.1	28.6	40	130	8.1	15.8	3.68	
Average	21.3	33.1	46	186	7.1	18.3	5.25	

Fig. 2 ET₀ estimation (for sample station) using Penman–Montieth method via the CROPWAT software

Table 2 Comparison of ET_0 estimated using PM and ANN model and error in the prediction for sample inputs

Method	1	2	3	4	5	6	7	8	9	10
Proposed	4.2587	4.2186	4.0887	4.1333	2.9356	3.9582	4.3922	4.7545	4.7364	4.3771
PM	4.3500	4.2300	4.0900	4.1900	2.9600	4.0000	4.3900	4.8100	4.8200	4.4600
Error	-0.0913	-0.0114	-0.0013	-0.0567	-0.0244	-0.0418	0.0022	-0.0555	-0.0836	-0.0829

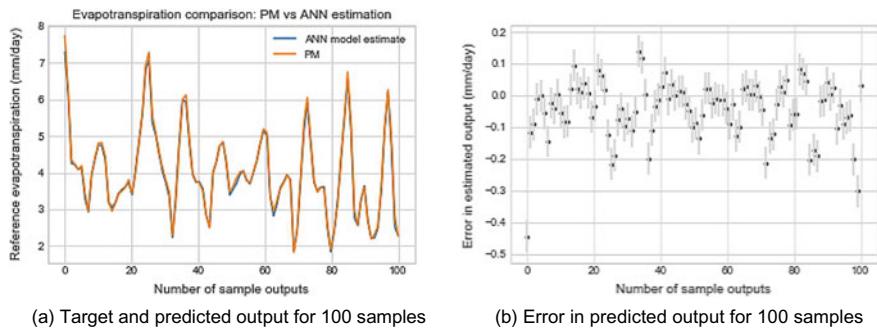


Fig. 3 Penman–Montieth versus ANN model: obtained value and error comparison of ET

values can also be seen in a plot between the target and obtained output from Fig. 3. From Fig. 3a, it can be observed that the predicted output follows the target with high accuracy and is almost same. The error for some 100 sample inputs is also plotted and is shown in Fig. 3b. Thus the model accurately predicts the evapotranspiration with infinitesimally small error.

4 Conclusion

Precise computation of crop water needs plays a pivotal role not only to reduce crop loss but also to avoid unnecessary water wastage. Reference or potential evapotranspiration estimation is one of the ways to technically compute the irrigation or water requirement. On this line, this paper has presented and discussed ET_0 prediction using an ANN model. The model uses sensor data and climatic data from CLIMWAT database to form samples for training and testing purpose. The target output, i.e., the Penman–Montieth method is obtained by using the inputs in CROPWAT software, whereas the estimated output is predicted using the ANN model. These outputs have been compared to compute the loss in predicted values which is found out to be infinitesimally small, which validates the usage of ANN model for ET_0 estimation.

Acknowledgements The work presented in this paper has been supported by Visvesvaraya Ph.D. scheme, Ministry of Electronics and Information Technology.

References

1. Brouwer C, Heibloem M (1986) Irrigation water management: irrigation water needs. <http://www.fao.org/docrep/s2022e/s2022e00htm#Contents>
2. Bogawski P, Bednorz E (2014) Comparison and validation of selected evapotranspiration models for conditions in Poland (central Europe). Water Resour Manage 28(14):5021–5038

3. Chauhan S, Shrivastava R (2009) Reference evapotranspiration forecasting using different artificial neural networks algorithms. *Can J Civ Eng* 36(9):1491–1505
4. Tabari H, Talaee PH (2013) Multilayer perceptron for reference evapotranspiration estimation in a semiarid region. *Neural Comput Appl* 23(2):341–348
5. Trajkovic S, Todorovic B, Stankovic M (2003) Forecasting of reference evapotranspiration by artificial neural networks. *J Irrig Drainage Eng* 129(6):454–457
6. Kumar M, Raghuwanshi N, Singh R, Wallender W, Pruitt W (2002) Estimating evapotranspiration using artificial neural network. *J Irrig Drainage Eng* 128(4):224–233
7. Pandorfi H, Bezerra AC, Atarassi RT, Vieira F, Barbosa Filho JA, Guiselini C (2016) Artificial neural networks employment in the prediction of evapotranspiration of greenhouse-grown sweet pepper. *Revista Brasileira de Engenharia Agrícola e Ambiental* 20(6):507–512
8. Kelley J, Pardyjak ER (2019) Using neural networks to estimate site-specific crop evapotranspiration with low-cost sensors. *Agronomy* 9(2):108
9. Feng Y, Peng Y, Cui N, Gong D, Zhang K (2017) Modeling reference evapotranspiration using extreme learning machine and generalized regression neural network only with temperature data. *Comput Electron Agric* 136:71–78
10. Han H, Bai J, Yan J, Yang H, Ma G (2019) A combined drought monitoring index based on multi-sensor remote sensing data and machine learning. *Geocarto Int* 1–16
11. Najmaddin PM, Whelan MJ, Balzter H (2017) Estimating daily reference evapotranspiration in a semi-arid region using remote sensing data. *Remote Sens* 9(8):779
12. Pandey P, Nyori T, Pandey V (2017) Estimation of reference evapotranspiration using data driven techniques under limited data conditions. *Model Earth Syst Environ* 3(4):1449–1461
13. Reis MM, da Silva AJ, Junior JZ, Santos LDT, Azevedo AM, Lopes EMG (2019) Empirical and learning machine approaches to estimating reference evapotranspiration based on temperature data. *Comput Electron Agric* 165(104):937
14. Kumar D, Adamowski J, Suresh R, Ozga-Zielinski B (2016) Estimating evapotranspiration using an extreme learning machine model: case study in North Bihar, India. *J Irrig Drainage Eng* 142(9):04016032
15. Patil AP, Deka PC (2016) An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. *Comput Electron Agric* 121:385–392
16. Muñoz G, Grieser J (2006) Climwat 2.0 for CROPWAT. *Water Resour Dev Manage Serv* 1–5
17. Smith M (1992) CROPWAT: a computer program for irrigation planning and management. *Food Agric Org* 46

Entropy: A New Parameter for Image Deciphering



Naveen Cheggoju, Neha K. Nawandar, and Vishal R. Satpute

Abstract Data and information, these two terms may look quite similar and also sometimes used with the same meaning. But in-depth, these terms are completely different from each other. This paper mainly focuses on this difference to extract the encrypted information from the data (importantly image data). Here, every aspect of the study and the way of formulation are explained using “Einstein’s theory of relativity” and “Arrow of time” for better understanding. Einstein’s theory is used to explain “how similar things can be seen in different ways?” (here data and information) and arrow of time is used to explain “how the correct information can be extracted from the data?”. To explain this concept, an image encrypted using a chaotic logistic map is used. The edge information present in this data is taken as the parameter to differentiate the unorganized data and organized data (correct information). Here, a new parameter called “Randomness parameter (R_p)” is introduced, which gives the entropy (randomness) of the data. The outcome of this parameter is used to differentiate the data and correct information.

1 Introduction

The “Einstein’s theory of relativity” [1, 2] states that time travel is possible through the concept of “Time-Dilation” [3, 4] which occurs due to the differences in motion and gravity. For a person moving faster or for a person near a very high gravitational field, the clock runs slower in relative with the stationary clock on the Earth. This

N. Cheggoju (✉) · N. K. Nawandar · V. R. Satpute

Image Processing and Computer Vision Lab, Department of Electronics and Communication Engineering, Visvesvaraya National Institute of Technology, Nagpur, India

e-mail: nani-naveen.ch@gmail.com

N. K. Nawandar

e-mail: nehanawandar@gmail.com

V. R. Satpute

e-mail: vrsatpute@ece.vnit.ac.in

phenomenon of time dilation makes the time travel possible in the current universe. But according to the theory of “Arrow of time,” travel to past is not possible because of the increase in the entropy [5–7]. To explain these two phenomenon of time dilation and arrow of time, let us consider an example of a glass of milk on table. This state of the glass on table is considered as its present. Now if somehow due to some force, the glass with milk has fell down from table and broke into pieces by spilling the milk out, this state of glass is now considered as future of the glass on table. To get back the glass into its present, i.e., past for the broken glass, one need to get back its original shape by reducing its entropy.

In general, the term “entropy” is used to measure the randomness in a given data [7, 8]. To get a good idea of this term entropy, one should be able to clearly differentiate between the data and information. Data can be defined as any text and/or numbers and/or symbols without any relationship among them. Information can be defined as the data with relationship, i.e., the structured arrangement of data which has some meaningful context [9–11]. Here, in this glass example, data corresponds to the broken pieces, information corresponds to the systematic arrangement of the broken pieces and the spilled milk and how that arrangement is broken.

The glass example is not a reversible process here, because we have only data, i.e., the broken pieces. We do not have any idea of how the glass looks and how it has broken into pieces, i.e., the information. It means any natural phenomenon is a irreversible process because we have only one parameter called data with us. Any natural event like waves in the sea, wind-blown by tree, sound generated by living beings, etc., are irreversible processes with respect to the respective events mentioned.

Now, if we replace the glass example with image encryption. The original image can be compared with glass of milk on table and the encrypted image can be compared with the broken glass. This comparison shows that we are left with only data, without any prior knowledge of it. This data can form meaningful information only if the data is arranged in a specific pattern. Otherwise, the data would become meaningless and cannot be used for any other applications. But, what happens if the concepts of data, information, and entropy are applied to image encryption? Will it work and get back the original information? If it is possible, what are the things we should know about the process? How that things should be applied on the data to get back the original information?

These questions are answered in the upcoming sections and the rest of the paper is organized as follows: image encryption and the process of applying the concept of data, information, and entropy to break the encryption is discussed in Sect. 2; results and discussions for the proposed algorithm are presented in Sect. 3, and finally, the conclusion is given in Sect. 4.

2 Proposed Algorithm and Its Justification

To begin with this section, let us have a look at what details does an encrypted image gives us? If we observe an encrypted image or ciphered image, it has only data, i.e., some text or numbers without any relationship between them. Along with the data, the algorithm used for encryption is also known to us. Algorithm tells us how the image is transformed to get into this present state but it cannot tell us what triggered it to come to this position or what is its initial position. It means that the algorithm can give us some part of information. In addition to data and some part of information, encrypted image can also tell us that the entropy has increased, which means the randomness has increased. But the encrypted image cannot give us any information regarding about the contents in the original image.

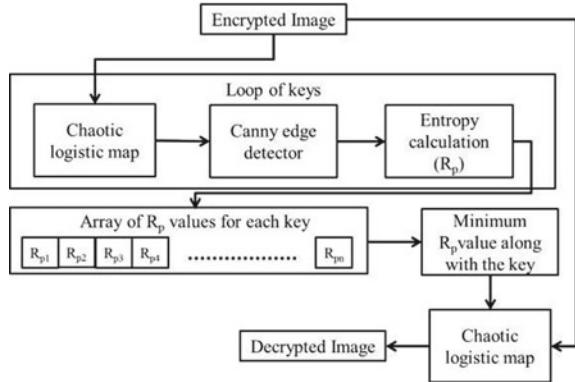
Before going into the details, one must know the aspects that are to be known for deciphering the encrypted image. In general, image is the visual representation of the data. Hence, it does not contain sharp discontinuity in the data. The data arrangement with the minimum discontinuity may lead us to retrieve the original image correctly. So, it is clear that the content of image does not matter in retrieving the image, but the amount of discontinuity matters. Now we are left with three parameters, i.e., (i) data, (ii) some part of information, and (iii) entropy which should be used for reversing the encryption process. To do so, first thing is to fill the gap between some part of information and full information. To fill this gap, we should know the trigger (starting point) for encryption and the initial state of the image. The initial state of the image can be confirmed by entropy, because in the initial state, all the data is systematically arranged which means that discontinuity or randomness or entropy is less. So, it is now an easy task to reverse the encryption and get back to initial state if we can try the possible triggers on the encrypted image. Here, the key can be treated as the trigger and the entropy can be used to get the initial state.

The theoretical study now clearly say that the “Arrow of time” can be reversed for the image encryption process provided we have data and some part of information, i.e., algorithm with us. The process for deciphering the encrypted image is discussed with details in Sect. 2.1.

2.1 Process of Image Deciphering

Let us consider present state of “Lena” image which is shown in Fig. 1a. After encryption through some algorithm, the image has changed its state to another state as shown in Fig. 1b. In this paper, the encryption algorithm used is “chaotic logistic map” [12]. From the encrypted image, we can get the data and some part of information as discussed earlier. If we apply a brute force attack on the encrypted image, it is possible to break the encryption algorithm, but only if the original image is known. If the original image is a unknown parameter, then one must have an indicator parameter, which tells this is the original image. Here, randomness or entropy of the image

Fig. 1 Original and the encrypted image



is used as the indicator parameter which gives accurate results. To calculate the entropy of the image, a parameter called as randomness parameter (R_p) is proposed in this paper. Equation for R_p is given in (1). Hence, image is a visual data with less discontinuities, value of R_p must be less for the original image (without any tampering). While processing the algorithm, output image with minimum value of R_p should have less randomness, which means it is the original image before encryption.

$$R_p = \frac{\sum_{i=1}^m \sum_{j=1}^n I_{\text{dec}}^e(i, j)}{mn} \quad (1)$$

Algorithm 1 Proposed algorithm for image deciphering Require: DecipheredImage, R_p^{\min}

```

Begin loop:
  Initialise the key ( $I_k$ );
  OutImage( $I_k$ ) = chaotic logistic map (Encrypted Image);
  EdgeOutImage( $I_k$ ) = edge detection(OutImage( $I_k$ ));
   $R_p = \text{RPCalculator}(EdgeOutImage(I_k))$ ;
  RPArray[counter]  $R_p$ ; counter++;
End loop
if counter != 0 then
  [ $R_p^{\min}, I_k^{\text{correct}}$ ] = min(RPArray);
  DecipheredImage = chaotic logistic map (Encrypted Image( $I_k^{\text{correct}}$ ));
else
  End the
  program; end if
  edge
  detection(OutImage); f

```

```

OutEdgeImage = cannyedge(OutImage);
Send OutEdgeImage;
g
End

```

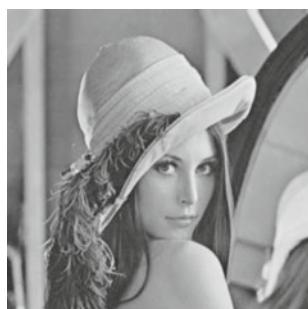
where ${}^0m^0$ represents height of the image, ${}^0n^0$ represents width of the image, ${}^0I^0$ represents the original image, I^e represents the edge detected image, I_{enc} represents the encrypted image, I_{dec} represents the decrypted image and I_{dec}^e represents the edge detected encrypted image.

Proposed algorithm is an iterative process which is presented in Algorithm 1 and the block diagram for the same is presented in Fig. 2.

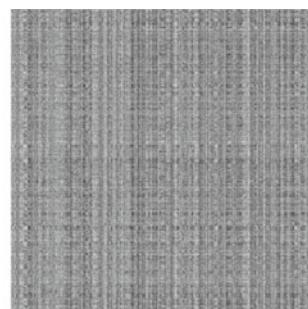
2.2 Justification for the Proposed Algorithm

As discussed in Sect. 2.1, entropy change in the image is chosen as the parameter to identify the original image. The reason for doing so, lies in the theory of “Arrow of Time,” which states that the entropy of the universe changes with increase in time as any moment which occurs in universe tends to create the disturbance. From this statement, it is clear that, any process before some manipulation has less disturbances than the manipulated process. The same concept is applied here on the image, original image before any manipulation is considered to have fewer disturbances, i.e., entropy and the manipulated image tend to have higher entropy. This is proven to be right from the discussions done in Sect. 3, the entropy calculations are giving us the correct estimate of the original image from the manipulated images.

The idea of using entropy as the parameter is adapted from “Einstein theory of relativity,” which states that any motion in the universe is different when some from



(a) Original ($R_p=0.086533$)



(b) Encrypted ($R_p=0.294742$)

Fig. 2 Block diagram for the proposed algorithm

different frame of references. It briefly tells that any action can be perceived in many other different ways depending on the way we are seeing it. Hence, by taking this statement into consideration, all images that are encrypted and unencrypted are perceived as the original images of that stage. And to differentiate the required image from them, entropy is used. To support this justification, results are presented in Sect. 3.

3 Results and Discussions

This section presents the results obtained by the proposed algorithm. Results are presented for Lena and Einstein image for various values of R_p . To find out the value of R_p , initially, image is passed through the chaos decryption algorithm and then the obtained result through canny edge detector to find the edges in the image. By using Eq. 1, value of R_p is found out. For the original image shown in Fig. 1a, value of R_p is obtained as 0.086533 and for the encrypted image shown in Fig. 1b, value of R_p is obtained as 0.294742.

To verify the correctness of R_p , a single value in original image ${}^0I^0$ is changed and value of R_p is calculated. The value of R_p has come out to be 0.086540, which is greater than that of the original image. Some of the decrypted images whose R_p value is near to that of the original image R_p value is presented in Fig. 3a–c. The effect of small change in the randomness can be clearly noticed from Fig. 3a–c. The decrypted image with minimum R_p value, which is same as that of the original represents the exact original image which is shown in Fig. 3d. This shows the importance of the differences between the fundamental terms data and information. Unless and until the data is not arranged in the ordered way, the information cannot be obtained. And thus, the justification given to the proposed algorithm is proved to be right.

The outcomes for the Einstein image is shown in Fig. 4 with their respective R_p values. Here, also one can clearly see that the value of R_p is high for the images which are encrypted and low for the original image. This clearly shows the accuracy of the new parameter introduced in this paper. This parameter can be a better estimate to differentiate between the original image and the manipulated image.

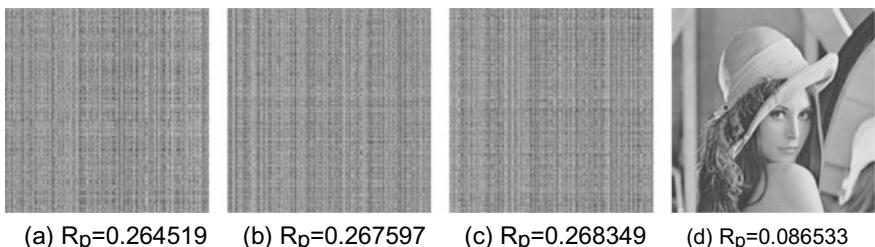


Fig. 3 Decrypted with different R_p values that are near to the original image

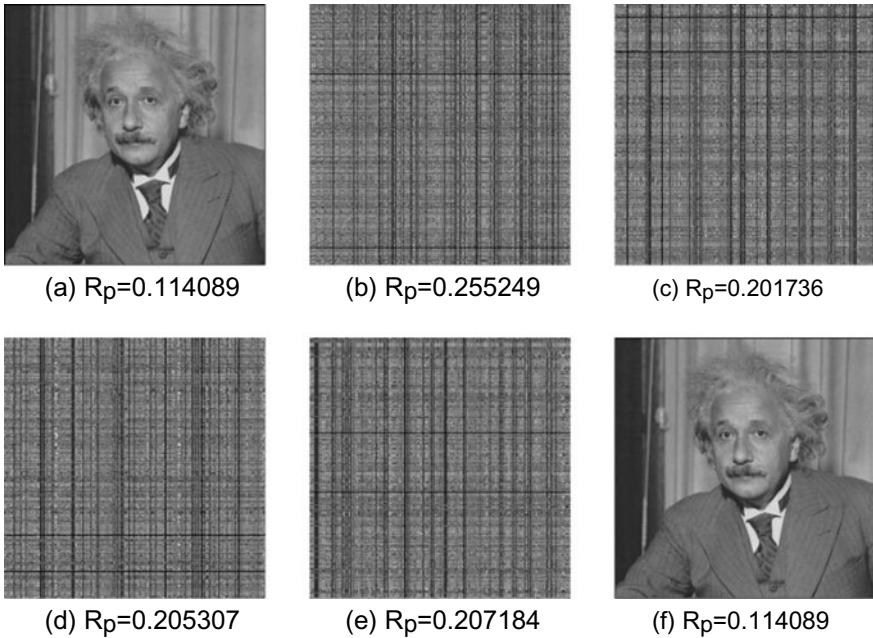


Fig. 4 **a** Original. **b** Encrypted. **c–f** Decrypted with different R_p values

4 Conclusion

The effect of the basic differences between the data and the information is studied with respect to the image encryption algorithms. The entire study has been compared with the real-time events or processes with respect to “Einstein theory of relativity and Arrow of time.” This comparison helps in better understanding of the physics behind the proposed algorithm. It is shown how a manipulation/forward motion of image can be brought back to its original position by using the basic data and the available information. The new parameter called as “Randomness parameter” is determined and proved that it works very well in determining how random the image is or how high the entropy is?. High value of R_p indicates the high entropy or heavy randomness and vice versa. The output image after decryption with lower entropy or R_p value is considered as the original image. The outcomes show that the correct perception toward any event can lead to the correct path if the data is understood properly. The results of “Lena” image and “Einstein” image show the accuracy of R_p in determining the randomness of the image.

References

1. Einstein (1916) The Foundation of the general theory of relativity. (1997) Princeton University Press, pp 146–200
2. Einstein A (1916) Relativity: the special and general theory, Methuen & Co Ltd. (trans: Lawson RW)
3. Buzzo D (2014) Time Travel: Time Dilation, Electronic Visualisation and the Arts (EVA 2014), British Computer Society, London, UK, 8–10 July 2014. London, UK: British Computer Society
4. Balazs B, Time Travel?!. Informatika 12(2)
5. Layzer D (1975) The arrow of time. Sci Am 233(6):56–69
6. Carroll SM, The origin of the universe and the arrow of time
7. Darrow KK (1942) Entropy. Bell Syst Tech J 21(1):51–74
8. Brockett RW, Willems JC (1978) Stochastic control and the second law of thermodynamics. In: 1978 IEEE conference on decision and control including the 17th symposium on adaptive processes
9. Pohl JG (2001) Transition from data to information. In: Collaborative agent design research center technical report-RESU72, pp 1–8
10. Chen M et al (2009) Data, information, and knowledge in visualization. IEEE Comput Graph Appl 29(1):12–19
11. Sanders J (2016) Defining terms: data, information and knowledge. In: SAI computing conference 2016, London, UK
12. Pentigen H, Jurgens H, Saupe D (2016) Chaos and fractals-new frontiers of science, Springer; 1st edn. 1992. Corr. 2nd printing edn. (4 March 1993). London, UK

A Systematic Review of Approximate Adders: Accuracy and Performance Analysis



M. Lakshmi Akhila, E. Jagadeeswara Rao, R. V. V. Krishna,
and Durgesh Nandan

Abstract In present years, integration technology is drastically changed and urgently needs the high-speed adder circuits for the image processing (IP) applications (APL) like image smoothing (IS) and image sharpening (ISH), but unfortunately any exact adder circuit did not use IS and ISH, so few researchers proposed approximate adders (AA) for IP APL and also improved the speed compare to exact adder circuits (CKT), and still there is no systematic literature survey about AA. This paper gives the one root map for AA, and this is more useful for some of the researcher's work in this area. Finally, the performance analysis of AA in terms of power consumption, delay, area, PDPm, and error is provided.

Keywords AA · DISP · AC · APPRC · RCA

1 Introduction

Approximate computing (AC) [1] is the promising method in the epoch when power is the supreme constraint because it can trade accuracy for power. To decrease the power, area, and delay. We mainly use the method of AC in VLSI design [2]. It provides the required space between the level of accuracy which is required by the APL. These APL are provided by the computing system for getting the required

M. Lakshmi Akhila · R. V. V. Krishna

Department of Electronics & Communication Engineering, Aditya College of Engineering & Technology, Surampalem, AP, India
e-mail: akhilamaireddy@gmail.com

R. V. V. Krishna

e-mail: rvvkrishnaece@gmail.com

E. Jagadeeswara Rao · D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India
e-mail: durgeshndano51@gmail.com

E. Jagadeeswara Rao

e-mail: emandi.jagadeesh@gmail.com

© Springer Nature Singapore Pte Ltd. 2021

689

V. K. Gunjan and J. M. Zurada (eds.), *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications*, Advances in Intelligent Systems and Computing 1245, https://doi.org/10.1007/978-981-15-7234-0_65

optimization. Nowadays, the AC is mainly used in many APL like search engine [3], synthesis [4], cognitive APL [5], machine learning [6], IP [7], signal processing [8], scientific computing [9], and some wireless communications [10].

Let us take an example, in IP, the required O/P video or required image with some errors which are produced by minute image quality loss or by dropping the particular frame. In computer system design, the AC has been successfully applied to all the abstractions [11], software [12], and APL levels [13].

In this paper, we mostly focus on the review of AA, accuracy, and performance analysis. Approximate CKT (aproxcir) has less cost than existing CKT. So, we prefer approximate CKT rather than other CKT. This paper starts with the introduction in Sect. 1 and gives the literature survey see in Sect. 2. Further, some recent AA proposed by the author's results in Sect. 3, and finally concludes this paper in Sect. 4.

2 Systematic Literature

Earlier research efforts have oppressed the error resilience of approximate adders and its applications (app) at different levels of design abstraction. To decrease the drawbacks in the previous papers, we introduce various techniques in this paper.

In AC, there are some errors to avoid those errors in 1997, Mei-Chen Hsueh et al. proposed an automatic error injection framework, and it has been widely studied in the error tolerance community that revises the app under errors [14].

In 2008, D. Shin et al. proposed a new method for data path modules in fault-tolerant app which is focused mainly on manually approximating (approx) specific circuits like adders [15], and in 2011, P. Kulkarni et al. focused on multipliers [16] and the variation in the impact of their output bits by taking benefit of their structural properties.

In 2010, A. B. Kahng et al. proposed that through cell sizing by raising the slack of the often-exercised paths, the delay path distribution of the circuit is redesigned [17]. In 2011, R. Venkatesan et al. proposed a new technique [11] for estimating and analyzing the faults that occurred due to approx. In 2010, D. Shin proposed error-tolerant applications, and it was improved in power. In this, they focused on two-level clippers [18]. For simplifying the circuit and error-tolerant in 2011, D. Shin proposed multi-level circuits [19], and the circuit is simplified by propagating the redundancy.

In 2012, Jinghang Liang introduced several new methods for evaluating probabilistic errors and approx based on consistency and power efficiency which is as shown in figure [20]. In 2012, Swagath venkataramani proposed systematic logic synthesis of approx circuits, and in terms of area and power savings, there is a lot of benefit for error restraint in the framework to synthesize aproxcir automatically [21].

In 2013, Vaibhav Gupta introduced several inexact methods for error-resilient digital signal processing (DISP) systems to trade-off power and quality AA which is used effectively. For error and power consumption of an approx ripple carry adder

(RCA), we derive simplified mathematical models using the approx full adder cells. To attain maximum power saving for a given quality restraint by using these models, we discuss how to apply these approx [8]. In 2013, Vinay K. Chippa introduced an application resilience characterization (APPRC) framework for the analysis and characterization of resilience [13]. In 2014, Cong Liu proposed an analytical framework for characterizing AA designs. It has many arithmetic ckts and APL, and it is used for time consuming [22].

In 2015, Babu M. Pranay modified a new booth multiplier. The accuracy can be calculated during the run time also by using an AA. Compared with the conventional booth multiplier, it gives the most efficiency, less delay, and low power which is mainly used, and there are many APL on the booth multiplier [23]. In 2015, Adnan Aquib Naseer proposed the design of approximate adders using parallelized genetic algorithms (GAs). The GA has been modified by designing large ckts, and these are implemented by chips [24].

In 2016, Sunghyun Kim proposed an adaptive approximate adder (A3) architecture to configure the bit length of the AA during runtime. By using XOR operations, the proposed adder improved in both error distance and error rate [25]. In 2016, Ashim Gogoi proposed an AA derived from 14T accurate adder. It is consuming low power than mirror AA and AA's based on transmission gate (TG) and complementary pass transistor logic (CPL) [26]. In 2017, Jorge Echavarria proposed a generic algorithm of order $O(n^2)$ to compute the arithmetic error rate (AER) for deterministic AA units [27]. In 2017, Xiaoliang Chen proposed about correlation for real I/P stream. Mainly, they have said about the spatial correlation which is exciting in the real I/P stream and also in the carry-in-signal. [28]. In 2017, Z. Vasicek used binary decision diagrams to obtain an average error, worst error, error rate, and average Hamming distance for approximate errors [29, 30].

In 2018, Masoud Pashaeifar has been proposed an approx reverse carry propagate full adder (RCPFA). In this RCPFA the carry has been propagated from the lower side bits which is nothing but from LSB. This RCPFA has more stability in delay [31]. In 2018, Tongxin yang et al. proposed an accuracy configurable AA's, and this adder can deliver more significant energy savings than the conventional ripple carry adder (RICA). When compared to previous adders, this adder delivers improvement in energy saving, design area, and accuracy [32]. In 2018, Ayad Dalloo proposed an optimized lower part constant OR adder which shows the best improvement in the error and cost of the component metric which is compared to the previous architecture [33]. In 2018, a methodology has been proposed by Amina Qureshi based on a probability theory based on a higher-order logic theorem for finding the errors for AA [34].

3 Methodology

In AAs, there are COMP and counters. A(M, 2) COMP is a logic CKT that takes the I/Ps M bits of the same features, and it generates carry bits and sum bits as

the O/P, even though a COMP gives carry and sum as O/P which is different from conventional adder. For example, COMP adds M bits of same precision, whereas an adder adds two operands of M bit numbers of different precision. A (M, 2) COMP operation is shown in Eq. (1).

$$\begin{aligned} I_1 + I_2 + \dots + I_M + (C_{in1} + C_{in2} + \dots + C_{in k}) \\ = \text{Sum} + 2 * (\text{Carry} + C_{out1} + \dots + C_{out k}) \end{aligned} \quad (1)$$

where $I_1, I_2 \dots$ and C_{in1}, C_{in2} are I/P for COMP. However, (L, M) are parallel counter in a CKT, and it generates an M bit count while providing an L-I/P where the logics are ones. However, a counter is different from COMP, and COMP consists of CARRY I/P and similarly CARRY O/P including the applied I/P and O/P. But counter does not contain any of them. An (L, M) bit counter operation is shown in Eq. (2).

$$I_0 + I_1 + \dots + I_L = 2^0 * S_0 + 2^1 * S_1 + \dots + 2^m * S_M \quad (2)$$

Difference between the AA and counters block diagrams is shown in Fig. 1a and b.

Fig. 1 a Block diagram of AA. b Block diagram of counter

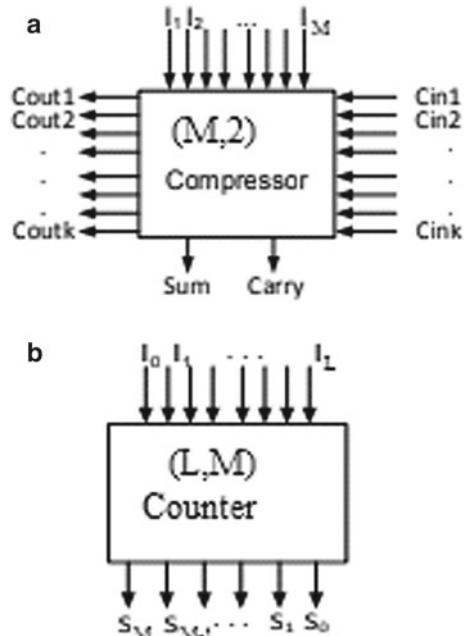


Table 1 Different types of hardware performance are compared by using

Design	Power (mw)	Area (μm^2)	Delay (ns)	PDP (fj)	Errors (%)
AA1 [28]	0.00112	26.5	0.27	0.0003	0.0757
AA2 [28]	0.00132	23.01	0.2	0.00026	0.159
AA3 [31]	0.0000043	7.224	0.164	0.0000007	–
AA4 [31]	0.0000035	5.705	0.116	0.0000004	80.08
AA5 [31]	0.0000031	4.35	5.3	0.000016	–
AA6 [35]	22.3	99.5	0.97	21.6	89.99
AA7 [35]	23	114.6	1.16	26.7	–
AA8 [35]	25.8	121	1.24	31.9	–
AA9 [35]	23.9	109.8	2.32	55.4	–
AA10 [35]	29.9	145.3	1.78	53.2	–
AA11 [36]	0.00152	341	1.11	0.0016	–

4 Results

Results for power, area, delay, power delay product (PDP), error, no. of transistors of different approximate adders from the above graphs, we can see that power is more for AA10 [35], i.e., 29.9%, and the power is less for AA5 [31], i.e., 0.0000031%. Area is more for AA11 [36], i.e., 341%, and area is less for AA5 [31], i.e., 4.35%. Delay is more AA5 [31], i.e., 5.3%, and the delay is less for AA4 [31], i.e., 0.116%. PDP is more AA9 [35], i.e., 55.4%, and PDP is less for 0.0000004%. When power is increasing, area increases, delay increases, and PDP also increases. If power is decreasing, the area decreases, delay decreases, and PDP also decreases (Table 1; Fig. 2).

5 Conclusion

Until now, many researchers have worked to conclude the various AA and also proposed which is the best AA, and till now it is partially proposed. In this paper, we discuss approximate adders, a systematic review of approximate adders, accuracy, performance analysis, and which one is the best approximate adder. In this paper, we discuss the literature review and methodology that this approximate adder is the best one among all the previous papers. As the power is increasing, area increases, delay increases, and PDP also increases, so the cost will be more. As the power is decreasing, the area decreases, delay decreases, and PDP also decreases, so cost will be less. In this proposed approximate adder, there is less power, area, delay, and PDP, so cost is also less for this proposed approximate adder.

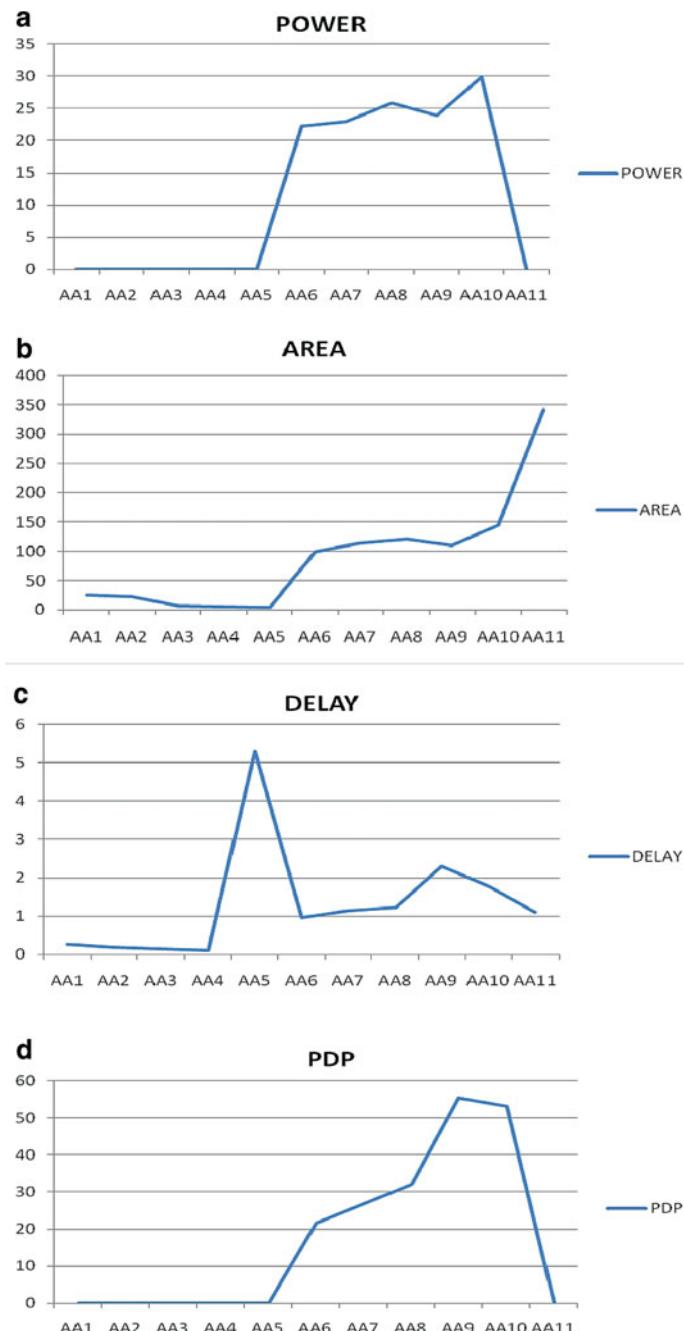


Fig. 2 **a** Comparison of power (mw). **b** Comparison of area (μm^2). **c** Comparison of delay (ns). **d** Comparison of PDP (fj)

References

1. Kosovichev AG, Severny AB (2018) On the stability of solar gravity mode oscillations and the structure of the sun. *Liege Int Astrophys Colloq* 25:278–282. <https://doi.org/10.1109/ETS.2013.6569370>
2. Mittal S (2016) A survey of techniques for approximate computing. *ACM Comput Surv* 48:4. <https://doi.org/10.1145/2893356>
3. Sidiropoulos S et al (2011) Managing performance vs. accuracy trade-offs with loop perforation. In: SIGSOFT/FSE 2011—proceedings of the 19th ACM SIGSOFT symposium and the 13th European conference on foundations of software engineering, pp 124–134. <https://doi.org/10.1145/2025113.2025133>
4. Zhang Q et al (2014) ApproxIt: an computing framework approximate for iterative methods. In: Proceedings of the 51st annual design automation conference. <https://doi.org/10.1145/2593069.2593092>
5. Zadeh LA (1994) Fuzzy logic, neural networks, and soft computing. *Commun ACM* 37(3):77–84. <https://doi.org/10.1145/175247.175255>
6. Amant RS et al (2014) General-purpose code acceleration with limited-precision analog computation. In: Proceedings in the international symposium on computer architecture, pp 505–516. <https://doi.org/10.1109/ISCA.2014.6853213>
7. Khudia DS et al (2015) Rumba: an online quality management system for approximate computing. In: Proceedings in the international symposium on computer architecture, pp 554–566 (2015). <https://doi.org/10.1145/2749469.2750371>
8. Gupta V et al (2013) Low-power digital signal processing using approximate adders. In: IEEE transactions on computer-aided design of integrated circuits and systems, vol 32(1), pp 124–137. <https://doi.org/10.1109/TCAD.2012.2217962>
9. Ringenburg M et al (2015) Monitoring and debugging the quality of results in approximate programs. *ACM SIGPLAN Not* 50(4):399–411. <https://doi.org/10.1145/2694344.2694365>
10. Intanagoniwat C et al (2002) Impact of network density on data aggregation in wireless sensor networks. In: Proceedings 22nd international conference on distributed computing systems, pp 457–458. <https://doi.org/10.1109/icdc.2002.1022289>
11. Venkatesan R et al (2011) MACACO: modeling and analysis of circuits for approximate computing. In: IEEE/ACM international conference on computer-aided design technical paper ICCAD, pp 667–673. <https://doi.org/10.1109/ICCAD.2011.6105401>
12. Sampson A et al (2011) EnerJ: approximate data types for safe and general low-power computation. In: Proceedings in the ACM SIGPLAN conference programming and language design implement, pp 164–174. <https://doi.org/10.1145/1993498.1993518>
13. Chippa VK et al (2013) Analysis and characterization of inherent application resilience for approximate computing. In: Proceedings of the 50th annual design automation conference. <https://doi.org/10.1145/2463209.2488873>
14. Hsueh MC et al (1997) Fault injection techniques and tools. *Computer (Long. Beach. Calif.)* 30(4):75–82. <https://doi.org/10.1109/2.585157>
15. Shin D, Gupta SK (2008) A re-design technique for datapath modules in error tolerant applications. In: Proceedings in the Asian test symposium, pp 431–437. <https://doi.org/10.1109/ATS.2008.75>
16. Kulkarni P et al (2011) Trading accuracy for power with an under designed multiplier architecture. In: Proceedings in the IEEE international conference on VLSI design, pp 346–351. <https://doi.org/10.1109/VLSID.2011.51>
17. Kahng AB et al (2010) Slack redistribution for graceful degradation under voltage overscaling. In: Proceedings in the Asia and South Pacific design automation conference, ASP-DAC, pp 825–831. <https://doi.org/10.1109/ASPDAC.2010.5419690>
18. Shin D, Gupta SK (2010) Approximate logic synthesis for error tolerant applications. In: Proceedings in the design, automation & test in Europe DATE, pp 957–960. <https://doi.org/10.1109/date.2010.5456913>

19. Shin, D, Gupta SK (2011) A new circuit simplification method for error tolerant applications. In: Proceedings in the design, automation & test in Europe, DATE, pp 1566–1571. <https://doi.org/10.1109/date.2011.5763248>
20. Liang J et al (2013) New metrics for the reliability of approximate and probabilistic adders. IEEE Trans Comput 62(9):1760–1771. <https://doi.org/10.1109/TC.2012.146>
21. Venkataramani S et al (2012) SALSA: systematic logic synthesis of approximate circuits. In: Proceedings in the design automation conference, pp 796–801. <https://doi.org/10.1145/2228360.2228504>
22. Liu C et al (2015) An analytical framework for evaluating the error characteristics of approximate adders. IEEE Trans Comput 64(5):1268–1281. <https://doi.org/10.1109/TC.2014.2317180>
23. Pranay BM, Jandhyala S (2016) Accuracy configurable modified booth multiplier using approximate adders. In: Proceedings in the 2015 IEEE international symposium on nanoelectronic and information systems, INIS 2015, pp 281–285. <https://doi.org/10.1109/iNIS.2015.50>
24. Mrazek V, Vasicek Z (2018) Evolutionary design of large approximate adders optimized for various error criteria. In: GECCO 2018 Companion—proceedings of the genetic and evolutionary computation conference companion, pp 294–295. <https://doi.org/10.1145/3205651.3205678>
25. Kim S, Kim Y (2016) Adaptive approximate adder (A3) to reduce error distance for image processor. In: ISOCC 2016 international soc design conference smart soc intelligent, pp 295–296. <https://doi.org/10.1109/ISOCC.2016.7799794>
26. Gogoi A, Kumar V (2016) Design of low power, area efficient and high speed approximate adders for inexact computing. In: 2016 international conference on signal processing and communication, ICSC 2016, pp 452–456. <https://doi.org/10.1109/ICSPCom.2016.7980623>
27. Echavarria J et al (2018) Efficient arithmetic error rate calculus for visibility reduced approximate adders. IEEE Embed Syst Lett 10(2):37–40. <https://doi.org/10.1109/LES.2017.2760922>
28. Chen X et al (2017) Low latency approximate adder for highly correlated input streams. In: Proceedings in the 35th IEEE international conference on computer design, ICCD 2017, pp 121–124. <https://doi.org/10.1109/ICCD.2017.26>
29. Chandrasekharan A et al (2016) Approximation-aware rewriting of AIGs for error tolerant applications. In: IEEE/ACM international conference on computer-aided design technical papers, ICCAD. <https://doi.org/10.1145/2966986.2967003>
30. Vasicek Z et al (2017) Towards low power approximate DCT architecture for HEVC standard. In: Proceedings in the 2017 design, automation & test in Europe, DATE 2017, pp 1576–1581. <https://doi.org/10.23919/DATE.2017.7927241>
31. Pashaeifar M et al (2018) Approximate reverse carry propagate adder for energy-efficient dsp applications. In: IEEE transactions on very large scale integration systems, vol 26(11), pp 2530–2541. <https://doi.org/10.1109/TVLSI.2018.2859939>
32. Yang T et al (2018) A low-power configurable adder for approximate applications. In: Proceedings in the international symposium on quality electronic design, ISQED, pp 347–352. <https://doi.org/10.1109/ISQED.2018.8357311>
33. Dalloo A et al (2018) Systematic design of an approximate adder the optimized lower part constant-OR adder. Journal.pdf 1:1–5
34. Qureshi A, Hasan O (2019) Formal probabilistic analysis of low latency approximate adders. IEEE Trans Comput Des Integr Circuits Syst 38(1):177–189. <https://doi.org/10.1109/TCAD.2018.2803622>
35. Zhou R, Qian W (2016) A general sign bit error correction scheme for approximate adders. In: Proceedings of the ACM Great Lakes symposium on VLSI, GLSVLSI, pp 221–226. <https://doi.org/10.1145/2902961.2903012>
36. Esposito D et al (2017) On the use of approximate adders in carry-save. In: Proceedings in the IEEE international symposium on circuits and systems, pp 6–9. <https://doi.org/10.1109/ISCAS.2017.8050437>

A Review Paper Based on Image Security Using Watermarking



V. Ch. S. Ravi Shankar, R. U. S. D. Vara Prasad, Rama Vasantha Adiraju, R. V. V. Krishna, and Durgesh Nandan

Abstract This paper reviews the performance analysis based on image security for digital watermarked images using different techniques in transform domains. Digital watermarking is a technique of embedding data like text or image which is called watermark into digital data. It can be done by using different methods depending upon the digital data. It is examined that the most robust method in providing the security for watermarked images can be achieved by combining DWT with SVD when compared to other methods. Though DWT itself maintains good authenticity, when SVD is added to DWT, it increases the robustness in extracting watermark image. Firstly, DWT is applied to the original image and then SVD is applied to the resultant of DWT image. The steps involved in the algorithm to embed and extract the watermark are explained. Finally, the performance analysis of DWT and DWT with SVD is compared and quality of the image is examined.

Keywords Watermarking · Cover image · DWT · DWT-SVD

V. Ch. S. Ravi Shankar · R. U. S. D. Vara Prasad · R. V. Adiraju · R. V. V. Krishna

Department of Electronics and Communication, Aditya College of Engineering and Technology,
Surampalem, Andhra Pradesh, India

e-mail: ravishankar251999@gmail.com

R. U. S. D. Vara Prasad

e-mail: 141dattu@gmail.com

R. V. Adiraju

e-mail: vasanthaadiraju@gmail.com

R. V. V. Krishna

e-mail: rvvkrishnaece@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India

e-mail: durgeshnandano51@gmail.com

1 Introduction

A watermark is first used in the thirteenth century in Italy. Like wax seal, it is an emblem of prestige. It is used as security for personnel corresponding. These watermarks are created by wire sewn on the paper mold. The term “Digital Watermark” is created by Andrew Terkel and Charles Osborne in December 1992 Steganographic spread spectrum watermark is the first successful embedded watermark demonstrated by Andrew Tirkel, Charles Osborne, and Gerard Rankin.

In the current world scenario, digital data like pictures, paintings, speech, audio, video through media and transfer has been increased drastically. Nowadays, cyber-crimes and illegal authentication of bank details has been increased. There is a necessity to provide security for our digital data to avoid illegal access. It is more important to protect our digital data that are exposed publicly and are pirated for using illegally and violating the rights of the owner. To provide security and illegal authentication of data, it is necessary to create copyright as a mark. This technique is called DIGITAL MARKING. Such type of watermark, signature, stamps, and seal are used from ancient time to identify the correct creator.

Digital watermarking is one of the techniques which hide digital content by attesting digital watermark. It is used in several applications like data authentication, security purposes in banknotes, copyright protection, source tracking, hidden communication broadcast tracking from news organizations, ID card security, fraud and tamper detection, video authentication, etc. Features of watermarking and undetectable, invisible/imperceptible, undetectable, security, robustness, universal, not destroyed while compressing.

Digital watermarking is of two types: (1) visible digital watermarking and (2) invisible digital watermarking. Visible digital watermarking is visible data like any logo or text denoting owner which is embedded as a watermark and invisible digital watermarking is embedded as invisible and in case of audio, it is inaudible.

Few methods of watermarking are three coding methods for hiding electronic marking, implementing in spatial domain by cropping the required portion may eliminate watermark on an image, implementing in frequency domain that uses algorithms like discrete cosine domain (DCT), discrete Fourier domain (DFT), digital signature which is based on cryptographic algorithm, SVD and DWT which is able to fill the gap of DCT-based watermarking on images.

The information which is embedded as the watermark is hidden in two ways: Steganography technique and cryptography technique. In steganography, technique information is hidden in writings, drawings, images, and even in speech. It is used in ancient times by governments, armies, and rulers. But in cryptography, it does not contain any secret information but makes undetectable by various encoding and decoding techniques. But there is a possibility of decoding the code. This process of embedding the watermark can be applied on text, images, and videos to hide the important information and used to avoid unethical access to the data.

Data is hiding in text: Duplicate content has a moderate absence of excess data contrasted with picture or sound documents. There are three methods in this: (1)

open space methods which encode messages through unused space on the page and manipulation of the white space (2) syntactic method which utilizes punctuation, and (3) semantic method which manipulate the words themselves. Data hiding in images: Steganography is the craft of concealing information in a harmless spread medium utilized while concealing information in images. Data hiding in videos: A video information implanting plan is proposed while concealing information in recordings in which the inserted mark information is reproduced without knowing the first host video.

2 Literature Review

Zhao and Koch [1] proposed a novel steganography method that secretly embeds strong labels in the image for identifying the owner to the image. This embedded label is irremovable and undetectable. It can survive many attacks like low pass filtering and image format conversions. The disadvantage is against few physical damages like cut a line, grab an area.

Choudhury et al. [2] designed architecture and a protocol is presented which make the appropriation of illicit duplicates troublesome. The key thought in this answer to utilize a showcase (or printing) customer which is liable for unscrambling the scrambled archive sent to the expected client. These customers have a key embedded in them. Upon execution, the application utilizes the client's private key (gave as a contribution) to get to the key, which is then used to decode the conveyed record. Accordingly, appropriating the customers fills no need without conveying the private key of the client too.

Cox et al. [3] recommended an algorithm which is tamper-resistant can be applied for watermarking images. Gaussian random vector can be applied. It makes watermark strong to signal processing operations. With this technique, the owner can be identified perfectly. Electronic watermark distribution of copyright became more general. With the above algorithm, it is robustness to attack, robustness to common signal and geometric operations and it is applicable to image, audio, and video.

Hsu and Wu [4] suggested audios and videos to which watermark is applied to survive standard signal transformations. If correct registrations patterns are used to watermark videos, images and audios, then watermark detection fails. Watermarks are robust to lossy compression, spatial filtering.

Cox [5] advised complex watermark embedding is the correct method to identify unauthorised copying. In this approach, visually recognizable patterns are embedded which can selectively modify the middle-frequency parts of the image. This technique successfully survives the image processing, image processing operations. This technique can also be applied to the multi-resolution image structures with the modification for the middle-frequency coefficients. The main disadvantage is for image resampling and image rotation.

Voyatzis and Pitas [6] stated Fundamental demands for watermarking are like watermarking algorithms, watermark generations, embedding, detection are stated.

GWF is such a method that protects the copyright. But there are several unsolved problems for a pire watermarking for this approach. Piracy is possible in public distribution networks. The demand for robustness is not satisfied correctly.

Bloom et al. [7] stated several challenges rise in protecting copyright from piracy for digital versatile disk(DVD). An analogue copy system protection is used in protecting NTSC/PAL output channel from piracy to VHS. Robust encryption protocol protects the digital transmission of content between two communicating devices. DVD video disks are encrypted and an analogue protection system is utilized to inhibit piracy.

Nikolaidis et al. [8] requested owner has to be careful in selecting different watermark methods. They have to check different scenarios while watermarking in order to define the appropriate parameters utilized in his method. The owner himself has to identify different attacks that could be done on his method. Firstly, possible attacks need to be divided and should refer to his method for identifying attacks.

Ho and Li [9] proposed a separate algorithm used based on block DCT transform where the original image is divided into 8×8 blocks and each block is transformed into DCT domain. This algorithm is robust to JPEG compression. This method uses invariant features of watermark and a quantized down-sampled approximation of a host image for recovering the watermark. It also uses spatial-frequency properties of DWT to make sure the good quality of the watermarked image. It is safe to use in image authentication.

Zhang et al. [10] stated mostly used image segmentation algorithms are region-based which fail correct segmentation results due to the intensity inhomogeneity. A separate novel-based methods can be used robustly to intensity inhomogeneity. This method is applied to MRI image segmentation. The proposed scheme is robust to JPEG operation but vulnerable to image malicious tamper. The disadvantage for this algorithm is sensitive to image malicious tampering and another disadvantage is it has a low rate of false detection and failure detection.

Zhou and Jin [11] urged with the help of DWT and SVD, a novel copyright protection zero watermarking scheme is done. The watermarking is applied without degrading the quality of the original image by DWT and SVD. With this scheme, we will obtain good robustness and the watermarked image quality without any distortion. The zero watermarking scheme is perceived through XOR operation. SWT-SVD scheme is applied to only some features of the original image but not to entire data of the original image.

Bin [12] requested to ensure a safe travelling and protection of image over Internet DWT method is suggested. This scheme results in excellent robustness and transparency. A semi-fragile watermarking method based on DWT is used. The algorithm used in this scheme has a strong anti-conventional attack capability. This results in a good image quality when we check against parameters such as PSNR and JPEG compression. It is invulnerable to many image manipulations.

Ye et al. [13] explained the combination of DWT and SVD schemes which use the energy concentration characteristic of DCT for applying the watermark. This method sets strong robustness and larger capacity so that the contradiction between capacity and robustness is removed. Chaotic mapping is used to guard the security of

data. Compared to DWT–SVD–DCT-based watermarking, this method sets strong robustness to attacks like cropping, filtering, noise, etc.

Kunhu et al. [14] proposed a blind digital multiwatermarking algorithm for medical images to make sure the image protection. The proposed algorithm uses DWT and hash functions. This scheme not only make sure the robustness but also secures the originality of the image so that information of the image not damaged. The DWT-based watermark can withstand several attacks like scaling and rotating attacks. The Hash-based watermark information is vulnerable to small modifications. So the combination of DWT and Hash functions is used.

Goli and Naghsh [15] offered two-step SUDOKU methods to withstand the cropping attack in digital images are proposed in this paper. Spatial domain methods used for watermarking are unprotected salt-pepper noise, compression, and cropping attacks. This method repeats the water image 81 times in the original image. The efficiency of this method is up to 98.8% and it can withstand many attacks. This method is also called a blind method as it does not require an original image when there is a need to extract the watermark.

3 Methodologies Used in Watermarking

3.1 Discrete Wavelet Transform (DWT)

DWT has been utilized in digital image watermarking, all the more oftentimes, because of its amazing spatial confinement and multi-goals attributes, which are like the hypothetical models of the human visual framework. It is valuable for preparing of non-stationary signals. The images decomposed into four sub-bands LL1, LH1, HL1, and HH1. For each successive level of decomposition, the LL subband of the previous level is used as the input. And now, DWT is applied to LL1. Now again four sub-groups LL2, LH2, HL2, and HH2 are framed. To perform third-level deterioration, the DWT is applied to LL2 band which break down this band into the four sub-groups—LL3, LH3, HL3, and HH3. This brings about ten sub-groups per segment (Fig. 1).

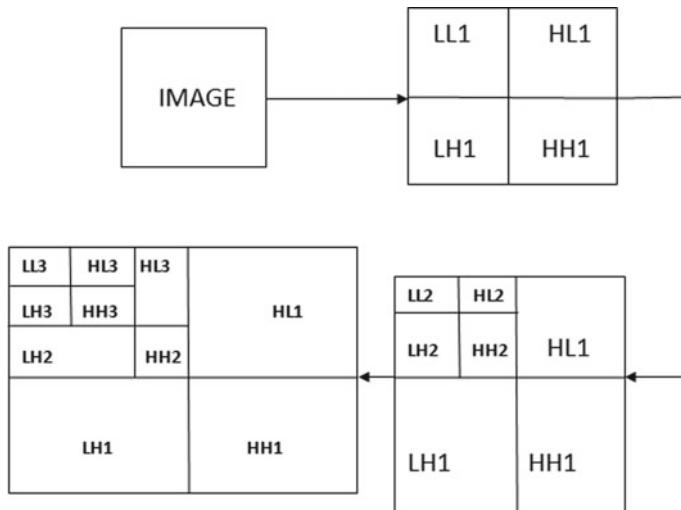


Fig. 1 Three-level DWT of image

3.2 SVD (Sinular Value Decomposition)

U, D, and V are the three matrices which are the result of decomposition of a matrix in the SVD transformation. The two steps involved in this scheme are (a) watermark embedding procedure and (b) watermark extracting procedure.

Steps involved in the watermark embedding procedure are:

Step1: Blocks of $n \times n$ pixels are obtained by partitioning the image and to the partitioned blocks SVD transformation is applied.

Step2: Complexity of blocks is determined in the D component of each block by calculating a number of nonzero coefficients.

Step3: Magnitude difference between neighboring coefficient is calculated by selecting the more fabulous complexity blocks in the first column of U using the feature of D component and PRNG [pseudo random number generator].

Step4: The non-zero coefficients are retained if the embedding watermark is matched to magnitude difference between the neighboring coefficient. For example, 1 is matched to a positive value and a negative value to 0. Or else the coefficient is modified.

Step5: Strong robustness of a watermarking scheme is provided in order to retain the image quality.

Steps involved in watermark extraction procedure are:

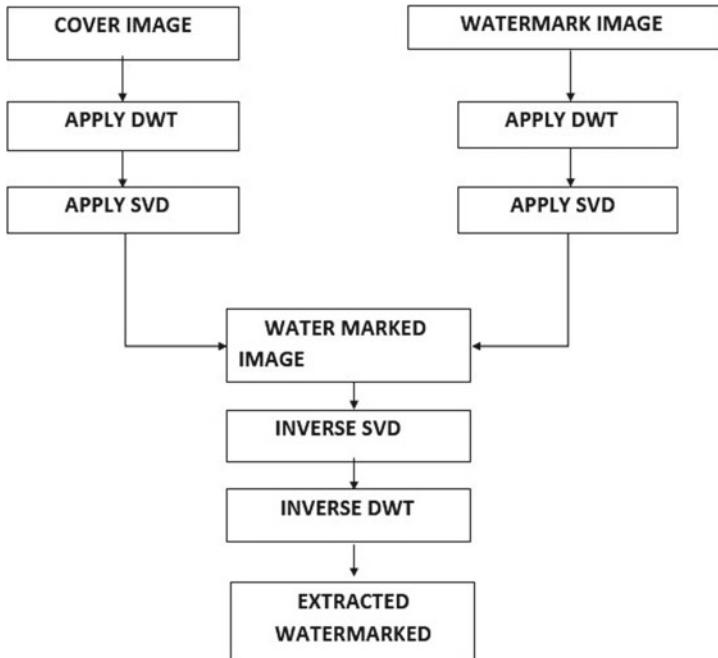
Step 1: By partitioning the watermarked image, blocks are obtained and SVD is applied to the partitioned image.

Step 2: In the block, complexity nature is resolved and the non-zero coefficient is determined in each block of the D component.

Step 3: The relationship of the U component is determined by using the component of D segment and PRNG.

Step 4: In the event that a negative relationship is determined, the extracted watermark is allocated a bit value of 0. Something else, the extricated watermark is given 1.

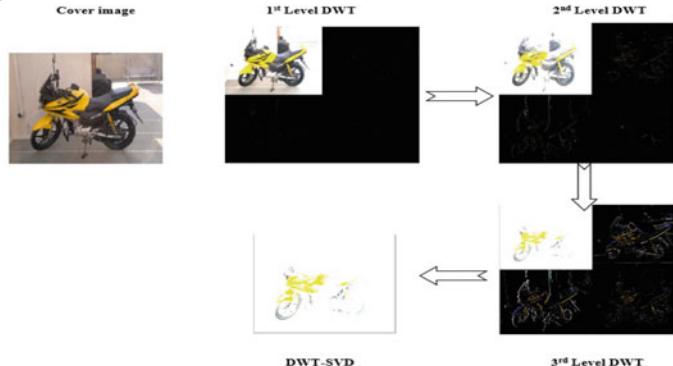
Flowchart representation of DWT + SVD



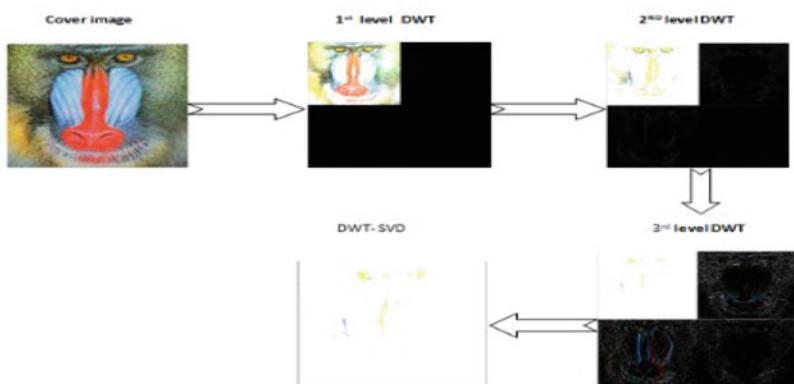
4 Result

In this paper, watermarking is done using DWT and SVD methods. In the image watermarking, the useful implementation of the image with respect to embedding the watermark into a host image using the DWT and SVD are the most significant

Cover image 1:



Cover image 2:



Cover image 3:

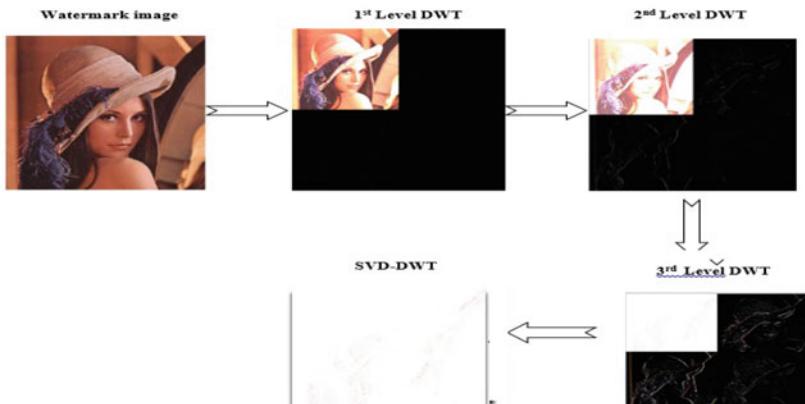
**Fig. 2** Results of watermarking using three-level DWT and SVD

Table 1 Comparison of PSNR values based upon different images

Cover image	Watermarked image	PSNR values DWT + SVD	PSNR values 3 level DWT + SVD
1	Bike	50.9	51.8
2	Baboon	50.7	51.7
3	Lena	54.25	55.47

part to achieve imperceptibility and better visibility. Higher PSNR results in a better quality of an image.

Figure 2 shows the results of watermarking using three-level DWT and SVD (Table 1).

5 Conclusion

In this paper, watermarking technique is based on DWT and SVD. DWT is a three-level discrete wavelet transform. The proposed methods present a robust digital watermarking and are able to attain good imperceptibility. With the help of DWT and SVD methods, we can apply to watermark without degrading the quality of the original image and also it increases the robustness in extracting watermark image.

References

1. Zhao J, Koch E (1995) Embedding robust labels into images for copyright protection. *KnowRight*, pp 242–251
2. Choudhury AK, Maxemchuk NF, Paul S, Schulzrinne HG (1995) Copyright protection for electronic publishing over computer networks 1 introduction. <https://doi.org/10.1109/65.386048>
3. Cox IJ et al (1998) Some general methods for tampering with watermarks. <https://doi.org/10.1109/49.668980>
4. Hsu C et al (1999) Hidden digital watermarks in images. *IEEE Trans image proc* 8(1):58–68
5. Cox IJ, Linnartz JPMG (1998) Some general methods for tampering with watermarks. *IEEE J Sel Areas Commun* 16:587–593. <https://doi.org/10.1109/49.668980>
6. Voyatzis G (1999) The use of watermarks in the protection of digital multimedia products. <https://doi.org/10.1109/5.771072>
7. Bloom JA et al (1999) Copy protection for DVD video. <https://doi.org/10.1109/5.771077>
8. Nikolaidis A et al (2001) A survey on watermarking application scenarios and related attacks June 2014. <https://doi.org/10.1109/ICIP.2001.958292>
9. Li CT (2004) Digital fragile watermarking scheme for authentication of JPEG images. *IEE Proc Vis Image Sig Process* 151:460–466. <https://doi.org/10.1049/ip-vis:20040812.8>
10. Zhang D et al (2009) A novel watermarking algorithm in DCT domain to authenticate image content. In: 2009 IEEE international conference on intelligent computing and intelligent systems, ICIS 2009, vol 3, pp 608–611. <https://doi.org/10.1109/ICICISYS.2009.5358112>

11. Zhou Y, Jin W (2011) A novel image zero-watermarking scheme based on DWT-SVD. In: 2011 international conference on multimedia technology, ICMT 2011, pp 2873–2876. <https://doi.org/10.1109/ICMT.2011.6002066>
12. Bin M (2011) Experimental research of image digital watermark based on DWT technology, pp 9–12
13. Ye X et al (2015) A Multiple-level DCT based robust DWT-SVD watermark method, pp 479–483 <https://doi.org/10.1109/CIS.2014.28>
14. Kunhu A (2016) A new multi watermarking algorithm for medical images using DWT and hash functions, pp 230–234
15. Goli MS, Student MS (2017) Two-step sudoku. IPRIA, pp 237–242

Smart Healthcare Analytics Solutions Using Deep Learning AI



K. P. Subiksha and M. Ramakrishnan

Abstract Machine learning is widely used in various applications such as business organizations, e-commerce, and healthcare industry, scientific and engineering for predicting and discovering relationships among data. In the healthcare industry, the predictive analytics in machine learning is mainly used for disease prediction. ML techniques help in predicting relationships in the electronic health record (EHR) data. The clinical process guidelines in the corpus may be considered as one of the inputs, and various healthcare parameters can be feature scaled, and the resultant architecture provides a positive impact in healthcare systems decision making. The objective of this work is to determine suitable features and optimal classifier design for a Deep Learning Healthcare Diagnosis system (DLHDS) to differentiate the endothelial dysfunction. It is used to predict under-perfused as well as over-perfused tissues during dynamic contrast material-enhanced magnetic resonance (MR) imaging of the peripheral vascular and muscular system. Early detection of disease and by mapping the drug side effects with patient histories is the needed approach for future prescriptions. The real-time applications of knowledge acquisition of healthcare data research require deep learning healthcare diagnosis systems. This paper presents deep learning service share model architecture for the generalizations of knowledge processing which is available in form of cloud and by using various parameters which enhances the assistive intelligence. By using semi-supervised machine learning techniques, features are reduced based on their nature of correlation in EHR and analyzed by ML techniques classifier and achieved an accuracy of 97%, and it outperforms the other existing prediction models, and this improves the throughput in the prediction of diseases.

Keywords Electronic health record · Predictive analytics · Semi-supervised approach · Machine learning techniques · Deep learning

K. P. Subiksha (✉)
Olive Tree Consultives and Freelancing, Bengaluru, India
e-mail: subiksha.kp@olivetreeconsultives.com

M. Ramakrishnan
School of Information Technology, Madurai Kamara University, Madurai, India

1 Introduction

Third wave AI technologies and contextual adaptation and Natural communication, and breakthrough of algorithms, had begun to have explosive causal models development [1]. EHR data and clinical text written by healthcare professionals to communicate the status and history of a single patient to other healthcare professionals or themselves and its time and visibility are eased [2] via machine learning. The major components of a healthcare system are the health professionals such as physicians or nurses, health facilities like clinics, hospitals for delivering medicines and other diagnosis or treatment technologies, and financing institution supporting the former two. The health professionals belong to various health sectors like dentistry, medicine, nursing, psychology, physiotherapy, and many others data will be uploaded in multi-cloud (Fig. 1) [2].

Cloud technologies help to keep the healthcare service for instance an algorithm as a service in cloud, and healthcare data can also be in cloud. Access to the healthcare service and data may be kept either as private or public; mainly based on the usage, the scope can be determined. With unified scalable healthcare systems and healthcare data, modern healthcare organizations can possibly revolutionize the medical therapies and medicine. Dynamic Susceptibility Contrast is capable of detecting under-perfused as well as over-perfused tissues. It has also been used on occasion to measure perfusion in animal experiments [3, 4]. MRI, however, provides the possibility for multiple and quantifiable parameters pertaining to tissue perfusion and microvascular status. Since its initial presentation, the approach has mainly been applied in studies of the myocardium, the brain and oncological research. Lately, DCEMRI has also been adopted to examine the peripheral vascular and muscular system [5, 6].

Some examples of commonly seen machine learning for medicine and healthcare research include diagnosis support [7] and outcome. Some causal models in Machine Learning helps in optimal workflow improvement [9], decision making [8], and outcome and risk [10], patient phenotyping [11].

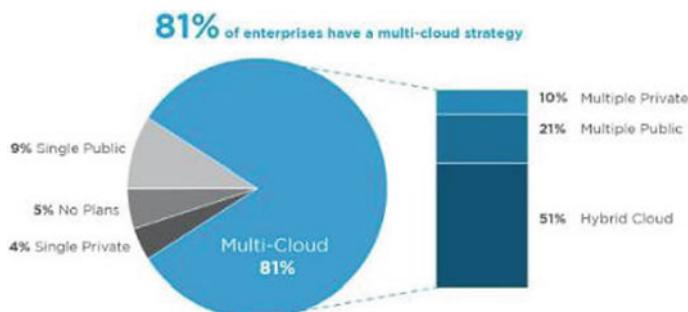


Fig. 1 Multi-cloud strategy enables to implement unified scalable healthcare systems

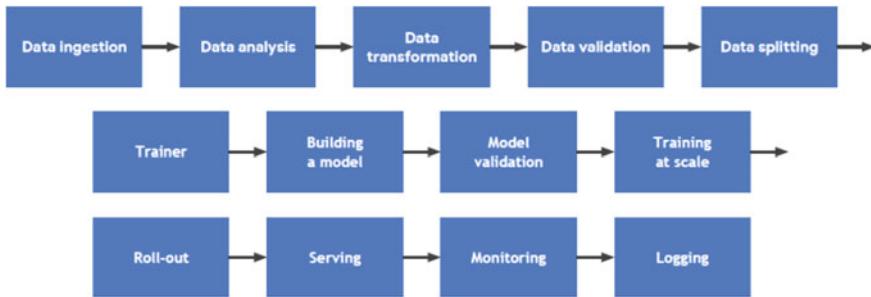


Fig. 2 Deep learning healthcare diagnosis systems workflow

2 Deep Learning Healthcare Diagnosis System (DLHDS)

The Deep Learning Healthcare Diagnosis System (DLHDS) is used for predicting the likelihood of endothelial dysfunction risk, and it uses the disease indicators [2] which are features mentioned below.

The relevant features used for the detection of endothelial dysfunction are listed below [1].

Age, sex, chest pain, blood pressure, cholesterol, fasting blood sugar, ECG Thalalch, EXANG, old peak, slope of ST segment, thallium, vessels colored, class.

2.1 Deep Learning Healthcare Diagnosis Systems Workflow

Enable adverse drug events like allergic and overdoses in care settings which helps in extracting the available information from the EHRs to acquire individual case safety processes in reports, and to avoid double data entry. Strengthening the current signal detection processes in Spontaneous Reporting System (SRS) for back tracking their case reports to the corresponding patient records, and to provide additional information on extended parts of the underlying medical history of the patient.

Both the events like ADE and SRS maintain the consistency for health record and then follow data ingestion. The gradient of a two-dimensional function is given. The derivative of the signal is evolved by convolving the signal with the filter kernel (Fig. 2).

2.2 Training the DLHDS Model

The gradient of a two-dimensional function is given. The derivative of the signal is evolved by convolving the signal with the filter kernel (Figs. 3 and 4).

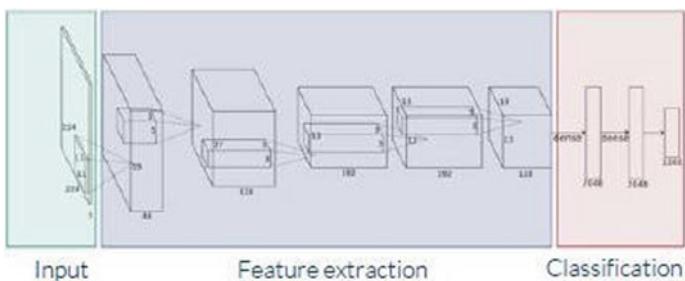


Fig. 3 Training DLHDS model

$$\nabla I(x, y) = \left[\frac{\partial I(x, y)}{\partial x} \frac{\partial I(x, y)}{\partial y} \right]^T \frac{I(x+h, y) - I(x-h, y)}{2h}$$

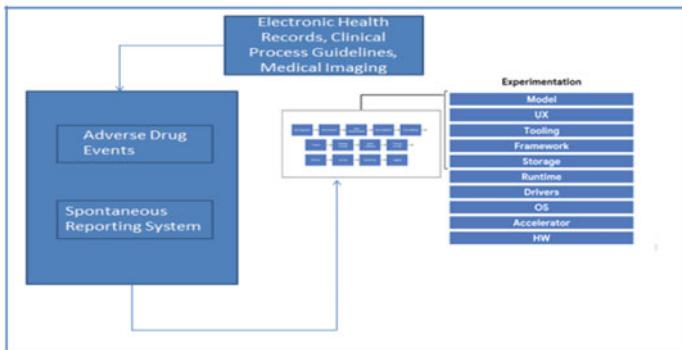


Fig. 4 Deep learning healthcare diagnosis systems architecture

$$\nabla I(x, y) = \left[\frac{\partial I(x, y)}{\partial x} \frac{\partial I(x, y)}{\partial y} \right]^T$$

Batch size—The batch size determines the number of training data points in each mini-batch.

Number of batches—The number of batches gives the total number of mini-batches in the entire training dataset.

Epochs—One epoch consists of one full pass of training over the entire dataset. To be more specific, one epoch is equivalent to a forward pass plus one backpropagation over the entire training dataset. So, one epoch would consist of n number of (forward pass + backpropagation) where n denotes the number of batches.

Algorithm for softmax has been used as the output layer.

Import the required libraries

Function to read the EHR dataset along with the labels

```
def readinfile():
```

```
ehr = inputdata.readdataets("EHRdata/","
```

```
onehot = True)
```

```
trainX, trainY, testX, testY = ehr.train.images,
```

```
ehr.train.labels, ehr.test.images, ehr.test.labels
```

```
return trainX, trainY, testX, testY
```

Define the weights and biases for the neural network

```
def weightsbiasesplaceholder(ndim,nclasses):
```

Define the forward pass

```
def forwardpass(w,b,X):
```

```
out = tf.matmul(X,w) + b
```

```
return out
```

Algorithm to activate the stochastic gradient descent

```
with tf.Session() as sess:
```

```
sess.run(init)
```

```
for i in xrange(epochs):
```

```
sess.run(optrain,feeddict = {X:trainX,Y:trainY})
```

```
loss = sess.run(cost,feeddict = {X:trainX,Y:trainY})
```

```
accuracy = np.mean(np.argmax(sess.run(out,feeddict=
```

```
{X:trainX,Y:trainY}),axis = 1) == np.argmax(trainY,axis=1))
```

```
losstrace.append(loss)
```

```
accuracytrace.append(accuracy)
```

```
if (((i + 1) >= 100) and ((i + 1) % 100 == 0)):
```

```
print 'Epoch:',
```

```
(i + 1),'loss:',loss,'accuracy:',accuracy
```

```
print 'Final training
```

```
result:','loss:','loss,'accuracy:','accuracy
```

```
losstest = sess.run(cost,feeddict={X:testX,Y:testY})
```

```
testpred = np.argmax(sess.run(out,feeddict=
```

```
{X:testX,Y:testY}),axis=1)
```

```
accuracytest = np.mean(testpred ==
```

```
np.argmax(testY,axis=1))
```

```
print 'Results on test
```

```
dataset:','loss:','losstest,'accuracy:','accuracytest
```

Multi-class classification with softmax function using full batch gradient descent

Epoch: 100 loss: 1.56331 accuracy: 0.702781818182

Epoch: 200 loss: 1.20598 accuracy: 0.772127272727

Epoch: 300 loss: 1.0129 accuracy: 0.800363636364

Epoch: 400 loss: 0.893824 accuracy: 0.815618181818

Epoch: 500 loss: 0.81304 accuracy: 0.826618181818

Epoch: 600 loss: 0.754416 accuracy: 0.834309090909
 Epoch: 700 loss: 0.709744 accuracy: 0.840236363636
 Epoch: 800 loss: 0.674433 accuracy: 0.845
 Epoch: 900 loss: 0.645718 accuracy: 0.848945454545
 Epoch: 1000 loss: 0.621835 accuracy: 0.852527272727
 Final training result: loss: 0.621835 accuracy: 0.852527272727
 Results on test dataset: loss: 0.596687 accuracy: 0.8614

Multi-class classification with softmax function using stochastic gradient descent

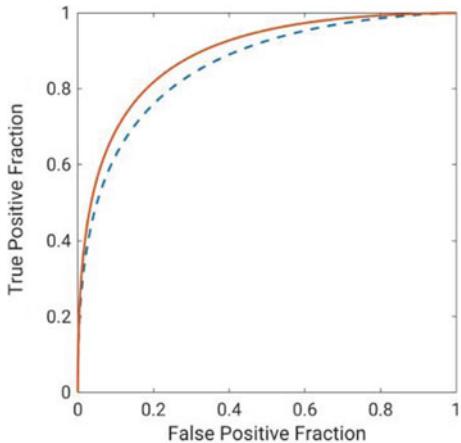
Epoch: 100 Average loss: 0.217337096686 accuracy: 0.9388
 Epoch: 200 Average loss: 0.212256691131 accuracy: 0.939672727273
 Epoch: 300 Average loss: 0.210445133664 accuracy: 0.940054545455
 Epoch: 400 Average loss: 0.209570150484 accuracy: 0.940181818182
 Epoch: 500 Average loss: 0.209083143689 accuracy: 0.940527272727
 Epoch: 600 Average loss: 0.208780818907 accuracy: 0.9406
 Epoch: 700 Average loss: 0.208577176387 accuracy: 0.940636363636
 Epoch: 800 Average loss: 0.208430663293 accuracy: 0.940636363636
 Epoch: 900 Average loss: 0.208319870586 accuracy: 0.940781818182
 Epoch: 1000 Average loss: 0.208232710849 accuracy: 0.940872727273
 Final epoch training results: Average loss: 0.208232710849 accuracy: 0.940872727273

Results on test dataset: Average loss: 0.459194 accuracy: 0.9155

Epoch 1/10 60,000/60,000—5 s 78 μ s/step—loss: 0.5929—accuracy: 0.8450
 Epoch 2/10 60,000/60,000—4 s 75 μ s/step—loss: 0.2804—accuracy: 0.9199
 Epoch 3/10 60,000/60,000—4 s 74 μ s/step—loss: 0.2276—accuracy: 0.9350
 Epoch 4/10 60,000/60,000—4 s 74 μ s/step—loss: 0.1933—accuracy: 0.9449
 Epoch 5/10 60,000/60,000—4 s 74 μ s/step—loss: 0.1682—accuracy: 0.9518
 Epoch 6/10 60,000/60,000—4 s 74 μ s/step—loss: 0.1490—accuracy: 0.9573
 Epoch 7/10 60,000/60,000—4 s 74 μ s/step—loss: 0.1332—accuracy: 0.9622
 Epoch 8/10 60,000/60,000—5 s 75 μ s/step—loss: 0.1202—accuracy: 0.9658
 Epoch 9/10 60,000/60,000—4 s 75 μ s/step—loss: 0.1090—accuracy: 0.9693
 Epoch 10/10 60,000/60,000—4 s 75 μ s/step—loss: 0.1000—accuracy: 0.9716
 Results on test dataset accuracy: 97.16. Results on test dataset accuracy: 97.16

3 Evaluation

The network was implemented in Python and Keras⁶⁸ with a TensorFlow backend using an NVIDIA P100 GPU with 16 GB GDDR5 RAM. The effect of network initialization was investigated by training the network using random initialization as well as initialization from another segmentation network that was developed using data from a 3T GE scanner. Some of the classifiers used are SVM, CNN. Metric used

Fig. 5 ROC curve

is F1, ROC. Threshold value varies from 0.15 to 0.22. Results are shown with F1 Value is 0.3 and ROC value is 0.78 which is shown in Fig 5.

The classifiers were trained and evaluated using nested fivefold cross-validation; here, all lesions from the same patient were kept together in the same fold in order to eliminate the impact of using correlated lesions for both training and testing. The cross-validation evaluation process was repeated with different random seeds, and the final prediction score for each lesion was averaged over the repetitions. Class relevance was held constant across the five cross-validation folds. Classifier performances were evaluated using receiver operating characteristic (ROC) curve analysis, with area under the ROC curve (AUC) serving as the figure of merit. The two classification methods were compared using the DeLong test standard errors, and 95% confidence interval (CI) of the difference in AUCs was calculated by bootstrapping the posterior probabilities of endothelial dysfunction [12].

4 Conclusion

In this paper, we focus on exploiting deep learning technique and its applications in health care. The prediction performance was boosted, so the learned space can help in predicting the endothelial dysfunction. DLHDS, a generation and discuss several techniques for learning such a model for EHR data. We demonstrate that the proposed model can produce realistic data samples by mimicking the input real data and the learned representation model by semi-supervised learning. Experimental results on two datasets show that the proposed model improves the generalization power and the prediction performance compared with improved baselines.

References

1. <http://www.technologystories.org/ai-evolution/>
2. <https://tw.leaderg.com/>
3. Versluis B, Backes WH, van Epen MG et al (2011) Magnetic resonance imaging in peripheral arterial disease: reproducibility of the assessment of morphological and functional vascular status. *Invest Radiol* 46:11–24 [PubMed] [Google Scholar]
4. Partovi S, Schulte AC, Jacobi B et al (2012) Blood oxygenation level-dependent (BOLD) MRI of human skeletal muscle at 1.5 and 3 T. *Magn Reson Imaging* 35:1227–1232 [PubMed] [Google Scholar]
5. Partovi S, Aschwanden M, Jacobi B et al (2013) Correlation of muscle BOLD MRI with transcutaneous oxygen pressure for assessing microcirculation in patients with systemic sclerosis. *Magn Reson Imaging* 38:845–851 [PubMed] [Google Scholar]
6. Partovi S, Schulte AC, Aschwanden M et al (2012) Impaired skeletal muscle microcirculation in systemic sclerosis. *Arthritis Res Ther* 14:R209 [PMC free article] [PubMed] [Google Scholar]
7. Miotto R, Wang F, Wang S, Jiang X, Dudley T (2017) Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*
8. Cheng Y, Wang F, Zhang P, Hu J (2016) Risk prediction with electronic health records: a deep learning approach. SIAM
9. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J et al (2016) Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *AMA* 316(22):2402–2410
10. Liu Y, Gadepalli K, Norouzi M, Dahl GE, Kohlberger T, Boyko A, Venugopalan S, Timofeev A, Nelson PQ, Corrado GS et al (2017) Detecting cancer metastases on gigapixel pathology images. arXiv preprint [arXiv:1703.02442](https://arxiv.org/abs/1703.02442)
11. Yala A, Barzilay R, Salama L, Griffin M, Sollender G, Bardia A, Lehman C, Buckley M, Coopey SB, Polubriaginof F et al (2017) Using machine learning to parse breast pathology reports. *Breast Cancer Res Treat* 161(2):203–211
12. DeLong ER, DeLong DM, Clarke-Pearson DL (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44:837–845

Design of 32—Bit MAC Unit Using Vedic Multiplier and XOR Logic



Aki Vamsi Krishna, S. Deepthi, and M. Nirmala Devi

Abstract Most of the computing circuits require multiply-accumulate (MAC) operation which involves the computation of product of the input data bits and subsequent addition or subtraction of that product to the accumulator. As it is seen that the multiplier is the most fundamental module of any MAC unit, it becomes necessary to design the multiplier such that it consumes less power and exhibits minimum delay. Initially, in our work, a 32 bit MAC unit is designed using xor gates and 16-bit Vedic multipliers. The design is then synthesized and implemented in Xilinx Vivado using Verilog HDL. Synopsys DC is also used to analyze parameters such as area and power. When compared to the existing methods, a reduction in power of around 3% is obtained with our proposed work.

Keywords MAC unit · Vedic multiplier · Synopsys DC · Verilog HDL · Xilinx vivado

1 Introduction

In this machinating world, multiplication is a frequently used arithmetic in many applications. Under this scenario, it is mandatory for the multiplier to be efficient in terms of power and speed. One such multiplier is the Vedic multiplier which incorporates the age-old Vedic mathematics. Though it is ancient, Vedic mathematics

A. V. Krishna (✉) · S. Deepthi · M. Nirmala Devi

Department of Electronics and Communication Engineering, Amrita School of Engineering,
Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: vamsikrishna2950@gmail.com

S. Deepthi
e-mail: sande.cbe@gmail.com

M. Nirmala Devi
e-mail: m_nirmala@cb.amrita.edu

proves to be a simple and powerful tool to date, for computing mathematical operations. Vedic mathematics includes 16 formulae known worldwide. Out of the 16 formulae, commonly called as ‘sutras’, Urdhva-Tiryagbyham sutra is used in our work to perform multiplication [1].

MAC unit is used extensively in modern processing systems which require high performance. This means that the MAC module determines the speed of the overall system. So, it is obvious that efficient multiplier is needed. Basically, a MAC unit consists of a multiplier along with an accumulator register which stores result in every clock.

This paper focuses on designing and implementing a 32 bit MAC with 16-bit Vedic multipliers using Urdhva-Tiryagbyham algorithm. The MAC unit is designed in such a way so as to store the values of repeated addition as well as subtraction of the product obtained from the Vedic multiplier. The process of addition or subtraction depends on the mode selected. So, it is obvious that the MAC has an inherent add submodule giving out a 64-bit output.

The paper shows a few associated works on this concept in Sect. 2 followed by a short introduction on Vedic mathematics in Sect. 3 and correspondingly, Sect. 4 illustrates 32-bit MAC design using a 16-bit Vedic multiplier. Section 5 shows the simulation output and compares power values. Section 6 provides inference and future scope of the proposed work.

2 Literature Survey

Based on the papers surveyed, the following observations are made. The authors in [1] proposed a 16-bit MAC module using Vedic multiplier. The implementation was done on Artix-7 FPGA and a power reduction of about 9.5% was obtained. The proposed parallel Vedic MAC unit in [2] showed a significant reduction in gate area and power dissipation while improving the speed of operation. The implementation of Vedic MAC units showed considerable improvement in total power, critical path delay and area of about 20–30% in case of 4 bit MAC unit and around 7–18% for 8 bit MAC unit. The design was then synthesized in 90 nm CMOS technology.

Gadakh and Khade [3] provided an architecture to design 16×16 bit Vedic multiplier based on Urdhva-Tiryagbyham sutra which used vertical and crosswise multiplication. The design was then optimized using carry save adders. Implementation was done using VHDL in Xilinx ISE Design Suite 14.5. A 33.26% reduction in delay was achieved when compared to other previously existing systems.

MAC unit in [4] consisted of 16×16 bit Vedic multiplier with carry look ahead (CLA) adder. The whole design was implemented in 16 nm technology. Though the number of transistors used were high, the results show progress in power when compared to power values in existing designs. The power dissipation was observed to be 0.17 mW and propagation delay time obtained is 27.15 ns.

In [5], the design of MAC with 8-bit Vedic multiplier and square root carry select adder was presented. It was then realized on FPGA and comparison was done in

terms of delay and LUT slices with Booth multiplier. It was determined that the latency of the neural network was increased.

The authors presented a MAC unit with an improved Kogge Stone adder and tested on FPGA in [6]. The design was modelled using Verilog HDL. Simulation and synthesis were done using Xilinx Vivado 2015.2. Improvements in the power of nearly 11.3% and power delay product of about 6.2% were observed when compared with traditional architecture.

In [7], a survey was made on different designs implemented in MAC units such as Wallace multiplier, Dadda multiplier, Booth multiplier, Carry Save Adder, Carry Select Adder, and Carry Look Ahead Adder with implementation done in 90 nm technology.[8] describes the use of Partial Product Reduction Block in multiplier to have better performance. When compared to conventional multiplier design, PPRB achieves a power reduction of nearly 39% and area is reduced by 17%.

The work in [9] focuses on MAC unit using an integrated hybrid multiplier and CLA network. The design is implemented using Verilog HDL. A 4 bit MAC is developed in [10] with a reduction in delay of about 25% when related with MAC unit having Wallace multiplier. In [11], MAC units with high speed and low power consumption are reviewed. The power of multiplier is 0.249 W and the area is 102 LUTs in [12].

3 Vedic Mathematics

Vedic mathematics is a list of ancient mathematical techniques that are considered to be powerful and directly exploited in numerous branches of mathematics. It is highly logical comprising of 16 sutras among which only two sutras are meant for multiplication.

As mentioned, the two sutras used for multiplication are Urdhva Tiryakbhyam (UT) and Nikhilam Navatashcaramam Dashatah (NND) for multiplication of any two numbers. Out of these two algorithms, UT sutra is chosen for this work as it involves the computation of two smaller bit numbers.

Urdhva Tiryakbhyam means ‘vertically and crosswise’. The basic concept is that partial products are obtained and parallel addition is done on these partial products.

An example of two four-bit numbers, namely, a [3:0] and b [3:0] is used to demonstrate multiplication using UT sutra. The following steps show the working algorithm.

$$\text{Step 1: } c0d0 = a0b0$$

$$\text{Step 2: } c1d1 = a0b1 + a1b0 + c0$$

$$\text{Step 3: } c2d2 = a0b2 + a1b1 + a2b0 + c1$$

$$\text{Step 4: } c3d3 = a0b3 + a1b2 + a2b1 + a3b0 + c2$$

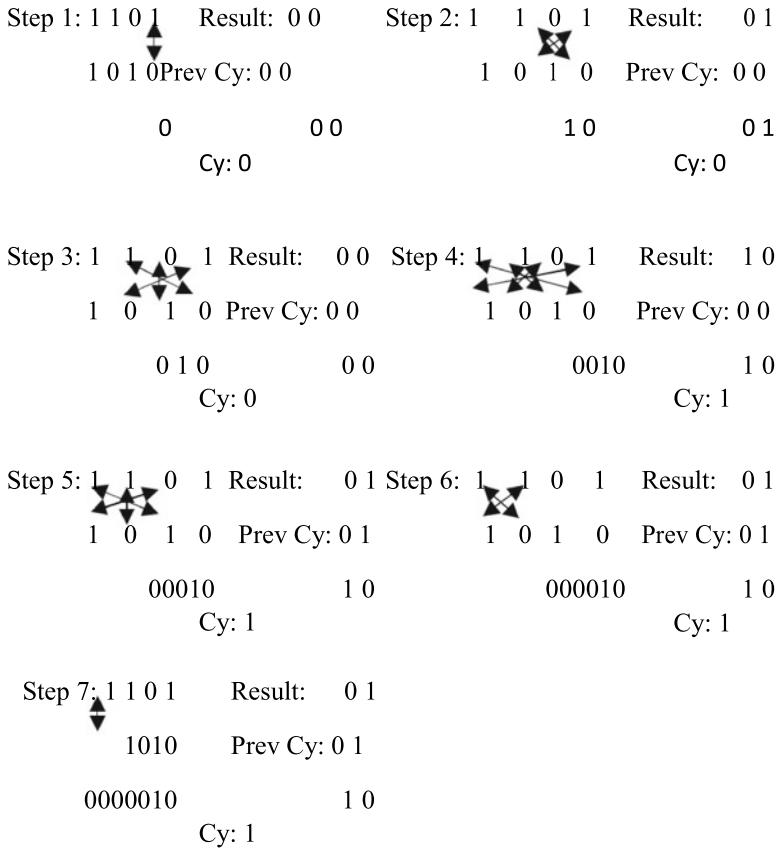
$$\text{Step 5: } c4d4 = a1b3 + a2b2 + a3b1 + c3$$

$$\text{Step 6: } c5d5 = a2b3 + a3b2 + c4$$

$$\text{Step 7: } c6d6 = a3b3 + c5$$

The result is finally obtained as c6d6d5d4d3d2d1d0 which is an 8-bit number. Here, c [3:0] represents the carry and d [3:0] represents the output. Similarly, this concept can be extended to any number of bits. The number of steps involved increases correspondingly as the number of bits involved in computation. Initially, UT sutra was designed using decimal number systems. It was then extended to represent binary number systems since most of the circuits used in this digital world require binary numbers for easy computation and storage.

For example, consider two four-bit numbers 1101 and 1010, the result of the product being 130 in decimal. Using UT sutra, the result is obtained as follows:



4 Design of MAC Unit

The structure of basic MAC module is shown in Fig. 1.

The multiply-accumulate unit consists of a multiplier, in this case, a Vedic multiplier is considered, the result of which is added to or subtracted from the data of the accumulator that stores the output. In the proposed work, a 32 bit MAC unit is designed using a 16×16 Vedic multiplier and xor gates along with an accumulator register. Figure 2 shows the block diagram of the proposed design.

Fig. 1 Basic structure of MAC unit

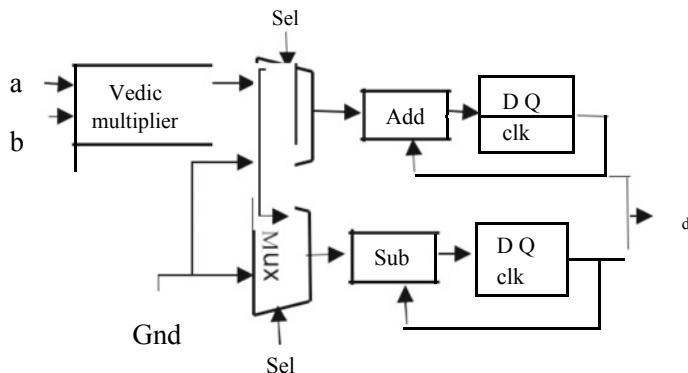
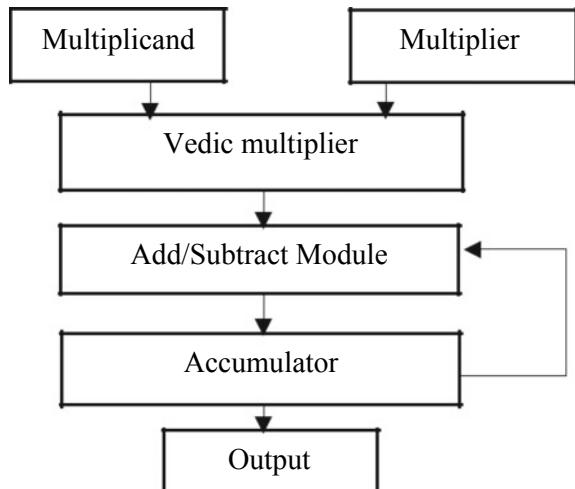


Fig. 2 The proposed design architecture

4.1 Vedic Multiplier

The design of the 16-bit Vedic multiplier consists of few stages and is initially started with developing a 2×2 Vedic multiplier using UT sutra as illustrated for two 4-bit number multiplication. Later a 4×4 multiplier is constructed using the 2×2 multiplier designed in the first stage. Similarly, 8×8 multiplier is made using 4×4 multiplier blocks. In the end, a 16×16 Vedic multiplier is done using 8×8 multiplier blocks. Mux is used to select the mode, that is, whether to add or subtract.

4.2 XOR Logic

In the initial stage of the design, a 2-bit multiplier using UT sutra was implemented with half adders to add the partial products. Later, for implementing a 4×4 Vedic multiplier, four 2×2 multiplier blocks were used along with xor logic applied to the partial products. This reduces the hardware complexity and at the same time understandable. Similar logic is applied to the design of 8-bit and 16-bit multipliers.

4.3 MAC Unit

A clock signal is used to enable the MAC operation. If the accumulator is initially loaded with, say, a value x , then the MAC unit is often said to implement functions with expression $x + ab$, where ab is the output of the 16-bit Vedic multipliers. The MAC is capable of doing repeated addition or subtraction based on the mode selected through select lines of the mux. It is essential to design such an efficient MAC module as it behaves to be the core of any DSP algorithm. The output of the proposed MAC is of 64-bit length.

The whole design is then executed using Verilog HDL which is simulated and synthesized in Xilinx Vivado. The simulation results are discussed in the following section.

5 Results and Discussion

The simulation results are as shown in Fig. 3 and Fig. 4. Figure 3 shows the result of 16-bit Vedic multiplier and Fig. 4 shows the output for MAC unit.

Figure 5 shows the output of the subtraction operation in the MAC unit.

The proposed design is synthesized in 90 nm technology with Synopsys Design Compiler apart from synthesizing and implementing in Xilinx Vivado. The area

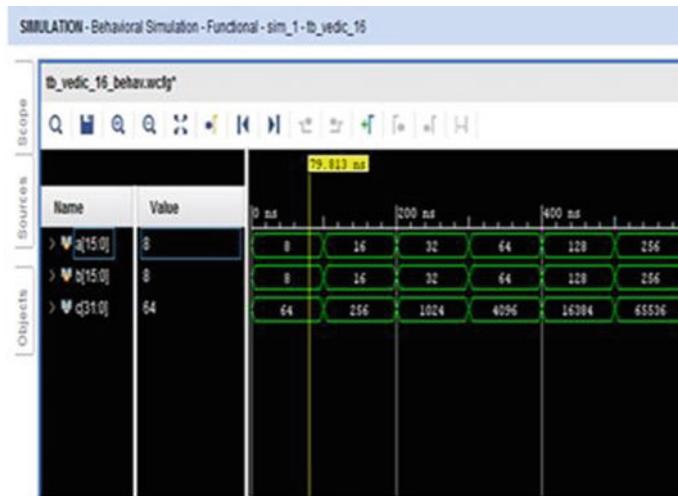


Fig. 3 16×16 vedic multiplier output

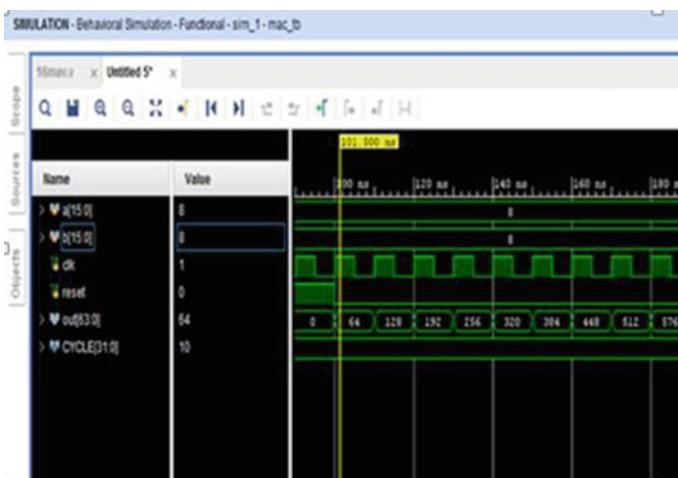


Fig. 4 MAC addition output

and power obtained for multiplier and accumulator modules in Synopsys Design Compiler are tabulated in Table 1.

The value of MAC power obtained for the proposed design in Vivado is as shown in Table 2 which is very much less when compared to the existing design as shown in [6]. It is known that if the number of bits at input increases, then the power consumed will also rise. In such a situation the existing MAC unit consumes more power than the proposed one for higher bit numbers.

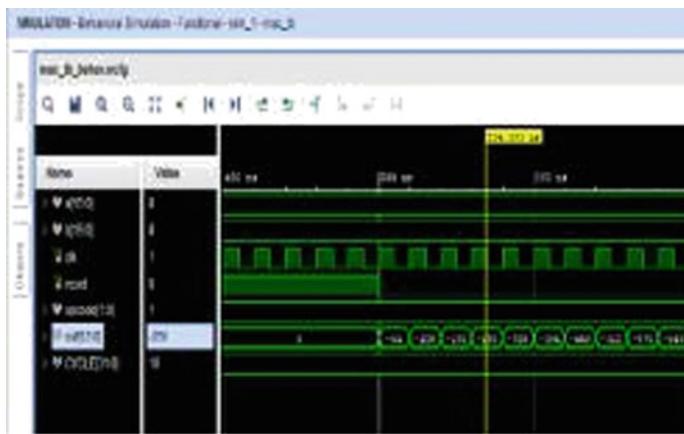


Fig. 5 MAC subtraction output

Table 1 Power and area values

Parameter	Circuit	
	Multiplier	Accumulator
Total power (μW)	10.7569	545.8040
Total area (μm^2)	8227.8909	5555.9737

Table 2 Comparison of power values of MAC

Parameter	Design	
	Existing [6]	Proposed
Bits	16	32
Power (W)	0.124	0.241

6 Conclusion

A MAC unit has been designed using 16-bit Vedic multiplier based on Urdhva-Tiryagbyham sutra. The design is then synthesized in Xilinx Vivado and Synopsys DC. It is shown that a significant improvement in power of about 3% has been observed. Hence the proposed MAC unit could be used in various signal processing applications to improve performance.

References

- Vamsi ASK, Ramesh SR (2019) An efficient design of 16 bit MAC unit using vedic mathematics. In: International conference on communication and signal processing, pp 0319–0322

2. Jithin S, Prabhu E (2015) Parallel multiplier—accumulator unit based on vedic mathematics. *ARPN J Eng Appl Sci* 10(8):3608–3613
3. Gadakh SN, Khade A (2016) Design and optimization of 16×16 bit multiplier using vedic mathematics. In: International conference on automatic control and dynamic optimization techniques (ICACDOT), pp 460–464
4. Bathija RK, Meena RS, Sarkar S, Sahu R (2012) Low power high speed 16×16 bit multiplier using vedic mathematics. *Int J Comput Appl* 59(6):41–44
5. Ranganath L, Jay Kumar D, Siva Nagendra Reddy P (2016) Design of MAC unit in artificial neural network architecture using verilog HDL. In: International conference on signal processing, communication, power and embedded system (SCOPES), pp 607–612
6. Rakesh S, Vijula Grace KS (2019) VLSI based low power multiply accumulate unit employing kogge stone adder with modified pre-processing and post-processing stages. *Int J Eng Adv Technol (IJEAT)* 8(4):295–299
7. Rakesh S, Vijula Grace KS (2017) A survey on the design and performance of various MAC unit architectures. In: Proceedings of 2017 IEEE international conference on circuits and systems (ICCS 2017), pp 312–315
8. Ahish S, Kumar YBN, Sharma D, Vasantham MH (2015) Design of high performance multiply-accumulate computation unit. In: IEEE international advance computing conference (IACC), pp 915–918
9. Dwivedi K, Sharma RK, Chunduri A (2018) Hybrid multiplier-based optimized MAC unit. In: 9th international conference on computing, communication and networking technologies (ICCCNT)
10. Yuvraj M, Bhaskar N, Kailath BJ (2017) Design of optimized MAC unit using integrated vedic multiplier. In: International conference on microelectronic devices, circuits and systems (ICMDCS)
11. Patil PA, Kulkarni C (2018) A survey on multiply accumulate unit. In: Fourth international conference on computing communication control and automation (ICCUBEIA)
12. Shawl MS, Singh A, Gaur N, Bathla S, Mehra A (2018) Implementation of area and power efficient components of a MAC unit for DSP processors. In: Proceedings of the 2nd international conference on inventive communication and computational technologies (ICICCT)

Deep Learning Model for Detection of Attacks in the Internet of Things Based Smart Home Environment



Raveendranadh Bokka and Tamilselvan Sadasivam

Abstract The number of devices getting connected to the internet is growing exponentially day by day with the advent of the Internet of Things (IoT), commensurately in every domain of IoT attacks and threats are also growing in alarming rate because of the lack of proper security measures in the devices and network. Proper methods are required in the IoT environment for detecting attacks thus by providing effective defence and security. In this paper, we developed Deep Learning (DL) based Deep Neural Network (DNN) to detect attacks like Denial of Service (DoS), spying, malicious control in IoT based smart homes. The model was evaluated using DS2OS (Distributed Smart Space Orchestration System) dataset for detection of DoS, Data Type Probing, Scan, Spying, Malicious Control, Wrong setup and Malicious Operation attacks which have given best accuracy as 99.42% and performance metrics like precision, F1-score and recall are evaluated. The proposed model was compared to the DL methods proposed by other authors.

Keywords Internet of things (IoT) · Security · Deep learning (DL) · Deep neural networks (DNN) · Attacks detection

1 Introduction

Nowadays IoT technology is widely used in common household applications [1] such as washing machines, smart light control, smart door lock and surveillance camera devices which are interconnected through the internet. But those devices are insecure and exposed to different types of attacks [2]. As more number of devices are interconnected as a system, the importance of network security is increasing [3]. Therefore, the IoT infrastructure has to be secured from the cyber-attacks. So, there is

R. Bokka (✉) · T. Sadasivam

Department of Electronics and Communication Engineering, Pondicherry Engineering College, Puducherry, India

e-mail: bravindra64@pec.edu

a need for developing a reliable, smart and secured system for detecting cyber-attacks and recover itself automatically.

To secure the IoT environment effectively the existing methods has to be enhanced to detect the attacks. Because of the huge data generated in IoT environment, the learning methods like Machine Learning (ML) and Deep Learning (DL) are powerful for identifying ‘normal’ and ‘abnormal’ behaviour of IoT devices [4]. DL having vital advantages over traditional ML because of its superior performance for the large datasets, so DL became imperative research in IoT system [5].

Here to detect the attacks in smart home when it is in abnormal state deep learning based solution was proposed. The DL is a subfield of ML, it mimics the human brain mechanism by the establishment of neural network to interpret data such as images, text and signals [6].

The remaining paper was organized to address the related work in the literature as Sect. 2. Section 3 describes DS2OS dataset and discussed about effects of attacks. Section 4 discusses about data preprocessing, DNN model implementation and performance metrics used to assess the model performance. Section 5 gives a description of experimental setup, results and methods in the literature are compared with the proposed method. Finally, Sect. 6 addresses the conclusion and future work.

2 Related Work

The studies show that the traditional methods are completely surpassed by deep learning [7]. In [8], the authors used a deep learning approach for flow-based anomaly detection with the deep neural network, the experimental results shows that deep learning can be applied for anomaly detection in Software Defined Network (SDN). In [9], the authors used Deep Belief Network (DBN) to construct the DNN with one input, three hidden and one output layers for detection of several attacks in IoT networks such as ‘sinkhole’, ‘Denial of Service (DoS)’, Opportunistic service, Wormhole and blackhole, they obtained an average precision of 0.95 and recall of 0.97. In [10], for detection of online network attacks in IoT-connected home environment the authors used a Dense Random Neural Network (DRNN) based on deep learning approach, the results show that DRNN detects attacks correctly when the attacks are inserted in the network packets. In [11], the authors designed a DL model for detecting the routing attacks in IoT system, the model detects three types of attacks namely rank, hello flood and version number attacks with low accuracy of 38.5% for softmax activation in the output layer. In [12], the authors evaluated the latest CICIDS2017 datasets with deep learning models for detection of Distributed Denial of Service (DDoS), which gives the highest accuracy of 97.16%. In [13], the authors discussed the comparison study of deep and shallow neural networks for detection of attacks in fog-to-things architecture by using open-source dataset, they designed models for detecting four classes of attacks for DNN model and got the better accuracy as 98.27% compared to shallow neural network of accuracy 96.75%.

In inference to the literature survey most of the authors used the deep learning models for detection of attacks and anomaly, they got comparative performance metrics with one another but accuracy, precision and recall average values are low in predicting the attacks. The objective of proposed work is to implement a Deep Neural Network (DNN) based on deep learning for recognition of attacks with enhancement in performance metrics like accuracy, precision and recall.

3 Description of Data Set

The data set was collected from Kaggle [14], it is an open-source dataset provided by Pahl and Aubet[15]. By using DS2OS synthetic data set was collected with the virtually created IoT environment. This dataset contains traces captured from different IoT simulated sites with multiple types of services like a light controller, thermometer, movement sensors values, washing machines, the status of batteries and thermostats, operation of smart doors and smartphones.

The data set consists of total 357,952 samples with 13 features, in that 347,935 are normal data and 10,017 samples contain eight classes of attack and anomalous data. Table 1 gives the description about 13 features with data type and detailed description of the different attacks and normal data distribution through whole data was described in Table 2. The features ‘Access Node Type’ of 148 and ‘Value’ of 2050 has missing data [16].

The effects of each attack described in Table 2 are explained as follows.

- *Denial of Service (DoS):* Flood out the target by sending too many ambiguous packets and make its services unavailable to the server or other devices [10].

Table 1 Feature description in DS2OS dataset

Name of the feature	Data type
Source ID	Nominal
Source address	Nominal
Source type	Nominal
Source location	Nominal
Destination services address	Nominal
Destination service type	Nominal
Destination location	Nominal
Accessed node address	Nominal
Accessed node type	Nominal
Operation	Nominal
Value	Continuous
Timestamp	Discrete
Normality	Nominal

Table 2 Frequency distribution of each class in whole dataset

Name of class/attack	Total number of samples in whole data set
Denial of service	5780
Data type probing	342
Malicious control	889
Malicious operation	805
Scan	1547
Spying	532
Wrong setup	122
Normal	347,924

- *Data Type Probing (D.T.P.):* The malicious nodes sends different data instead of the original data in these type of attacks [15].
- *Malicious Control (M.C.):* With the help of software inabilities the attackers can capture network traffic and control the whole system [17].
- *Malicious Operation (M.O.):* Malware distracts the original operation by a snare. Malicious operation negatively affects the performance of the Device's [16].
- *Scan:* Sometimes the data can be corrupted in the process of scanning the system by hardware to acquire the data [16].
- *Spying:* The attackers use a backdoor channel to capture important information by using the vulnerabilities of the system [13].
- *Wrong Setup (W.S.):* Because of the wrong system setup sometimes the data may also get disordered [18].
- *Normal:* The normal data is entirely correct and accurate [16].

The number of samples contained for each attack in whole data set are ‘DoS’ is 5780, ‘Data Type Probing’ is 342, ‘Malicious Control’ is 889, ‘Malicious Operation’ is 805, ‘Spying’ is 532, ‘Scan’ is 1547, ‘Wrong Setup’ is 122 and ‘Normal’ is 347,924.

4 Deep Learning for Attacks Detection

4.1 Deep Learning Model

In this paper, the proposed Deep Neural Network (DNN) model based on deep learning uses the supervised learning and multiclass classification for identifying the eight-category class of attacks. The model uses backpropagation algorithm to update the weights bypassing the residuals that were updated. The deep learning model design with five-layers consists of one input layer, three hidden layers, and one output softmax layer for the multiclass classification as shown in Fig. 1.

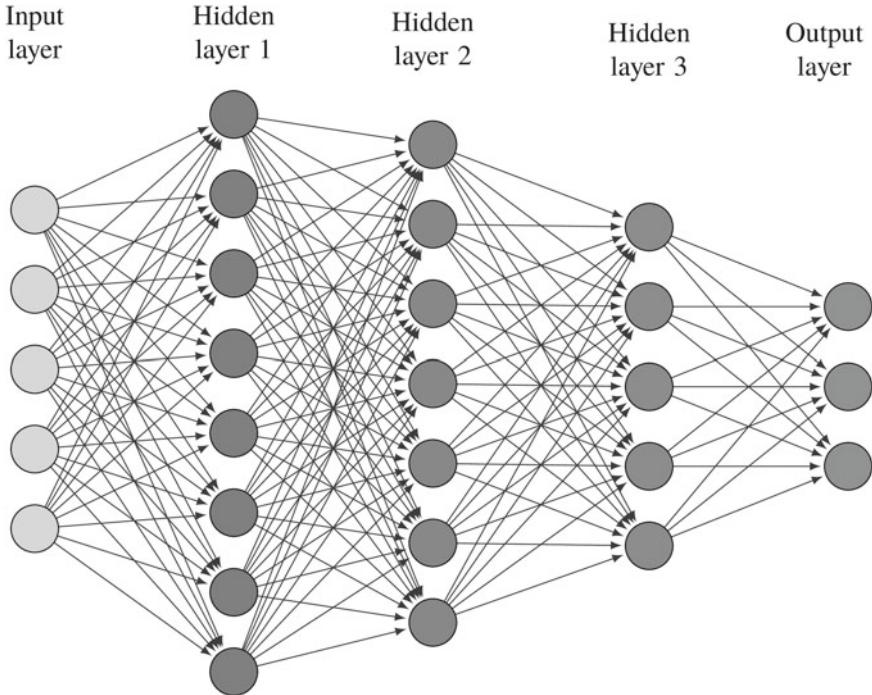


Fig. 1 Deep neural network (DNN) model architecture

The input layer consists of 11 nodes for fed the 11 input traffic traces features to Deep Neural Network (DNN). The features fed into the DNN passes through the first hidden layer which consists of 40 nodes for filtering the most substantial features and they are passed to the second hidden layer which consists of 25 nodes for further filtering out the features. The same features are fed to the third hidden layer which consists of 15 nodes for filtered out the eight outputs. The result of the last hidden layer is passed to the output layer consists of eight nodes for representing the classification as normal or attacks specified in the data description.

In the training mechanism of DNN model, the features input fed to the DNN is passing each layer. At each DNN layer neural nodes uses activation function to calculate the filtered output and passes that into the next layer.

In each layer of deep learning model, neural nodes use different activation functions. In this proposed model the first hidden layer developed with the ‘sigmoid’ activation, the second and third hidden layers uses the Rectified Linear Unit (ReLU) and output layer developed with the ‘Softmax’ activation function for multiclass classification. The activation functions used are defined as follows.

Sigmoid activation function defined as (1)

$$f(x) = \frac{1}{1 + e^{-x}}, \text{ where } x \text{ is an input to the node} \quad (1)$$

ReLU function is defined as (2):

$$f(x) = \max(0, x) \quad (2)$$

The optimizer ‘Adam’ was tied with the loss function ‘sparse_categorical_crossentropy’ because of the integer targets. The loss function used to guide the optimizer to move in the right or wrong direction to for shape and mould the models into its most accurate possible form by updating its weights.

4.2 Data Preprocessing

In data processing, the feature ‘Timestamp’ is removed from the dataset and that was not considered for analysis and also 11 rows of ‘normal’ class data which contains the missing data were removed from the data set. The deep learning networks will take only the numerical data as feature inputs, so the nominal categorical data to be converted into numerical data. The categorical data can be converted into numeric data using many ways in that most of the researchers are using the Label Encoding or One Hot Encoding. In this paper, categorical data are converted to feature numeric data using a Label Encoding technique. After the data preprocessing, the data set consists of Normal data of 347,924 and 10,017 samples of data attacks. These data were separated into 60% of the train and 40% of test data sets. To train the model train dataset is used and test data set is used for validating the final model.

4.3 Performance Metrics

To measure the performance of DNN model for detecting the attacks and normal traffic traces, the performance metrics used are Accuracy, Precision, Recall and F1-score. Table 3 shows the representation of confusion matrix with the notations used to calculate the performance metircs.

- True Positive (TP): the number of attack records are classified as attack.
- False Positive (FP): the number of normal records are classified as attack.
- True Negative (TN): the number of normal records are classified as normal.
- Flase Negative (FN): the number of attack records are classified as normal.

Table 3 Representation of confusion matrix

	True class/label	Predicted class/label	
		Attack	Normal
Attack	TP	FN	
Normal	FP	TN	

Accuracy: It is ratio of the correctly predicted records to the total number of records in the given data set shown in (3)

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

Precision: It is the ratio of the correctly classified records to all actual classified records shown in (4)

$$\text{Precision} = \frac{TP}{(TP + FP)} \quad (4)$$

Recall: It is the ratio of the correctly classified records to all records that should be classify shown in (5)

$$\text{Recall} = \frac{TP}{(TP + FN)} \quad (5)$$

F1-Score: It is the harmonic mean of precision and the recall shown in (6). It is an important parameter for measuring the performance of the model when the data set is imbalanced.

$$\text{F1 - Score} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \quad (6)$$

5 Experiment Results and Discussion

The experiment was done using Dell Laptop where the operating system was Windows 1064-bit running on Core i7 processor with 8 GB RAM. The model was designed and evaluated in Jupyter Notebook environment using Keras, it is written in Python and capable of running high-level neural networks API's on top of TensorFlow [19]. The data cleaning and feature engineering was done using NumPy and Pandas framework, Matplotlib and Seaborn libraries used for data visualization and plotting, for data analysis Keras packages and for measuring the performance metrics Scikit-Learn framework was used.

The DNN model was implemented as discussed in the Sect. 4 was compiled with the learning rate 0.1, beta_1 and beta_2 values 0.9 and 0.999, respectively, for the optimizer ‘Adam’ along with the ‘loss function’ explained in Sect. 4. The model was trained with 150 epochs and the training time taken to each epoch is 2 s. The training and testing data sets were used to evaluate the trained model to measure the average value of history keys like ‘Training _Accuracy’, ‘Testing _Accuracy’, ‘Training _Loss’ and ‘Testing _Loss’ the same keys are shown in Figs. 2 and 3 with average values as ‘99.42%’, ‘99.42%’, ‘0.58%’ and ‘0.58%’ respectively. The proposed DNN model

Fig. 2 DNN model training and testing accuracy

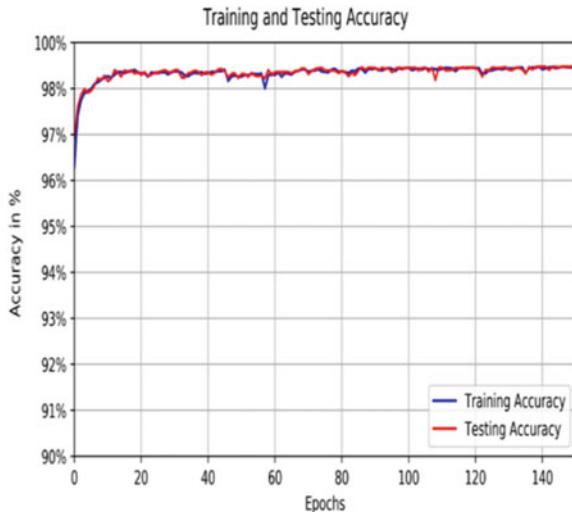
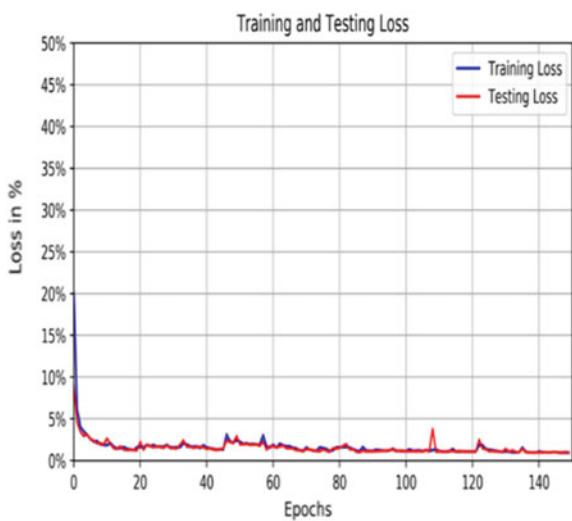


Fig. 3 DNN model training and testing loss

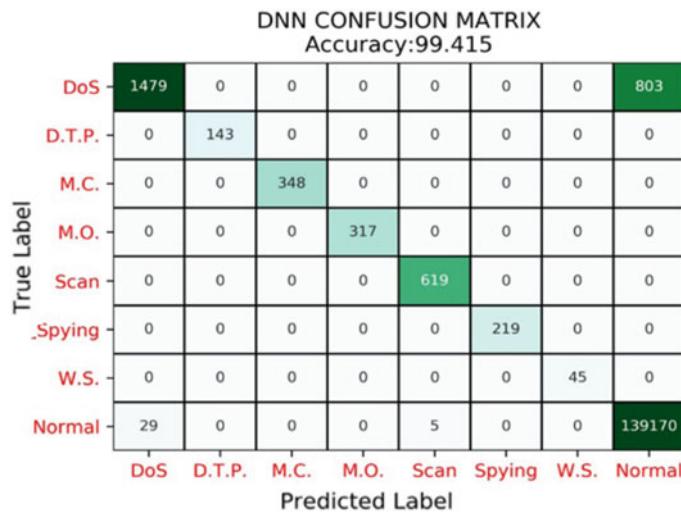


gives the best accuracy for detecting the attacks as compared to the other techniques used in the literature as shown in Table 4.

Figure 4 shows the confusion matrix that was generated for the test data using a trained model. From the confusion matrix, it was observed that except DoS and Normal classes every class correctly classified. For the DoS class, 803 samples are misclassified as Normal. In normal class 29 and 5 samples are misclassified as DoS and Scan, respectively. The accuracy of the model for prediction shown as 99.42%. Individual class detection accuracy is high compared with the methods proposed in [16].

Table 4 Performance comparisons

Author and year	Dataset used	Type of classification	Techniques used	Performance metrics evaluated
Geethapriya et al. 2019 [9]	Own	Multiclass	Deep Neural network	Precision = 0.95 Recall = 0.97
Brun et al. 2018 [10]	Own	Multiclass	Random dense neural network (DRNN)	Compared with threshold values not specified any metrics
Yavuz et al. 2018 [11]	Own IRAD	Multiclass	Deep neural network	Accuracy = 38.5%
Monika et al. 2019 [12]	CICIDS 2017	Binary class	Deep learning	Accuracy = 97.16%
Diro et al. 2018 [13]	NSL-KDD	Multiclass	Shallow neural network Deep neural network	Accuracy = 96.75% Accuracy = 98.27%
Our method	DS2OS	Multiclass	Deep neural network	Accuracy = 99.42% Precision = 0.99 Recall = 0.99 F1-score = 0.99

**Fig. 4** Confusion matrix of eight-category classification

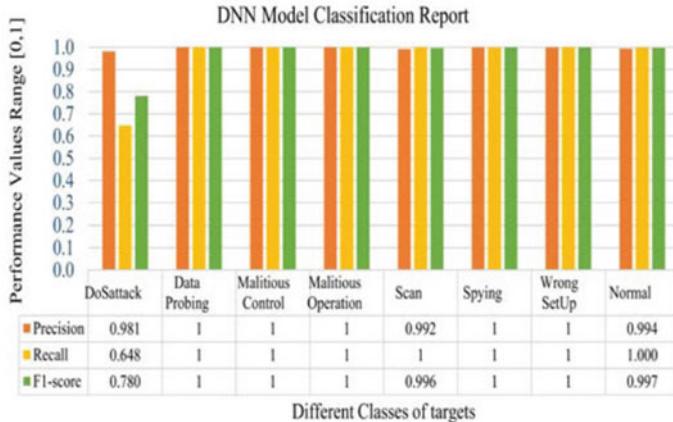


Fig. 5 Precision, recall, F1-score of DNN for testing data

The classification report of performance metrics Precision, Recall and F1-score measured using the designed DNN model by giving the testing data as input and compared the predicted class values with the test class values is shown in the Fig. 5. From the graph, it was observed that except the DoS attack all are classified exactly and its performance values are nearly equal to 1. The average precision, recall and F1 score values are 0.98, 0.99 and 0.99, respectively for detecting the attacks. The performance values are best when compared with the method proposed by [9].

Table 4 shows the comparative study of different methods in literature with our proposed deep learning model. As compared to all other implementations our DNN model gives a better accuracy of 99.42% with improvement in other performance metrics like precision, recall and F1 score.

6 Conclusion

The deep learning model proposed in this paper was implemented for the detection of attacks and anomaly in the IoT environment with DS2OS dataset. The model has predicted the attacks and anomaly like DoS, Scan, Spying, M.C., D.T.P, M.O and W.S. with an accuracy of 99.42% and the model has predicted all the attacks correctly except DoS and Normal with average performance measures of Precision, Recall and F1-Score as 0.99, 0.99 and 0.99, respectively. Compared to other IoT attack detection models based on DNN, our model has given best performance. The only disadvantage of this model is, DoS attack was not predicted correctly. So, in future work, the model parameters can be enhanced for detection of zero-day attacks and DoS attack in real-time IoT network environment.

References

1. Zhang Y, Li P, Wang X (2019) Intrusion detection for IoT based on improved genetic algorithm and deep belief network. *IEEE Access* 7:31711–31722. <https://doi.org/10.1109/ACCESS.2019.2903723>
2. Doshi R, Aphorpe N, Feamster N (2018) Machine learning DDoS detection for consumer internet of things devices
3. Ioannou L, Fahmy SA (2019) Network intrusion detection using neural networks on FPGA SoCs. In: 2019 29th international conference on field programmable logic and applications, pp 232–238. <https://doi.org/10.1109/fpl.2019.00043>
4. Al-garadi MA, Mohamed A, Al-ali A, Du X, Guizani M (1932) *Surv Polit Q* 3:581–589. <https://doi.org/10.1111/j.1467-923X.1932.tb01141.x>
5. Li H, Ota K, Dong M (2018) Learning IoT in edge: deep learning for the internet of things with edge computing. *IEEE Netw.* <https://doi.org/10.1109/MNET.2018.1700202>
6. Lecun Y, Bengio Y, Hinton G (2015) Deep learning
7. Yin C, Zhu Y, Fei J, He X (2017) A deep learning approach for intrusion detection using recurrent neural networks. *IEEE Access* 5:21954–21961. <https://doi.org/10.1109/ACCESS.2017.2762418>
8. Niyaz Q, Sun W, Javaid AY, Alam M (2015) A deep learning approach for network intrusion detection system. In: EAI international conference on bio-inspired information and communications technologies (BICT)
9. Thamilarasu G, Chawla S (2019) Towards deep-learning-driven intrusion detection for the internet of things. *Sensors* 19. <https://doi.org/10.3390/s19091977>
10. Brun O, Yin Y, Gelenbe E (2018) Deep learning with dense random neural network for detecting attacks against IoT-connected home environments
11. Yavuz FY, Ünal D, Güllü E (2018) Deep learning for detection of routing attacks in the internet of things. *Int J Comput Intell Syst* 12:39–58. <https://doi.org/10.2991/ijcis.2018.25905181>
12. Roopak M, Yun Tian G, Chambers J (2019) Deep learning models for cyber security in IoT networks. In: 2019 IEEE 9th annual computing and communication workshop and conference, CCWC 2019
13. Diro AA, Chilamkurti N (2018) Distributed attack detection scheme using deep learning approach for internet of things. *Futur Gener Comput Syst* 82:761–768. <https://doi.org/10.1016/j.future.2017.08.043>
14. DS2OS traffic traces (2018) Kaggle. <https://www.kaggle.com/francoisxa/ds2ostraffictraces>
15. Pahl MO, Aubet FX (2018) All eyes on you: distributed multi-dimensional IoT microservice anomaly detection. In: 14th international conference on network and service management, CNSM 2018 and workshops, 1st international workshop on high-precision networks operations and control, HiPNet, SR+SFC 2018
16. Hasan M, Islam MM, Zarif MII, Hashem MMA (2019) Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches. *Internet Things* 7:100059. <https://doi.org/10.1016/j.iot.2019.100059>
17. Zhang ZK, Cho MCY, Wang CW, Hsu CW, Chen CK, Shieh S (2014) IoT security: ongoing challenges and research opportunities. In: Proceedings—IEEE 7th international conference on service-oriented computing and applications, SOCA 2014, pp 230–234. Institute of Electrical and Electronics Engineers Inc
18. Leister W, Schulz T (2012) Ideas for a trust indicator in the internet of things. In: SMART 2012—the first international conference on smart systems, devices and technologies
19. Home-Keras Documentation. <https://keras.io/>. Online Access

Smart Irrigation Using Decision Tree



Chinmay Patil, Shubham Aghav, Sagar Sangale, Shubham Patil, and Jayshree Aher

Abstract Indian climate which primarily consists of four seasons and the majority of agriculture in India is practised in monsoon which is rainfed agriculture. The lack of irrigation makes India a potent place for the experiment of efficient water irrigation. Internet of Things (IoT) have emerged has ubiquitous technology which is being used in many fields. Machine Learning has proven to be helpful in data-intensive tasks. IoT and machine learning would prove helpful in agriculture. This paper proposes a system based on the research of machine learning and IoT to increase the efficiency in irrigation. With the help of data gathered through sensors and using machine learning algorithm the amount of water the needed for irrigation is calculated and used for irrigation.

Keywords Internet of things · Agriculture · Machine learning

1 Introduction

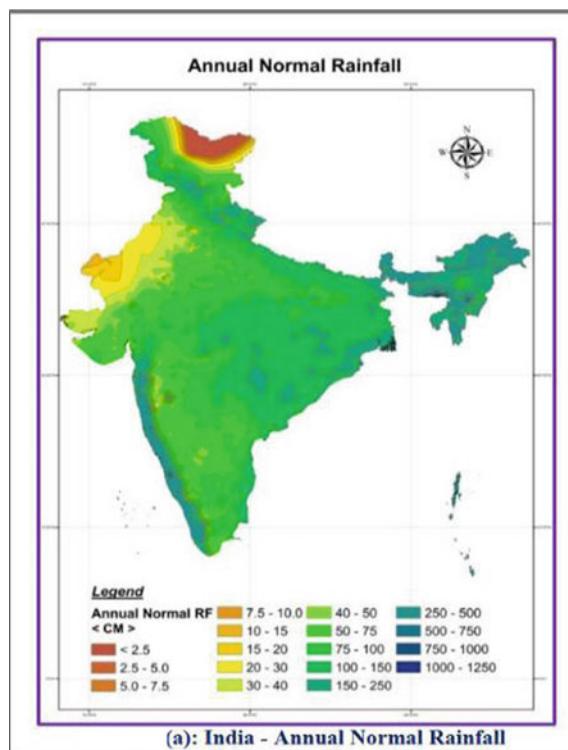
India is predominately an agriculture-based country. There is a dire need of saving the agriculture sector which makes 15.82% of our current GDP making it one of the major sectors for the economy. Irrigation forms the backbone of the agriculture sector In India, majority of farmers use rainfed type agriculture style. In which it is predominately dependent upon the rainfall and the weather surrounding the rainfall.

Figure 1a shows the Annual Normal Rainfall. And the other Fig. 2 shows the Withdrawal of Monsoon. According to Revitalizing Rainfed Agriculture Network's report, India is ranked first in crop arear under rainfed agriculture in the world. Around 60% of farmers rely on it. And around 55% of the gross cropped area is

C. Patil (✉) · S. Aghav · S. Sangale · S. Patil · J. Aher
MIT-COE Pune, Pune, India
e-mail: chinmaypatil8@gmail.com

S. Aghav
e-mail: s.aghav410@gmail.com

Fig. 1 Annual normal rainfall. *Source* Indian Meteorological Ministry report no: ESSO/IMD/HS/Rainfall Report/01(2018)/24



under it. The lack of efficiency in irrigation causes wasteful use of water, which is an important resource.

Depending on the weather, water available and crops the requirement of water needs to be adjusted accordingly. Collecting data from sensors and analyzing it, we make a decision on what amount of water is needed and can be used for irrigation. This can increase the efficiency of water irrigation.

This paper is divided into sections, Sect. 2 contains Literature Survey, Sect. 3 contains various Technique's studied, and Sect. 4 describes the proposed system.

2 Literature Survey

This section contains the survey of various paper, which uses varying approaches in agriculture using IoT.

The paper authored by Suciu et al. [1] analyzes different types of IoT platforms present in the current market. It proposes a system for disease management of crops. The System Under Analysis (SUA) uses the MQ Telemetry Transport (MQTT) network protocol, and Grafana for monitoring and management.

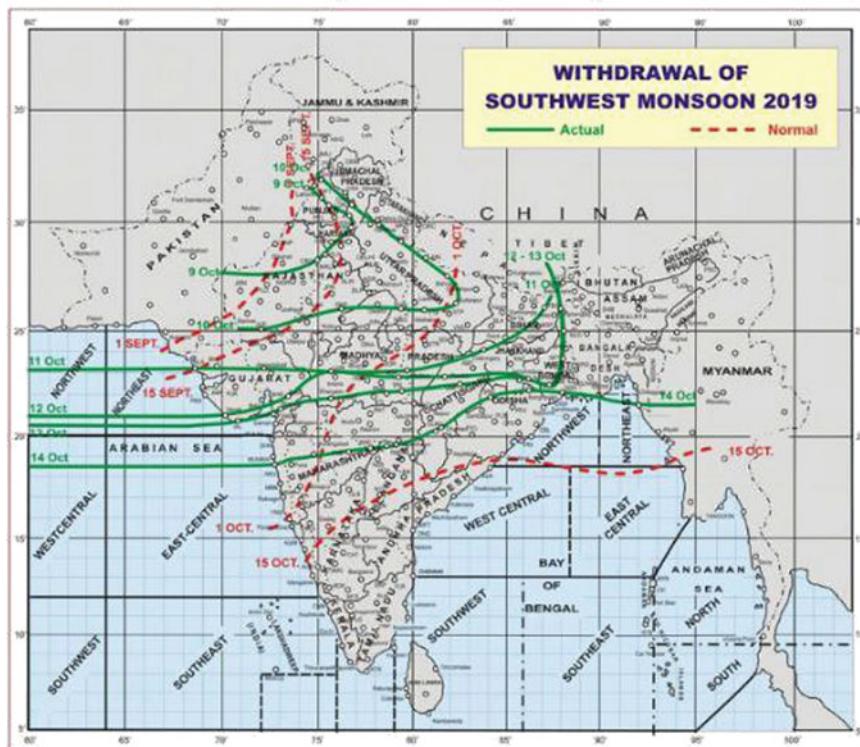


Fig. 2 Withdrawal of southwest monsoon. Source Indian Meteorological Ministry report no: ESSO/IMD/HS/Rainfall Report/01(2018)/24

The paper authored by Hamouda et al. [2] the system proposed uses Fuzzy Logic using Wireless Sensor Network for communication. It excellently integrates the humidity and temperature but limitations of this system are it fails to recognize the amount rainfall and the then calculates the amount of the time the field should be irrigated.

In the system proposed in the paper authored by Alomar and Alazzam [3] uses Fuzzy Logic and communication through WSN. This proposes a detailed approach towards the management of irrigating using Fuzzy Logic. The water is managed through a controlled valve. By using soil moisture and temperature the amount of water to be released is calculated.

In the paper authored by Srinivasulu et al. [4] cloud service-oriented architecture is proposed. Data is collected from the sensor and sent to the server, after processing the data the system takes action such as irrigation. Actions are done automatically without any human intervention. The services provided as per the paper are Web Portal Service, Voice-Based Service, SMS Based Service, Interactive Video Conferencing Service, Farming Community Online Chatting Service.

In the paper authored by Baranwal et al. [5], the system proposed monitors the threat to crops. It integrates electronic security devices for efficient food preservation. WSN is used for transmitting data. The system in paper proposes Things-as-a-service. The IoT architecture proposed is divided into Perception Layer, Network Layer, Application Layer and Middleware, which is present between Application Layer and Network Layer.

The paper authored by Zhao et al. [6] uses cost-effective RFID for connection. Transmission of data is through a wireless network. The system is made up of terminal link, business link and M2M support platform. System software includes data acquisition software and web application software.

3 Technique

An algorithm reaches solution for given output following sequence of instruction. The solution which is economical in computing resources and time is a good solution. Some tasks or problems we don't have a definite algorithm, there machine learning proves helpful. Machine Learning uses past experiences or training data to optimize performance. Machine Learning using Statistics to build a Mathematical Model. For the purpose of decision-making in IoT various methodology are available. For an efficient irrigation system, some of the methodology studied are Bayesian Models, Fuzzy Logic, and Decision Tree.

3.1 Bayesian Models

In Bayesian Models, Bayesian theorem is used for inference. It is a mathematical procedure that applies probabilities to statistical problems, providing the tools to update the beliefs based on the new evidence. It is a supervised learning algorithm. Bayes' Theorem is given as:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)} \quad (1)$$

The $P(H|E)$ is the probability of hypothesis “ H ” holds given evidence “ E ” is true also called Posterior. $P(H)$ is the probability of “ H ”, which is also called Prior. $P(E|H)$ is probability of evidence given hypothesis is true also called Likelihood and $P(E)$ is probability of evidence “ E ”.

3.2 Fuzzy Logic

In Fuzzy Logic, fuzzy refers to things which are vague, ambiguous. When modeling the real world, the world is complex we encounter various situations where the results are not in binary, true or false. Here Fuzzy Logic prove to be helpful and flexible in reasoning. The reasoning is approximate than precise. In Standard Logic, result is either true or false, but in Fuzzy Logic, we have a degree of truth.

3.3 Decision Tree

It is a type of hierarchical data structure using divide and conquer strategy. It is inexpensive to compute and are popular since it has good accuracy [7]. It consists of nodes and leaves. Nodes are used for making a decision, and leaf nodes representing result or output. Decision Tree is nonparametric model, the tree structure is not fixed. It grows according to the problems of data and complexity.

Based on the research the machine learning algorithm chosen is a decision tree. Decision Tree is easy to understand compare to Bayesian model. Decision Trees are also easier to debug.

4 Proposed System

4.1 Architecture Design

This paper proposes an Intelligent Irrigation system. This automation of irrigation is being developed with the help of the decision tree. Sample data required for the decision tree is taken from the experienced farmer. The main purpose of the proposed system is to determine the irrigation delay. Sample data is collected from the farmers for training decision tree model for particular region and particular crop. Input to the proposed system is the data collected from sensors and the decision tree model which was previously built. Various sensors such as Humidity Sensor, Soil Moisture Sensor, Temperature Sensor are used.

System then collects the amount of moisture present in the soil, humidity present at the time, and predicted weather. The system will determine the time after which irrigation should be started. Signal is then passed to the module which controls motor switch, which is responsible for the actual irrigation. As it will be the one to take action based on the decision made by the system.

The sensor used is DHT11 which give temperature and Humidity in digital format. SEN-13322 would give us soil moisture detail. The sensors data will be collected from a distinct part of land in Arduino. The data will be pre-processed in Arduino

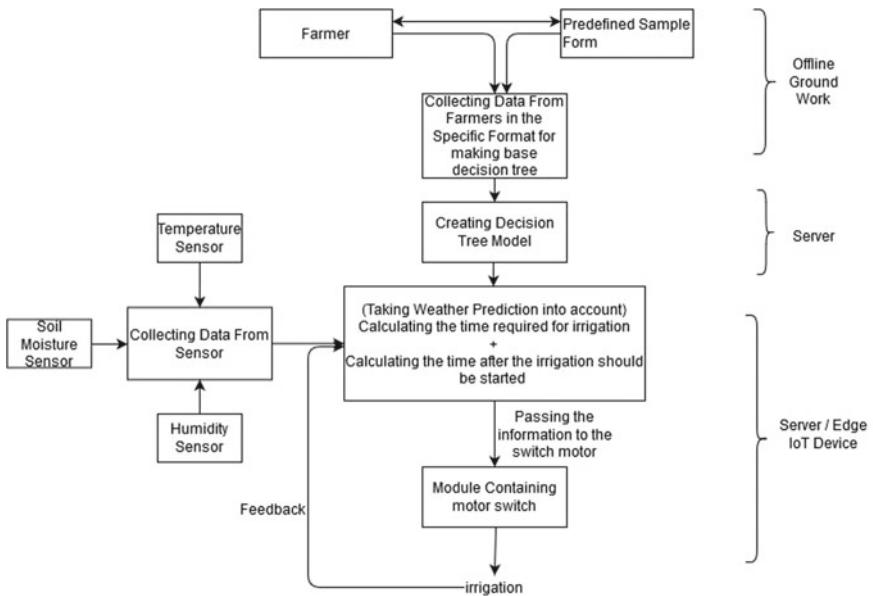


Fig. 3 System architecture of the proposed system

Table 1 Sample table used for collecting data from farmers

Temperature sensor value (°C)	Humidity sensor value (%)	Soil moisture sensor value (%)	Predicated weather	Predicated delay (output hour)	Predicated on time (output hour)
22	69	55	Cloudy	2	1
26	55	60	Cloudy	1	3

and it will send the pre-processed data to Raspberry Pi after a certain amount of time. Raspberry Pi will act as a local server and decision-making system (Fig. 3; Table 1).

4.2 Experimental Analysis

This Section covers the categorical conversion and Mathematical Model of the proposed system.

4.2.1 Categorical Conversion

The decision tree will be made by using the categorical attributes. Table 2 presents

Table 2 Conversion of numerical attributes to categorical attributes

Humidity sensor values		Soil moisture sensor value		Temperature (DHT 11) values		Predicted weather
Numerical attributes (%)	Categorical attributes	Numerical attributes (%)	Categorical attributes	Numerical attributes (°C)	Categorical attributes	
00–20	VL	00–20	VL	00–10	VL	Clear
20–40	L	20–40	L	10–20	L	Sunny
40–60	M	40–60	M	20–30	M	Cloudy
60–80	H	60–80	H	30–40	H	Scatter rain
80–90	VH	80–100	VH	40–50	VH	Rainy

the conversion of the numerical attributes coming from the sensor to the categorical attributes for further operation.

4.2.2 Mathematical Model

Input to the decision tree model:

$$t_{PW} = \{\text{Sunny, Clear, Scatter Rain, Cloudy, Rainy}\}$$

$$t_{SM} = t_{TE} = t_{HU} = \{VL, L, M, H, VH\}$$

where

SM Soil Moisture

TE Outside Temperature

HU Outside Humidity

PW Predicted weather

L Low

VL Very Low

M Medium

H High

VH Very High

The numerical attributes are the sensors output and the symbols in parenthesis indicate units.

Output of decision tree model:

PRO Predicted On time

PRD Predicted Delay

$$t_{PRO} = t_{PRD} = \{1, 2, 3, 4, 5\}$$

5 Conclusion

The paper proposes a smart irrigation system, which uses sensors to collect data, Arduino for pre-processing and Raspberry Pi for decision-making. Using a decision tree algorithm, we calculate amount of water to be irrigated. Decision Tree uses categorical data to make decision. The accurate automation of irrigation enables the farmer to schedule his works efficiently. The proposed system is user friendly, low cost and automated process encourage water conservation. The precise water supply helps in proper water utilization by plants, which in turn grow healthy crops and give optimum yield. The proposed system can be further expanded for creating a Software as a Service (SaaS) for data mining. Further research can give accurate information about the water usage of a particular region. The use of solar power can deepen the autonomy of this system.

References

1. Suciu G, Istrate C, Dițu M (2019) Secure smart agriculture monitoring technique through isolation. In: 2019 global IoT summit (GIoTS), pp. 1–5. <https://doi.org/10.1109/GIOTS.2019.8766433>
2. Hamouda YEM (2017) Smart irrigation decision support based on fuzzy logic using wireless sensor network. In: 2017 international conference on promising electronic technologies (ICPET), pp. 109–113. <https://doi.org/10.1109/icpet.2017.26>
3. Alomar B, Alazzam A (2018) A smart irrigation system using IoT and fuzzy logic controller. In: 2018 fifth HCT information technology trends (ITT), pp. 175–179. <https://doi.org/10.1109/ctit.2018.8649531>
4. Srinivasulu P, Babu MS, Venkat R, Rajesh K (2017) Cloud service oriented architecture (CSoA) for agriculture through internet of things (IoT) and big data. In: 2017 IEEE international conference on electrical, instrumentation and communication engineering (ICEICE), pp 1–6. <https://doi.org/10.1109/ICEICE.2017.8191906>
5. Baranwal T, Nitika Pateriya PK (2016) Development of IoT based smart security and monitoring devices for agriculture. In: 2016 6th international conference—cloud system and big data engineering (confluence), pp 597–602. <https://doi.org/10.1109/CONFLUENCE.2016.7508189>
6. Zhao J, Zhang J, Feng Y, Guo J (2010) The study and application of the IOT technology in agriculture. In: 2010 3rd international conference on computer science and information technology, pp 462–465. <https://doi.org/10.1109/iccsit.2010.5565120>
7. Safavian SR, Landgrebe D (1991) A survey of decision tree classifier methodology. IEEE Trans Syst Man Cybern 21:660–674. <https://doi.org/10.1109/21.97458>

Contextually Aware Multimodal Emotion Recognition



Preet Shah, Patnala Prudhvi Raj, Pragnya Suresh, and Bhaskarjyoti Das

Abstract Detection of emotion in a conversation has a lot of applications such as humanizing chatbots, understanding public opinion through social media, medical counseling, building security systems and interactive computer simulations, etc. Since humans express emotion not only from what they speak but also from their tone and facial expressions, we have used features from three modes—text, audio and video and tried out different fusion techniques to combine the models. We have proposed a new architecture specially designed for dyadic conversation where each individual is modelled using a separate network that exchanges emotion context and seems to have a conversation with the other network. We have refined this model using teacher force.

Keywords Multimodal agent · Teacher force · Dyadic conversation · Emotion recognition

1 Introduction

Emotions are vital for perception, human experience, cognition and some tasks such as learning, communication, and rational decision making. Human-computer interaction (HCI) systems which sense a user's state and give adequate feedback are

P. Shah (✉) · P. P. Raj · P. Suresh · B. Das

Department of Computer Science and Engineering, PES University, RR Campus, Bengaluru, 560100, India

e-mail: shahh.preet@gmail.com

P. P. Raj

e-mail: pruthvipatnala@gmail.com

P. Suresh

e-mail: pragnyasuresh@gmail.com

B. Das

e-mail: bhaskarjyoti01@gmail.com

perceived to be more natural, trustworthy and persuasive. Thus, it's no surprise that emotion recognition is gaining a lot of attention in the research field. Applications are spread across different fields like Medicine, E-learning, Monitoring, Marketing, Entertainment and Law.

Many current emotion recognition systems use speech transcripts to identify sentiment and emotion. Humans are very expressive by nature. They give out several behavioral cues that can supplement emotion recognition. Significant information to better identify affective states of a person are provided by facial expressions and vocal modifications in addition to textual data. Therefore, a combination of audio, video and text helps to create a good emotion or sentiment analysis model. We have built models which individually take features from each mode and performed various combination methods to compare and contrast the effectiveness of each mode in identifying emotion.

Current systems also do not take care of different entities in a conversation. A single network is used to model the states of different individuals. We propose a method to model individuals using different networks and mimic a real life conversation by interactions between the networks. Three separate aspects of a conversation: the speaker, context of the conversation and the emotion response of the previous utterance are taken care of by our model. We use multilayer LSTM cells to maintain context of the conversation and concatenate emotion information of the previous utterance of a network with the current utterance in the other network to maintain the flow of emotion throughout the conversation.

2 Related Work

The art of combining modalities goes a long way in determining the accuracy of emotion recognition. The effective combination or merging of modalities should result in a single network being able to substitute an ensemble model consisting of a network for each mode.

Earlier work on emotion recognition from text involved an emotion estimation module [1] that assessed the affective content of text based messages using the relationship between subject, verb and object in a sentence. Emotional Movie Transcript Corpus (EMTC) [2], a multi-label corpus that is claimed to be closely related to real life conversations as opposed to tweets is being widely used for emotion recognition and conversational models.

Realizing that along with "What is said?", "How it is said?" is also important, research started inclining towards using more than just text to identify emotions.

A scalable methodology for fusing multiple affect sensing modules [3], allowing the subsequent addition of new modules without having to retrain the existing ones by the use of continuous evaluation-activation space has shown an enhancement in an Instant Messaging application.

Emotion recognition by considering audio and video [4] has been explored using a Visual and Speech Network. Discarding the LSTM layers of the individually trained

networks resulted in visual and speech features which were then concatenated. The Multimodal network used these features for emotion recognition.

A deep neural framework called Conversational Memory Network [5] addresses utterance-level emotion recognition in dyadic conversational videos by leveraging contextual information from the conversation history. Interactive Conversational Memory Network [6] was proposed to detect emotion from multimodal features that were extracted from conversational videos to hierarchically model inter-speaker and self emotional influences. Gesture recognition using a multi-modal approach [7] has overcome the difficulty in classifying similar motion patterns by considering speech and face recognition systems. A similar approach can be used to overcome the shortcomings of text-only models like sarcasm detection. Multimodal Transformer Networks (MTN) [8] can be used to encode videos and incorporate information from various modalities. In order to extract query-aware features from audio and video, an auto-encoder was used to implement the so called query-aware attention. The ability of neural networks to estimate complex functions [9] has been exploited to effectively recognize emotion from different modalities using the IEMOCAP dataset. Some major issues frequently ignored in multimodal sentiment analysis research [10] include the role of speaker-independent models, importance of different modalities, and generalizability. The proposed framework illustrates different facets of analysis to be considered while performing multimodal sentiment analysis.

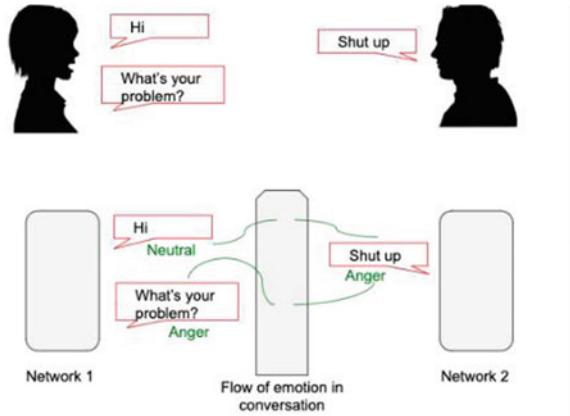
A new method to keep track of individual party states throughout the conversation using recurrent neural networks [11] was used for emotion classification. The emotion representation for the current utterance was modeled as a function of state of the present speaker and previous utterance using different states for each party.

Next came the problem of combining the modes effectively so as to increase the overall accuracy. Principle Component Analysis (PCA) [12] was tested for feature selection and an intermediate-level feature fusion technique on the MOSI dataset was proposed. The paper [13] addresses the lack of context in unimodal features. The approach proposed incorporates the inter-dependencies between utterances and effectively captures the contextual information to better classify sentiment from user generated videos. A hierarchical approach to fusion of modalities has been used employing LSTMs to capture contextual information. A novel feature fusion strategy that proceeds in a hierarchical fashion [14] was presented where modalities were fused two at a time and then all three modalities. Unimodal features were considered pairwise and passed as input to bimodal models whose output was passed to the trimodal model. Their technique has shown a significant improvement over concatenation of features.

3 Methodology

We propose a new split-set architecture that is inspired by how humans have a conversation (Fig. 1). When two people converse, there is a topic of conversation and each person adds points according to what he/she has said earlier and as a reaction to what the other person has just said.

Fig. 1 Inspiration behind our architecture



We propose that each individual can be modeled as a separate network (Fig. 2). Thus, for a dyadic conversation we have two networks—network A and network B. Network A will receive input from individual A from all three modes and the emotion predicted by network B for the previous utterance of individual B. This way the two networks seem to have a conversation where they exchange emotion and context information.

To implement this, we would need to wait for a network to output an emotion and then send this along with individual data to the other network. Thus, training time will increase drastically since each network has to wait to get input. Also, the loss of one network would be required to be propagated to the other network, increasing complexity. So, we used a technique called teacher force.

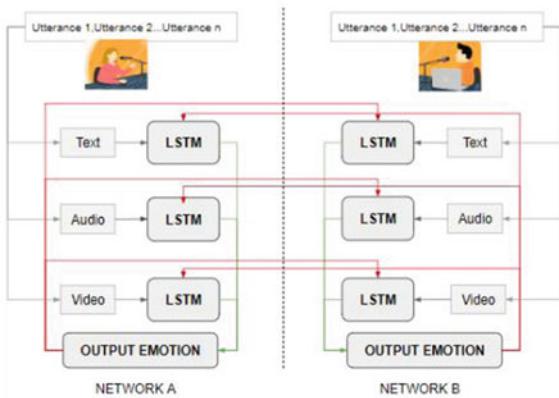
In teacher force (Fig. 3), we use the expected output directly from the dataset as the input to the next time step instead of the output generated by the network. In the architecture if we assume that network B has 100% accuracy, it would be equivalent to replacing it with the dataset comprising of output emotions of the sentences spoken by the second person (Figs. 1 and 2).

3.1 Feature Extraction

Text from transcripts is converted to integers using a tokenizer that creates the vocabulary index using word frequency. This is padded so that each sentence is represented by the same length vector. Emotion of the previous utterance is encoded with the text vector. We use pre-trained GloVe embeddings of dimension 300, along with the maximum sequence length of 26 to obtain a vector for each utterance.

Audio For the audio data our preprocessing follows the work of Chuang [15]. A total of 34 features have been extracted using energy-based features Mel-frequency Cepstral Coefficients (MFCC) and Fourier frequencies. The features comprise of

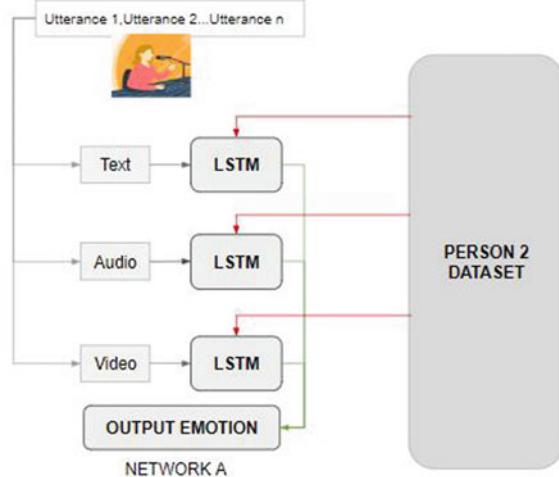
Fig. 2 Architecture for dyadic conversation



8 Time Spectral Features like short-term entropy of energy, spectral centroid and spread, spectral rolloff, spectral entropy, short-term energy, spectral flux and zero crossing rate. A 0.2 s window with a stride of 0.1 s and 16 kHz sample rate has been used to calculate features. Each utterance is represented as a (100, 34) sized vector by considering 10 s of input which is roughly 100 frames with zero padding as applicable. Delta and double delta features of MFCC were also experimented. There was no enhancement in performance but for increase in computation (Fig. 3).

Video In order to capture video features, a MOCAP device was used. The device monitors different parts of a face. In addition, position of the hands and the rotation of the head are also taken into account. The obtained data is directly used to capture video features.

Fig. 3 Refined model using teacher force



3.2 Dataset

The IEMOCAP dataset has been used. The dataset is composed of 12 h of audiovisual data. It includes video, speech, text transcriptions and motion capture of face. The dataset is based on dyadic sessions where the actors role-play a scripted scenario to elicit emotional expressions.

3.3 Model

The unimodal text model architecture comprises of an embedding layer followed by two bidirectional LSTM layers and two dense layers with Relu and Softmax activations. Categorical cross entropy is the loss function and adam optimizer is used. The audio model uses a similar architecture. The video model employed CNN network with 2D convolutions.

Since we are dealing with dyadic conversation, we split the data into two—utterances of person A and person B. While giving in the utterance of person A, the emotion of person B for the previous utterance in the conversation is encoded along with it. This is the principle of teacher force. So the input to each unimodal model is the utterance $x_A(t)$ along with emotion $e_B(t - 1)$.

Also the IEMOCAP data is highly skewed towards anger emotion and hence upscaling of other emotions is done. The technique of label smoothing is applied where the expected output of 1 is replaced with 0.9 and 0 with 0.02. The sum of all values in target output vector is still 1.

Following the hfusion method proposed in [14], for bimodal models, the output of the penultimate layer of the unimodal model is taken as features. Features of two models are concatenated pairwise and given as input to the bimodal model. Thus there are three combinations: Text+Audio, Audio+Video, Text+Video.

Finally the features from the penultimate layers of the bimodal models are given as input to the trimodal model. Thus this model effectively gives weights to each mode of speech.

4 Results

We have implemented and contrasted between full-set architecture (a single network handling all entities of a conversation) and our split-set architecture accuracies. Experiments with two fusion techniques—Concatenation and HFusion are done on the full-set models to contrast their performance. All of the following models have been trained for 30 epochs (Tables 1, 2, 3 and 4).

Table 1 Unimodal classification accuracies while using the full-set architecture

Modes	Full-set architecture (%)
Text	61
Audio	56.68
Video	40.87

Table 2 Accuracies after applying concatenation fusion technique on unimodal models

Modes	Full-set architecture (%)
Text and audio	61.85
Audio and video	40.84
Text and video	23.4
Text and audio and video	40.84

Table 3 Accuracies after applying hierachal fusion technique on unimodal models

Modes	Full-set architecture (%)
Text and audio	28.79
Audio and video	27.73
Text and video	13.86
Text and audio and video	62.78

Table 4 Accuracy of unimodal split-set architecture

Modes	Split-set architecture (%)
Text	72.38

5 Conclusions

We have observed that text gives the most information followed by audio and video respectively. Fusing video along with other modes has visibly reduced accuracy with the exception of trimodal fusion model.

Though concatenation fusion technique has higher bimodal accuracies, its training time is much higher than hfusion model. After combining all modes, hfusion gave a better accuracy than concatenation method.

We have implemented split-set architecture on text mode and there is a drastic increase in accuracy as compared to the normal model. Our hypothesis is that if the unimodal model accuracies are higher, then a fusion technique(concatenation or hfusion or any other) should improve upon this accuracy.

6 Future Work

We need to try out different fusion techniques on our split-set architecture. Also, the proposed architecture is specifically designed for dyadic conversations. It needs to be extended to adapt to conversations with more than two entities.

The individual models are rather simplistic and more variations like bidirectional LSTMs and GRUs along with attention layers need to be tried out.

References

1. Ma C, Osherenko A, Prendinger H, Ishizuka M (2005) A chat system based on emotion estimation from text and embodied conversational messengers. In: Proceedings of the 2005 international conference on active media technology, (AMT 2005). IEEE, pp 546–548
2. Phan DA, Matsumoto Y (2018) Emtc: multilabel corpus in movie domain for emotion analysis in conversational text. In: Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)
3. Hupont I, Ballano S, Baldassarri S, Cerezo E (2011) Scalable multimodal fusion for continuous affect sensing. In: 2011 IEEE workshop on affective computational intelligence (WACI). IEEE, pp 1–8
4. Tzirakis P, Trigeorgis G, Nicolaou MA, Schuller BW, Zafeiriou S (2017) End-to-end multimodal emotion recognition using deep neural networks. *IEEE J Sel Top Signal Process* 11(8):1301–1309
5. Hazarika D, Poria S, Zadeh A, Cambria E, Morency LP, Zimmermann R (2018) Conversational memory network for emotion recognition in dyadic dialogue videos. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, vol 1(Long Papers), pp 2122–2132
6. Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R (2018) Icon: interactive conversational memory network for multimodal emotion detection. In: Proceedings of the 2018 conference on empirical methods in natural language processing, pp 2594–2604
7. Akrouf S, Belayadi Y, Mostefai M, Chahir Y (2011) A multi-modal recognition system using face and speech
8. Le H, Sahoo D, Chen NF, Hoi SC (2019) Multimodal transformer networks for end-to-end video-grounded dialogue systems. Preprint at [arXiv:1907.01166](https://arxiv.org/abs/1907.01166)
9. Tripathi S, Beigi H (2018) Multi-modal emotion recognition on iemocap dataset using deep learning. Preprint at [arXiv:1804.05788](https://arxiv.org/abs/1804.05788)
10. Cambria E, Hazarika D, Poria S, Hussain A, Subramanyam R (2017) Benchmarking multimodal sentiment analysis. In: International conference on computational linguistics and intelligent text processing. Springer, pp 166–179
11. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E (2019) Dialoguernn: An attentive RNN for emotion detection in conversations. *Proc AAAI Conf Artif Intell* 33:6818–6825
12. Williams J, Comanescu R, Radu O, Tian L (2018) Dnn multimodal fusion techniques for predicting video sentiment. In: Proceedings of grand challenge and workshop on human multimodal language (Challenge-HML), pp 64–72
13. Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency LP (2017) Context-dependent sentiment analysis in user-generated videos. In: Proceedings of the 55th annual meeting of the association for computational linguistics, vol 1 (Long Papers), pp 873–883

14. Majumder N, Hazarika D, Gelbukh A, Cambria E, Poria S (2018) Multimodal sentiment analysis using hierarchical fusion with context modeling. *Knowl Based Syst* 161:124–133
15. Chuang ZJ, Wu CH (2004) Multi-modal emotion recognition from speech and text. In: International journal of computational linguistics & chinese language processing, vol 9, Number 2, August 2004: Special Issue on New Trends of Speech and Language Processing, pp 45–62

A Machine Learning Based Approach for Prediction of Actual Landing Time of Scheduled Flights



S. Deepudev, P. Palanisamy, Varun P. Gopi, and Manjunath K. Nelli

Abstract In India, the Central Air traffic Flow Management system (C-ATFM), New Delhi has introduced the Ground Delay (GD) program to manage the airport capacity constraints of three metro airports (New Delhi, Mumbai and Bengaluru). The predictability of arrival(landing time) has a direct impact on the GD program, tactical air traffic control (especially Conflict Detection, Flight level allocation) and Sector capacity. The predictive analysis study used in this paper uses a range of statistical techniques from supervised machine learning and data mining technique on historical data to make predictions for landing time from departure information. The predictive modelling developed using Multi Linear Regression (MLR) model proposed in this study can lead to an accurate prediction of the actual landing time at the time of departure using minimum attributes. Exponential moving average of historical flying time traces non-stationary flight time variation. The proposed MLR model gives lesser Root Mean Square Error (RMSE) for predicting landing time comparing to Estimated Landing Time (ELDT) prediction by existing ATM (Air Traffic Management) automation system. In addition to the highlighted significant factors, the study gives an insight into the root cause for the early arrival of the aircraft and congestion at the capacity-constrained airport, which is one of the perpetual problems of Indian air traffic management. This analysis uses data collected by C-ATFM Delhi, India, during the period of October–November 2018.

S. Deepudev (✉) · P. Palanisamy · V. P. Gopi

Department of Electronics and Communication, National Institute of Technology Tiruchirappally, Tiruchirappally, Tamil Nadu, India

e-mail: deepudevs@aai.aero

P. Palanisamy

e-mail: palan@nitt.edu

V. P. Gopi

e-mail: varun@nitt.edu

M. K. Nelli

Integrated Planning Group-ANS Airports Authority of India, New Delhi, India

e-mail: mknelli@aai.aero

keywords Machine learning · Landing time · Prediction · Exponential moving average · Regression

1 Introduction

Scheduled airlines have become the backbone of the worldwide transportation system, bringing significant socio-economic utility by enabling cheaper and more comfortable mode of travel. After the introduction of the Regional Connectivity Scheme (RCS) in India by MoCA (Ministry of Civil Aviation) and the development of various airports by the Airports Authority of India, the domestic airline movement increased tremendously in recent years. An increase in air traffic will increase the revenue and at the same time poses challenges of efficient tactical air traffic management and capacity constraints which has become a significant hurdle for the Air Navigation Service Provider (ANSP) and Airport Operator (AO). Presently, three of the major metro airports (Delhi, Mumbai and Bengaluru) are facing capacity constraints during the peak hours of the traffic. Airports Authority of India has implemented Central Air Traffic Flow Management (C-ATFM) system for the efficient use of airports, airspace and support tactical air traffic service. Scheduled airlines commit to their customers and service providers a service that is economical, safe, predictable, dependable with reliable trip times and delays managed within acceptable limits.

Flight Delays from schedule time has become a common and complex phenomenon; it occurs due to the problems at the origin airport due to inbound aircraft delay, turn around delay, air delays due to different factors like en route weather, traffic congestion at destination airport due to airport capacity (limitations in the runway, taxiway, parking position, etc.) or a combination of these different factors. Delay may also due to specific airline operational procedures, which may vary during hours of a day due to available infrastructure and manpower [1]. Flight schedule delay is complicated to model, but it is still measurable with decent accuracy and can be predicted on detailed analysis. Several works in the literature focus on optimising various phases of the flight delay. Those works mainly focuses on the delay factors and their dependencies on turnaround time prediction, taxi time optimisation and air delay predictions. Different algorithms were proposed to optimise aircraft taxi movements on the ground by reducing aircraft taxi-times which include Ant algorithm [2] and Genetic algorithm [3]. Based on an evaluation of all the traffic in the airport, Aircraft start their push back process from the gate within a given time-slot, to minimise taxi-times [3]. Qin et al. [4] studied the periodicity of flight delay rate and pointed out the influence of time factor for flight delays by analysing changes in the delay rate with different models. European airport capacities and the correlation between levels of delays were computed by Reynolds et al. [5]. They suggested different approaches to deal with the air traffic congestion problem with merits and demerits.

Different prospective of flight delay analysis was done in [6–10] to find the correlations (regression approach) and thereby understand the principal causes of

delays from historical data. These works were mainly focused on causes of delay and optimising the same from airline point of view or passenger point of view for different airports. Developments in Trajectory-based operations and artificial intelligence (machine learning) were turning points for a more accurate prediction of arrival time. Several attempts [11, 12] were there made for modelling air traffic flow and delay patterns using historical data. Rebello and Balakrishnan [13] have used machine learning techniques to predict network-related delays of the future by utilising the system-level dependencies among airports. Machine learning was also used for the occurrence prediction of the ground delay and possibility of the on-time flight about the meteorological conditions [14, 15]. Kim et al. [16] studied prediction tasks of air traffic delay and its effectiveness by using the deep learning models. They analysed the patterns in air traffic delays by combining multiple models based on the deep learning paradigm. Deep architectures to categorise flight delay or no delay using neural nets were studied by Venkatesh et al. [17]. A statistical evaluation of the operational efficiency of scheduled flights defined by average differences of fuel consumption, flight time, and flight distance between the original and the optimised flight of domestic flights in Japan was done by Harada et al. [18].

Regression analysis and machine learning model is used to predict the arrival schedule variation. The predictability of change in scheduled arrival time and prediction of landing time is examined using Multi Linear Regression (MLR) model. The objectives of this paper include analysis of the scheduled departure time variation of domestic scheduled flights for the trial period and develop a better predictive model to predict actual landing time using departure information. In Sect. 2, problem definition, data description and data source are discussed. Data analysis and predictive modelling techniques using multiple linear regressions are analysed in Sect. 3. Test data results and prediction analysis presented in Sect. 4. This paper is concluded in Sect. 5.

2 Problem Definition

Air traffic congestion and air delay are two significant issues faced by the capacity-constrained airports. Growth of scheduled flights and increased number of passengers per year will lead to an increase in air traffic congestion. Currently, inadequate information about flight profile and schedule reliability prevents us from fulfilling the future requirement for accurate in 4D (3 Dimensional space with time) trajectory prediction. Novel accurate methods are to be tested and used for predicting the aircraft position with respect to time until the necessary information from aircraft is received in real-time. Based on the problem which we present, we propose an alternative method based on MLR models with minimal attributes to predict aircraft landing time using machine learning from historical data of previous flight operations. This can be used for prediction of arrival (Gate-In) time.

2.1 Data Set and Features

In order to train and test our models, data collected at C-ATFM Delhi during the trial period was used. The list of variables and definitions [19] are listed below.

T_{SO} : The time that an aircraft is scheduled to depart from the parking position

T_{AO} : The time the aircraft pushes back/vacates parking position.

T_{AT} : The time that an aircraft takes off from the runway. T_{AL} : Actual time an aircraft lands on a runway.

T_{AL} : Actual time an aircraft lands on a runway.

T_{EL} : The estimated time that an aircraft will touch-down on the runway (equivalent to ETA).

T_{SI} : The Time that an aircraft is scheduled to arrive at its first parking position.

T_{AI} : The time that an aircraft arrives in-blocks.

A brief of basic aggregate statistics of schedule time variation of this exercise is shown in Table 1. All the departures schedule variation and arrivals schedule variation has been classified into three categories based on the time variation. Flight departing within 5 to +10 min of scheduled departure and 5 to +10 for scheduled arrival is categorised as ‘On time’. Departure taking place 6 or less minutes and arrival reaching 6 min or earlier is classified as ‘Early’. The departure happening 11 min or later than scheduled departure time and Arrival reaching 11 min or greater than scheduled time is classified as ‘Delay’. One of the objectives of the trial was to find the effect of early departure to the capacity-constrained airports. Our study was limited for this paper to one of the capacity-constrained airports among three, which is Delhi International airport. Due to the restrictions like a bay shortage, operational constraints at departure airports, there were early departures to New Delhi (VIDP) during the trial period. However, all the early push back have not resulted in early arrival. On-time and delayed push back also contributed to early arrival. So a detailed analysis of predictability of actual landing time from historical data using multilinear regression analysis was conducted.

Table 1 Schedule time variation

Gate out variation	Gate-in variation			
	Early	On-time	Delay	Grand total
Early	6716	1943	556	9215
On-time	12,261	11,173	4964	28,398
Delay	544	1634	8106	10,284
Grand total	8981	28,361	10,555	47,897

2.2 Multi Linear Regression (MLR) Model

In general regression equation can be written as

$$Y_n = f \{X_n\} + e_n \quad n = 1 - N$$

where N denotes the number of instance, X_n is input random variable consists of K random attributes $x_{n1}, x_{n2}, x_{n3}, \dots, x_{nK}$. The e_n is the error coefficient of the model. Assume that the function f as linear and Multi linear regression equation can be written at time n as

$$Y_n = \beta_0 + X_{1n}\beta_1 + X_{2n}\beta_2 + \dots + X_{Kn}\beta_K + e_n \quad (1)$$

The parameters $\beta_1, \beta_2, \dots, \beta_k$ are the regression coefficients associated with X_1, X_2, \dots, X_K , respectively, β_0 is the y -intercept (constant term) and e_n is the model error, component reflecting the difference between the observed and fitted linear relationship.

Assume that e_n is independent and identically distributed (i.i.d.) random variables, such that the mean $E(e_n) = 0$ and $\text{var}(e_n) = s^2$.

The Multi Linear Regression is a parametric technique and the following assumption are made while building the model

1. The dependent variable and the error terms must possess a normal distribution.
2. The error terms (e_n) must possess constant variance. Absence of constant variance leads to heteroscedasticity.
3. There exists a linear and additive relationship between a dependent (DV) and independent variables (IV).
4. No correlation exists between independent variables. If any, the presence of correlation in independent variables will lead to Multicollinearity.
5. The error terms must be uncorrelated, i.e. error at e_n must not indicate the at error at e_{n+1}

3 Prediction of Actual Landing Time (TAL) Using Multiple Linear Regression

Predictability of actual landing time at the time of departure is one of the key requirements in ATFM. Detailed analysis and trails were done for identifying the parameter which has more influence on the landing time prediction. The observation was comparing to taxi out and taxi in time the flying time delay having more effect on the actual landing time predictability. In order to improve the predictability of landing time, random variation in flying time has to be traced accurately. The flying time variation of the same scheduled flight on different days is plotted in Fig. 1.

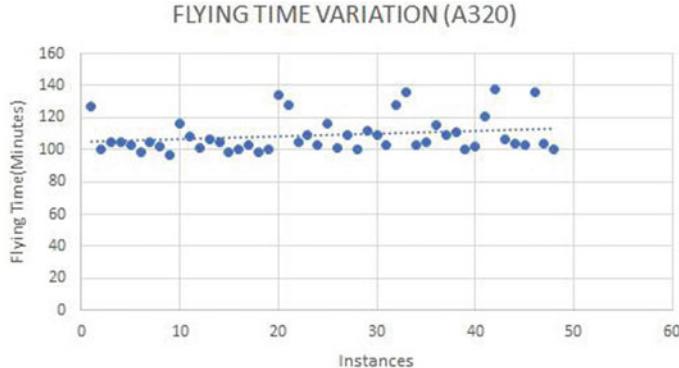


Fig. 1 Flying time variation

We can infer from the graph that the flying time is increasing and decreasing randomly by a small margin, such that the average remains almost constant. Many a times we are provided with a data set with different types of aircraft, which though varies by a small margin throughout its time period, but the average at each time period remains constant. In such a case we can predict the variation of the following day somewhere similar to the average of the past days or previous movements.

We analysed improving prediction accuracy by adding the moving average value of historical flying time on input attributes of the MLR model. Moving averages tend to smooth out short term irregularity in the data series based on an average of weighted observations. They are effective if the data series remains fairly steady over time. Further analysis was done on the data using Simple Moving average and Exponential Moving Average (EMA) of flying time. In Simple Moving average, since all the data points in the moving average process are given equal weight, this method fails to deal with non-stationary data. Exponential Moving weighted Average methods are the techniques that place more weights on the recent observations. Holt [20] proposed exponentially weighted moving averages(EMA) in dealing with forecasts of seasonal and trends. EMA's reaction directly depends on the pattern of the data. The exponentially weighted average of the forecast is an exponentially weighted (i.e. discounted) moving average with reducing factor $(1-\alpha)$:

$$FT_{(n)} = T_{AL(n)} - T_{AT(n)} \quad (2)$$

where $FT_{(n)}$ is the flying time of n^{th} flight movement and $T_{AL(n)}$, $T_{AT(n)}$ are corresponding Actual landing time and Actual Take off time data. Here the exponential moving average of the previous flying time used as an input attribute comparing to the previous model. It can be written as

$$\widehat{FT}_{(n)} = \alpha FT_{(n1)} + (1 - \alpha)FT_{(n2)} + (1 - \alpha)^2 FT_{(n3)} + (1 - \alpha)^3 FT_{(n4)} \quad (3)$$

where α denotes a ‘smoothing constant’ (α number between 0 and 1). Current flying time calculated by the sum of the exponentially weighted average of remaining historical value in the window. Here we have taken a window length of 5. Here all the flights departed from Mumbai to Delhi are considered. The flights include different airlines, different time and different type of aircraft. The flights are grouped according to the type of aircraft and the Exponential Moving average of flying time for each group is calculated. The data (2655) was randomly split into 80% (2124) training data 20% (531) test data. The regression model for T_{AL} is given in Eq. 3

$$T_{AL} = -24.5409 + 0.9988 * T_{AT} + 1.2442 * EMA \quad (4)$$

The performance matrix of the proposed model for training data shows excellent regression statics with Rsquare = 0.9998 and Adjusted Rsquare = 0.9998. The ‘p’ values of both the independent variable are less than 0.05 and approximately zero which indicates that Null hypothesis not valid and these variables dependent on T_{AL} . The RMSE for the training data 4.6 The Mean Absolute Error(MAE) is 3.5.

4 Results and Analysis

The proposed model in the above section was analysed using test data. The prediction accuracy, RMSE on test data is 4.8 and MAE is 3.5. Thus, the model gives excellent prediction accuracy.

4.1 Residual Analysis

Residual analysis was conducted for assessing the effectiveness of the proposed model. Once the assumptions are taken in Sect. 2.2 get violated, regression makes biased and erratic predictions. In order to check performance metrics further, the following test was done. The observed value of the dependent variable n^{th} flight is denoted by T_{ALn} (Actual Landing Time) and the predicted value is denoted by (T_{pALn}) . The residual e_2 can be calculated by Eq. (12)

$$e_1 = T_{ALn} - \hat{T}_{pALn} \quad (5)$$

Residuals versus Fitted plot (Fig. 2) is a graph which shows the presence of any nonlinear patterns in the data as well as in residuals. The values are centred at 0 and the red line indicates the trend of the residuals. It is also observed that all the values are scattered randomly along the 0^0 line. An Optimum Least Square (OLS) model was built with mathematical assumptions that a line can fit the data. Hence

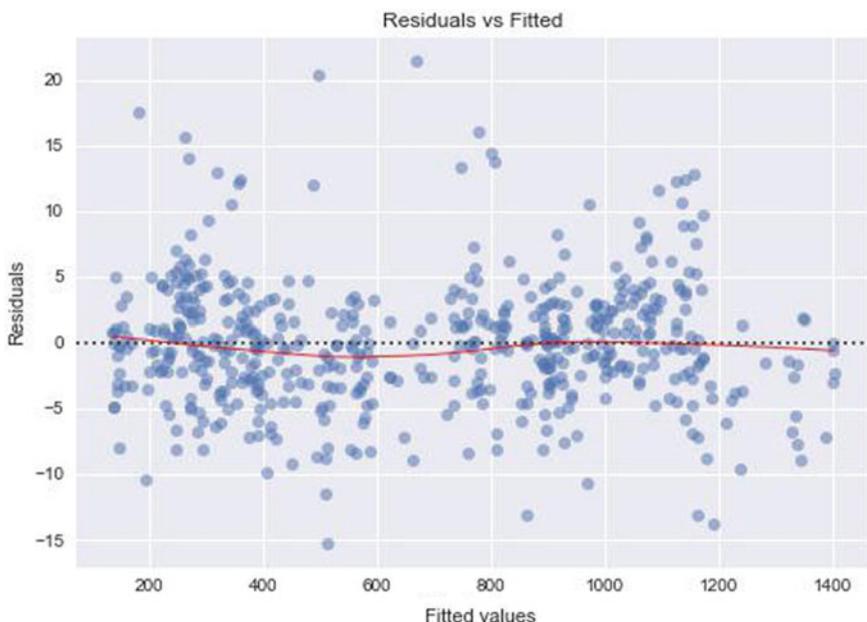


Fig. 2 Residual versus fitted values

the assumption no: 3 holds and the data can be fit by a linear model and a relatively flat line which indicates that the data is almost fitted to a linear model.

4.2 Normal Q-Q Plot

This plot indicates whether the residuals are normally distributed or not. The *Q-Q* plot follows the 45° line $y = x$, when the two distributions compared are identical. The red line indicates $y = x$. If the residuals lie on or very close to the red line, it indicates a good normal *Q-Q* plot. In Fig. 3, some points were located slightly away from the red line, which indicates that the errors are not absolutely normally distributed, and this occurred at the tails of the distribution. Table 2 shows sample results obtained on test data. In the table, D. D denotes schedule departure delay ($T_{AO} - T_{SO}$) and A.D denotes schedule arrival delay ($T_{AI} - T_{SI}$). The term S.E ($T_{AL} - T_{EL}$) denotes the difference between estimated landing time calculated by automation system and actual landing time. The estimated landing time (T_{EL}) is based on Terminal area boundary estimate. ABT denotes actual block time, i.e. difference between actual Inblock time and Actual Off Block time ($T_{AI} - T_{AO}$). EMA indicates extended moving average of previous movements. T_{AL} and E_{AL} denoted actual landing time predicted by proposed method and error in prediction, respectively. E_{AL} denotes system prediction error and is calculated by $T_{AL} - T_{pAL}$. From the samples in Table 2, it can be

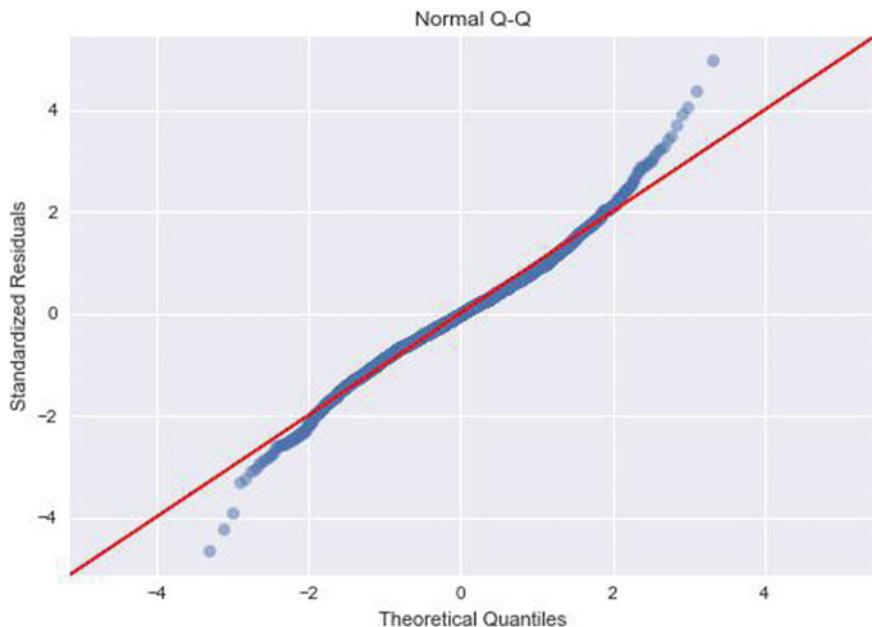


Fig. 3 Normal Q - Q plot

observed that even though considerable variation (121–142 minutes) occurs in the block time, the model was able to capture most of them. In the overall test results, the model predicted 75% the block time within -5 to $+5$ min of variation. The prediction accuracy variation for different types of aircraft is shown in Fig. 4.

It can be observed that different types of aircraft the model were able to predict the landing time very much accurately. The prediction result was compared with the estimated landing time predicted by the automation system. The RMSE and MAE for proposed method on test data are 4.8 and 3.5. The RMSE and MAE of the existing system predicted landing time were 10.1 and 6.4, respectively. It was observed that the large prediction error(greater than 15 minutes) occurred for the flights which took more than 30 min of average flying time.

5 Conclusion

By analysing the data of thousands of flights that operated during the trial period, one of the important findings was that the reason for traffic congestion at the constrained airport was due to a large flying time window(Block time) for the same type of aircraft. This large flying time variation is not accurately predicted by the existing ATM Automation system. Another aim of this study was to develop a model to predict flight Actual Landing Time (T_{AL}). The model is based on multiple linear

Table 2 Examples of prediction using MLR model (using EMA flying time)

Date	A/L	ACT	T_{SO}	T_{AO}	T_{AT}	T_{AL}	$D:D$	$A:D$	$S:E$	ABT	EMA	T_{AL}	EAL
08 = 10 = 18	VTI	A32W	10:10	10:10	10:30	12:14	12:20	0	10	2	130	100:9	12:10
26 = 11 = 18	IGO	A320	15:30	16:06	16:25	18:15	18:22	36	37	5	136	106:8	18:12
30 = 11 = 18	JAI	B739	07:30	07:50	08:07	09:41	09:51	20	6	9	121	96:5	09:41
14 = 11 = 18	JAI	B739	10:30	10:32	10:49	12:47	12:54	2	1	22	142	109:5	12:39
03 = 11 = 18	VTI	A320	13:00	13:04	13:25	15:05	15:11	4	1	0	127	100:8	15:04
26 = 11 = 18	SEJ	B738	00:35	00:47	01:07	02:48	02:57	12	12	3	130	101:0	02:48
08 = 11 = 18	GOW	A320	04:25	04:17	04:45	06:34	06:38	8	7	0	141	105:4	06 :31
30 = 11 = 18	IGO	A320	20:30	20:19	20:36	22:17	22:21	11	24	0	122	103:7	22 :19
15 = 10 = 18	IGO	A320	20:30	20:26	20:53	22:26	22:31	4	14	5	125	94:6	22 :24
05 = 10 = 18	AIC	A32W	10:30	11:27	11:43	13:27	13:33	57	43	1	126	103:3	13 :26

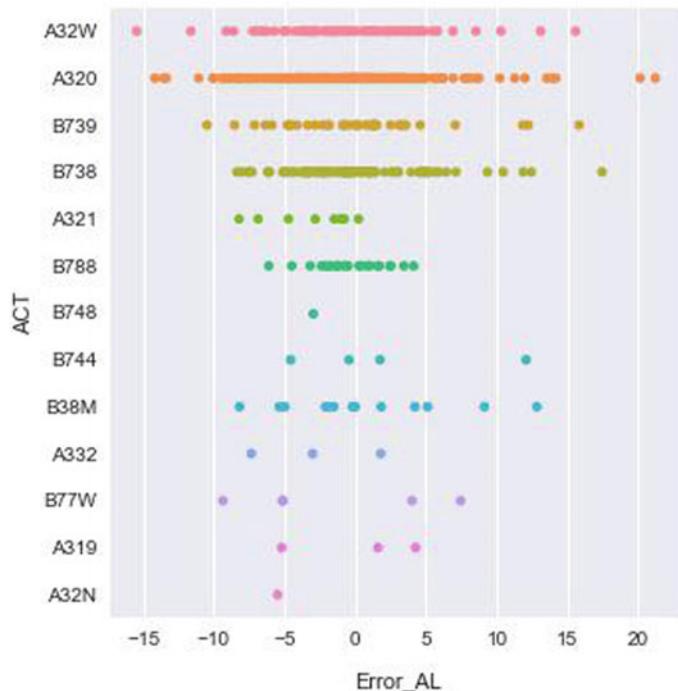


Fig. 4 Actual landing time prediction error distribution

regressions which uses a limited number of attributes to actual landing time based on historical data. By introducing an exponential moving average of historical flying time in MLR attribute in the machine learning process is able to trace the variation flying time more accurately, which in turn improved prediction accuracy. The model gives better prediction in arrival time for the flights (between Mumbai and Delhi) than the existing prediction used for air traffic flow Management in India. Using the proposed model different types of aircraft can be predicted using the same model and which predicted landing time based on departure information. The residual analysis shows some non-linearities in the tails. The future scope of this study involves various approaches that can be used to further analyse the data and using linear and nonlinear models for predicting landing of multiple departure and destination.

References

1. Small KA (1982) The scheduling of consumer activities: work trips. Am Econ Rev JSTOR 72(3):467–479
2. Nogueira KB, Aguiar PH, Weigang L (2014) Using ant algorithm to arrange taxiway sequencing in airport. Int J Comput Theory Eng 6(4):357

3. Jiang Y, Xu X, Zhang H, Luo Y (2015) Taxiing route scheduling between taxiway and runway in hub airport. *Mathematical Problems in Engineering*, Hindawi
4. Qin QL, and Yu H (2014) A statistical analysis on the periodicity of flight delay rate of the airports in the US. *Adv Transp Stud* 3:93–104
5. Reynolds-Feighan AJ, Button KJ (1999) An assessment of the capacity and congestion levels at European airports. *J Air Transp Manag* 5(3):113–134
6. Abdel-Aty M, Lee C, Bai Y, Li X, Michalak M (2007) Detecting periodic patterns of arrival delay. *J Air Transp Manag* 13(6):355–361
7. Rong F, Qianya L, Bo H, Jing Z, Dongdong Y (2015) The prediction of flight delays based the analysis of random flight points. In: IEEE 34th Chinese Control Conference (CCC), pp 3992–3997
8. Ryerson MS, Hansen M, Bonn J (2014) Time to burn: flight delay, terminal efficiency, and fuel consumption in the national airspace system. *Transp Res Part A: Policy Pract* 69:286–298
9. Sternberg A, Carvalho D, Murta L, Soares J (2016) Ogasawara and Eduardo: an analysis of Brazilian flight delays based on frequent patterns. *Transp Res Part E Logist Transp Rev* 95:282–298
10. Antonio AS, Juan AA, Calvet L, Guimaraes D others (2017) Using simulation to estimate critical paths and survival functions in aircraft turnaround processes. In: IEEE press proceedings of the 2017 winter simulation conference, pp 278
11. Zonglei L, Jiandong W, Guansheng Z (2008) A new method to alarm large scale of flights delay based on machine learning. In: IEEE international symposium on knowledge acquisition and modeling, pp 589–592
12. Sternberg A, Soares J, Carvalho D, Ogasawara E (2017) A review on flight delay prediction. Preprint at [arXiv:1703.06118](https://arxiv.org/abs/1703.06118)
13. Rebollo JJ, Balakrishnan H (2014) Characterization and prediction of air traffic delays. *Transp Res Part C: Emerg Technol* 44:231–241
14. Mukherjee A, Grabbe SR, Sridhar B (2014) Predicting ground delay program at an airport based on meteorological conditions. In: 14th AIAA aviation technology, integration, and operations conference, pp 2713
15. Choi S, Kim YJ, Briceno S, Mavris D (2016) Prediction of weather-induced airline delays based on machine learning algorithms. In: IEEE/AIAA 35th digital avionics systems conference (DASC), pp 1–6
16. Kim YJ, Choi S, Briceno S, Mavris D (2016) A deep learning approach to flight delay prediction. In: IEEE/AIAA 35th digital avionics systems conference (DASC), pp 1–6
17. Venkatesh V, Arya A, Agarwal P, Lakshmi S, Balana S-J (2017) Iterative machine and deep learning approach for aviation delay prediction. In: 4th IEEE Uttar Pradesh section international conference on electrical, computer and electronics (UPCON), pp 562–567
18. Harada A, Ezaki T, Wakayama T, Oka K (2018) Air traffic efficiency analysis of airliner scheduled flights using collaborative actions for renovation of air traffic systems open data. *Transp Res Part C: Emerg Technol* 2018
19. ICAO (2015) The fifth meeting of ICAO Asia/Pacific air traffic flow management steering group ATFM/SG/5: ATFM Terminology And Communications. *Int Civ Aviat Organ*
20. Holt CC (2004) Forecasting seasonals and trends by exponentially weighted moving averages. *Int J Forecast* 20(1):5–10
21. Kuhn N, Jamadagni N (2017) Application of machine learning algorithms to predict flight arrival delays. cs229AUTUMN

Data Integrity and Security in Distributed Cloud Computing—A Review



Abdullatif Ghallab, Mohammed H. Saif, and Abdulqader Mohsen

Abstract Data storage of cloud services has increased rates of acceptance due to their flexibility and the concern of the security and confidentiality levels. Many of the integrity and security problems raised based on the differences between client and service provider for resolution of third-party auditor. This review paper gives a brief view of current data integrity and security issues in the distributed cloud computing environment. The paper compared eight different models of the cloud data integrity and security. It highlights nearly solutions for some of the current cloud security risks and challenges by summarizing the key schemes of the privacy-preserving public auditing, particularly access control, attribute-based access control, and public key encryption. Moreover, the paper assigning the existing models, algorithms, and methodologies of data integrity and security had done in the literature of distributed cloud security. It suggested further research in cloud security domain regarding many of the security and data integrity issues.

Keywords Cloud security · Distributed cloud · Data integrity · Auditing schemes · Privacy features

1 Introduction

The rapid development in networking technology and the variation of computing resources requirements have enforced companies to outsource their need of storage and computing services. In the new economic, cloud computing encompasses

A. Ghallab (✉) · A. Mohsen
University of Science and Technology, Sana'a, Yemen
e-mail: ghallab@ust.edu

A. Mohsen
e-mail: a.alabadi@ust.edu

M. H. Saif
University of Science and Technology, Taiz, Yemen
e-mail: m.naji@ust.edu

different kinds of services. With the infrastructure as a service (IaaS) mode, clients use computing services from a provider through Internet. They are in charge of storage as well as for the networking infrastructure. In the platform as a service (PaaS) mode, clients use the resources of the provider for running their custom apps, whereas in the software as a service (SaaS) clients use the software, which runs on the infrastructure of the providers.

Cloud infrastructures may fall in the private group or in the public group. A private cloud refers to one in which the customer does management of the infrastructure. The customer is the owner. At the same time, the location is on premise. It also implies that the client is capable of controlling access to data. Access can be given to the people who are trusted. When it comes to public cloud, the company providing the service is the one, which owns and manages the infrastructure. The location of the infrastructure is on premise of the company providing the service. Generally, a different party is managing client information, whereas untrusted parties are capable of gaining access to the data.

Storage services like Microsoft's Azure and Amazon's S3 offer clients through storage and can be scaled dynamically. Moving the data of clients into the cloud required different kinds of cost with ensuring a proper maintenance of their private storage infrastructure. This made them to resort to other service providers at a fee to meet their storage needs. For a number of the customers, it generates numerous advantages, which generally include the fact that they are readily available. The clients are capable of accessing the data at any time from any point. The other major advantage is the fact that they are highly reliable. This generally implies that the customers do not have to be worried concerning anything like backups.

2 Review of the Literature

A model proposed by [1] is called “provable data possession (PDP).” The key feature of this model is verification of data on an untrusted server without retrieving. This is done by generating probabilistic proofs of blocks and maintaining metadata of the proof; also, the response protocol is small and constant reducing network communication. The PDP model with two schemes can support huge data in distributed systems with lower overheads at server level and the performance being dependent on disk I/O. In [2], a proposed model emphasizes on third-party auditing to enable customers assessing risks and the associated insurance risk mitigation. The focus was on both internal and external auditing of storage service offered online. This model is aimed at enabling customers to make informed choices give the service providers, and auditors to develop approaches for auditing and overcome the challenges.

In [3] improvement of PDP scheme of [1] as “dynamic provable data possession (DPDP).” The PDP scheme works with only the static files the DPDP scheme by the usage of rank information also supports the updates (apart from static) to the data stored on CSP. The scheme used “Merkle hash tree (MHT)” for verification but works only for single file copy and is not encrypted. By adopting protocol, a model

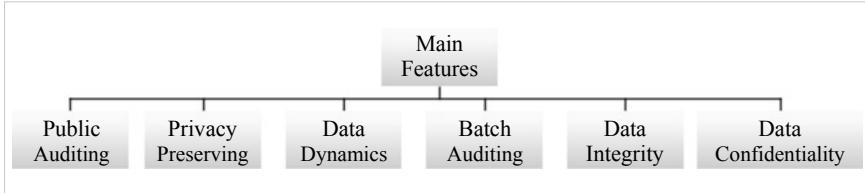


Fig. 1 Main features of the privacy preserving and public auditing

is proposed by [4] using PDP scheme. The key feature of this scheme is public verifiability for the data stored on CSP. The public verifiability is done by a TPA without exposing the information of the data owner. This model does not consider data encryption and is limited to single data files. A PDP scheme is proposed by [5]. The key aspect of this model was the usage of FHE algorithm for data file encryption. The benefit of this feature is, it generated multiple copies on the CSP, and whenever the file copies were updated, they did not require re-encryption. Two PDP schemes were proposed in this paper that enabled CSP to store lesser and fewer copies of files and adopt the dynamic behavior on cloud servers for data copies modifications. The model provided a solution to reduce storage costs and storage space requirements.

The key features of the privacy preserving and public auditing which summarized from the literature are shown in Fig. 1.

In [6] proposed a “proofs of retrievability (POR)” technique based on two schemes. The first is modeled with shortest query and response with public verifiability built from BLS signatures. The second is modeled on shortest response with private verifiability built on pseudorandom functions (PRFs). The model enables the client’s data to any prover passing the authentication check. In [1] proposed PDP technique for integrity of storage. The key focus is a third-party auditor (TPA) acting for a client to verify the data on the cloud dynamically. Cloud is not only limited to data backup but also involves block modification, insertion, and deletion. The model is improvised on Merkle hash tree (MHT). The model works for block authentication and dynamic public verifiability. The [7] as compared to [6] introduced a model of compact “proofs of retrievability (POR)” scheme. The TPA’s role has been better focused by eliminating the need for local copy of data, thereby reducing the role of cloud user and removing vulnerabilities of user data privacy. The model integrates homomorphic authenticator along random masking. The feature of simultaneous multiple auditing is also added with a multi-user setting. Public verification schemes have been proposed in [8, 9] based on [7] model with a homomorphic signature, a trusted TPA, and certificate-based cryptography.

The work [9] with a solution of POR for dynamic storage focused on a scheme against malicious auditor by the technique of oblivious RAM. In this model, the client is enabled to execute arbitrary reads/writes and audits in their data with a protocol on the server to check/ensure the latest version. In [10], a dynamic POR scheme ensuring client storage is proposed in a cost-effective way as compared to Merkle hash tree (MHT). The model outperformed two dynamic POR schemes, namely

(ACSAC 2012) and (EUROCRYPT 2013). Further, [11] proposed a PDP scheme to support the dynamic authentication. Further, the dynamic authentication schemes have been proposed in [12–14].

In [15] introduced a protocol by utilizing the services of a TPA. The method called as privacy-preserving auditing employs homomorphic linear authenticator (HLA) and random masking. The shortcomings in this model noticed are message attacks and external attacks. To overcome the shortcomings, in [15, 16] introduced an improvised scheme built from “Boneh–Lynn–Shacham signature (BLS).” Though improvised the scheme is not as efficient due to computationally intensive pairing operation. The scheme was implemented on Amazon EC2 but not tested on commercial public cloud making it unsuitable for handling large-scale data.

In [17] introduced a protocol with “Merkle hash tree (MHT)” along with BLS-based HLA. The model works for data dynamics and public auditing. While the model ensured integrity of data, it lacked in ensuring confidentiality of cloud stored data. In [18] introduced a design to get the intended blocks from various servers. The model used homomorphic token pre-computation and subsequently coded technique for erasure. In [19] introduced a design that collects signatures on blocks as a bundle. While the security aspect in the model is similar to [16] and ensured better efficiency as compared, there was an increase of overhead in communication and computation. In [20] developed a model using Merkle hash tree algorithm for TPA of the user’s data. While the data dynamics were supported, it lacked in ensuring confidentiality of cloud stored data.

In [21] proposed a model of MHT and RSA-based cryptography. The model ensured both integrity and privacy of data. In [22] proposed a different model to monitor the data changes on cloud. They placed an attacking module a code on the cloud server that performs the function of monitoring on cloud server while the confidentiality is ensured by employing AES algorithm. In [23] introduced a method of “Hash Message Authentication Code (HMAC)” along with homomorphic tokens.

By using a secret key, the integrity of data shared between two entities is ensured. The shortcoming in this model is fraud messages created by malicious attackers if the secret key is compromised. In [24, 25] introduced a model using a TPA in a privacy-preserving public auditing scheme. The user creates the blocks by AES algorithm, assigning hash, sequencing the hashes, and generating RSA signature. The TPA ensures the verification of data integrity by signature matching.

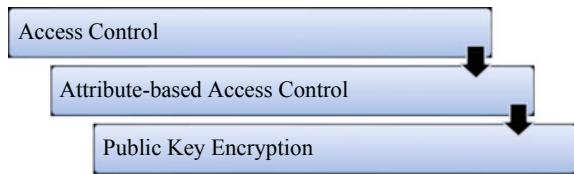
A comparison of the privacy-preserving and public auditing models and schemes based on key features of the privacy preserving and public auditing are represented in Table 1.

Among the eight models included in the comparison, two models satisfy five among six features, HLA with random masking and AES algorithm [22], HLA with BLS signature along with MHT [17], all features except data confidentiality and data dynamic, respectively. The next rank of four models achieved four features. All the four models satisfied the same three features, public auditing, privacy preserving, and data integrity, whereas two of them satisfied data dynamic. A different single one feature only, batch auditing and data confidentiality, was satisfied by HLA with BLS signature [16] and MHT and RSA algorithm [21], respectively. Finally, both

Table 1 Comparison of the privacy-preserving and public auditing schemes

Models	HLA with random masking [15]	HLA with BLS signature [16]	HLA with BLS signature along with MHT [17]	Homomorphic tokens with erasure code [18]	Merkle hash tree [20]	MHT and RSA algorithm [21]	HLA with random masking and AES algorithm [22]	HMAC algorithm [23]
Features								
Public auditing	✓	✓	✓	✓	✓	✓	✓	✓
Privacy preserving	✓	✓	✓	✓	✓	✓	✓	✓
Data dynamics					✓	✓		
Batch auditing					✓	✓	✓	
Data integrity	✓	✓	✓	✓	✓	✓	✓	✓
Data confidentiality					✓	✓	✓	

Fig. 2 Key schemes of the privacy preserving and public auditing



HLA with random masking [15] and HMAC algorithm [23] models satisfied the lowest rank with similar three features: public auditing, privacy preserving, and data integrity.

Figure 2 illustrates three key schemes of the privacy-preserving public auditing. It is noted well that the majority of models for privacy-preserving public auditing, exposed in Table 1, depends on the three schemes: access control, attribute-based access control, and public key encryption. The following sections review the main features of each scheme.

2.1 Access Control

Access control is a key feature for trusted security in cloud storage services. This requirement has evinced research interest from the academia and the industry. In [26], the researchers used a combination of three encryptions for cloud access security. In the model, the access rules were defined based on data characteristics, and the owner of the data can assign tasks in cloud servers without opening the actual content. In [27], as compared to “hierarchical attribute-based encryption (HABE),” a low communication and computing cost attribute-based system was developed. The data access control is through user attribute rules, and authentication is through identity-based signature.

In [28], broadcast encryption approach was adopted with a focus on smaller enterprises that are constrained by tight budgets and aided in cost savings by productivity enhancements. The model uses a combination of “hierarchical identity-based encryption (HIBE)” system and the “ciphertext-policy attribute-based encryption (CP-ABE)” system. In [29], a model of proxy re-encryption was adopted with patient-centric framework. The model leveraged multi-authority “attribute-based encryption (ABE)” for patient health records. The model is also built with dynamic modification of user, attributes, and access policies.

In [30], a model of role-based encryption is built called as “hierarchical attribute-set-based encryption (HASBE).” This model overcomes the shortcomings of attribute-based encryption (ABE) like lack of flexibility or executing complex access policies. HASBE has the ability to employ multiple values for user access management.

In [31], a model of “multi-message ciphertext-policy attribute-based encryption (MCP-ABE)” is employed for sharing consumer data attributes excluding the actual

names. The benefit of MCP-ABE is it enables the content provider to specify the access policy giving the data only to the intended and approved users.

In [32], a model of “ciphertext-policy attribute-based encryption (CP-ABE)” is employed for data sharing. The model is enabled to overcome the shortages of key escrow and fine-grained user revocation for each attribute.”

2.2 *Attribute-Based Access Control*

In [33], the model combined a method of ciphertext delegation enabling it to be “re-encrypted” and provides security in the standard ABE framework. This model enabled in dynamically disqualifying revoked users. In [34], a dynamic policy update is implemented for big data. The access policies in this model are designed for minimal computation for data owners, use of old data and access policies, algorithmic update of policies, and check mechanism for update of ciphertexts. In [35], the authors have proposed a scheme where “ciphertext-policy attribute-based proxy re-encryption (CP-ABPRE)” supports the attribute-based re-encryption. The model is built to overcome “chosen-ciphertext attack (CCA)” securely enabling the scheme to handle the problem.

In [36], a model “public key encryption (PKE)” is proposed to verify whether two ciphertexts are encryptions of an identical message. The scheme eliminates the need for bilinear map operations except for equality test. The applications where PKE is useful are searchable encryption and encrypted data partitioning. Similar kind of PKE schemes have been proposed in [37, 38].

In [39], a model of identity-based distributed provable data possession (ID-DPDP) is proposed. ID-DPDP protocol is developed on a multi-cloud storage and is secure under computational Diffie–Hellman problem (CDH). The model allows remote data checking without downloading the whole data and reduces the costs. This model is applied for patient records in public cloud under KP-ABE.

CP-ABE characteristics due to their flexibility are more preferred in applications of cloud access control. In [40], to overcome the problem of complicity of data storage, a “multiple-replica provable data possession (MR-PDP)” is proposed. The storage is done and authenticated by challenge-response protocol. The scheme is economical as compared to single-replica PDP scheme. In [41], a proof of retrievability (PoR) is proposed to reduce the computational load by outsourcing files on low-power client’s verifications on high-end servers and supporting dynamic updates. Performance analysis was also done for this scheme giving it an upper hand to the compared ones [42]. Also focused on CP-ABE application. In [43–45], studies were done on policy updates. The studies had their own shortcomings where they used proxy re-encryption. This does not really update or extend the access policy and lacks integrity of linking to the actual data.

2.3 Public Key Encryption

In [46], the author introduced a public key encryption scheme (PKE). In this scheme, the bilinear map operations are required only in the case of equality test of encrypted messages between two ciphertexts. The scheme is useful for encrypted applications like search or partition. The shortcoming observed in PKEET is the lack of integrity check. In [47], Tang introduced a model enabling two users with public/private key to issue token(s) for equality test between ciphertexts. The model incorporates fine-grained authorization policy. The model is useful for TPA operations. In [48], Tang improvised on [47] fine-grained authorization (FG-PKEET) by working on flaws on equality test, compare with AoN-PKEET by Tang and PKEET and make FG-PKEET function on a two-proxy setting.

In [49], Wang, keeping in view the aspects of verification, multi-cloud storage, and costs, proposed an identity-based distributed provable data possession (ID-DPDP) in multi-cloud storage. ID-DPDP protocol is made on bilinear pairings and is secure under standard CDH problem. The ID-DPDP protocol functions for private, delegated, or public verification. The shortcoming of this model is it does not work in multiple-replica settings. In [50], keeping in view the problem of server collusion and no evidence on storage of multiple copies of data proposed multiple-replica PDP. The scheme empowers the client to store replicas of files with a challenge-response protocol for verification. The MR-PDP scheme is better as compared to single-replica PDP scheme in computational aspect and can generate further replicas at lesser costs. The shortcoming of this scheme is it cannot perform public auditability.

In [51] used a model of indistinguishability obfuscation technique for remote data integrity auditing and reduced the computational burden of generation of signature for user. The model is useful in scenarios of outsourcing files by low-power client and verifications by cloud servers. In [52] proposed a model of protecting the privacy of the user to generate signatures by using third-party medium (TPM). The TPM is employed to develop a simple model for auditing integrity remotely. The TPM has an expiration time for authorization with a valid period. In [53–55] and in [56], Yu and Wang focused on reducing the damage of key exposure. They introduced remote auditing schemes which are key-exposure resilient and based on key update techniques in various scenarios.

Information sharing is an important aspect in cloud storage. In [57], keeping in view data sharing as a key aspect introduced a privacy-preserving approach that enables public auditing on cloud. The scheme focuses on modifying the ring signature for secured cloud storage. The scheme has the capacity to perform multiple auditing tasks at once. In [58] designed a public auditing scheme to store identity confidentiality for a group of members at once. The scheme uses blind signature technique for authentication.

In [59] proposed a privacy-based public auditing method. The scheme is modeled for shared cloud data by generating a homomorphic verifiable group signature. The model needs a minimum of t group managers avoiding single-authority abuse, and the users can track data changes in an assigned binary tree. In [60], keeping in

view the risk of modification and sharing of data for a revoked user introduced an approach tailored for the shortcoming. It is a public auditing mechanism to ensure integrity where the cloud server re-signs data blocks of the revoked user. The scheme supports multiple auditing tasks verification at once. In [61] designed a scheme supporting user revocation in shared data integrity auditing. This scheme is designed to avoid compromise keeping in view the complacency between revoked users and malicious cloud servers. The scheme is tailored on secret sharing and polynomial-based authentication tags.

In [62] introduced identity-based proxy-oriented data uploading with remote integrity in public cloud (ID-PUIC). The system and security model are defined, and ID-PUIC protocol is based on bilinear pairings. Further, the ID-PUIC protocol is secure on hardness of CDH. In [63] introduced identity-based remote data integrity checking (RDIC) protocol. The scheme uses homomorphic cryptography. It reduces the costs for the management of PKI modeled RDIC protocols. In [64], the author introduced incentive and unconditionally anonymous identity-based public PDP scheme. IAID-PDP system and security model are defined, and the protocol is based on bilinear pairings. IAID-PDP is secure and eliminates the certificate management. In [65] introduced a scheme of user revocation without affecting the blocks held by the revoked user. Instead of focusing on the verifiers of the revoked user, the model focused on updating the non-revoked group keys. The scheme is made on ID cryptography; it does not need certificate management as needed in public key infrastructure (PKI) systems. Further, many other aspects were focused on such as privacy-preserving authentication in [66] and data deduplication [67, 68] in remote data integrity auditing. Despite all these approaches, the remote data integrity approaches mentioned above cannot completely support data sharing with information hiding.

In [69], based on earlier works developed an improvised model of PDP data checking remotely. The model focused on reducing the I/O costs by random sampling of blocks from the server. The challenge/response protocol reduces network communication making the model lightweight and more suitable for distributed storage scenarios. The authors presented two PDP schemes better than previous approaches. The shortcoming noticed was the model is not suitable for public audit.

In [70], Merkle introduced protocols for public key systems. The paper focused on unique properties and protocols on public keys and digital signatures along with comparisons. In [71] focused on a paper of cryptographic cloud storage. The paper focused on developing a secure cloud on a public cloud. Various architectures are described at a higher level and the benefits that accrue for customers and service providers. In [72], the author introduced a lesser energy-consuming protocol in the integrity of storage services on mobile cloud. The model focused on reducing mobile energy consumption while supporting dynamic operations. The authors used the concepts of incremental cryptography and trusted computing.

In [73] proposed a mechanism of message authentication code (MAC) for two parties communicating across an insecure channel. The model focused on authentication tag and shared key approach between two parties for data communication. In [74] introduced data access control for multi-authority cloud storage (DAC-MACS). The model is developed as new CP-ABE scheme. The key features are competent

decryption and feature revocation for forward and backward security. In [75] investigated on [74] and proposed that there is a security vulnerability in the model where a revoked user can decode new ciphertexts based on an attack method revoke.

In [76], the author proposed identity-based remote data possession checking (ID-RDPC) protocols. The protocol is secure assuming CDH. The key benefit of this approach is bypassing the process of certificate management. Further, the model performs better as compared to RDPC protocols in PKI framework on: computation, communication, and associated costs. In [77] constructed a protocol where they combined ID-based signature and public verification. This model enables the TPA to bypass the user task checking and focuses purely on integrity of data. In [78], the author proposed a cancelable identity-based encryption (IBE) model that reduces the various tasks related to key management by enabling key update on cloud. This is done by introducing outsourcing computation into IBE to handle identity revocation. The model reduces the operational tasks for “private key generator (PKG)” and users.

In [79] addressed the key management by proposing fuzzy identity-based auditing. In this model, a user identity is a set of descriptive characteristics built as a protocol through biometrics. The protocol has been proven on CDH and discrete logarithm. In [80], the author addressed the issues of verifying public key certificates and their management. The author proposed “identity-based cloud data integrity checking protocol (ID-CDIC).” The model is proposed to eliminate certificate administration in out-of-date cloud checking.

Integrity verification in cloud storage is a topic of interest in recent times for researchers. In [81] introduced the concept of checking of files for integrity. They based the model on challenge-response protocols, and the challenge is generated randomly. The main shortcoming of this model was it is unsuitable for large amount of data load for verification. This is improvised by [1] with a scheme for PDP with RSA signatures. The RSA signatures had a drawback of tags of 1024 bits increasing costs, and the scheme is incapable of privacy preserving if there is a TPA. Further, [7] used BLS signatures over [1] RSA signatures limiting the length to 160 bits with security. In addition, work [8] on privacy-preserving public auditing for cloud joining HLA and random masking. In [82] proposed a signature scheme on CDH assumption. The secure signature length is 50% of DSA signature and suitable in cases of human typing or communicated simple bandwidth. In [83] proposed a public auditing scheme based on hash table dynamic in nature (DHT). It is a 2D structure present at a third-party auditor (TPA). The scheme reduces computational and communication aspects by transferring the information from the CSP to the TPA. The scheme has a good updating efficiency, supports privacy preservation, and enables batch auditing through BLS signatures. Over this, [13] improvised the dynamic verification scheme with multiple owners.

In [84, 85] introduced a scheme with critical information hiding. It is a remote auditing scheme that uses a cleanser to mask critical information on the blocks while enabling remote integrity auditing. The scheme is based on ID cryptography. In [86], keeping in view multiple cloud service providers working in tandem proposed a cooperative PDP scheme. The scheme is based on homomorphic verifiable response and hash index, and the model proved to have lesser cost and overhead aspects in

comparison with non-cooperative approaches. This model resonates with [49] that focused on ID-DPDP. In [84] introduced ID multi-replica PDP (IDPMR-PDP). The scheme provides TPA with multiple replicas without PKI. The scheme is protected against malicious servers and attackers. In [87] proposed a scheme named MuR-DPA that is an authenticated data structure (ADS) based on MHT. The scheme enabled for the authentication of active datasets with multiple replicas on the cloud by including values in computation of MHT nodes in a top-down order as replica sub-tree. All the approaches for integrity verification are epoch-based auditing having time periods, and the attacks can be detected at the completion of each period. Also, the ambiguity in the real verifier being the user or third party, trusted, and authorized is also a concern.

Motivated by integrity audit shortcomings, some schemes comprising real-time assessment and fair mediation have been proposed. In [88] a scheme where checking is performed with each file used in operation. A data structure has been developed called FBH-tree that stores the hash values. A file in operation requests a part which is the hash value for real-time authentication. The scheme suffers from drawbacks in efficiency, and computation overhead is directly proportional to incremental FBH-tree. In [89] and in [90] proposed a similar kind of model where the motive to cheat is taken into account with party being either client or CSP. To overcome this, they introduce a third-party arbitrator based on signature exchange idea. The limitation to these models is that an exchange implies consent between parties with dispute resolution if arises is postponed to a later date.

For overcoming the problem of authentication, reversible watermarking is a novel technique on which few models are discussed below. In [91] investigated high capacity no loss data embedding for images where the authentic image can be restored from the watermarked image. They presented two techniques: (i) least significant bit prediction and Sweldens' lifting scheme and (ii) improvement of Tian's technique of difference expansions. They also compared the techniques with various other embedding methods. In [92] proposed changeable image masking approach over encrypted field. In this model, using an SVM classifier by decoder, the distinction is made between encrypted and non-encrypted image patches and gets the embedded message and original image. In [93] proposed reversible hiding scheme based on Shamir's sharing. The information is distributed in random shares with the embedded information key shared to the correct owner. Using the key, the data can be extracted either directly or by media authentication. In [94] proposed a changeable watermarking algorithm. In this method, the authentic image is embedded with digital meta-data with removal at a later time. The lossless recovery of original image enables a digital signature of image to be embedded in the image itself only to be recovered later for authentication. In [95] presented a reversible hiding algorithm. After extracting the data, the authentic image is secured without disruption from the marked image. The algorithm alters the pixel values to implant the information in a histogram shifting modulation.

In [96], Tian proposed a DE modulation-based algorithm. The difference expansion algorithm is capable to overcome overflow and underflow problems. This is achieved by calculation of the variance of adjacent pixel values to select some

for DE to embed watermark. In [97] improvised in [95] skewed histogram shifting where the model uses a set of extreme predictions. By this, the distortion problem is addressed in a better way by embedding the skewed structure histogram pixels from peak and short tail. In [98] used a method of lossless watermarking where parts of image are reversibly watermarked with message embedding by conventional Haar wavelet transform coefficients. The approach is one of the most competitive with high capacity and low distortion. In [99] experimented with identification of areas in an image considered most ideal for watermarking and embedded the area by histogram shifting.

In [100] introduced a prediction-error expansion (PEE) method. This model is derived from DE and histogram shifting (HS) approaches. The variance between the pixel and its estimate is used for data implanting. The models introduced also need to embed auxiliary information overhead.

In [101], Coltuc aimed at reduction in embedding distortion of prediction error. The method used here is not embedding the entire stretched difference but split the variance of current pixel and its calculation context. Testing is done on various changeable watermarking schemes. SGAP yielded the best results. In [102], pairwise prediction-error expansion (PEE). The sequence results in a 2D prediction-error histogram improve performance due to a better embedding approach. In [103] improvised on [102] and proposed a familiar kind of pixel pairing. In this approach, only pixels with similar prediction errors are paired and embedded, thereby decreasing the number of shifted pixels. In [104] introduced a segmented data-embedding method for efficient RDH. In this method, the host is not considered as a whole but partitioned into multiple sub-hosts. Each sub-host can have its own embedding enabling to apply varied RDH algorithms as an ensemble. The major shortcomings in all the schemes of reversible watermarking are to provide stable capacity and exposure of images to be checked.

3 Conclusion

This review paper explores the most ideas of data integrity and security problems in the distributed cloud computing environment along with some models, challenges, and limitations involved in this field. It presented many of the data security concepts on cloud servers such as schemes, protocols, algorithms, access policies, storage scenarios, access services, and a third-party auditor.

A comparison for eight models of data integrity and security models was done based on common six features, namely public auditing, privacy preserving, data dynamic, batch auditing, data integrity, and data confidentiality. Besides the two models, homomorphic tokens with erasure code and HMAC algorithm, several enhanced models of HLA and MHT combinations were compared too. HLA and MHT are used with a diversity of schemes and algorithms, HLA is used with

random masking, BLS signature, BLS signature along with MHT, and AES algorithm, whereas MHT is used individually and/or combined either with BLS signature or with RSA algorithm.

The three security features, public auditing, privacy preserving, and data integrity, were satisfied by all models. Both, data dynamics and batch auditing integrity, were satisfied by three models, whereas data confidentiality was satisfied in two models only. Two models of data integrity and security, HLA with random masking and AES algorithm and HLA with BLS signature along with MHT, satisfied (83%) of the required features. All features except data confidentiality and data dynamic. Four models satisfied with (66.6%), and two models satisfied with only (33%) of the features.

This review investigated a lack of data integrity and security models with certain required features like data confidentiality, data dynamics, and batch auditing integrity. The work on data confidentiality and data dynamics of the cloud security can be extended by adding these features in different models. So, more research can be done to improve HLA with random masking and HMAC algorithm models by adding more features of data integrity and security.

Studying possibilities of adding advanced security features for models, such as HLA with BLS signature, MHT and RSA algorithm, and HMAC algorithm, can be investigated in future research.

References

1. Ateniese G, Burns R, Curtmola R et al (2007) Provable data possession at untrusted stored. In: Proceedings of the 14th ACM conference on computer and communications security. ACM, New York, pp 598–609
2. Shah MA, Baker M, Mogul JC, Swaminathan R (2007) Auditing to keep online storage services honest. In: HOTOS'07: proceedings of the 11th USENIX workshop on hot topics in operating systems, Berkeley, CA, USA, pp 1–6
3. Erway C, Küpcü A, Papamanthou C, Tamassia R (2009) Dynamic provable data possession. In: Proceedings of 16th ACM conference on computer and communication security (CCS), New York, NY, USA, pp 213–222
4. Hao Z, Zhong S, Yu N (2011) A privacy-preserving remote data integrity checking protocol with data dynamics and public verifiability. IEEE Trans Knowl Data Eng 23(9):1432–1437
5. Barsoum AF, Hasan MA (2011) On verifying dynamic multiple data copies over cloud servers. In: Cryptology ePrint Archive, Report 2011/447. <http://eprint.iacr.org/>
6. Juels A, Kaliski BS Jr (2007) PORs: proofs of retrievability for large files. In: Proceedings of CCS. ACM, pp 583–597
7. Shacham H, Waters B (2008) Compact proofs of retrievability. In: Proceedings of ASIACRYPT. Springer, pp 90–107
8. Wang C, Chow SS, Wang Q, Ren K, Lou W (2013) Privacypreserving public auditing for secure cloud storage. IEEE Trans Comput 62(2):362–375
9. Zhang Y, Xu C, Yu S, Li H, Zhang X (2015) SCLPV: secure certificateless public verification for cloud-based cyber-physical-social systems against malicious auditors. IEEE Trans Comput Soc Syst 2(4):159–170

10. Sookhak M, Gani A, Talebian H, Akhunzada A, Khan SU, Buyya R, Zomaya AY (2015) Remote data auditing in cloud computing environments: a survey, taxonomy, and open issues. *ACM Comput Surv* 47(4):159–170
11. Ateniese G, Pietro RD, Mancini LV, Tsudik G (2008) Scalable and efficient provable data possession. In: Proceedings of SecureComm. ACM
12. Shi E, Stefanov E, Papamanthou C (2013) Practical dynamic proofs of retrievability. In: Proceedings of CCS. ACM, pp 325–336
13. Yang K, Jia X (2013) An efficient and secure dynamic auditing protocol for data storage in cloud computing. *IEEE Trans Parallel Distrib Syst* 24(9):1717–1726
14. Sookhak M, Gani A, Khan MK, Buyya R. Dynamic remote data auditing for securing big data storage in cloud computing (to appear). <https://doi.org/10.1016/j.ins.2015.09.004>
15. Wang C, Wang Q, Ren K, Lou W (2010) Privacy-preserving public auditing for data storage security in cloud computing. In: INFOCOM, 2010 proceedings IEEE. IEEE, pp 1–9
16. Wang C, Chow SSM, Wang Q, Ren K, Lou W. Privacy preserving public auditing for secure cloud storage. <http://eprint.iacr.org/2009/579.pdf>
17. Wang Q, Wang C, Ren K, Lou W, Li J (2011) Enabling public auditability and data dynamics for storage security in cloud computing. *IEEE Trans Parallel Distrib Syst* 22(5):847–859
18. Wang C, Wang Q, Ren K, Cao N, Lou W (2012) Toward secure and dependable storage services in cloud computing. *IEEE Trans Serv Comput* 5(2):220–232
19. Worku SG, Xu C, Zhao J, He X (2014) Secure and efficient privacy-preserving public auditing scheme for cloud storage. *Comput Electr Eng* 40(5):1703–1713
20. Meenakshi IK, George S (2014) Cloud server storage security using TPA. *Int J Adv Res Comput Sci Technol (IJARCST)*. ISSN: 2347-9817
21. Tejaswini KS, Prashanth SK (2013) Privacy preserving and public auditing service for data storage in cloud computing. *Indian J Res PARIPEX* 2(2)
22. Santosh J, Nandwalkar BR. Privacy preserving and batch auditing in secure cloud data storage using AES. In: Proceedings of 13th IRF international conference. ISBN: 978-93-84209-37-72014
23. Ezhil Arasu S, Gowri B, Ananthi S (2013) Privacy-preserving public auditing in cloud using HMAC algorithm. *Int J Recent Technol Eng (IJRTE)*. ISSN: 2277, 3878
24. Wang C, Wang Q, Ren K, Cao N, Lou W (2011) Towards secure and dependable storage services in cloud computing. *IEEE Trans Serv Comput* 5(2):220–232
25. Morea S, Chaudhari S (2016) Third party public auditing scheme for cloud storage. *Int J Procedia Comput Sci* 79:69–76
26. Berger S, Garion S, Moatti Y, Naor D, Pendarakis D, ShulmanPeleg A, Rao JR, Valdez E, Weinsberg Y (2016) Security intelligence for cloud management infrastructures. *IBM J Res Dev* 60(4):11:1–11:13
27. Secure access control for cloud storage. <https://www.research.ibm.com/haifa/projects/storage/cloudstorage/secureaccess.shtml>
28. Boneh D, Gentry C, Waters B (2005) Collusion resistant broadcast encryption with short ciphertexts and private keys. In: CRYPTO 2005. LNCS, vol 3621, pp 258–275
29. Ateniese G, Fu K, Green M, Hohenberger S (2006) Improved proxy re-encryption schemes with applications to secure distributed storage. *ACM Trans Inf Syst Secur* 9(1):1–30
30. Zhou L, Varadharajan V, Hitchens M (2013) Achieving secure rolebased access control on encrypted data in cloud storage. *IEEE Trans Inf Forensics Secur* 8(12):1947–1960
31. Goyal V, Pandey O, Sahai A, Waters B (2006) Attribute-based encryption for fine-grained access control of encrypted data. In: Proceedings of the 13th ACM conference on computer and communications security, CCS 2006, pp 89–98
32. Hu VC, Kuhn DR, Ferraiolo DF (2015) Attribute-based access control. *IEEE Comput* 48(2):85–88
33. Attrapadung N, Libert B, de Panafieu E (2011) Expressive key-policy attribute-based encryption with constant-size ciphertexts. In: PKC 2011. LNCS, vol 6571, pp 90–108
34. Bethencourt J, Sahai A, Waters B (2007) Ciphertext-policy attribute based encryption. In: 2007 IEEE symposium on security and privacy (S&P 2007), pp 321–334

35. Waters B (2011) Ciphertext-policy attribute-based encryption: an expressive, efficient, and provably secure realization. In: PKC 2011, LNCS, vol 6571, pp 53–70
36. Yu S, Wang C, Ren K, Lou W (2010) Achieving secure, scalable, and fine-grained data access control in cloud computing. INFOCOM 2010:534–542
37. Huang J, Chiang C, Liao I (2013) An efficient attribute-based encryption and access control scheme for cloud storage environment. In: Grid and pervasive computing GPC 2013, LNCS, vol 7861, pp 453–463
38. Wang G, Liu Q, Wu J (2010) Hierarchical attribute-based encryption for fine-grained access control in cloud storage services. In: Proceedings of the 17th ACM conference on computer and communications security, CCS 2010, pp 735–737
39. Li M, Yu S, Zheng Y, Ren K, Lou W (2013) Scalable and secure sharing of personal health records in cloud computing using attribute based encryption. IEEE Trans Parallel Distrib Syst 24(1):131–143
40. Wan Z, Liu J, Deng RH (2012) HASBE: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing. IEEE Trans Inf Forensics Secur 7(2):743–754
41. Wu Y, Wei Z, Deng RH (2013) Attribute-based access to scalable media in cloud-assisted content sharing networks. IEEE Trans Multimedia 15(4):778–788
42. Hur J (2013) Improving security and efficiency in attribute-based data sharing. IEEE Trans Knowl Data Eng 25(10):2271–2282
43. Sahai HS, Waters B (2012) Dynamic credentials and ciphertext delegation for attribute-based encryption. In: CRYPTO 2012. LNCS, vol 7417, pp 199–217
44. Yang K, Jia X, Ren K (2015) Secure and verifiable policy update outsourcing for big data access control in the cloud. IEEE Trans Parallel Distrib Syst 26(12):3461–3470
45. Liang K, Fang L, Wong DS, Susilo W (2015) A ciphertext-policy attribute-based proxy re-encryption scheme for data sharing in public clouds. Concurrency Comput Pract Experience 27(8):2004–2027
46. Yang G, Tan CH, Huang Q, Wong DS (2010) Probabilistic public key encryption with equality test. In: Topics in cryptology—CT-RSA 2010. LNCS, vol 5985, pp 119–131
47. Tang Q (2011) Towards public key encryption scheme supporting equality test with fine-grained authorization. In: Information security and privacy—16th Australasian conference, ACISP 2011. LNCS, vol 6812, pp 389–406
48. Tang Q (2012) Public key encryption schemes supporting equality test with authorisation of different granularity. IJACT 2(4):304–321
49. Wang H (2015) Identity-based distributed provable data possession in multicloud storage. IEEE Trans Serv Comput 8(2):328–340
50. Curtmola R, Khan O, Burns R et al (2008) MR-PDP: multiple-replica provable data possession. In: The international conference on distributed computing systems. IEEE Computer Society, pp 411–420
51. Guan C, Ren K, Zhang F, Kerschbaum F, Yu J (2015) Symmetric key based proofs of retrievability supporting public verification. In: Computer security—ESORICS. Springer, Cham, Switzerland, pp 203–223
52. Shen W, Yu J, Xia H, Zhang H, Lu X, Hao R (2017) Light-weight and privacy-preserving secure cloud auditing scheme for group users via the third party medium. J Netw Comput Appl 82:56–64
53. Yu J, Ren K, Wang C, Varadharajan V (2015) Enabling cloud storage auditing with key-exposure resistance. IEEE Trans Inf Forensics Secur 10(6):1167–1179
54. Yu J, Ren K, Wang C (2016) Enabling cloud storage auditing with verifiable outsourcing of key updates. IEEE Trans Inf Forensics Secur 11(6):1362–1375
55. Yu J, Wang H (2017) Strong key-exposure resilient auditing for secure cloud storage. IEEE Trans Inf Forensics Secur 12(8):1931–1940
56. Yu J, Hao R, Xia H, Zhang H, Cheng X, Kong F (2018) Intrusion resilient identity-based signatures: Concrete scheme in the standard model and generic construction. Inf Sci 442–443:158–172

57. Wang B, Li B, Li H (2012) Oruta: privacy-preserving public auditing for shared data in the cloud. In: Proceedings of IEEE 5th international conference on cloud computing (CLOUD), pp 295–302
58. Yang G, Yu J, Shen W, Su Q, Fu Z, Hao R (2016) Enabling public auditing for shared data in cloud storage supporting identity privacy and traceability. *J Syst Softw* 113:130–139
59. Fu A, Yu S, Zhang Y, Wang H, Huang C. NPP: a new privacy-aware public auditing scheme for cloud data sharing with group users. *IEEE Trans Big Data* (to be published). <https://doi.org/10.1109/tbdata.2017.2701347>
60. Wang B, Li B, Li H (2015) Panda: public auditing for shared data with efficient user revocation in the cloud. *IEEE Trans Serv Comput* 8(1):92–106
61. Luo Y, Xu M, Fu S, Wang D, Deng J (2015) Efficient integrity auditing for shared data in the cloud with secure user revocation. In: Proceedings of IEEE Trustcom/BigDataSE/ISPA, pp 434–442
62. Wang H, He D, Tang S (2016) Identity-based proxy-oriented data uploading and remote data integrity checking in public cloud. *IEEE Trans Inf Forensics Secur* 11(6):1165–1176
63. Yu Y et al (2017) Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. *IEEE Trans Inf Forensics Secur* 12(4):767–778
64. Wang H, He D, Yu J, Wang Z. Incentive and unconditionally anonymous identity-based public provable data possession. *IEEE Trans Serv Comput* (to be published) <https://doi.org/10.1109/tsc.2016.2633260>
65. Zhang Y, Yu J, Hao R, Wang C, Ren K. Enabling efficient user revocation in identity-based cloud storage auditing for shared big data. *IEEE Trans Depend Secure Comput* (to be published). <https://doi.org/10.1109/tdsc.2018.2829880>
66. Shen W, Yang G, Yu J, Zhang H, Kong F, Hao R (2017) Remote data possession checking with privacy-preserving authenticators for cloud storage. *Future Gener Comput Syst* 76:136–145
67. Li J, Li J, Xie D, Cai Z (2016) Secure auditing and deduplicating data in cloud. *IEEE Trans Comput* 65(8):2386–2396
68. Hur J, Koo D, Shin Y, Kang K (2016) Secure data deduplication with dynamic ownership management in cloud storage. *IEEE Trans Knowl Data Eng* 28(11):3113–3125
69. Ateniese G, Burns R, Curtmola R (2011) Remote data checking using provable data possession. *ACM Trans Inf Syst Secur* 14(1):12
70. Merkle RC (1980) Protocols for public key cryptosystems. In: IEEE symposium on security & privacy, issue 3, pp 122–122
71. Kamara S, Lauter K (2010) Cryptographic cloud storage. In: International conference on financial cryptography and data security. Springer, pp 136–149
72. Itani W, Kayssi A, Chehab A (2010) Energy-efficient incremental integrity for securing storage in mobile cloud computing. In: International conference on energy aware computing. IEEE, Cairo, pp 1–2
73. Bellare M, Ran C, Krawczyk H (1996) Message authentication using hash functions—the HMAC construction. *Cryptobytes* 2
74. Yang K, Jia X, Ren K (2013) DAC-MACS: effective data access control for multi-authority cloud storage systems. In: INFOCOM, 2013 proceedings IEEE. IEEE, Turin, pp 2895–2903
75. Hong J, Xue K, Li W (2017) Comments on “DAC-MACS: effective data access control for multiauthority cloud storage systems”/Security analysis of attribute revocation in multi-authority data access control for cloud storage systems. *IEEE Trans Inf Forensics Secur* 10(6):1315–1317
76. Wang H, Domingo-Ferrer J, Wu Q, Qin B (2014) Identity-based remote data possession checking in public clouds. *IET Inf Secur* 8(2):114–121
77. Tan S, Jia Y (2014) NaEPASC: a novel and efficient public auditing scheme for cloud data. *Front Inf Technol Electron Eng* 15(9):794–804
78. Li J, Li J, Chen X (2015) Identity-based encryption with outsourced revocation in cloud computing. *IEEE Trans Comput* 64(2):425–437
79. Li Y, Yu Y, Min G (2017) Fuzzy identity-based data integrity auditing for reliable cloud storage systems. *IEEE Trans Dependable Secure Comput* (99):1

80. Yu Y, Xue L, Man HA, Susilo W, Ni J, Zhang Y et al (2016) Cloud data integrity checking with an identity-based auditing mechanism from RSA. *Future Gener Comput Syst* 62(C):85–91
81. Deswarte Y, Quisquater JJ, Saïdane A (2004) Remote integrity checking. In: Proceedings of 5th working conference on integrity international control in information system (IICIS), pp 1–11
82. Boneh D, Lynn B, Shacham H (2004) Short signatures from the weil pairing. *J Cryptol* 17(4):297–319
83. Tian H et al (2017) Dynamic-hash-table based public auditing for secure cloud storage. *IEEE Trans Serv Comput* 10(5):701–714
84. Peng S, Zhou F, Wang Q, Xu Z, Xu J (2017) Identity-based public multi-replica provable data possession. *IEEE Access* 5:26990–27001
85. Shen W, Qin J, Yu J, Hao R, Hu J (2019) Enabling identity-based integrity auditing and data sharing with sensitive information hiding for secure cloud storage. *IEEE Trans Inf Forensics Secur* 14(2):331–346
86. Zhu Y, Hu HX, Ahn G-J, Yu M (2012) Cooperative provable data possession for integrity verification in multicloud storage. *IEEE Trans Parallel Distrib Syst* 23(12):2231–2244
87. Liu C, Ranjan R, Yang C, Zhang X, Wang L, Chen J (2015) MuRDPA: top-down levelled multi-replica Merkle hash tree based secure public auditing for dynamic big data storage on cloud. *IEEE Trans Comput* 64(9):2609–2622
88. Hwang G-H, Chen H-F (2016) Efficient real-time auditing and proof of violation for cloud storage systems. In: Proceedings of IEEE 9th international conference on cloud computing (CLOUD), pp 132–139
89. Jin H, Jiang H, Zhou K (2018) Dynamic and public auditing with fair arbitration for cloud data. *IEEE Trans Cloud Comput* 6(3):680–693
90. Küpcü A (2015) Official arbitration with secure cloud storage application. *Comput J* 58(4):831–852
91. Kamstra L, Heijmans HJAM (2005) Reversible data embedding into images using wavelet techniques and sorting. *IEEE Trans Image Process* 14(12):2082–2090
92. Zhou J, Sun W, Dong L, Liu X, Au OC, Tang YY (2016) Secure reversible image data hiding over encrypted domain via key modulation. *IEEE Trans Circuits Syst Video Technol* 26(3):441–452
93. Singh P, Raman B (2018) Reversible data hiding based on Shamir's secret sharing for color images over cloud. *Inf Sci* 422:77–97
94. Honsinger CW, Jones PW, Rabbani M, Stoffel JC (2001) Lossless recovery of an original image containing embedded data. U.S. Patent 6 278 791 B1, 21 Aug 2001
95. Ni Z, Shi Y-Q, Ansari N, Su W (2006) Reversible data hiding. *IEEE Trans Circuits Syst Video Technol* 16(3):354–362
96. Tian J (2003) Reversible data embedding using a difference expansion. *IEEE Trans Circuits Syst Video Technol* 13(8):890–896
97. Kim S, Qu X, Sachnev V, Kim HJ. Skewed histogram shifting for reversible data hiding using a pair of extreme predictions. *IEEE Trans Circuits Syst Video Technol* (to be published). <https://doi.org/10.1109/tcsvt.2018.2878932>
98. Pan W, Coatrieux G, Cuppens N, Cuppens F, Roux C (2010) An additive and lossless watermarking method based on invariant image approximation and Haar wavelet transform. In: Proceedings of annual international conference IEEE engineering in medicine and biology (EMBC), 2010, pp 4740–4743
99. Coatrieux G, Pan W, Cuppens-Boulahia N, Cuppens F, Roux C (2013) Reversible watermarking based on invariant image classification and dynamic histogram shifting. *IEEE Trans Inf Forensics Secur* 8(1):111–120
100. Thodi DM, Rodriguez JJ (2007) Expansion embedding techniques for reversible watermarking. *IEEE Trans Image Process* 16(3):721–730
101. Coltuc D (2011) Improved embedding for prediction-based reversible watermarking. *IEEE Trans Inf Forensics Secur* 6(3):873–882

102. Ou B, Li X, Zhao Y, Ni R, Shi Y-Q (2013) Pairwise prediction-error expansion for efficient reversible data hiding. *IEEE Trans Image Process* 22(12):5010–5021
103. Dragoi I-C, Coltuc D (2016) Adaptive pairing reversible watermarking. *IEEE Trans Image Process* 25(5):2420–2422
104. Wu HZ, Wang W, Dong J, Wang HX (2018) Ensemble reversible data hiding. In: Proceedings of 24th international conference on pattern recognition (ICPR), pp 1–6

Independent Learning of Motion Parameters for Deep Visual Odometry



Rahul Kottath, Rishab Kaw, Shashi Poddar, Amol P. Bhondekar,
and Vinod Karar

Abstract Vision-based localization is one of the major aspects of industrial and space robotics. Though many sensing modalities exist for motion estimation, cameras have been used widely due to its availability and reduced cost. Visual odometry estimates the motion parameters of a camera through the images it captures. Multiple sensing modalities are fused to improve estimation accuracy with increased cost. With the success of deep learning architectures in the area of computer vision, one of the recent paradigm shift occurred in visual odometry is estimating motion using non-geometric schemes by the end-to-end manner. The different stages of the traditional visual odometry pipeline are estimated as a single function mapping input images to output 6 DoF pose of the camera. There are many ways to apply deep learning in visual odometry, one of the common techniques is through transfer learning. In this work, analysis has been done on traditional DeepVO and ResNetVO by incorporating a novel architecture splitting and independent learning scheme. The estimation results show the efficacy of the proposed algorithm.

Keywords Motion estimation · Visual odometry · DeepVO · Localization

R. Kottath (✉)

School of Electrical and Electronics Engineering, VIT Bhopal University, Bhopal, India
e-mail: rahulkottath@gmail.com

Academy of Scientific and Innovative Research (AcSIR), Ghaziabad, India

R. Kaw

Center for Nanotechnology Research, VIT University, Vellore, India

S. Poddar · A. P. Bhondekar · V. Karar

CSIR-Central Scientific Instruments Organisation, Chandigarh, India

1 Introduction

Vehicle/Robot localization is an essential aspect of dead reckoning systems and a significant feature of a rescuing robot or an extra-terrestrial robot. Motion estimation techniques used in the localization of a vehicle can be broadly classified based on the sensing modality used; they are a sensor-based estimation and camera-based estimation. One of the popular systems used for localization is Global Positioning System (GPS) and is available with most smartphones. The unavailability of GPS signals in indoor environments, bias accumulation problems of IMU sensors [1], etc. necessitates the requirement of low-cost camera-based navigation systems. The introduction of cameras for navigation is inspired by bird navigation, which heavily relies upon its eyes [2]. Visual odometry estimates the camera motion between successive time instants by using the images alone. Though such systems were existing since the 1980s, the term ‘visual odometry’ has been coined for camera-based motion estimation by David Nister in 2004 [3].

The works in visual odometry can be classified as geometry-based and non-geometric learning-based. The early 2000s saw the development of geometrical visual odometry and which was improved further with the introduction of multiple techniques such as feature-based and direct schemes [4]. Geometrical techniques have the advantage of not using complicated estimation pipeline, instead, it can estimate motion in the end-to-end manner. Memisevic proposed a way to relate two images that were previously done by geometric schemes [5]. Multiplicative interaction among models is explored for learning the relation between images. Traditional techniques for visual odometry and SLAM has many difficulties by using hand-engineered features, dense matching, etc., and obtaining reasonable estimates with varying environmental conditions such as lighting variation, motion blur, repetitive structures, degenerate configurations [6]. The requirement of camera calibration for estimating ego-motion is another challenge in geometrical schemes. The development of deep learning techniques addressed this problem to an extent with its ability to learn motion parameters irrespective of camera used. Earlier works in learning VO was developed without using camera calibration parameters for estimation, but its accuracy was relatively low; also, it required colossal ground truth data for training. But most of the recent works require camera parameters also for geometrical aiding, which improves the estimates. Machine learning has evolved a lot in the area of visual odometry and SLAM after the introduction of learning VO by Roberts et al. which used a neural network for estimating motion. With the introduction of AlexNet, deep learning has taken a leap in various areas of computer vision. Most of the recent work uses transfer learning in ResNet, VGG Net, etc., which reduces the use of a large number of parameters for the task. The performance of ResNet for the classification problems leads to its wide popularity [7]. A large corpus of ground truth data is required for supervised algorithms, which leads to the development of unsupervised/self-supervised techniques. Unsupervised techniques generally incorporate some geometrical aiding such as left-right consistency, forward-backward consistency [8], etc. These additional terms in cost function improve the training and

overall performance of the system. Though learning techniques are said to be independent of camera parameters, if one trains the model using data collected by a single camera, the network may overfit to the used camera which is still an open problem in learning-based VO. The concept of inertia of motion has also been introduced in visual odometry to remove some outlier points. This scheme was introduced to improve the traditional geometric VO pipeline [9]. This concept recently extended to learning inertia for self-supervised odometry by Wang et al. [10]. Though learning-based VO is developing, one cannot say the geometrical methods are saturating. It has proved its ability to estimate under challenging conditions with improved accuracy.

In this paper, a novel architecture splitting or independent parameter learning scheme has been proposed to improve the estimation results. The proposed method removes the necessity of a large number of parameters in the learning framework. The remaining part of the paper is organized as follows: Sect. 2 provides a brief overview of the theoretical background required for learning VO, Sect. 3 discusses the proposed methodology. Section 4 analyzes the performance of the proposed algorithm on the visual odometry, and finally, Sect. 5 concludes the paper.

2 Theoretical Background

Geometrical estimation computes pose using projective geometry of the camera, whereas, learning-based techniques learn the representation of the input at different levels to map to the required output. In this section, some basic theoretical concepts of transfer learning and loss functions are discussed for a better understanding of Deep VO methods.

2.1 Transfer Learning

Deep networks require a very large corpus of data to obtain reasonable performance. The main reason behind this is the requirement of learning millions of weight factors for these networks. One of the effective ways of utilizing the resources is reusing a network which trained for some other tasks and is called transfer learning. Transfer learning has been applied in the VO problem as well, where deep networks trained on ImageNet are used as a base model for training. There is a trend in the research community to go deeper by adding layers for obtaining good performance. But the performance saturates after some depth due to vanishing gradient problem. The basic approach for transfer learning is to freeze some weights which are already learned on large data like ImageNet and fine-tune the network for a specific task. ResNet is used here as the base framework for applying the proposed method.

2.2 Loss Functions

Loss functions are an essential ingredient of deep neural networks. Carefully designed loss functions make the network function well for a specific task. The mainly used loss function for classification task is cross-entropy loss, whereas for regression, Euclidian loss, or mean squared error loss is used. The most straightforward loss function which can be used at the output layer is mean square error (MSE). It is calculated between predicted and the actual values of the output. The norm of difference is calculated and is mainly used to train the network. The error computation may contain multiple parameters based on the output vector length.

Considering the output vector as a single unit, the loss can be defined as,

$$L = \frac{1}{N} \sum_{i=1}^N \|f(x_i) - y_i\|_2 \quad (1)$$

Many of the recent works use unsupervised learning, and most of them incorporate geometric cues for estimating the loss. The consistency check is one of the stages which was introduced to enhance the performance compared to state-of-the art techniques. This includes left-right consistency [11] used in the case of stereo odometry and forward-backward consistency [12, 13] by estimating forward and backward optical flow.

3 Proposed Methodology

Deep learning based methods estimate the pose of the camera in end-to-end manner. Different stages of the traditional VO pipeline can be merged into a single learnable model that learns the required 6DoF pose of the camera. A wide range of machine learning/deep learning solutions are proposed for visual odometry. The supervised learning scheme is used in this work with CNN as the base architecture. Instead of learning the motion vector as a whole, this work proposes to learn the parameters of the vector independently. The proposed method is compared with the results of base architecture. Figure 1 depicts how the architecture splitting helps to reduce the number of weights to be learned and hence the complexity. The features are learned from the optical flow images are used for predicting the ego-motion values. The architecture used has the following layers;

Conv (7×7)—Conv (5×5)—Conv (3×3)—Maxpool—Conv (3×3)—Conv (3×3)—Maxpool—Conv (3×3)—Conv (3×3)—Maxpool—Flatten—Dense layers.

Modifications are done on dense layers which include splitting the layer into six separate parallel layers and predict the individual parameters. The experimentation is done on deep CNN based visual odometry architecture, where after the flatten layer, parallel architecture is used for learning individual parameters of ego-motion separately (Fig. 2).

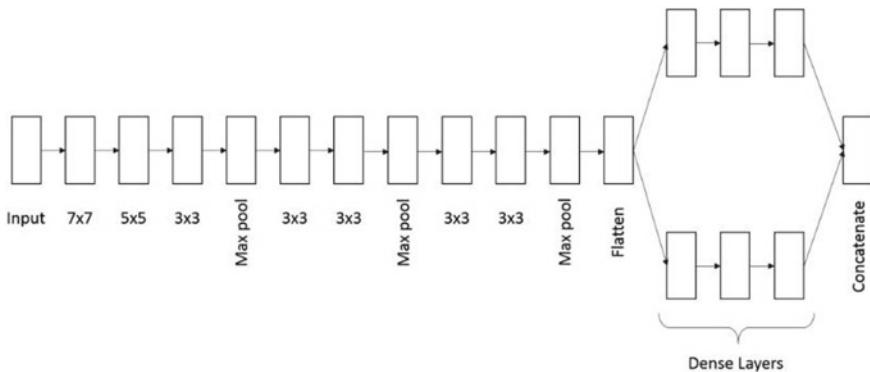


Fig. 1 DeepVO–rotation and translation vector learned separately

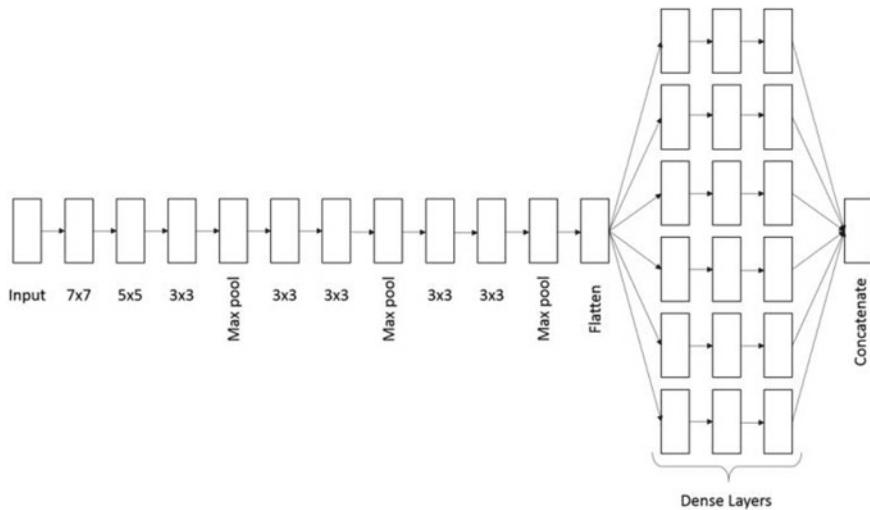


Fig. 2 DeepVO–individual parameters of rotation and translation learned separately

The ResNet50 architecture is used as a base for transfer learning for deep visual odometry. The network is trained for ImageNet classification challenge and the weights of the same are adapted for this task. The fully connected layer of the network is not included for the proposed architecture instead a split dense layer is incorporated. Splitting into two parts predicts the rotation and translation as 3×1 vector whereas splitting into six parts predicts the individual components of rotation and translation (Figs. 3 and 4).

Fig. 3 ResNetVO with rotation and translation vectors learned separately

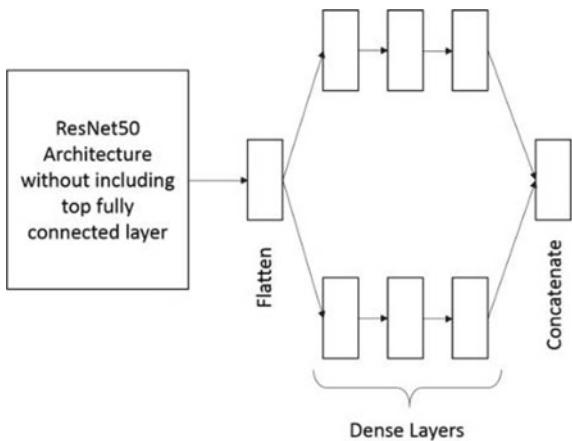
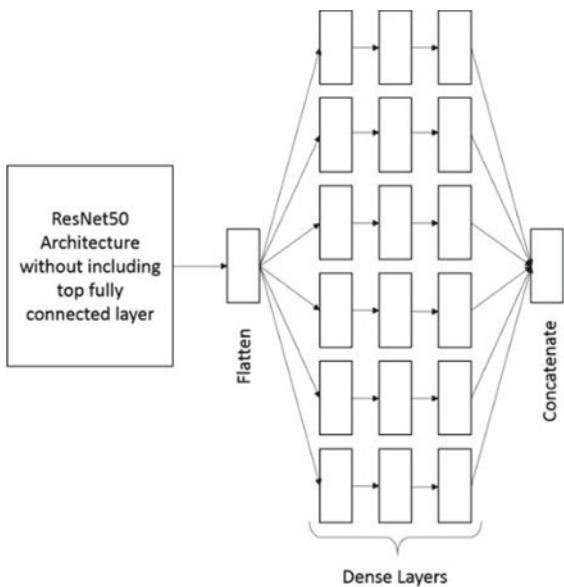


Fig. 4 ResNetVO with an individual component of pose learned separately



4 Results and Discussion

The learning problem is addressed using two types of frameworks. The first one uses deep convolutional architecture, and the second uses transfer learning on ResNet50 architecture. Implementation has been done in Python with Keras deep learning library. The KITTI Vision benchmark [14] is used for training the network. Sequences 0–7 are used for training and 8–10 are used for testing. The optical flow is computed from the KITTI dataset images given as input to the deep neural network. A total of 16,330 optical flow images were used for training the model and 6860 images

are used for testing the performance of the model. Optical flow is computed offline through available techniques and hence OF estimation framework is not separately incorporated in the model. The estimated optical flow images are cropped to obtain a common size of 1226×370 and are resized to 307×93 for reducing the memory requirements. The pose vector is of six dimensional which comprises of 3×1 rotation vector in Euler format and 3×1 translation vector [15]. The predicted vector is converted into the KITTI pose format for plotting. The predicted pose is between two frames which is concatenated with the previous pose for obtaining global position. A sample optical flow image obtained is shown in Fig. 5.

The initial experiments were done by using deep CNN model with output as a single vector of 6×1 dimensions. The model was not able to predict the motion path accurately and further experiments were conducted by assuming independence between rotation and translation part. The improvement obtained after this single split give motivation to try more splits to the architecture and the final results were improved even further. The final model learns the individual component of pose vector separately. The result of deep CNN model-based architecture is shown in Fig. 6. The deep CNN architecture shows significant improvement by learning the



Fig. 5 Optical flow input used for learning ego-motion (a), (b) consecutive images in KITTI dataset, (c) optical flow computed

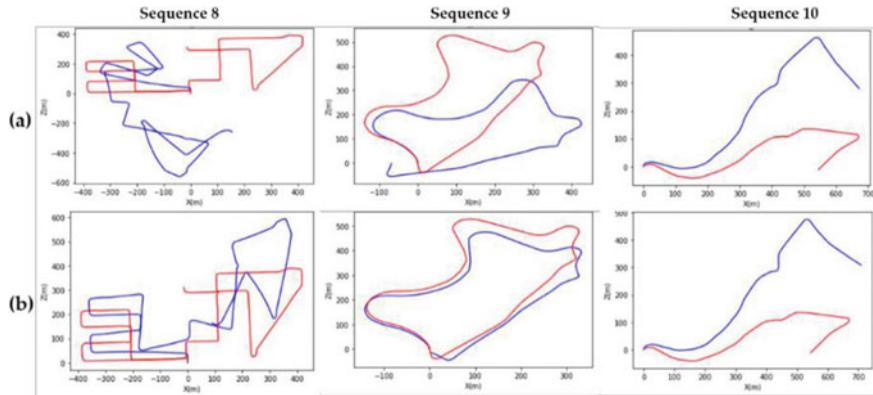


Fig. 6 Deep VO results (a) rotation and translation vectors learned separately, (b) individual parameters of rotation and translation vectors learned separately

three Euler angles and three translational values separately. The networks are implemented in Keras/Tensorflow and trained using an Nvidia GTX 1060. The results of the test sequences of ResNetVO is shown in Fig. 7. The performance of ResNetVO is relatively better than deep CNN as the number of parameters is relatively high.

This is a novel learning scheme for neural networks and can be used for various applications where the parameters can be assumed to be independent of each other. Some of the latest advancements of machine learning include multi-task learning schemes where the same model can be used for multiple related tasks. The architecture splitting scheme proposed here can be extended for such applications.

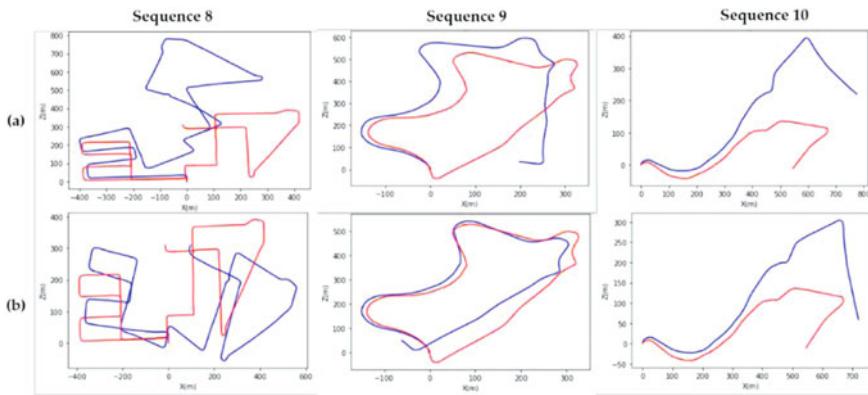


Fig. 7 Results of test sequences in case of ResNet VO (a) learning R and t Separately, (b) learning individual components of R and t

5 Conclusion

In this work, an attempt is made to give a broader idea of different deep learning architectures used in visual odometry and improve the existing technique by incorporating a novel architecture splitting scheme. The prediction accuracy improved when the model is trained to predict the parameters independently. The performance of the algorithm is compared with conventional learning techniques without architecture splitting. The results show that the traditional learning results can be improved by incorporating the proposed independent learning scheme which learns the individual parameters of 6DoF pose separately. The performance is analyzed in both traditional learning and transfer learning schemes. ResNet version of the results is comparatively better than the traditional learning scheme. In the future, it can be incorporated in different established architectures of ego-motion estimation. A future study can also be performed for investigating the effect of a varying number of weights used for learning with the accuracy of estimation.

References

1. Kottath R, Narkhede P, Kumar V, Karar V, Poddar S (2017) Multiple model adaptive complementary filter for attitude estimation. *Aerospace Sci Technol* 69:574–581
2. Ettinger SM (2001) Design and implementation of autonomous vision-guided micro air vehicles. University of Florida
3. Nistér D, Naroditsky O, Bergen J (2004) Visual odometry. In: Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, 2004. CVPR 2004. IEEE
4. Roberts R et al (2008) Memory-based learning for visual odometry. In: 2008 IEEE international conference on robotics and automation. IEEE
5. Memisevic R (2013) Learning to relate images. *IEEE Trans Pattern Anal Mach Intell* 35(8):1829–1846
6. Poddar S, Kottath R, Karar V (2019) Motion estimation made easy: evolution and trends in visual odometry. In: Recent advances in computer vision. Springer, pp 305–331
7. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
8. Sardana R, Kottath R, Karar V, Poddar S (2019) Joint forward-backward visual odometry for stereo cameras. arXiv preprint [arXiv:1912.10293](https://arxiv.org/abs/1912.10293)
9. Kottath R et al (2017) Inertia constrained visual odometry for navigational applications. In: 2017 Fourth international conference on image information processing (ICIIP). IEEE
10. Wang C, Yuan Y, Wang Q (2019) Learning by inertia: self-supervised monocular visual odometry for road vehicles. In: ICASSP 2019–2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE
11. Godard C, Mac Aodha O, Brostow GJ (2017) Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition
12. Zhou T et al (2017) Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE conference on computer vision and pattern recognition
13. Xu Y, Wang Y, Guo L (2018) Unsupervised ego-motion and dense depth estimation with monocular video. In: 2018 IEEE 18th international conference on communication technology (ICCT). IEEE

14. Geiger A, Lenz P, Urtasun R (2012) Are we ready for autonomous driving? The kitti vision benchmark suite. In: 2012 IEEE conference on computer vision and pattern recognition. IEEE, pp 3354–3361
15. Costante G, Ciarfuglia TA (2018) LS-VO: learning dense optical subspace for robust visual odometry estimation. *IEEE Robot Autom Lett* 3(3):1735–1742

Smart Street Lights to Reduce Death Rates from Road Accidents



Rajat, Naresh Kumar, and Manoj Sharma

Abstract Road accidents are a principal cause of demises, detriments and property damage every year. The fatal crashes are the primary reason for death among young people aged between 5 and 29 years. Road fatalities can be minimized by using smart street lights. This paper reviews the existing street light systems, including solar energy, light detection and motion detection systems. Furthermore, this paper proposes a system to detect and prevent road accidents and improve post-crash care using smart street lights. In case of an accident, the proposed system immediately assists in sending a message to the emergency services using GSM module and changing the colour of the street light LED to red, which acts as a warning signal for the vehicles approaching from behind on the road. The system uses solar energy to power the connected LED. Hence, this makes the system energy-efficient also.

Keywords Smart street lights · Road accidents · Image processing · Post-crash care · Solar energy

1 Introduction

As indicated in Fig. 1 by the World Health Organization, street traffic wounds cause an expected 1.35 million deaths globally for each year. Tens of millions of individuals are hurt or disabled by road accidents each year, and more than 3700 individuals die on the world's roads consistently [1].

Rajat (✉) · N. Kumar
UIET, Panjab University, Chandigarh, India
e-mail: rajatblogger11@gmail.com

N. Kumar
e-mail: naresh_uiet@pu.ac.in

M. Sharma
Giani Zail Singh Campus College of Engineering & Technology, MRSPTU, Bathinda, India
e-mail: neelmanoj@gmail.com

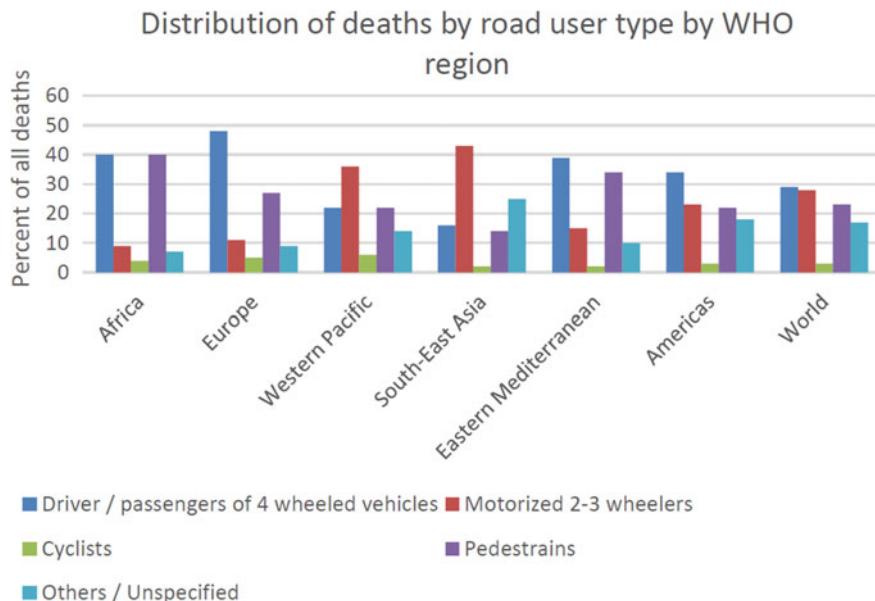


Fig. 1 Distribution of deaths by road user type, by WHO Region. *Data from Global status report on road safety 2018*

Road accidents are the principal cause of death for youngsters and grown-ups aged 5–29 years. There are many risk factors involved in road accidents including speeding, driving under the influence of alcohol and other psychoactive substances, distracted driving, etc., but one of the major risk factors is insufficient post-crash care. Latencies in identifying and giving considerations to those engaged in a road accident raise the seriousness of wounds. Care of wounds after an accident has happened is incredibly urgent: postponements of minutes can have the effect among life and demise [2]. This paper focuses on improving post-crash care by assuring access to convenient prehospital care and alerting incoming vehicles about the accident spot on the road.

2 Causes of Road Crashes During Night-Time

Because your vision accounts for nearly 90% of your reaction while driving, night time driving dramatically decreases your ability to effectively respond to potential hazards on the road.

- Reduced visibility: At night, we no longer have natural light to help us see road signs, other drivers, pedestrians, debris in the road, animals, and other obstacles. It also makes it more difficult to judge the distance between your car and another

car. Driving at night means relying on headlights and street lights, which do not provide the same visibility that natural light does.

- Age factors: Unfortunately, as we age, our ability to see at night deteriorates. In addition, older people may have undermined vision as a result of cataracts and increasing eye diseases.
- Rush hour: Any time of the year, rush hour can be dangerous driving time. As the days get shorter and darkness comes earlier, the drive time becomes more dangerous especially when driving in stop-and-go or bumper-to-bumper traffic.
- Drowsy or fatigued driving: A study published by the National Sleep Foundation tells us that sleep-deprived drivers are the cause of 6400 deaths and 50,000 serious injuries annually on the US roads. A drowsy driver's reaction times are greatly reduced. Fatigued drivers can be on the road any time of the day, but night time hours (especially from 3 a.m. to 7 a.m.) are the prime time.

As can be seen in Fig. 2, a higher incidence of accidents was seen from 0:00–7:00 and most of the accidents happened during 4:00–6:00 [3].

- Driving under the influence: Impaired drivers are more likely to be on the road after dark, between the hours of midnight to 3 a.m. on weekends. There is a higher risk of sharing the road with an impaired driver at night as people leave restaurants and bars. According to the National Safety Council, weekend nights are the worst part of the week for fatal accidents. As shown in Fig. 3, road accidents are the major cause of death among young people aged between 18 and 24 in countries such as the USA, Austria, France, and Ireland.

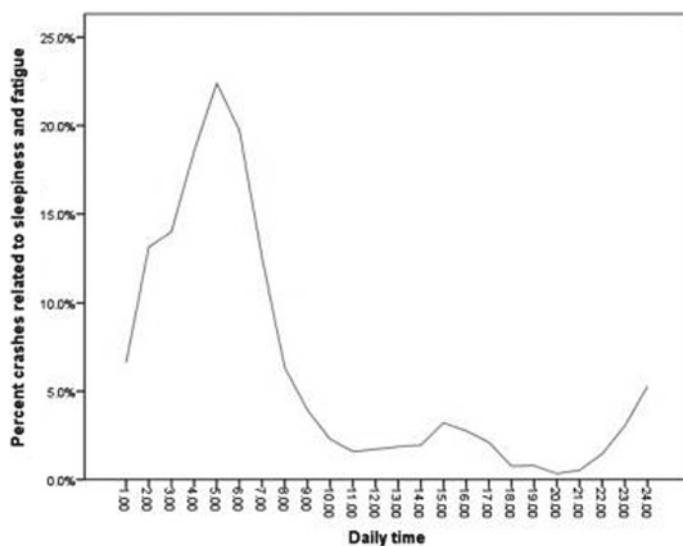


Fig. 2 Hourly distribution of car accidents due to overturning of the vehicle or hitting with various roadside features [3]

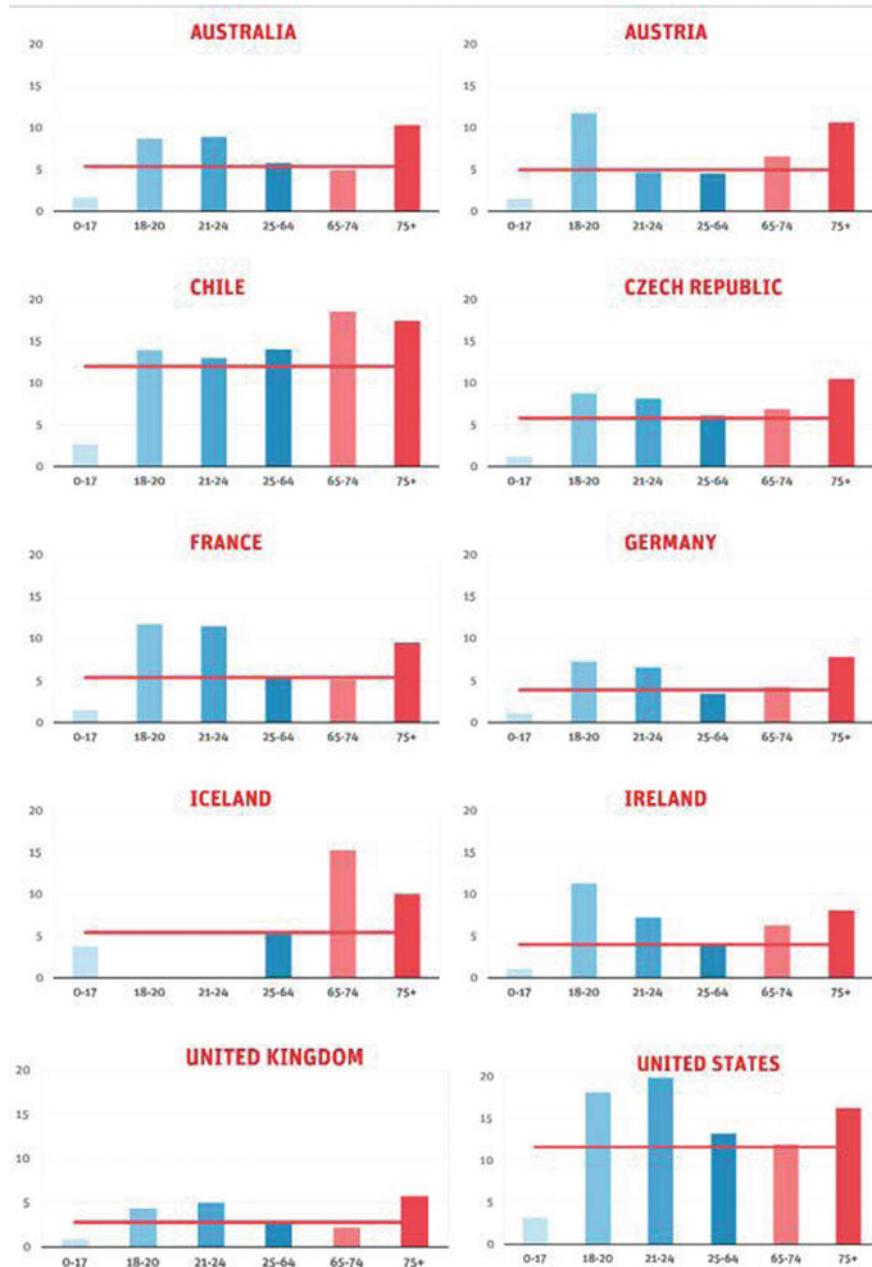


Fig. 3 Mortality rate by age group 2016. Source Road safety annual report 2018 by International Transport Forum

- Distracted drivers: Anything that takes your hands off the wheel, eyes off the lane and brain off driving is a distraction. This can be an even deadlier combination at night
- Construction activity: Often, road construction happens in the evening hours. With poor light and other factors, it can be difficult to see construction work zones and you can get blinded by the brightwork lights being used.

3 The Need for Smart Street Lights

During heavy rain or fog, it is very hard to see what is in front of you while driving on the road. It is even worse that unless you are wearing reflective clothing, it is even hard for the drivers to see you while crossing the road at night. This cannot only put other vehicles at risk but also the place pedestrians and even pets at risk of injury. As shown in Fig. 4, car accident statistics are harsh on the roads with no street lighting at night. Over 40% of all deadly auto crashes happen at night, even though there is 60% less traffic on the roads. [4] The drivers get very less time to respond while driving at higher speeds, even with high-beam headlights on, because perceptibility is restricted to about 500 feet (250 feet for normal headlights) [5]. Smart street lights hold the ability to dispense with this issue altogether. Cities can use smart street lights equipped with wireless sensors and connectivity for every situation, including responding to emergency situations, increasing the brightness level during dangerous time periods and changing the colour of the street light in case of an accident.

4 Review of Existing Systems

See Table 1.

5 Improving Visibility During Dangerous Time Periods

Diminished perceptibility is the most evident risk of late evening driving. The visibility of the person driving the vehicle reduces drastically at night; hence, there are high risks that can often happen suddenly. In addition, it requires some moments for the eyes to change in accordance with the dimness after being in a lit structure or after driving on a sufficiently bright highway.

This can be an issue, especially for older drivers. As we get older, our eyes become less ready to respond rapidly to changes in light and we can experience issues with colours and differentiation in dim light. Between the ages of 15 and 65, the time it takes to recuperate from glare rises from 1 to 9 s. This could be the cause of why some individuals find driving at night increasingly troublesome.

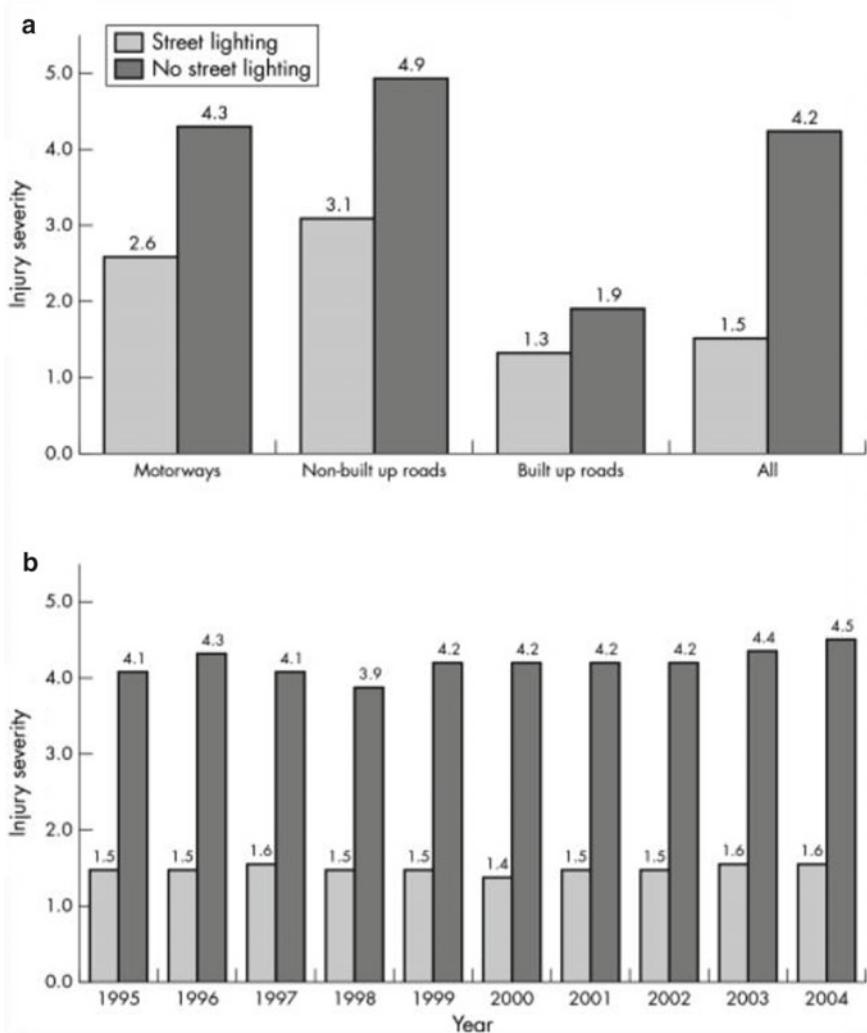


Fig. 4 UK mean (years 1995–2004) injury severity at night by street lighting and different road types (a), and as analysed across different road types for years 1995–2004 (b) [6]

Unsafe drivers may not be as simple to see at night, particularly if they are wearing dark apparel. Hence, it is suggested that exposed road users wear reflective clothes and that all road users take more care at dark hours. Pedestrians may not have recognized that drivers have not seen them. Cyclists are also tough to spot, as their lights are not as strong as cars. It can likewise be increasingly hard to recognize the single front light of a motorcyclist around evening time if they are around vehicles. This could be the reason for more serious injuries during night time, as shown in Tables 2 and 3.

Table 1 Review of existing systems

Focused	References	Methods used	Remarks
Motion sensing	[12–14, 17]	1. Motion sensor	1. The street lights will be turned on when the sensor detects any moving object and will get turned off in case of no motion
		2. Brightness sensor	2. If stray animals sit in the middle of the road during night time, then the motion sensor will not be able to detect the animal because the animal is stationary 3. Slow-moving objects are often not detected by the motion sensor. For example, old people or special children walking very slowly on the zebra crossing may not be detected as a moving object by the motion sensor
Solar energy	[13, 14, 17, 18]	1. Solar panel	1. The solar panels will convert the energy of light into electricity
		2. Photovoltaic cells	2. The efficiency of the solar panels drops during cloudy and rainy days 3. We cannot rely upon solar panels for energy generation in some countries. For example, some parts of Norway, Iceland, Sweden, Finland, Canada and Alaska experience no sunlight during winters
Light sensing	[15–19]	1. Light-dependent resistor	1. The street lights will turn on and off depending upon the input from the light sensor
		2. Photoresistor	The lights can only be turned on and off using this sensor, no changes can be made in the brightness levels of the street light

Table 2 Number of fatalities during night time hours (between 7 pm and 6 am) [7]

	Fatal injuries	Serious injuries	Slight injuries	Total injuries
Motorcyclists	88	1100	3085	4273
Car occupants	303	2710	24,777	30,803
All road users	629	5821	34,835	41,285

Table 3 Reported road casualties by severity and road user group, 2016 [8]

	Killed	Seriously injured	All severities
Pedestrians	448	5140	23,550
Pedal cyclists	102	3397	18,477
Motorcycle users	319	5533	19,297
Car occupants	816	8975	109,046
Total	1792	24,101	181,394

At night, it is difficult to accurately determine the velocity and distance between the objects. Some objects can be nearer than what they appear to human eyes or travelling more rapidly than what the driver expects.

The most effective way smart street lights improve safety on the roadways is by increasing the brightness level. Smart technology surpasses the current, traditional street lights are either programmed to turn on at a specific time or must be turned on manually. Traditional street lights can only be switched on and off, whereas smart street lights utilize remote sensors to automatically turn on and change the light intensity according to changing weather circumstances. This ensures that a sufficient amount of light is provided for better visibility on the roads for the vehicles as well as pedestrians.

The modules that are connected to the sensors can send data to the cloud platform via a cellular network. These sensors provide information on changing weather circumstances, the number of vehicles on the road and movement recognition. This information is sent to the cloud, which is associated with a robotized system that controls different parameters of the street lights. The framework will change, dim or brighten, the light intensity level depending on the received data. For instance, on account of haze, the street light will change its colour to yellow and decrease the light intensity level to prevent light reflection and glare. This will adjust as per the ideal setting for conditions in a specific region.

6 Providing Warning to the Traffic Approaching from Behind by Changing the Colour in Case of an Accident Detection

In case, if there is no one on the road at night and the person is severely injured, in no condition to call the emergency services. Health outcomes for victims of collisions rely on the ability of the emergency medical care system to quickly locate and provide emergency first responder care to stabilize the victim and transport them to hospital for the appropriate care and treatment. The proportion of injured people who die before reaching a hospital in developing nations is over twice that in developed nations. Ideally, there would be at least a simple prehospital system. We can open the door to more efficient accident alertness by connecting the smart street lights to a cloud platform. This could provide timely care at the scene with equipped ambulances staffed with certified providers. For instance, a smart street light can detect when there is a vehicle-to-vehicle collision, vehicle-to-pedestrian accident, or any other accident and instantly turn its colour to red. This will help in signalling the incoming vehicles on the road to slow down and reduce further collisions with the crashed vehicle. According to the Global status report on road safety 2018 by WHO, 109 countries have a telephone number with national coverage to activate the emergency care system, but even so, it can take some time to report an accident. Sensors send the accident data directly to the cloud platform; from there, the platform sends a warning to the emergency services such as ambulance service and police control room. This will help in quick post-crash response that is essential to provide effective care for the injured. With smart street lights, the injured person won't need to rely upon people nearby to report the accident to officials and emergency care can be given more quickly. Emergency medical transport service can be instantaneously alerted by the cloud platform.

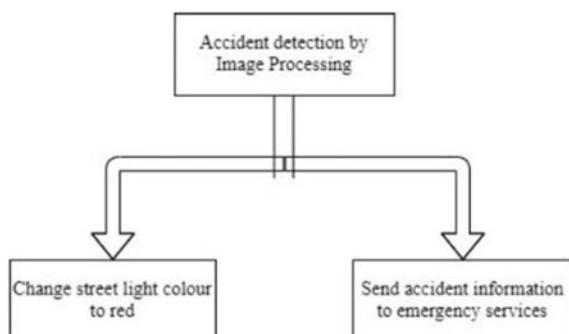
7 System Model and Components

According to research [9], it takes about 8 min to acknowledge the accidents and about 22 min before an urgent car comes to the spot where the accident occurs. Therefore, if we can detect accidents or stopped vehicles immediately, we can quickly alarm to the following cars and emergency services, thereby preventing secondary accidents, responding to accidents quickly, and reducing jammed time. The system architecture of the smart street lights consists of IP66 CCTV camera, control unit, LED control unit, LDR sensor, solar panels with maintenance-free battery, colour changing LED and a GSM module.

- Light Sensor: when the sunlight falls on the LDR sensor, its resistance decreases and makes the street light to switch off. When the sun sets, light does not fall on the sensor, then the resistance drops and the light is switched on.

- LED: it uses less energy than high-intensity discharge lamps. It is used to alert the incoming traffic about the accident by changing the colour to red. This will help in preventing further collisions.
- Solar Panel: it converts the solar energy into electricity, which is powers LED lamp during dark hours.
- Maintenance-free battery: the tabular rechargeable maintenance-free battery is used to store the power generated by solar panels.
- LED Control Unit: this is the main controller of the street light LED. It is responsible for changing the colour of the street light LED, adjusting the intensity level using a programmer, and turning the light on and off by using a relay.
- GSM module: it is used to send the short message service (SMS) to the emergency services in case of an accident detection.
- IP66 CCTV camera with IR night vision: it is used to capture the traffic movements. The camera is water, dirt and corrosion-resistant. The IR cut feature illuminates the footage in a dark environment. The one '6' indicates the level of dust protection and another '6' is fluid protection. The footage recorded by the CCTV camera is monitored in real-time.
- Control Unit: it uses a high-speed processor which runs the real-time collision and stopped vehicle/object algorithm. Moreover, it also controls the intensity of the LED by giving instructions to the LED control unit.
- Image Processing: it is possible to detect sudden unexpected events such as accidents, using image processing. Cameras can be installed on smart street lights at a curved area on an expressway, where motor vehicle accidents occur frequently, to detect accidents by processing images taken by the installed cameras. We can use accident detection algorithms that detect abrupt changes in the state of motion and stopped vehicles, by performing a tracking process on each vehicle. If certain conditions are met, an abnormal state is determined to exist and immediate further steps are taken as shown in Fig. 5. Such conditions include changes in traffic flow before an accident occurs and changes in individual motor vehicle speed. For the best detection accuracy, cameras could be installed at a height of 6.5 m and with a range of 40–50 m. Image processing can be used for accident information needed

Fig. 5 Information flow



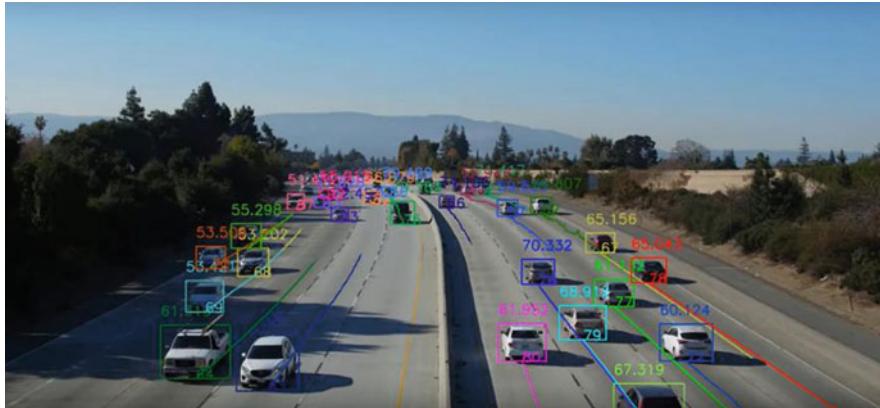


Fig. 6 A still from a demo video [10], which shows the estimated speed of each vehicle in miles/hour

for both, the drivers approaching from behind and emergency services to ensure rapid response to accidents and prevent secondary disasters [9] (Fig. 6).

8 Methodology

The working of the proposed model is as follows:

The camera captures the video and streams it on the control unit. The control unit receives the real-time video and divides it into frames. The high-speed processor in the control unit runs the algorithm, which detects a sudden change in motion, stopped vehicles and objects on the road. The algorithm runs real-time vehicle recognition, tracking, and collision detection. If certain conditions are met in the consequent frames, then an unusual state is determined to exist. In case of an accident, the control unit sends a signal to the GSM module, which sends a message to the emergency services and also changes the colour of the LED to red. This helps to improve post-crash care and avoids further collisions from vehicles coming from behind. If no accident is detected, then it the system goes back to take input from the camera. The LED is connected with a solar panel powered maintenance-free battery. The solar panel converts sunlight into electricity, which is then stored in the battery. The battery powers the LED during dark hours. This makes the proposed system energy-efficient.

Expected outcomes:

- If an accident is detected, then the control unit will send a signal to the GSM module and change the colour of the LED to red. The GSM module will send a message to the emergency services. This will help to reduce the delay, which can be deciding factor between life and death, after an accident has happened. The red colour of the LED will serve as an alert to the vehicles approaching from behind.
- If no accident is detected, then the system goes back to take input from the camera.

- If the resistance of the light-dependent resistor (LDR) sensor increases above the set threshold value, then the control unit will analyse the video frames to check the real-time lighting condition. Depending on the current lighting conditions, the control unit will send the data to the intensity control unit. The intensity control unit will then set the intensity level of the LED and turn it on.

The procedure of implementation of smart street lights is as follows:

1. Capturing the video
2. Processing the video frames
3. Running the algorithm
4. Determining the state of an accident
5. Activating the sensors
6. Wireless communication module.

- Capturing the video

The IP66 camera with IR night vision is installed on the street lights at highly accident-prone areas such as sharp curves. The camera is dust, water and corrosion-resistant that captures the video of live traffic. It is capable of capturing high-resolution frames at 24 frames per second. The high-quality video makes it easier to identify the vehicle's registration number, which can be further used by police control room to track the vehicle.

- Processing the video frames

The video is streamed in real-time from the camera, and then it is divided into frames. The frames make it easier and faster for the algorithm to identify a vehicle. The algorithm detects the edges of the vehicles and the distance between the traffic by processing each frame. If an uncommon situation is detected, then it is compared with preceding and succeeding frames in order to accurately identify the abnormal condition.

- Running the algorithm

A high-speed processor runs the algorithm, which detects a sudden change in motion and stopped vehicles and objects. If certain conditions are met, then an unusual state is determined to exist. The algorithm can be trained by feeding different collision situations and images, using machine learning. It is also capable of predicting the real-time speed of the vehicle, as shown in Figs. 7 and 8. This can also help in reducing the use of other devices such as radar speed gun (also known as radar gun or speed gun) by police for speed limit enforcement

- Determining the state of an accident

If the algorithm detects a road accident by vehicle recognition, tracking, and collision detection as shown in Fig. 8, then it immediately sends a signal to the GSM module, and also to the LED control unit to change the colour of the street light LED to red. The LED control unit changes the colour of the particular street light to red, which acts as a warning signal for the traffic approaching from behind to avoid further collisions.

- Activating the sensors

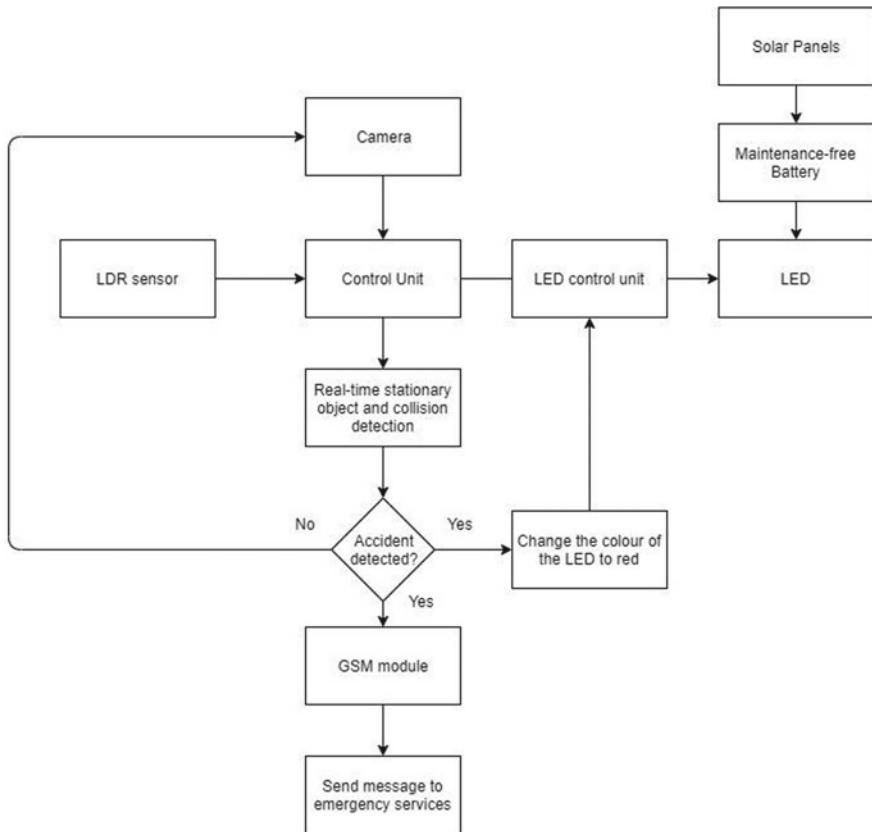


Fig. 7 Block diagram of smart street light

The LDR or photoresistor senses the lighting conditions of the surroundings and sends the resistance value to the control unit. As the resistance value passes the threshold value, the control unit sends a signal to the connected LED control unit to turns the street light on or off by using the relays in the LED control unit. The control unit also senses the darkness level by processing the video frames and running the algorithm and then sends the data to the LED control unit to set the intensity level of the street light LED according to the changing weather conditions.

- **Wireless communication module**

As soon as the GSM module receives the signal from the control unit, it sends a preset message, which also contains the location of the street light where an accident is detected to the emergency services for immediate post-crash care. This service helps to reduce the delay between the moment when a collision happened and the arrival of the emergency services at the accident spot. A small delay of a few minutes can be a deciding factor between life and death of the

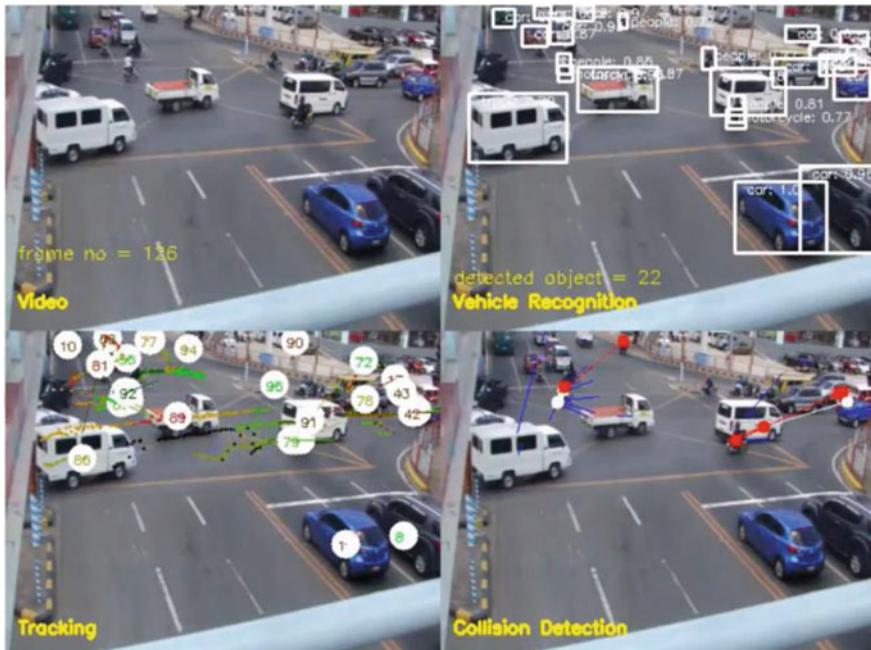


Fig. 8 A still from the video [11] by Kardi Teknomo, showing real-time vehicle recognition, tracking and collision detection

injured person. This will also eliminate the dependency on the nearby people to report the accident to the police and call for an ambulance.

9 Conclusion

Smart street lights are a good way to decrease the number of deaths due to road accidents. Moreover, the lights are powered by solar panels, which make them energy efficient. When combined, these factors emergency signalling using GSM module, real-time accident detection using image processing, increased brightness during dangerous time periods, using solar power to charge the maintenance-free battery and changing the colour in case of an emergency, make the roads safer and improves the post-crash response.

References

- WHO Road Traffic Injuries. https://www.who.int/violence_injury_prevention/road_traffic/en/. Last accessed on 6 Dec 2019

2. WHO Factsheets. <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>. Last accessed on 6 Dec 2019
3. Sadeghniat K, Yazdi Z, Moradinia M, Aminian O, Esmaili A (2015) Traffic crash accidents in Tehran, Iran: its relation with circadian rhythm of sleepiness. Chin J Traumatol 1. <https://doi.org/10.1016/j.cjtee.2014.09.001>
4. Pines Salomon Injury Lawyers Legal advice. <https://seriousaccidents.com/legal-advice/top-causes-of-car-accidents/nighttime-driving/>. Last accessed on 6 Dec 2019
5. National safety council night driving. <https://www.nsc.org/road-safety/safety-topics/night-driving> [Last accessed on 7 December 2019]
6. National center for biotechnology information (2019) U.S national library of medicines. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2564438/>. Last accessed on 7 Dec 2019
7. Royal society for the prevention of accidents driving at night factsheet June 2017. <https://www.rospa.com/rospaweb/docs/advice-services/road-safety/drivers/driving-at-night.pdf>. Last accessed on 8 Dec 2019
8. Department for transport (2017) Table RAS30001: Reported road casualties by road user type and severity, Great Britain. <https://www.gov.uk/government/statistical-data-sets/ras30-reported-casualties-in-road-accidents>. Last accessed on 8 Dec 2019
9. Tsuge A, Takigawa H, Osuga H, Soma H, Morisaki K (1994) Accident vehicle automatic detection system by image processing technology. In: Proceedings of VNIS'94—1994 vehicle navigation and information systems conference, Yokohama, Japan, pp 45–50. <https://doi.org/10.1109/vnis.1994.396868>
10. YouTube Demo of vehicle tracking and speed estimation at the 2nd AI City Challenge Workshop in CVPR 2018. https://www.youtube.com/watch?v=_i4numqv7Y. Last accessed on 8 Dec 2019
11. YouTube Near Miss Traffic Accident Detection Tool. https://www.youtube.com/watch?v=2_hG69S6GM0. Last accessed on 8 December 2019
12. Yoshiura N, Fujii Y, Ohta N (2013) Smart street light system looking like usual street lights based on sensor networks. In: 2013 13th International symposium on communications and information technologies (ISCIT), Surat Thani, pp 633–637. <https://doi.org/10.1109/iscit.2013.6645937>
13. Velaga R, Kumar A (2012) Techno-economic evaluation of the feasibility of a smart street system: a case study of rural India. Procedia Soc Behav Sci 62:1220–1224
14. El-Faouri FS, Sharaiha M, Bargouth D, Faza A (2016) A smart street lighting system using solar energy. In: 2016 IEEE PES innovative smart grid technologies conference Europe (ISGT-Europe), Ljubljana, pp 1–6. <https://doi.org/10.1109/isgteurope.2016.7856255>
15. Abinaya B, Gurupriya S, Pooja M (2017) IoT based smart and adaptive lighting in street lights. In: 2017 2nd International conference on computing and communications technologies (ICCCT), Chennai, pp 195–198. <https://doi.org/10.1109/iccct.2017.7972267>
16. Badgelwar SS, Pande HM (2017) Survey on energy efficient smart street light system. In: 2017 International conference on I-SMAC (IoT in social, mobile, analytics and cloud) (I-SMAC), Palladam, pp 866–869. <https://doi.org/10.1109/i-smac.2017.8058303>
17. Bhairi MN, Kangle SS, Edake MS, Madgundi BS, Bhosale VB (2017) Design and implementation of smart solar LED street light. In: 2017 International conference on trends in electronics and informatics (ICEI), Tirunelveli, pp 509–512. <https://doi.org/10.1109/icei.2017.8300980>
18. Maguluri LP, Sorapalli YSV, Nakkala LK, Tallari V (2017) Smart street lights using IoT. In: 2017 3rd International conference on applied and theoretical computing and communication technology (iCATccT), Tumkur, pp 126–131. <https://doi.org/10.1109/icatct.2017.8389119>
19. Suseendran SC, Nanda KB, Andrew J, Bennet Praba MS (2018) Smart street lighting system. In: 2018 3rd International conference on communication and electronics systems (ICCES), Coimbatore, India, 2018, pp 630–633. <https://doi.org/10.1109/cesys.2018.8723949>

An Improved Approach for Face Detection



C. A. Rishikeshan, C. Rajesh Kumar Reddy,
and Mohan Krishna Varma Nandimandalam

Abstract Face recognition from an arbitrary image has been a standout amongst the most considered topic in image processing and computer vision. The human face is a convoluted multidimensional visual model and henceforth it is exceptionally hard to build up a computational model to recognize the face. This paper presents an approach depending on the attributes extracted from the image to identify the human face. The proposed approach combines both morphological image processing techniques and cascade object detector capabilities. This method is effective in face detection for arbitrary images.

Keywords Face detection · Image processing · Object detection

1 Introduction

Face detection is a process of detecting the faces in a given image. The goal is to find face and to draw a rectangle around each face. There are many face detection algorithms available in this field. For instance, the template-matching techniques are applied for face localization and detection by calculating the correlation of a given image to a standard face pattern [1, 2]. The feature invariant methods are applied for facial elements detection of eyes, nose, mouth, ears, etc. [3, 4]. The appearance-based approaches are applied for face detection using eigen-face [5–7] neural network [8, 9] and information theoretical approach [10, 11]. A deep learning based pipeline is described for unconstrained face detection and verification which attains state-of-the-art performance on several standard datasets [12].

Several approaches for face recognition has been discussed over the decades, most of these approaches uses techniques such as Principal Component Analysis (PCA),

C. A. Rishikeshan (✉) · C. Rajesh Kumar Reddy
Department of CSE, Madanapalle Institute of Technology & Science (MITS), Madanapalle, India
e-mail: drrishikeshanca@mits.ac.in

M. K. V. Nandimandalam
Oceanic IT Convergence Technology Research Center, Hoseo University, Asan, South Korea

Hough Transform (HT), Artificial Neural Networks (ANN), machine learning, information theory, geometrical modelling, template matching. The NN based and view based methods necessitate a large quantity of face and non-face training samples [13].

Among those algorithms, Viola and Jones face detection is the one of the fast and accurate face detection algorithm [14]. The process followed in Viola-Jones algorithm is from the image input to scan a sub-window to detect faces [15]. Usually input image is rescaled for the process of face detection, but in the Viola-Jones algorithm, instead of the input image detector is rescaled with different sizes and detector runs repeatedly through the input image. Viola-Jones detector uses same number of calculations for any size of input images. More details about the Viola-Jones algorithm is given in the next section. This work compares Viola-Jones algorithm with cascade classifier by testing and improving the performance.

This paper is comprised of the following sections. In Sect. 2 deals with not only the technique used to construct both the complete and sub-images of a face, but also the approaches used to extract the face features from the constructed whole and sub-images. The face recognition performance of the present approach for various face image datasets is presented in Sect. 3 and finally Sect. 4 concludes the paper.

2 Data and Methods

The flowchart of the present study is shown in the Fig. 1. Initially, various pre-processing operations are done to the input image to improve brightness and contrast levels. A set of few MM operations were applied, from few fundamental MM opening, MM erosion operations. Morphological reconstruction is used to recover size and shapes of image which got affected during any MM operations. Top-hat and bot-hat operations performed upon the image to identify the human faces easily from imagery.

3 Results and Discussion

The modified algorithm is tested with different imageries with varying size and resolution, images acquired from different cameras and different with dissimilar acquisition dates. The results are shown visual and quantitative assessments comparison in terms of false positives.

Figures 2, 3, and 4 illustrates different input images and result obtained from each of these images. In Fig. 2, cascade classifier couldn't detect the face. In Fig. 4, there are no false positives for the cascade classifier. The results of the proposed approach depends upon the size and quality of the input image provided. If the image is with higher in pixel density and size it will take a more time to process as compared to images with lighter in size. The quality of the image also plays a major role in

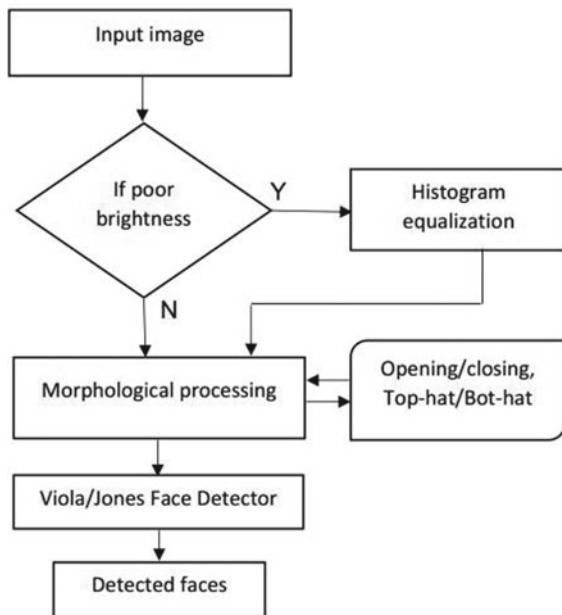


Fig. 1 The proposed face detection method



Fig. 2 Image 1

reducing false positives. The output image shown in Fig. 3 and 5 there are few false positives pixels are found due to its lower quality. In Fig. 5, proposed approach could detect more faces than the cascade classifier. Figure 6 is with higher quality provides good accuracy there is no false positives.

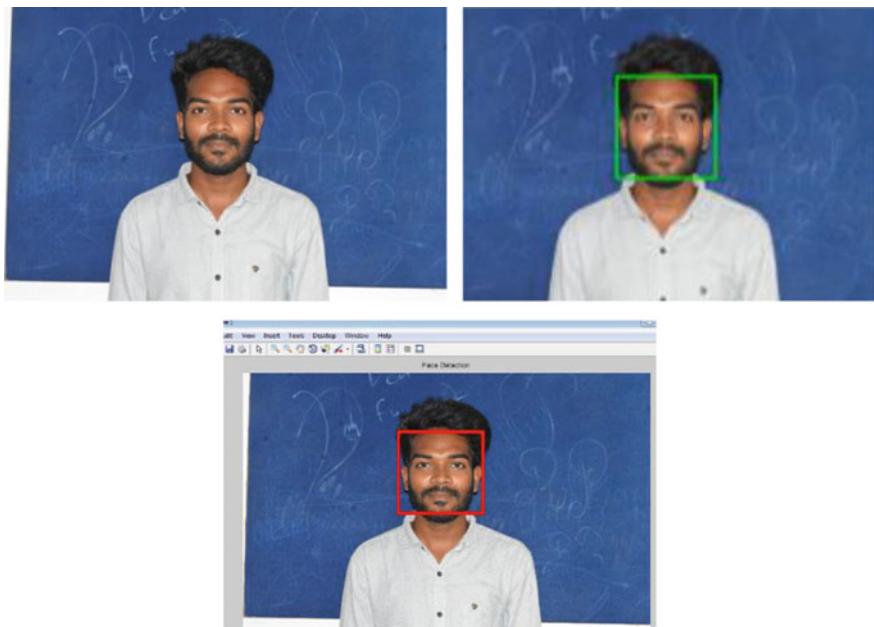


Fig. 3 Image 2

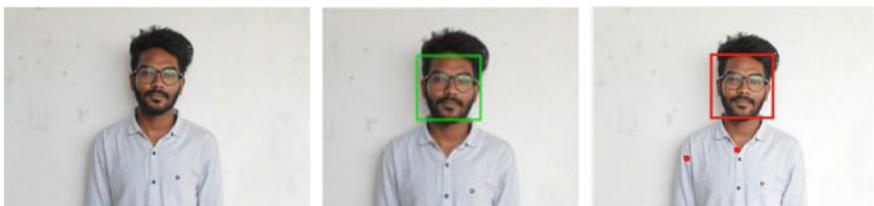


Fig. 4 Image 3

4 Conclusion

Face detection is very important and useful for many image processing and computer vision applications. This paper presents an approach for recognizing face depending image features. New approach combines both morphological image processing techniques and cascade object detector capabilities. This technique has more success for face as compared to cascade classifier.

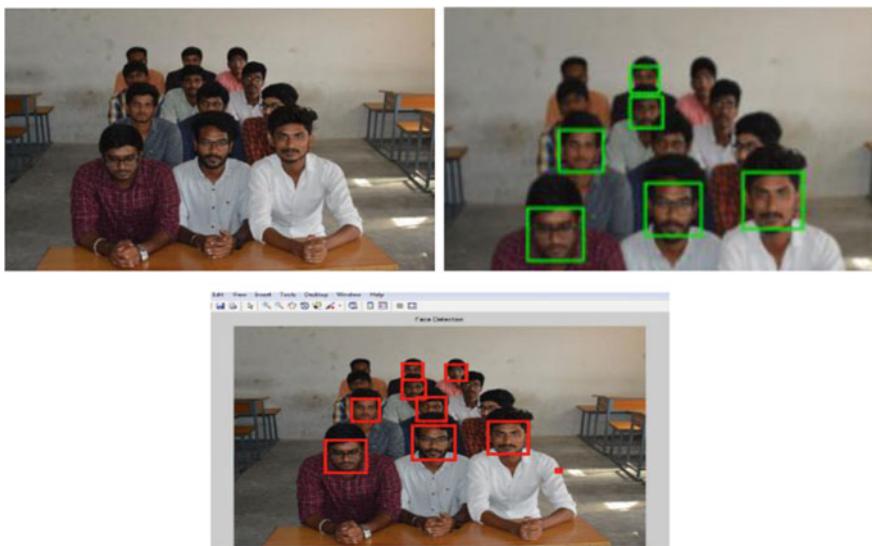


Fig. 5 Image 4, students groups



Fig. 6 Image 5, group of NASA astronauts

References

1. Craw I, Tock D, Bennett A (1992) Finding face features. In: Proceedings of 2nd European conferences on computer vision, pp 92-96
2. Lanitis A, Taylor CJ, Cootes TF (1995) An automatic face identification system using flexible appearance models. *Image Vis Comput* 13(5):393-401
3. Leung TK, Burl MC, Perona P (1995) Finding faces in cluttered scenes using random labeled graph matching. In: Proceedings of 5th IEEE international conferences on computer vision, pp 637-644
4. Moghaddam B, Pentland A (1997) Probabilistic visual learning for object recognition. *IEEE Trans Pattern Anal Mach Intell* 19(7):696-710
5. Turk M, Pentland A (1991) Eigenfaces for recognition. *J Cogn Neurosci* 3(1):71-86
6. Kirby M, Sirovich L (1990) Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Trans Pattern Anal Mach Intell* 12(1):103-108
7. Jolliffe IT (1986) Principal component analysis. Springer, New York
8. T. Agui, Y. Kokubo, H. Nagashi, and T. Nagao, "Extraction of face recognition from monochromatic photographs using neural networks," Proc. 2nd Int'l Conf. Automation, Robotics, and Computer Vision, vol.1, pp. 18.81-18.8.5, 1992

9. O. Bernier, M. Collobert, R. Feraud, V. Lemaried, J. E. Viallet, and D. Collobert, "MULTRAK: A system for automatic multiperson localization and tracking in real-time," Proc, IEEE. Int'l Conf. Image Processing, pp. 136–140, 1998
10. A. J. Colmenarez and T. S. Huang, "Face detection with information-based maximum discrimination," Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 782–787, 1997
11. M. S. Lew, "Information theoretic view-based and modular face detection," Proc. 2nd Int'l Conf. Automatic Face and Gesture Recognition, pp. 198–203, 1996
12. R. Ranjan, A. Bansal, J. Zheng, H. Xu, J. Gleason, B. Lu, ... & R. Chellappa, "A fast and accurate system for face detection, identification, and verification". IEEE Transactions on Biometrics, Behavior, and Identity Science, 1(2), 82–96, 2019
13. Hsu RL, Abdel-Mottaleb M, Jain AK (2002) Face detection in color images. IEEE Trans Pattern Anal Mach Intell 24(5):696–706
14. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: CVPR 2001
15. Viola P, Jones M (2004) Robust real-time face detection. IJCV 57(2)

Efficient Band Offset Calculation Method for HEVC and Its VLSI Implementation



I. Manju, K. S. Srinivasan, E. Rohith Kumar, and R. Haresh

Abstract Sample adaptive offset (SAO) adapted in high efficiency video coding (HEVC) standard is an in-loop filtering technique to reduce mean sample distortion occurring due to artifacts introduced by using larger transforms. The SAO involves classifying each sample into edge and band, calculates and adds appropriate offset values to the samples. In VLSI implementation of SAO, band offset (BO) calculation consumes a comparatively larger area and gate count. Some previously proposed algorithm for reducing the total bands from 32 to lower number of bands to reduce BO area compromises the proper prediction and selection efficiency of the four consecutive band group for which offset is needed to be calculated and signaled. This paper proposes a preselection technique using which from the total number of 32 bands, only 16 bands contributing majority samples in given coding tree unit (CTU) are selected and passed on for further processing. The usage of 16 bands provides sufficient space for making better prediction and selection of the required four consecutive band group along with optimum reduction in the consumed area.

Keywords BD rate · CABAC · CTB · CTU · HEVC · Predecision · RD cost · SAO

1 Introduction

The high efficiency video coding (HEVC) [1] standard is the successor of advanced video coding (AVC) which aims at reducing the bit rate by half the amount of AVC while preserving the video quality. The HEVC standard (Fig. 1) is drafted by the Joint Collaborative Team on Video Coding (JCT-VC), which is established by ISO/IEC

I. Manju · K. S. Srinivasan

Mohamed Sathak A.J College of Engineering, Chennai, India

E. Rohith Kumar (✉) · R. Haresh

Chennai Institute of Technology, Chennai, India

Moving Picture Experts Group (MPEG) and ITU-T Video Coding Experts Group (VCEG).

The major changes made in HEVC include the expansion of the pattern comparison and difference-coding areas from 16×16 pixel to sizes up to 64×64 . HEVC uses up to 32×32 transforms. A larger transform could introduce more artifacts which causes distortion in the reconstructed image at the decoder end. Therefore, in order to reduce the mean sample distortion the sample adaptive offset (SAO) has been introduced as an in-loop filter in HEVC standard. The SAO works by first classifying the given sample of the reconstructed picture broadly into edge and band sample. In edge classification, the sample is classified using predefined four different 2-D edge patterns, and each 2-D edge is further classified into four subclassifications. Therefore, totally 16 classifications of edges are available. The band classification of sample is done based on the magnitude of each sample. Based on magnitude value of samples, the entire magnitude range of 0–255 is classified into 32 bands (band categories). Then, the offset values which produces minimum distortion is calculated individually for all 48 categories (16 edge categories + 32 band categories). The offsets of 4 sub categories of each 2D edge are grouped together (4 edge groups) as edge offset groups and offsets of 4 adjacent band categories are grouped together (29 band groups) as band offset groups and the best group of four offset values among these groups for a given coding tree block (CTB) is chosen. The best group of four offsets are decided using rate distortion (RD) cost. For edge offset, the four offsets corresponding to the subcategories of each major 2-D edge patterns are compared with each other, and the group of four offsets corresponding to a major 2-D edge pattern possessing minimum RD cost together is chosen. For band offset, offsets of four consecutive bands possessing minimum RD cost together are chosen. The works over effective VLSI implementation of the SAO in order to reduce complexity have boosted up since the release of HEVC. A high-throughput pipelined VLSI architecture for sample adaptive offset and de-blocking filter [2] have been proposed which provides a five-stage pipelined architecture. It supports $4\text{ K} \times 2\text{ K}$ at 60 fps. Zhu et al. [3] have proposed changes of using bitmaps in statistics collection stage and to find offsets directly, avoiding iterations of multiple values in order to determine best offset. The method proposed by Chen et al. [4] of considering only seven band groups for finding best four consecutive band offsets based on the presence of the band which contains most sample of given CTU which has reduced the search space from 29 to 4 but significantly compromising the ability to choose the best band group.

The work by Zhou et al. [5], which proposed a dual-clock architecture for SAO utilizing the difference of critical path between statistics collection and parameter determination stages, has contributed an significant reduction in area of VLSI architecture of SAO to about threefold, and they also contributed some of the secondary optimizations in reducing BO calculation and SRAM.

The above methods even though contributed significant reduction in area of SAO architecture and BO classification part, and they possess some compromises in accurate prediction of best band offsets predicted by original BO calculation algorithm.

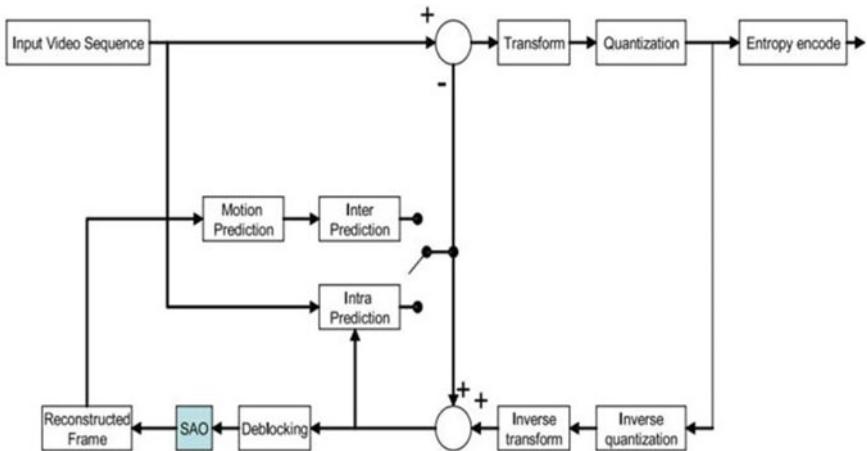


Fig. 1 Block diagram of HEVC encoder diagram

In this paper, a band offset calculation method for HEVC with good prediction efficiency possessing optimum reduction in area has been discussed.

2 SAO Implementation in HEVC

Sample adaptive offset has been implemented in two stages, namely statistics collection (SC) and parameter determination (PD). The SAO parameters are being calculated for luma and chroma samples sequentially one after another and are being signaled along with the bit stream for each coding tree block (CTB). The SAO parameters have been designed in HEVC such that they contribute a very less side information for signaling and optimum line buffer requirements, on consideration of minimization of the bit rate and buffer requirements.

2.1 Statistics Collection

The statistics collection block performs the processes of classifying the samples into 48 classifications (16 edges and 32 bands), calculating the gross error present in the samples of each category and counts the number of samples belonging to each category. The gross error of each category is termed as "Sum or S," and the count of number of samples belonging to each category is termed as "Count or C." Then, the calculated Sum and Count of each category is forwarded to the parameter determination stage.

2.2 Parameter Determination

The parameter determination stage on receiving the sum (S) and count (C) of each category finds value S/C for that category. It then iterates some values between 0 and S/C as offsets and finds the respective distortion for that offset value using Formula (1).

Then, the RD cost of each offset value is found by Formula (2).

$$\text{Distortion} = \text{Count} * \text{offset} * \text{offset} - 2 * \text{offset} * \text{Sum} \quad (1)$$

$$\text{Cost} = \text{Distortion} + \lambda * \text{rate} \quad (2)$$

where “rate” is the number of bits to code the parameters calculated from CABAC results, and λ is the Lagrange multiplier used in RD cost calculation. The offset value possessing the lowest RD cost is selected as the offset of the given category. After calculating the offset value for each 48 categories, the gross RD cost [6] for each 2-D edge and the best four consecutive bands are calculated. Then, these five groups of offset values (four 2-D edge groups and one set of best four consecutive bands) are compared, and the group possessing the lowest RD cost is chosen as the best one. After choosing the best group of offsets for each CTU, there are three options—(1) Reuse SAO parameters of the left CTU by setting the syntax element sao-merge-left-flag to true, (2) reusing SAO parameters of the above CTU by setting the syntax element sao-merge-up-flag to true, and (3) or transmitting new SAO parameters. When the current CTU selects SAO merge-left or SAO merge-up, all SAO parameters of the left or above CTU are copied for reuse, so no more information is sent. This CTU-based SAO information sharing is mainly aimed at reducing side information effectively. In some improvements made in [2, 3], it was proposed using the rounded value of S/C as offset value of corresponding category without iterating with multiple values which is time and resource consuming. They also propose using distortion as parameter to select best group of offsets rather than RD cost to select best group of offsets. These two proposals are followed and implemented in this work.

2.3 Edge Offset

In HEVC in order to reduce the code and time complexity, only four 2-D edge patterns representing four classes for EO (edge offset, EO_0: horizontal; EO_1: vertical; EO_2: diagonal 135; EO_3: diagonal 45) are used. There are four subcategories for each of EO class as shown in figure. Therefore, edges are classified totally into 16 types. Then, the offsets are calculated for each of the 16 edge classifications, and the corresponding distortion and RD cost for each of those classifications are calculated. After this, the RD cost of the four subcategories of each 2-D edge is added in order to find the gross RD cost of each 2-D edge. Then, the 2-D edge which possesses the

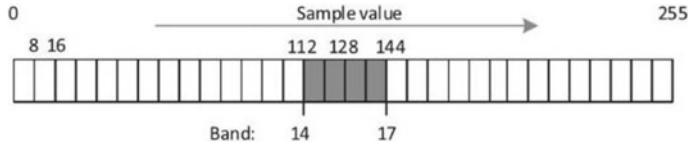


Fig. 2 Band classifications of entire sample range

lowest RD cost is selected, and the offsets of its four corresponding subcategories are forwarded for the successive stages.

2.4 Band Offset

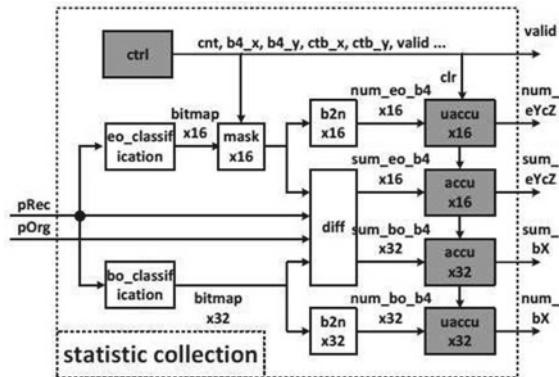
In HEVC for band offset calculation, the samples of a given CTB are classified into one of the 32 bands based on the targeted sample value. For example for sample range of 0–255, the samples with magnitude 0–7 are grouped to band 1, samples with magnitude 8–16 are grouped to band 2, and so on (Fig. 2). The entire range of the sample values is grouped evenly into 32 bands by grouping a fixed number of band values into a single band. Then, the offsets are calculated for each band group, and the corresponding distortion and RD cost are calculated for each band. The RD cost of a band group considering four consecutive bands is calculated by adding the individual cost of the consecutive four bands considering only four consecutive bands at a time. Therefore, this gives rise to 29 band group combinations of four consecutive bands. Then, best band group among all 29 possible band group combinations is found as the one which possesses lowest RD cost. Then, the offsets of this best four consecutive bands are being forwarded for the successive stages.

2.5 Hardware Implementation

The hardware implementation of statistics collection block using $4 * 4$ bitmaps for each 32 band classification and 16 edge classification to achieve 16 samples parallelism is proposed by [2].

Here at each cycle, 16 samples are taken for processing simultaneously. There will be 32 BO and 16 EO classification blocks—each use individual $4 * 4$ bitmaps to map position of samples in their respective classification out of 16 samples considered for processing. Here, the presence of 1 in bitmap in particular position represents that sample belongs to that classification and 0 represents that sample not belongs to that classification. The difference between original and reconstructed pixels is stored in a separate $4 * 4$ storage array. Here, the count of samples belonging to particular classification is obtained by adding number of 1's in the bitmap of particular classification and is accumulated in count accumulator. The “S” (Sum) of a classification

Fig. 3 Hardware implementation of statistic collection using bitmaps for 16-bit parallelism



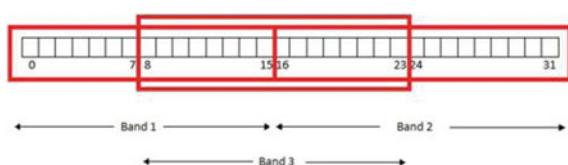
is obtained by multiplying the position bitmap array and difference storage array by corresponding positions and adding the elements in the resultant matrix.

3 Proposed Algorithm

In images, mostly CTU's containing smooth regions without edges consists of sample values mostly distributed over only 4–12 bands, whereas it may extend above 20 bands for a CTU containing edges. Therefore, in statistics collection phase for the band offset calculation, it is enough if we collect statistics of only a part of the bands where the image samples are concentrated more and eliminating the calculation of the remaining bands which contains no significant amount of samples to consider (Fig. 4).

In the proposed method, the 32 available bands are grouped into three broad groups in which each broad group contains 16 bands as shown in the figure (Fig. 3). In order to choose the best broad group among these three broad groups, a preselection stage is used before the statistics collection stage for band offsets. This preselection stage counts only the total number of samples present in these three broad band groups and chooses the band group of 16 bands containing maximum number of samples. Then, the sum and count are calculated only for those 16 individual bands present in the selected broad band, and the calculation for remaining 16 bands is eliminated. This proposed method eliminates unnecessary calculations for those unwanted bands containing very low number of samples, which reduces the amount of data to be

Fig. 4 Proposed preselection of bands containing most samples



processed for statistic collection stage reducing the circuitry required for processing it. It also additionally reduces the storage buffer, accumulators, adders, multipliers, and memory requirements significantly proving to an efficient hardware implementation of the algorithm. In this method, the distortion is used as the parameter to choose the best four offsets as proposed in [2].

4 Experimental Results

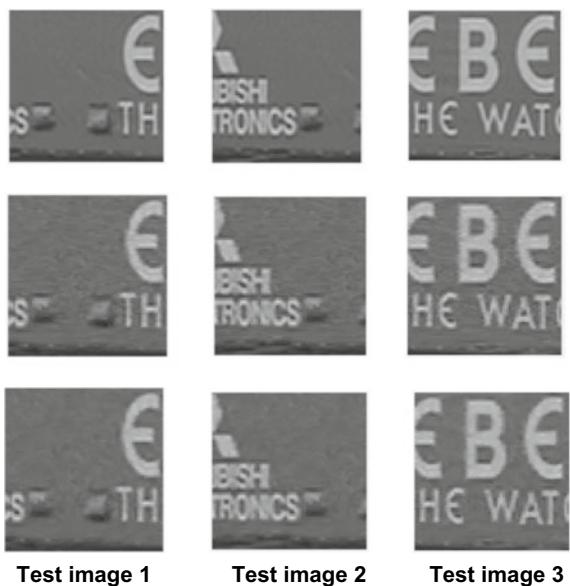
The existing and proposed algorithm for selection of band offset selection has been implemented using MATLAB, and the band offsets selected for each individual CTB's in the given image were compared to obtain the following results (Table 1; Fig. 5).

On comparing the results of the original BO algorithm and proposed BO algorithm, the proposed algorithm shows good efficiency for the CTB's containing smooth regions without significant edges. In this work, the prediction efficiency of the proposed work is tested with different kinds of input images, and the results are tabulated. The original algorithm always chooses the band group which produces the least distortion among all the 26 available band groups of four consecutive bands. Here, the proposed algorithm shows good efficiency when the best band group of four consecutive bands calculated by proposed algorithm is allowed to be one among the top three band groups producing the least distortion among the 29 available band groups calculated by the original algorithm. The band group calculated by proposed algorithm is allowed to be anyone of the top three band groups producing least distortion calculated by original algorithm because we are restricting the search range of bands only to 16 from 32 due to preselection, so there will be some restrictions in calculating the exact top one band by the proposed algorithm. So the top second and top third bands are also considered. Also the proposed algorithm works best for CTB's containing smooth regions without significant edges. The proposed algorithm

Table 1 Calculation efficiency of proposed band offset algorithm with respect to original algorithm for various inputs

Test video sequence	No. of CTB's predicted within top 3 band groups by proposed algorithm out of 20 CTB's	Prediction efficiency of proposed algorithm (%)
Soccer.yuv	17	85
Flower.yuv	18	90
Coastguard.yuv	17	85
Hall.yuv	13	65
Bus.yuv	13	65
Foreman.yuv	11	55
Stephan.yuv	9	45

Fig. 5 Top row consists of original input image, middle row consists of reconstructed images, and bottom row consists of the decoded images after applying band offset selected by proposed algorithm to reconstructed picture. *Image source* <http://trace.eas.asu.edu/yuv>; image: stephan.yuv



mostly predicts outside the top three bands when the CTB contains more edges than smooth regions. It is due to the fact that the regional samples usually be concentrated over a particular range of bands only, and it can be found out by the preselection stage easily. But for CTB containing edge samples, the sample distribution will not be concentrated but will be randomly distributed over the entire 32 bands. Therefore, the prediction by preselection stage is not efficient here. But this defect is negligible because for a CTB containing edge samples, the RD cost of band offsets will be higher than cost of edge offsets or the distortion is less when edge offsets are selected. Therefore, naturally these defective band offsets are mostly eliminated by selecting corresponding edge offsets for those CTB's by the parameter determination block of SAO. Therefore, the proposed algorithm provides good efficiency for predicting the band offsets.

Table 2 Hardware requirement for proposed band offset algorithm with respect to original algorithm

Element	No. of hardware elements required for original algorithm	No. of hardware elements required for proposed algorithm
BO classification stages	32	16 + 3 (for preselection)
EO classification stages	16	16
4 * 4 bitmaps for count	32	16 + 3 (for preselection)
4 * 4 bitmap multiplier	32	16
Sum accumulator	32	16
Count accumulator	13	16 + 3 (for preselection)

When the original BO block and the proposed BO block have been synthesized as ASIC or FPGA, the amount hardware elements utilized have been reduced as shown in Table 2. Therefore as shown in Table 2, the proposed BO algorithm utilizes less elements in ASIC and FPGA when compared to original BO algorithm due to reduction of bands taken for processing from 32 to 16 (Table 2).

5 Conclusion

Therefore, the proposed algorithm proves to be efficient for calculation of band offsets for images with CTB's containing smooth regions without significant edges. Even though the prediction by preselection stage is futile for CTB's with edges, this defect is negligible since band offsets are chosen as best offsets only for smoother regions without edges. Also as discussed, the proposed approach reduces the amount of data needed to be processed alongside reducing the hardware requirement in case of VLSI implementation in FPGA or ASIC.

References

1. Sullivan GJ, Ohm J-R, Han W-J, Wiegand T (2012) Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans Circuits Syst Video Technol* 22(12):1649–1668
2. Zhu J, Zhou D, Kimura S, Goto S (2014) Fast SAO estimation algorithm and its VLSI architecture. In: Proceedings of IEEE international conference on image processing (ICIP), Oct 2014, pp 1278–1282
3. Zhu J, Zhou D, Kimura S, Goto S (2014) Fast SAO estimation algorithm and its implementation for 8 k × 4 k 120 FPS HEVC encoding. *IEICE Trans Fundam Electron Commun Comput Sci* E97-A(12):2488–2497
4. Zhu J, Zhou D, Wang S, Zhang S (2017) A dual-clock VLSI design of H.265 sample adaptive offset estimation for 8 K ultra-HD TV encoding. *IEEE Trans Very Large Scale Integr (VLSI) Syst* 25(2). <https://doi.org/10.1109/tvlsi.2016.2593581>
5. Chen G, Pei Z, Liu Z, Ikenaga T (2010) Low complexity SAO in HEVC base on class combination, pre-decision and merge separation. In: IEEE 19th international conference on DSP. <https://doi.org/10.1109/icdsp.2014.6900840> (to be published). Liao K, Yang J (2010) Rate-distortion cost estimation for H.264/AVC. *IEEE TCSVT* 20(1). <https://doi.org/10.1109/icip.2014.7025255>
6. Shen S, Shen W, Fan Y, Zeng X (2013) A pipelined VLSI architecture for Sample Adaptive Offset and deblocking filter of HEVC. *IEICE Electron Express* 10(11)

Clinical Skin Disease Detection and Classification: Ensembled VGG



Gogineni Saikiran, G. Surya Narayana, Dhanrajnath Porika,
and Gunjan Vinit Kumar

Abstract Our research work has succeeded in integrating ensembles into VGG image classification technique aiming higher accuracy and performance than existing models. The convolutional layers are apt for feature extraction from images. In VGG classifier three fully connected dense layers are used for classification of class from these outputted features. But we have integrated ensemble methods immediately after convolutional layers for purpose of better classification output. Thus, the output (features of image) of convolutional layers is passed as a separate input to both ensemble methods and fully connected layers of VGG for obtaining the class of image. Final class of image is determined by specific strategy after analyzing outputs of ensemble and VGG fully connected layers. All earlier works focused on skin disease classification. Here, we have also experimented with yolo for detection of location and class of diseases. Skin is considered as the most significant part of the body. But this most significant part of the body is easily subjected to various kinds of diseases that spread throughout skin at a faster pace. Early detection and prevention are needed. Our research work aimed at detecting top 10 common skin diseases with higher accuracy. User can upload a pic in a mobile or cloud application and inbuilt AI algorithms will detect the type of skin disease with higher accuracy and thus offering prevention suggestions at an early stage without doctor intervention.

Keywords Yolo · VGG · Object detection · Bagging · Boosting · Random forest · XGBOOST

G. Saikiran (✉) · G. Surya Narayana · D. Porika
CMR College of Engineering and Technology, Hyderabad, India
e-mail: goginenisaikiran31677@gmail.com

G. Surya Narayana
e-mail: gsuryanarayana@cmrcet.org

D. Porika
e-mail: porikadhanrajnath@gmail.com

G. Vinit Kumar
CMR Institute of Technology, Hyderabad, India
e-mail: vinitkumargunjan@gmail.com

1 Introduction

The classification of disease must be more accurate than classifying cars or dogs because as a result of classification associated medicine is suggested. Intake of wrong medicine may lead to death sometimes. So, classification in this case became more important. So, we decided to carry on further research in achieving higher accuracy compared to existing classifiers in the task of classification. We have chosen ensembles as a means to achieve our aim. Ensembles constitute of bagging and boosting techniques. Later we also experimented with yolo to perform object detection in case of skin diseases. There are hundreds of skin issues that affect humans. However, every skin issue is identified with certain symptoms. Skin is considered as the most significant part of the body. It discharges its responsibilities such as maintaining body moisture, body temperature regulation, ultraviolet protection, immune system, and so on. It also protects the body from various viruses and bacterial attacks. It acts as a shield by preventing us from the sun's rays. Skin produces vitamin D (essential for the vital body functions) when exposed to the sun. It forms a boundary separating external environments. But this most significant part of body is easily subjected to various kinds of diseases that spread throughout the skin at a faster pace. Early detection and prevention are needed. References [1, 2] used support vector machines to classify skin cancer images. But we understood machine learning algorithms are not powerful in context of complex data such as images. References [3, 5, 6] took a step ahead and used artificial neural networks to extract features and to classify images. But convolutional layers are a wise choice to deal with images rather than fully connected layers. Chaithanya Krishna and Ranganayakulu [4] haven't succeeded in obtaining a decent score on skin disease classification as he used clustering techniques to perform classification. Ahmed and Jesmin [7] used data mining concepts to classify skin diseases. References [8, 9] used different types of deep convolutional architectures to classify images. They succeeded in achieving a decent score in performance. In contrast, here we are proposing an ensemble architecture which considers different architectures decisions and will declare output on careful analysis. We succeeded in implementing an ensembled VGG without much rise in parameters. References [10, 11] were the base for our paper as they have implemented object detection and segmentation tasks efficiently. All previous work has focused on skin disease classification and none worked on skin disease detection. Here we report the experimental results of skin disease detection using various advanced object detection techniques. We have considered the following 10 skin diseases to carry on further research work as they are most common and typically dangerous. They have to be detected in time with the use of proper technology.

1. Accessory Nipples
2. Anetoderma
3. Bowens Disease
4. Chicken pox
5. Coxsackie
6. Degos Disease

7. Fixed Drug Eruption
8. Malignant melanoma
9. NecrobiosisLipoidicaDiabeticorum
10. NevusSpilus.

Skin diseases are associated with significant impairment in the quality of a patient's daily life if not cured properly in time. They also result in decreased self-esteem, poor relationships, stress, anxiety, suicidal tendency, depression, lower mean cortisol levels, restricted clothing choice, and even psychological functioning disorders. Skin Disease statistics for Australia are illustrated below. It's no more widely different for other countries and almost similarly applicable to all nations with a little change. Some countries face worse situation than described statistics.

- 2 male people died per 150,000 population in India 2014, because of skin diseases as per (statistics of diseases in India from 2010 to 2014)
- 2 female people died per 140,000 population in India 2014, as per (statistics of diseases in India from 2010 to 2014)
- 3 male people died per 120,000 population in India 2013 (statistics of diseases in India from 2010 to 2014)
- 213,566 patients spend their daily days in public and private hospitals as a result of viral skin diseases and subcutaneous skin tissue diseases in India 2011–12 as per (statistics of diseases in India from 2010–2014)
- The number of anti-aging and cosmetic treatment procedures is expected to reach a high count and are going to involve high cost.

2 Data and Proposed Work

The data has been split into training and test data sets. Because of the availability of limited images, validation set is not considered. In addition to dataset collected from kaggle and university of iowa health care, various skin disease images are collected from the internet. Out of entire data, 80% is taken for training and rest 20% of data is taken for testing purposes. Image augmentation technique is adapted to increase data. Pytorch and tensorflow frameworks are used to code.

Experimentation: Detection With Yolo

Two types of works are done. As a part of the experimentation first part is focused on the detection of skin disease in an image using yolo. As a result of research work, the second part of the paper is focused on integrating ensembles at different layers of vgg architecture. Among the several object detection techniques, yolo is preferred as a result of its capability to achieve significant accuracy and to track tiny and large objects in an image at one go. But also, the results and analysis of other several object detection techniques such as RCNN, FASTER RCNN, SSD are also presented.

Object detection in an image is quite simpler and straight forwarded with yolo. Yolo is definitely fast. The fast version of yolo will run at more than 150 frames per

second and the base version of yolo will run at nearly 45 fps. Fps stands for frames per second or relative speed of the model in processing number of images per second. The detection process of yolo is focused on detection of tiny resolution objects also. So, yolo increased the input size of the image to from 224×224 (preferred for vgg) to standard 448×448 . The architecture resizes the given image into 448×448 sized image and runs a mixture of convolutional, pooling layers on it and finally nonmax suppression to obtain final output. The entire resized image is divided into $G \times G$ grid and each grid cell predicts B bounding boxes, confidence for those boxes, and C class probabilities if assumed there are total C classes. The yolo architecture outputs a tensor of size $G \times G \times (B \times 5 + C)$. B is a bounding box. Each and every bounding box has five parameters associated namely x, y, w, h coordinates, and confidence of class. The confidence of a class is probability value that lies between 0 and 1 and ensures how confident the bounding box is, regarding presence of an object in it. The (x, y) point represents center of the box. The height and width are predicted in association to the complete image. Finally, the confidence outputted will represent the intersection over union between the predicted box and ground truth of the object. The probability (class i object) is calculated for every grid of image. Only one set of this probability is calculated per each grid regardless of the number of boxes each grid holds. The final layer of this architecture is responsible for prediction of object coordinates and class of object. We generally scale the anchor box height and weight by input image height and weight so that they are always constrained to appear within zero and one. Linear activation function is used for the last layer of yolo system that is involved in delivering the final output coordinates and class probabilities. All the intermediate layers of the system use leaky rectified linear activation function which is described below.

$$f(x) = x, \text{if } x > 0 \text{ else } (0.1 * x)$$

Sum squared error is used as a metric to measure coordinate regression loss and thus to mitigate loss by back propagating. But this sum squared metric weights localization error and classification error equally which is not highly suitable or preferred. It pushes the class score of a grid to zero if that particular grid doesn't hold any object. This can cause high instability as it overpowers the gradient from these grids. In order to deal with this, the yolo introduces two new terms ($\lambda_{\text{noobject}}$ and $\lambda_{\text{coordinate}}$) to increase error from (x, y, w, h) predictions and eventually mitigate loss from label predictions for predicted (x, y, w, h) bounding boxes that doesn't hold objects.

The network architecture has twenty-four convolutional layers (composed of kernel weights) that are followed by two fully dense or connected layers. This simple architecture preferably uses 1×1 kernel weights followed by kernel weights of 3×3 in contrast to inception model of google net (Fig. 1).

For skin disease detection problems, we have decided to work on 10 different diseases as mentioned above. So, the number of classes $C = 10$. We used $S = 7$ grid size and $B = 2$ number of bounding boxes or anchor boxes. The final output tensor of yolo architecture is $(7 \times 7 \times 20)$. $(7 \times 7 \times 20)$ can be written as $(7 \times 7 \times 2 \times 5$

$$\begin{aligned}
& \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[(x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
& + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left[(\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{obj}} \left(C_i - \hat{C}_i \right)^2 \\
& + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^B \mathbb{1}_{ij}^{\text{noobj}} \left(C_i - \hat{C}_i \right)^2 \\
& + \sum_{i=0}^{S^2} \mathbb{1}_i^{\text{obj}} \sum_{c \in \text{classes}} (p_i(c) - \hat{p}_i(c))^2
\end{aligned}$$

Fig. 1 The total loss function of yolo, a combination of regression coordinates loss and classification loss is described above with parameter set

+ 10). Total there are 49 grids in an image and 20 predictions associated with each and every grid. These 20 predictions for each grid include [pc, bx, by, bh, bw] for 1st anchor box +[pc, bx, by, bh, bw] for second anchor box +[c1, c2, c3, c4, c5, c6, c7, c8, c9, c10], the class confidence of an object in that grid. pc will be 1 if the corresponding anchor box contains object else 0 if it doesn't contain the object. If pc is zero then other elements of vector such as x, y, w, h coordinates can be simply don't care. If the particular grid contains anetoderma disease symptom, then class confidence vector is [0, 1, 0, 0, 0, 0, 0, 0, 0, 0] and if grid contains disease malignant melanoma then ground truth confidence vector will be [0, 0, 0, 0, 0, 0, 0, 1, 0, 0] and ground truth box coordinates vector can be assumed as [1, 0.2, 0.6, 0.6, 0.8]. For each grid cell get two predicted bounding boxes. Get rid of low probability predictions. The aspect ratios and sizes of anchors are to be decided in consideration with types of objects being detected. The anchor used for the prediction of pedestrian can't be used for prediction of a car. We are generally aware of the fact that humans can be fitted into vertical boxes rather than square boxes. However, the initial yolo versions haven't taken this fact into consideration.

The prediction is always done without any assumptions on the shape of target objects. But YOLOV2 put some constraints on anchor sizes taking into consideration of target object sizes. This has driven model by a significant increase in performance.

anchors = [(1.8744, 2.0625), (0.5727, 0.67738), (3.3384, 5.4743), (9.7705, 9.1682), (7.8828, 3.5277)]

For anchor pair (1.87446, 2.06253)

Width = 1.87446 × 32 = 59.98 pixels

Height = 2.06253 × 32 = 66.0 pixels.

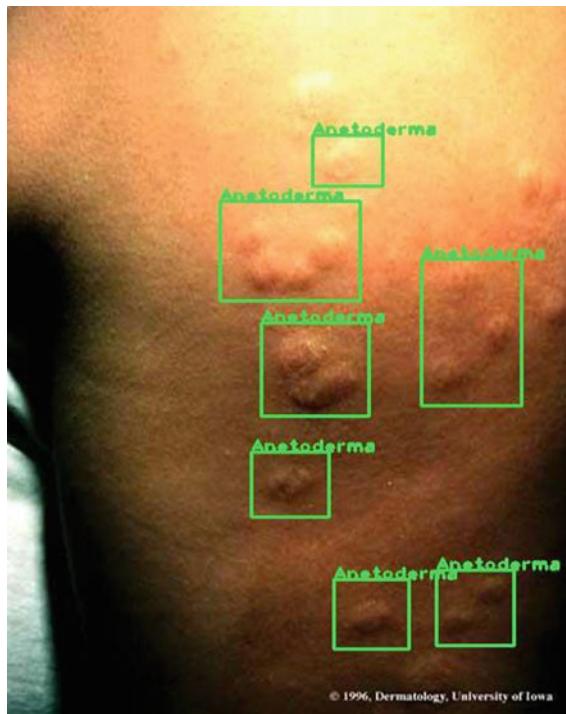


Fig. 2 Yolo algorithm detecting anetoderma disease locations

In above-illustrated example, (0.5727, 0.67738) are coordinates of one of the anchor boxes. They usually represent the height and width coordinates of that particular anchor box. The above list represents five different anchor boxes. Remember that, these are always chosen in consideration with output object shape. YOLOv2 generates a 13×13 output tensor. So, you can obtain actual values by performing multiplication with 13 by anchor box coordinates. In yolov2 the image $416 * 416$ is divided into 13 grids where $G = 13$ instead of 7 as in yolo. Yolov2 divides the image of size into $13 * 13$ grid each containing 32 pixels (Figs. 2, 3, 4, 5, and 6).

3 Results of Detection

The mean average precision is slightly higher in Fast RCNN using RESNET backbone classifier. However, yolo v2 also achieved comparable performance on skin diseases image dataset. Flips per second are higher in yolov2 than any other object detection algorithm (Table 1).



Fig. 3 Yolo algorithm detecting accessory nipples disease locations



Fig. 4 Yolo algorithm detecting chickenpox disease locations

4 Image Classification

Over the past few years, deep convolutional layers are capable of performing both detection and classification tasks successfully in the domain of computer vision. Indeed, as each year passes, the performance of these classification systems is increasing significantly with an increase in depth of layers. Priority is always given

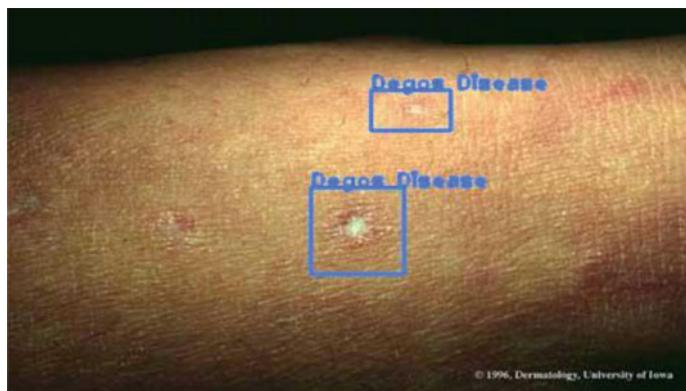


Fig. 5 Yolo algorithm detecting Degos disease locations

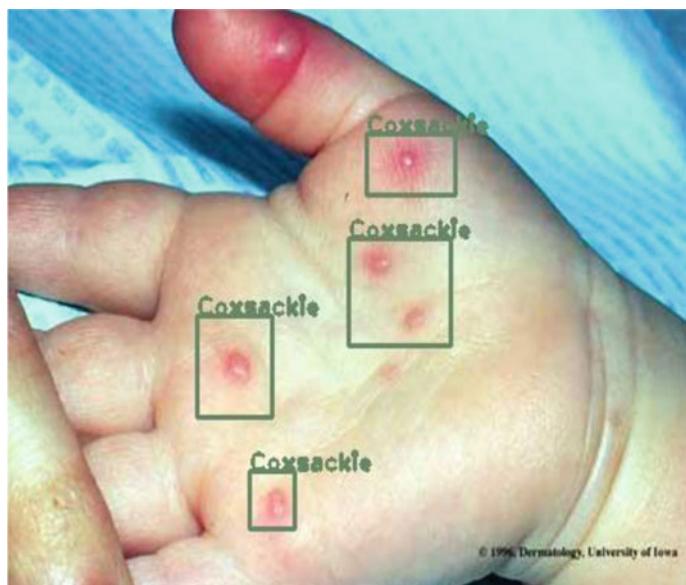


Fig. 6 Yolo algorithm detecting Coxsackie disease locations

to performance of the model regardless of depth of layers and high computational space it holds.

VGG16 vgg model is an improvement over Alex-Net. The difference is vgg replaces large-sized kernel weights with 3×3 kernel weights. The significant improvement was the increase in depth of the model. Standard 224×224 images is passed as input to standard vgg model, it is taken through five blocks of deep convolutional layers. Here, every block consists of 3×3 kernel weights in an increasing

Table 1 Illustrating performance of various object detection algorithms on skin disease dataset

Algorithm	mAP	FPS
YOLO448 × 448	62.8	45
YOLO V2416 × 416	75	67
FAST RCNN	70.1	0.5
FAST RCNN-VGG	72.7	7
FAST RCNN-RESNET	75.8	5
SSD 300	73.7	46
SSD 500	74.6	19

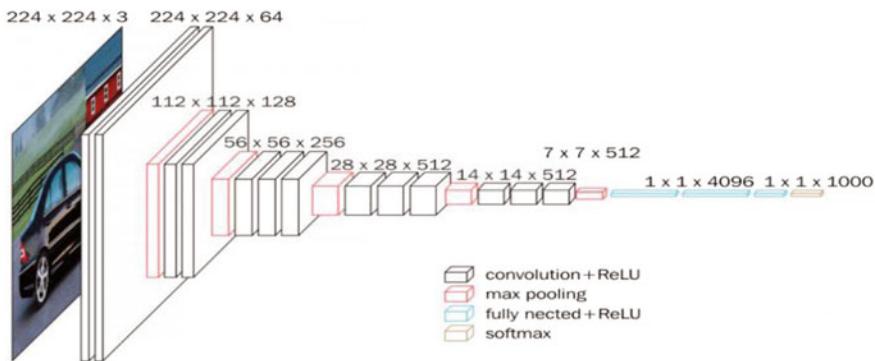


Fig. 7 VGG16 architecture

fashion. Stride is taken as 1. Deep convolutional layer inputs are actually padded so that, input size of the image is not altered even after convolution operation is performed in certain blocks of VGG16. The five blocks of VGG16 are separated by maxpooling layers. These maxpooling layers are capable of feature extraction without involving weights or filters.

Maxpooling is performed with stride 2. Five blocks of deep convolutional layers are usually followed by three dense layers or fully connected layers for the sake of classification (Fig. 7).

5 RESNET

This network model is based on the concept of deep residual block, proposed in recent research paper. It uses short cut connections to improve model performance. Resnet-k means deep residual network consisting number of layers. For example, ResNet-50 means Resnet consisting of total 50 layers. The issue with deep networks is the training error increases with an increase in depth. As a result, performance decreases. But RESNET provides a mechanism through which only performance

increase with an increase in depth but not training error. The current layer output is passed as concatenated input to the layer ahead of the next two layers. The phenomenon continues to happen throughout the model.

6 Integrating Ensemble into VGG16

Here we come up with significantly more accurate Convolutional Neural Networks. They not only achieve the state-of-the-art accuracy on skin disease classification and localization tasks but are also applicable to other custom image classification datasets, where they achieve more accurate performance compared to existing models. During the training phase, the input to our convolutional networks is a standard 224×224 image.

Ensemble An ensemble is a collection of separately trained learners (might be a group of decision trees or simply neural networks) whose decisions are combined in order to classify an instance. All the earlier research work has clearly proven that an ensemble is generally more powerful and correct when compared to individual learners. Bagging technique is always more correct than a single classifier. But it is sometimes, much less correct compared to Boosting. These techniques depend on “resampling or sampling with replacement” strategy to generate different training data sets from actual data to different learners.

Bagging Each individual learner’s training data is obtained by drawing data points randomly from D data points with replacement, where D is the size of actual training data. Many of the instances in actual training data may occur repeatedly in the construction of resultant training data while few of them are left. One classifier is independent of another classifier.

Boosting The aim of boosting is to actually generate a series (one after another) of dependent classifiers. Training data required for the classifier of the series is obtained as per the performance of the previous learner on dataset in the series. In this method instances that are incorrectly classified by earlier classifiers in the series are chosen more often than instances that were correctly classified. Thus, Boosting is capable of producing new excellent classifiers that are better able to classify instances for which the current classifier performance is poor.

1. The bias concept measures the closeness of classifier produced by the learning algorithm to the actual target function required to map input to output.
2. The variance term measures the disagree between classifier produced by learning algorithm and actual target function. It is a measure of how their decisions differ.

With the help of the above strategy, the bagging and boosting models try to bring down both bias and variance. Many scientists feel that boosting actually attempts to decrease the miss classification rate or error in the bias term as it is focused on miss classified examples. However, the same focus can force the individual learner to produce an ensemble that differs highly from the actual single learner. Bagging can also be used to mitigate the error but highly useful in reducing the variance (Fig. 8).

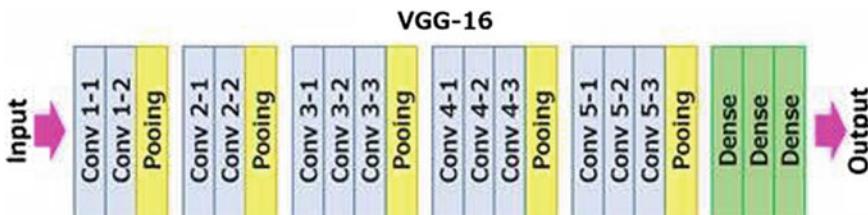


Fig. 8 VGG16 block diagram

The first five blocks of VGG16 are convolutional layers that are meant for extracting features from the images. The last (sixth block) dense block is meant for classification of class from the extracted features. We have considered the output of fifth block as the required significant outcome, which typically contains extracted features of the image. The output tensor of conv 5-3 is of shape $7 \times 7 \times 512$. It is flattened simply to a vector of size $[1 \times 25,088]$ or $[25,088]$. Now we will integrate ensemble after this conv 5-3 layer. Many classifiers are immediately built after this layer to achieve excellent accuracy. One of the classifier will be vgg dense block itself. Other classifiers will be decision trees or ADABOOST or simply yet powerful another neural network.

The issue is, the input vector of size 25,088 is fine for dense block of vgg and another neural network classifier but not aptly suitable for decision trees because of large input space which are no more discrete bit highly continuous. There must be some strategy to bring down this large input space to medium input space without losing any significant information. So, we used PCA to achieve this task. The main idea of principal component analysis (PCA) is to reduce the dimensionality of training set composed of many features usually correlated with one another. But it aims at, retaining or preserving the variation present in data set to a possible extent. It is done by transforming the features of original data set to a new feature space, known as the principal components. These principal components are orthogonal, ordered such that the retention of variation present in the original features reduces as we move down in the order. So, in this manner, the $k + 1$ th principal component retains minimum variation that was present in the original features compared to k th principal component. These principal components are known as eigenvectors of covariance matrix which are usually orthogonal.

Importantly, the training set on which PCA technique is to be applied must be normalized or scaled properly. Always the results are sensitive to the normalization applied to data set. In the dataset, normalization is done by subtracting the column mean from the respective numbers of that column.

If we consider $x_1, x_2, x_3, \dots, x_n$ as features of dataset. These features are typically output of VGG16 conv5-3 layer of an image with $n = 25,088$. Since we are dealing with the dataset containing 25,088 features, the PCA will construct a square matrix of size $25,088 \times 25,088$. The above is the covariance matrix. Kindly remember that always $\text{var}[x_n] = \text{covariance}[x_n, x_n]$. Further, we need to calculate the Eigenvectors and Eigenvalues of the above covariance matrix from equation $\det(\lambda I - w) = 0$ where

I is identity matrix. For each eigenvalue, a corresponding eigenvector is calculated using the equation $(\lambda I - w)v = 0$. ***Out of n Eigenvalues, we choose some d values to reduce feature dimension***

$$w = \begin{pmatrix} 2, 0 \\ 0, 0 \end{pmatrix} \quad (1) \quad \begin{pmatrix} 1, 2 \\ 0, 0 \end{pmatrix} \quad \text{or } \begin{pmatrix} 1, 0 \\ 0, 1 \end{pmatrix}$$

Consider the skin diseases image dataset contains N images. If there are E epochs, for each epoch there will be I iterations.

For each iteration, a batch size of images is processed in VGG16. Instead of parallel processing, we adapt to serial processing.

Suppose there are N images, in the Table 2, each row indicates conv5-3 layers output of VGG16 when that particular image is given as training input. It is directly given as input to another neural network (classifier in ensemble) without pca. However, this neural network performance will vary as it uses another random weight initialization method. But, pca is applied before it is supplied as input to the decision tree. PCA has reduced 25088-dimensional data space to much smaller dimensional space. There is loss in information but not much significant loss. Though there is a loss in information, the new set of features will be able to differentiate one category of image from the other. Here Batch wise processing is not needed. While VGG16 training is going on, the last convolutional layer output of all images in the last epoch is stored as separate data frame. This data frame can be input to the rest of classifiers in the ensemble. The output of last convolutional layer in the last epoch will be highly stable as it already captured a lot of features of the images as training is near to end. Classifiers working on this stable input are expected to yield much more accurate results. The class of the image is individually decided by the classifiers but the final class of the image in the test set is determined by either voting method or weighted saying method as we need to analyze all the outputs generated by all the classifiers of the ensemble (Fig. 9).

While VGG16 training is going on, the last convolutional layer output of all images in the last epoch is stored as a separate data frame. This data frame can be

Table 2 Illustrating output of VGG16 block5 before and after applying PCA

CONV5-3 layer output of VGG16 before PCA and CONV5-3 layer output of VGG16 After PCA

Image	Feature1	Feature2	Feature 25,088	Image	Feature1	Feature2	FEATURE2000
Image 1	F1-1	F1-2	...	F1-25,088	Image 1	F1-1	F1-2	...	F1-2000
Image 2	F2-1	F2-2	...	F2-25,088	Image 2	F2-1	F2-2	...	F2-2000
Image 3	F3-1	F3-2	...	F3-25,088	Image 3	F3-1	F3-2	...	F3-2000
Image 4	F4-1	F4-2	...	F4-25,088	Image 4	F4-1	F4-2	...	F4-2000
.....
.....
Image N	Fn-1	Fn-2	Fn-25088	Image N	Fn-1	Fn-2	Fn-2000

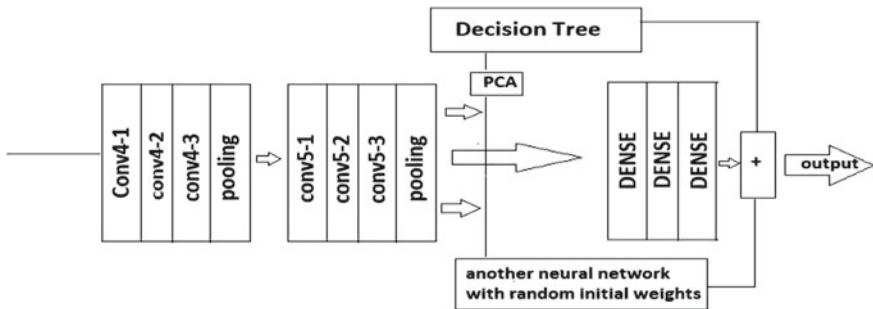


Fig. 9 Pipe line of workflow of ensembled VGG16

input to the rest of classifiers in the ensemble. The output of last convolutional layer in the last epoch will be highly stable as it already captured a lot of features of the images as training is near to end. Classifiers working on this stable input are expected to yield much more accurate results. Now the issue arises in determining the class of an image in test set.

(A) Voting

For the task of skin disease classification, we opted voting. suppose there are k classifiers in the ensemble. Each classifier will output a particular class for an image in test set. Suppose the classifiers be VGG16 dense block, another neural network, decision tree, random forest, and adaboost. Their combined outputs be a vector of classes $[c_0, c_1, c_1, c_2, c_1, c_7]$. c_0 predicted by VGG16 dense block, c_1 predicted by another neural network, c_2 predicted by decision tree, and so on. The final output will be the maximum occurring element of the above vector. This strategy of voting is quite simple and efficient also.

(B) Weighted Saying

Here we consider the loss function of each classifier in ensemble. After the complete training is done, the loss or error of each classifier is noted down. This is used in determining the final class of test image.

$A = \frac{1}{2} \log(1 - L/L)$ where L is the loss made by the classifier. A is the say of the classifier or weightage of the classifier in determining the final class of an image. Generally, dense blocks and neural networks use binary or cross-entropy loss function.

$$\text{If } L = 0.9, A = \frac{1}{2} \log(0.1/0.9)$$

$$A = -0.47$$

$$\text{If } L = 0.1, A = \frac{1}{2} \log(0.9/0.1)$$

$$A = 0.47$$

If classes predicted by classifiers in ensemble are $[1, 0, 0, 1, 1]$ and their respective says are $[0.67, 0.54, -0.75, -0.44, 0.87]$. It makes predictions by having each

classifier classify the sample. Then, we split the classifiers into groups according to their decisions. For each group, we add up the say of every classifier inside the group. The final classification made by the ensemble as a whole is determined by the group with the largest sum.

7 Classification Results

Ensemble VGG16 achieved a significant increase in accuracy compared to deep RESNET101 architecture.

As an example, consider all the images from 1 to 5 which are mentioned in the above classification results. The output of first convolutional layer of VGG16 aim at extracting certain features from an image. For better understanding the purpose of feature extraction by convolutional layers, which are input to our integrated ensemble the following figures are illustrated. The output of first convolutional layer of VGG16 is of size $224 * 224 * 64$. The 64 channels can be arranged as a $8 * 8$ square grid which is shown (Figs. 10, 11, 12, 13, 14, 15, and 16).

Features extracted at convolutional layer 1 (64 features)

An ensemble can be formed in n number of ways. It has a choice of choosing individual classifiers. The way of choosing individual classifiers will have a high notable performance impact on an ensemble. We tried and experimented with several individual classifiers in the skin disease classification task. This ensemble is integrated at the last possible convolutional layer of VGG16. The technique can be expanded to other custom datasets also (Figs. 17, 18, 19, 20, 21 and 22 and 23; Table 3).

Ensembled VGG16 achieved a significant increase in accuracy compared to deep RESNET101 architecture on skin disease image dataset

Fig. 10 Accuracy comparison among several models on skin disease image dataset

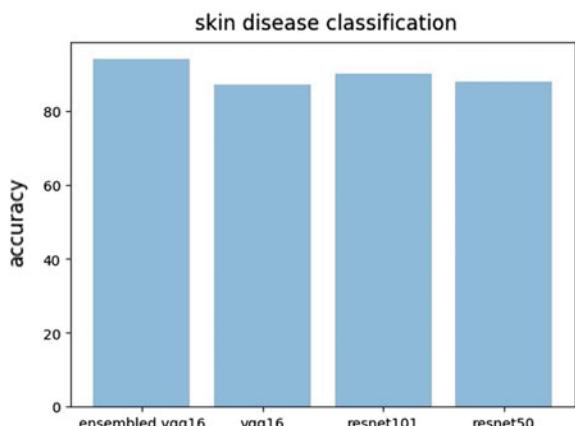


Fig. 11 Ensemble VGG16 predicts disease type



Fig. 12 Ensemble VGG16 predicts disease type



Fig. 13 Ensemble VGG16 predicts disease type



Fig. 14 Ensemble VGG16 predicts disease type



Fig. 15 Ensemble VGG16 predicts disease type

8 Conclusion

The work proposed is a significant contribution in the domain of dermatological skin diseases. People can use the above yolo experimentation and proposed ensembles as a simple cloud application to detect skin diseases in initial stages and can try to prevent them without much doctor intervention. By using multiple classifiers, the accurate prediction ability of an ensemble can be much better than that of a single classifier or single model.

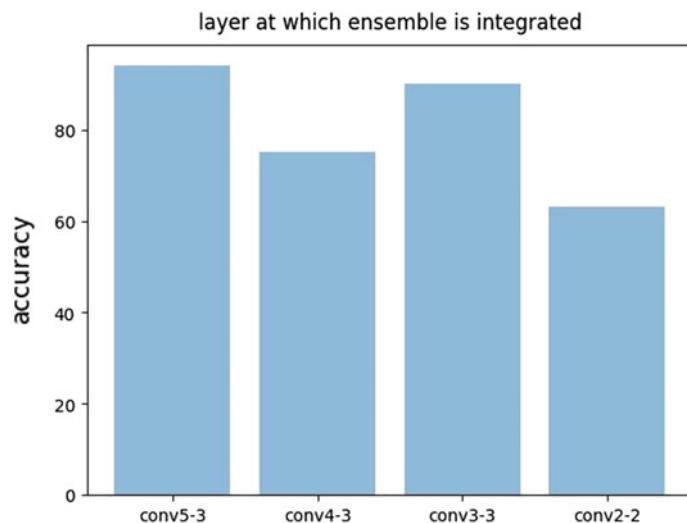


Fig. 16 Accuracy obtained as a result of integrating ensemble at various convolutional blocks of VGGG16

Fig. 17 Features extracted by first convolutional layer of VGG16 for Bowens disease image

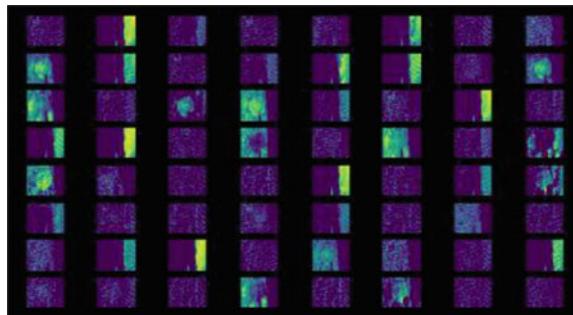


Fig. 18 Features extracted by first convolutional layer of VGG16 for nevus spilus disease

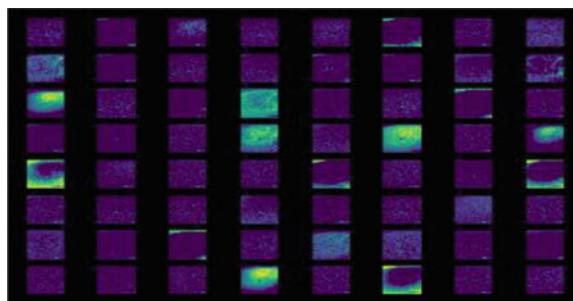


Fig. 19 Features extracted by first convolutional layer of VGG16 for fixed drug eruption disease

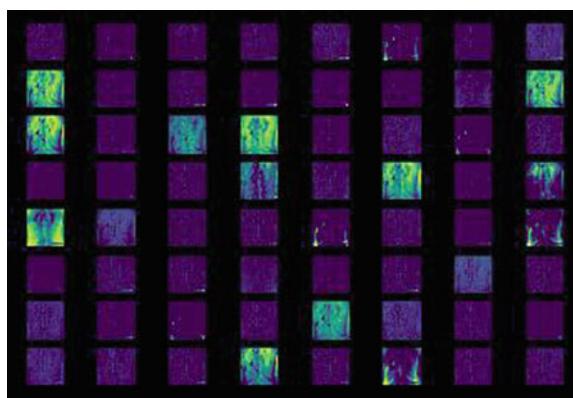


Fig. 20 Features extracted by first convolutional layer of VGG16 for malignant melanoma disease

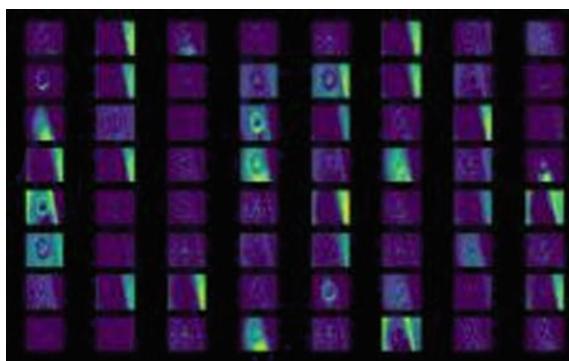
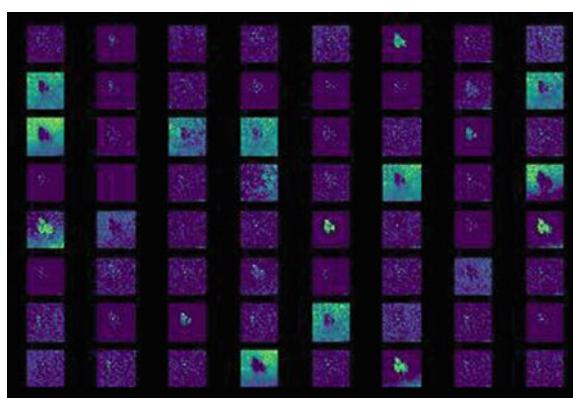


Fig. 21 Features extracted by first convolutional layer of VGG16 for necrobiosis lipoidica diabetorum

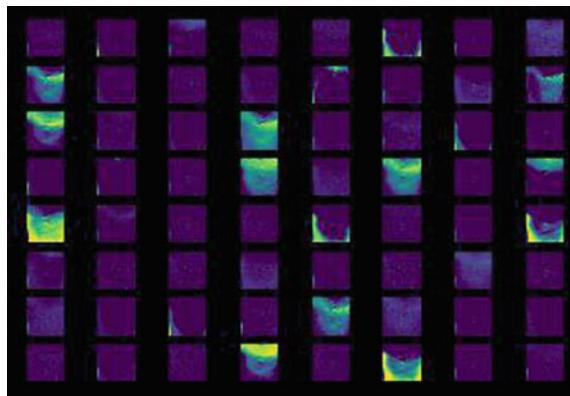


Fig. 22 Features extracted by first convolutional layer of VGG16 for Nevus Anetodermadisease

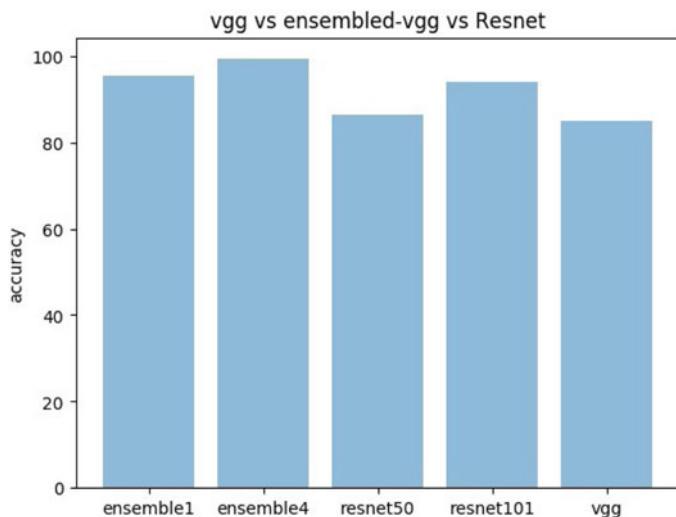


Fig. 23 Accuracy comparison among different VGG16 ensembles and RESNET

Table 3 Describing various ensembled VGG16 accuracies over RESNET

ResNET 50 accuracy	86.4	Individual classifiers	Vgg Ensemble1
ResNET 101 accuracy	94	Vgg dense block	
Vgg accuracy on skin disease dataset	85	Random initialized neural network	
		Decision tree after applying pca	
		Reported ensembled VGG16 accuracy	95.3
Individual classifiers	Vgg Ensemble2	Individual classifiers	Vgg Ensemble3
Vgg dense block		Vgg dense block	
Xavier initialized neural network		Zero initialized neural network	
Random forest		Random initialized neural network	
Reported ensembled VGG16 accuracy	96	Reported ensembled VGG16 accuracy	94.3
Individual classifiers	Vgg Ensemble4	Individual classifiers	Vgg Ensemble5
Vgg dense block		Vgg dense block	
Xavier initialized neural network		Zero initialized neural network	
He initialized neural network		Random forest	
Decision tree with PCA			
Reported ensembled VGG16 accuracy	96	Reported ensembled VGG16 accuracy	94.3

Illustrating how an ensemble if formed by choosing its classifiers

Bold differentiate different architectures

References

1. Maurya R, Surya KS (2014) GLCM and multi class support vector machine based automated skin cancer classification. IEEE J 12
2. Anshu bharadwaj, "Support vector machine", Indian Agriculture Statistics Research Institute
3. Sheha MA (2012) Automatic detection of melanoma skin cancer. Int J Comp Appl
4. Chaithanya Krishna M, Ranganayakulu S (2016) Skin cancer detection and feature extraction through clustering technique. Int J Innovative Res Comp Commun Eng 4(3) March 2016
5. Amarathunga AALC (2015) Expert system for diagnosis of skin diseases. Int J Sci Technol Res 4(01)
6. Bajaj L, Kumar H, Hasija Y (2018) Automated system for prediction of skin disease using image processing and machine learning. Int J Comp Appl (0975–8887) 180(19), February 2018
7. Ahmed A, Jesmin T (2013) Early prevention and detection of skin cancer risk using data mining. Int J Comp Appl 62(4) January 2013
8. Srivastava RK, Greff K, Schmidhuber J (2015) Training very deep networks. 1507.06228

9. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In CVPR, 2015
10. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014
11. Gidaris S, Komodakis N (2015) Object detection via a multi-region & semantic segmentation-aware cnn model. In ICCV, 2015

A Review of Smart Greenhouse Farming by Using Sensor Network Technology



D. Chaitanya Kumar, Rama Vasantha Adiraju, Swarnalatha Pasupuleti, and Durgesh Nandan

Abstract This paper mainly discusses the smart greenhouse farming at low cost. This farming is based on wireless multi-sensor network technology using ZigBee. It helps in the growth of plants in any climate by providing plant-friendly environmental conditions. A greenhouse environment measurement in GSM-SMS was known in this paper. GSM-SMS is a technique that gives a notification to our mobile through sensors. This GSM-SMS system includes mobile dialogue and microcontroller. The greenhouse environment is a thing that plants grow at a certain place under highly protected conditions such as a glass-covered area. The greenhouse environment is a special application of wireless sensor networks, wireless sensor network control is the key point. By monitoring and controlling the greenhouse environment, we can see many major improvements, such as it provides good quality of crops. The key factors for the good quality and productivity of crop growth in the greenhouse are monitoring and controlling them. Plant growth is facilitated by maintaining humidity, temperature, CO₂ concentration, and light intensity. This event could be happened easily by using a multi-sensor technique. The temperature sensor, humidity sensors are present to control and monitor. This will lead by giving more outcomes and fewer costs when comparing it with regular farming techniques. Here we use LoRa technology which means LoRa (short for long range) is a spread spectrum modulation technique derived from chirp spread spectrum (CSS) technology.

D. Chaitanya Kumar · R. V. Adiraju · S. Pasupuleti

Department of Electronic and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, East Godavari, Andhra Pradesh, India

e-mail: chaitanyakumar513@gmail.com

R. V. Adiraju

e-mail: vasanthaadiraju@gmail.com

S. Pasupuleti

e-mail: swarnalathapasupuleti@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd., CL Educate Ltd., New Delhi, India

e-mail: durgeshnandano51@gmail.com

Keywords LoRa · ZIGBEE · IoT · Green-house farming · WSN

1 Introduction

For a long time, we have confronted an issue that plants are developed in some specific climatic conditions, either cold or hot conditions as it were. Because of the expanded interest for sustenance, individuals are attempting to put additional exertion on an increment of the creation of nourishment. In this condition, the nursery is utilized. As a cutting edge some portion of farming, the greenhouse is acquainted with making the agribusiness procedure a brilliant horticulture process [1]. The Wireless sensor innovation implies the data preparing innovation comprises of little size. The sustenance issue has made numerous researchers take a shot at the theme of greenhouse utilizing remote sensors. The utilization of sensors is supporting farming in an extremely positive way. In our present age, the little estimated sensors are taking a noteworthy part in IoT [1, 2]. The sensors can understand the remote controlling, and this remote controlling causes the nursery to keep up proficiently. Furthermore, it additionally has an incredibly helpful incentive in improving space use. To understand the greenhouse condition parameters deductively, this theme gives information about the utilization of minimal effort remote multi-sensor arrange innovation system into greenhouse. There is another kind of sensor innovation utilized in the nursery is called ZigBee [3]. The remote system can control greenhouse electromechanical gear for remote control. These remote sensor innovation focal points are low control utilization and ease. These sensors can ready to give a warning about the base helpful data of the horticulture which is valuable for the greenhouse. In this remote framework, the information gathered by the sensors is transmitted legitimately to associated android versatile or any PC [1–3]. Zig-Bee is a kind of innovation that transmits information starting with one sensor then onto the next with high proficiency. This ZigBee is profoundly helpful in the agrarian field.

2 Literature Review

Remote estimation of the nursery was created utilizing GSM-SMS. The complete framework comprises of two stations focal and base. The remote checking framework can likewise react to remote continuous information handling. A nursery situation estimation dependent on GSM-SMS was created in this paper. The framework incorporates versatile correspondence and microcontroller, Greenhouse condition is a unique use of remote sensor system, and control is the key point in it. This paper contains a web-based administration framework with inserted control. This framework contains a human-based screen and control stage for the remote system sensors for nursery [1].

An embedded frameworks way to deal with screen nursery has a preferred position in nowadays, particularly for checking of the nursery frameworks. “An Embedded frameworks way to deal with screen nursery” because of the estimating of parameters like Humidity, Water pH, Soil wetness, Light power, and temperature by sensors are situated at better places, were estimated, and refreshed to the client through SMS utilizing GPS modem. A portion of the yield an incentive from the different sensors that show by the flickering of LED. After built up, the pack for nursery observing framework, it has been put on testing. The pack will distinguish wrong parameters and updates in the versatile beneficiary or their PC through GSM modem.

This article is intended for the new refreshed sensor organize. High dampness compelled to the potential harms. While doing the tests, the board harming element was coming into action [2].

This paper structured a greenhouse remote checking framework dependent on GSM. The temperature sensor, stickiness sensor establishes the discovery module. Proprietors can set the standard qualities and sends the data to the mobile phone or PC which was associated with it. Simultaneously, the framework can associate the client’s PC for remote information [3].

To keep away from the wasteful manual accumulation, wiring of link system checking, and troublesome support, the plan of wise greenhouse observing arrangement of remote sensor system was created, which joined sensors with ZigBee remote innovation. As the idea of the web wound up viral, the ZigBee arrange innovation turned out to be all the more proficiently utilized. This paper applies the ZigBee innovation to the greenhouse observing framework and presents the general structure of the greenhouse, and acknowledgment of remote sensor arrange. The key elements for the great quality and profitability of harvest development in the greenhouse are observing and controlling. Most examination centers the impact of each light source change, and it has been infrequently explored. In this paper, another age of greenhouse checking and controlling framework dependent on powerful supplemental light source control with remote sensors are discussed [4].

Plant development is encouraged by looking after stickiness, temperature, CO₂ fixation, and light force. These variables ought to be checked and kept up for any nursery framework. A noteworthy piece of the framework is the utilization of LED light rather than fluorescent lighting because of its low control utilization, long life, and utilize the thin band. We have utilized work arrange because it has numerous advantages. To fast plant development, we have utilized an LED lighting framework that adds to the vitality protection of the framework, empowering to deliver plants at an ideal worth. The utilization of control calculations utilized in the insightful farming framework. This has an alluring thought for future work to be actualized in this nursery framework [5].

Driven lighting framework gives a proficient and conservative lighting framework that helps plant development by changing light force and recurrence as per the light conditions and plant development necessities likewise helps in diminishing creation expenses and speed development. The model of the framework proposed has been introduced in a modest piece of the nursery. Information procurement and remote administration have demonstrated extremely acceptable execution [5].

The structure of the condition-dependent on remote system applies the innovation of remote sensor organize and receives the control capacity of Single Chip Micyoco (SCM). This structured framework utilizes the SHT11 sensors of temperature and dampness as discovery components, and the ongoing estimation information is transmitted through a remote handset module. Through structure up dispatching test and useful framework, it reasons that the framework status, i.e., the framework can understand implies the remote checking of the nursery would stay in the best condition for plant development additionally a programmed caution which is utilized to guarantee that the temperature and dampness are adequate for this procedure. This structure is utilization of remote sensors to organize innovation in the plan of a control framework which is proficient. This plan has a specific criticalness to the improvement of an observing framework for the nursery environment [6, 7].

The WSN conveyed modules have high conditions on their battery's utilization. Then again, occasion-based frameworks are spreading quickly step by step. This sort of framework makes to perform nonconcurrent activities dependent on standards. The WSN appropriate a few modules that are utilized to permit playing out a procedure before transmitting the information to any associated gadget. Along these lines, these module's vitality utilization can be diminished to a degree, expanding the battery life strongly. In this manner, consolidating the two innovations improves the Wireless sensor innovation self-rule. In this work, these thoughts are performed in the atmosphere observing structure. The most intriguing arrangement with the execution of this work was the decrease of transmitted parcels (around to 67%) and the battery utilization of this framework. It ought to be streamlined for vitality sparing to expand the battery life [8].

Remote sensor system is one of the most encouraging advances in the twenty-first century. Applying the WSN generation condition has been bringing about the best and effective outcomes. In the feeling of joining the earth detecting capacity of remote sensor systems into portable observing frameworks can give the best control towards the nursery for our versatile or pc anyplace whenever we can screen its results [9].

Here, we proposed a nursery domain versatile checking framework that recognizes temperature and dampness sensor, light sensor into a gadget to gather the data about the development of the nursery condition. Clients can check and control the status of nursery condition progressively through their android mobiles or their PCs. With the improvement of PC innovation, an assortment of checking frameworks has been applied compared to the nursery. Be that as it may, the majority of the present checking frameworks are wired, with the downsides of wiring rock-solid, enormous forthright speculation. As of late, when the remote sensor systems (WSN), comes into the image, normally, the WSN will be utilized regularly utilized in our day by day needs like buyer gadgets, home robotization, home security, individual social insurance such things [10, 11].

The remote sensor multi-hubs in the nursery are blended or combined, which acquires high accuracy information and decreases the information that has retransmitted. It contemplates the key innovation about information combination for Greenhouse Wireless Sensor Network. As indicated by the topological structure of the

nursery remote sensor arrange, the weighted calculation needs to advance and is kept up and broke down by the combination model. The weighted calculation needs to intertwine information from a similar kind of remote sensor hubs inside the nursery arrange, which acquires high exactness information, time decreases the transmitting time of the information, all are done simultaneously [12, 13].

The WSN dependent on Wi-Fi is utilized for short-separation correspondence and GSM for worldwide framework correspondence purposes. Every one of the parameters in the framework could be made simple as per the plant type and atmosphere necessities for the specific condition. To control the ecological components, the microcontroller is utilized to control the parameters as per present qualities or manual values [14].

The IoT-based savvy nursery gives a model of a keen nursery, which encourages the ranchers to convey the work in a homestead naturally without the manual troubles. Utilizing an ultrasonic sensor Proper water, the executive's tanks are developed and are loaded up with water after estimating the present water level. Temperature and air moistness are constrained by dampness and temperature sensors. The benefit of Smart Greenhouse in cultivating is that we had the option to deliver the best sans pesticide crops and make an appropriate development cordial atmosphere for plants. We can associate ranchers legitimately to customers utilizing IoT, which can spare him from the obstacles of a middleman [15].

To tackle the issues of overwhelming wiring, troublesome upkeep of that, and simple line erosion in nursery observing framework, a nursery savvy data checking framework dependent on remote sensor system was created. The framework has been applied in the class base for a long time, and it demonstrates that it is precise, dependable. Also, it can fulfill the necessities of a sun-based nursery [16]. A sort of control strategy dependent on-time control, manual control, programmed control, canny control and the remote control was came into the ascent, which can understand the change of nursery condition parameters and support for yield growth [17]. The Design of a programmed water system framework for nursery dependent on LoRa innovation comprises of Sensor Nodes that gather the information of soil dampness, temperature, and mugginess. This information will be transmitted to the core of the focal station named concentrator. The concentrator will use for the control of the water system process through the control hubs which are situated at the field level. Clients can set up water system mode, get to information, and oversee crops through both PC and Web Interface. At last, we displayed some outcomes [18–20].

3 Technique to Maintain Greenhouse

Pictorial representations of maintenance of greenhouse have been shown in Fig. 1.

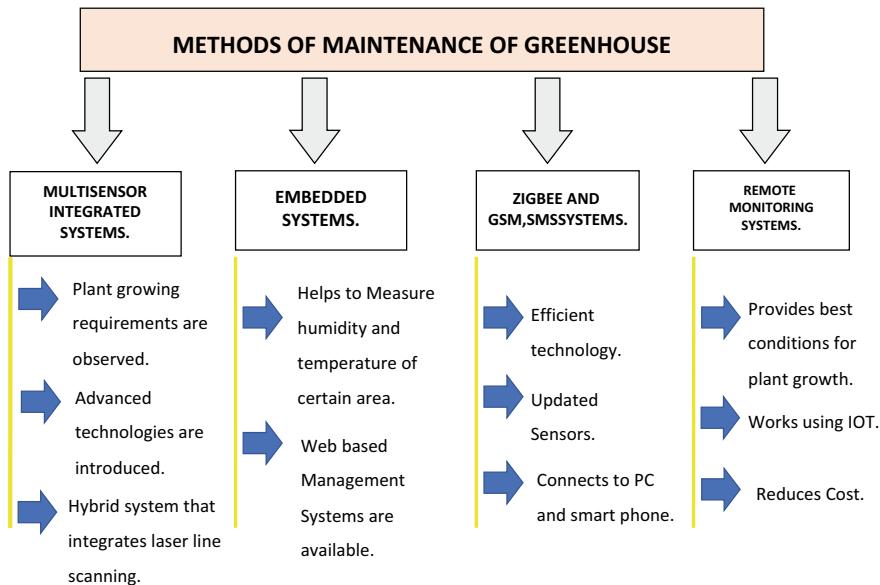


Fig. 1 Pictorial representations of maintenance of greenhouse

3.1 *Multi-Sensor Integrated Systems*

Sensor information got from multiple sources is melded and then connected with observational (learning rules, learning models, and so forth from every particular area) in the framework under test, after which the intertwined data is assessed by certain information rules. Comparing data is put away in the database framework for learning disclosure [3].

3.2 *Embedded Systems*

An embedded framework is a blend of PC equipment and programming, fixed incapacity, or programmable, intended for a particular capacity or capacities inside a bigger framework. Modern machines, farming and procedure industry gadgets, vehicles, therapeutic gear, cameras, family apparatuses, planes, candy machines, and toys, just as cell phones, are potential areas for an embedded framework [5].

3.3 Zigbee and Gsm Techniques

ZigBee is particular for a suite of systems administration, security and application programming layers utilizing little, and low-control, ease, low information rate correspondence innovation dependent on IEEE.

GSM framework was created as an advanced framework utilizing time division multiple entrance (TDMA) method for correspondence reason. A GSM digitizes and decreases the information, at that point sends it down through a channel with two distinct surges of customer information, each in its specific schedule vacancy [6–13].

3.4 Remote Monitoring Systems

Remote Monitoring is a standard detail that encourages the monitoring of system operational exercises using remote gadgets known as screens or tests. RMON helps organize directors with proficient system foundation control and the board [7].

4 Conclusion and Future Scope

For the better support of the nursery, we have discovered numerous ways and innovations. We have significantly observed a portion of the methods like multisensory coordinated frameworks, installed frameworks, Zigbee innovation GSM, SMS procedures, and remote checking framework strategies that are useful to keep nursery shrewd and minimal effort. The temperature sensor, moistness sensors are available in it. Clients can set standard qualities and sends the data to the phone. GSM framework was created as an advanced framework utilizing time division numerous entrance system for correspondence reason. This sort of framework makes to perform nonconcurrent activities dependent on standards. So, the IOT's are utilized. The IoT-based shrewd nursery gives a model of a brilliant nursery, which causes the ranchers to convey the work in a homestead consequently without the manual troubles. An assortment of checking frameworks has been applied compared to the nursery. In any case, a large portion of the present checking frameworks are wired, yet the wired frameworks are convoluted and much cost, so we use WSN (remote sensor organizes) that could make our work on nursery simpler and quicker.

References

1. Li LL, Yang SF, Wang LY, Gao XM (2011) The greenhouse environment monitoring system based on wireless sensor network technology. In: 2011 IEEE international conference on cyber technology in automation, control, and intelligent systems, pp 265–268. IEEE

2. Changqing C, Hui L, Wenjun H (2018) Internet of agriculture-based low-cost smart greenhouse remote monitor system. In: 2018 chinese automation congress (CAC), pp 3940–3945. IEEE
3. Gupta GS, Quan VM (2018) Multi-sensor integrated system for wireless monitoring of greenhouse environment. In: 2018 IEEE sensors applications symposium (SAS), pp 1–6. IEEE
4. Jin S, Jingling S, Qiuyan H, Shengde W, Yan Y (2007) A remote measurement and control system for greenhouse based on gsm-sms. In: 2007 8th international conference on electronic measurement and instruments, pp 2–82. IEEE
5. Rangan K, Vigneswaran T (2010) An embedded systems approach to monitor green house. In: Recent advances in space technology services and climate change 2010 (RSTS & CC-2010), pp 61–65. IEEE
6. Fu M, Yang L, Zhang J (2011) Study of light emitting diodes for the application of plant growth in green house. In: 2011 12th International conference on electronic packaging technology and high density packaging, pp 1–5. IEEE
7. Huang H, Bian H, Zhu S, Jin J (2011) A greenhouse remote monitoring system based on GSM. In: 2011 international conference on information management, innovation management and industrial engineering, vol 2, pp 357–360. IEEE
8. Luo Q, Qin L, Li X, Wu G (2016) The implementation of wireless sensor and control system in greenhouse based on ZigBee. In: 2016 35th chinese control conference (CCC), pp 8474–8478. IEEE
9. Le, J, Kang H, Bang H, Kang S (2012) Dynamic greenhouse supplemental light source control with wireless sensor network. In: 2012 international conference on ICT convergence (ICTC), pp 23–27. IEEE
10. Ijaz F, Siddiqui AA, Im BK, Lee C (2012) Remote management and control system for LED based plant factory using ZigBee and Internet. In: 2012 14th international conference on advanced communication technology (ICACT), pp 942–946. IEEE
11. Min Z (2013) Design of environment monitoring system based on wireless sensor network. In: 2013 IEEE 11th international conference on electronic measurement & instruments, vol 2, pp 547–551. IEEE
12. Ferre JA, Pawlowski A, Guzman JL, Rodriguez F, Berenguel M (2010) A wireless sensor network for greenhouse climate monitoring. In: 2010 Fifth international conference on broadband and biomedical communications, pp 1–5. IEEE
13. Li RA, Sha X, Lin K (2014) Smart greenhouse: a real-time mobile intelligent monitoring system based on WSN. In: 2014 international wireless communications and mobile computing conference (IWCMC), pp 1152–1156. IEEE
14. Yinghui L, Genqing D (2015) Study on data fusion of wireless monitoring system for greenhouse. In: 2015 8th international conference on intelligent computation technology and automation (ICICTA), pp 864–866. IEEE
15. Mekki M, Abdallah O, Amin MB, Eltayeb M, Abdalfatah T, Babiker A (2015) Greenhouse monitoring and control system based on wireless sensor network. In: 2015 international conference on computing, control, networking, electronics and embedded systems engineering (ICCNEEE), pp 384–387. IEEE
16. Kodali RK, Jain V, Karagwal S (2016) IoT based smart greenhouse. In: 2016 IEEE region 10 humanitarian technology conference (R10-HTC), pp 1–6. IEEE
17. Tsai CF, Hung KC (2016) Campus greenhouse monitoring with a simple ZigBee-based sensor network. In: 2016 international conference on advanced materials for science and engineering (ICAMSE), pp 305–308. IEEE
18. Trinh DC, Truvant TC, Bui TD (2018) Design of automatic irrigation system for greenhouse based on LoRa technology. In: 2018 International conference on advanced technologies for communications (ATC), pp 72–77. IEEE
19. Reka SS, Chezian BK, Chandra SS (2019) A novel approach of IoT-based smart greenhouse farming system. In: Druck H, Pillai R, Tharian M, Majeed A (eds) Green buildings and sustainable engineering. Springer, Berlin
20. Meah K, Forsyth J, Moscola J (2019) A smart sensor network for an automated urban greenhouse. In: 2019 international conference on robotics, electrical and signal processing techniques (ICREST), pp 23–27. IEEE

A Study on Low-Frequency Signal Processing with Improved Signal-to-Noise Ratio



G. Pavan Avinash, P. Ramesh Kumar, Rama Vasantha Adiraju, and Durgesh Nandan

Abstract The main objective of the signal processing is to acquiring higher-order derivatives (HOD) of low-frequency (LF) signals with higher signal-noise-ratio (SNR). The proposed method comprises of the oversampling to convert the signal from analog to digital by using multi-stage downsampling (DS). In the multi-stage processing (MSP), the digital-differentiator FIR follows the decimation by two. In the cascaded multi-stage processing, each stage contains a simple half-band low-pass FIR filter followed with decimation by two. In the majority of the research works recently done here, the important goal was to accomplish better exactness (SNR) even by using quite high-order algorithms. It is not only for accurate values but also in the simplicity of the algorithm with requiring significantly less memory space and computational complexity in comparison with the single stage.

Keywords Oversampling · Higher-order derivative · Signal-to-noise ratio · FIR digital differentiator · Multi-rate signal processing

G. Pavan Avinash · P. Ramesh Kumar · R. V. Adiraju

Department of Electronics and Communication Engineering, Aditya College of Engineering and Technology, Surampalem, Andhra Pradesh, India

e-mail: pavanavinashgedda@gmail.com

P. Ramesh Kumar

e-mail: ramesh.padala@acet.ac.in

R. V. Adiraju

e-mail: vasanthaadiraju@gmail.com

D. Nandan (✉)

Accendere Knowledge Management Services Pvt. Ltd, CL Educate Ltd, New Delhi, India

e-mail: durgeshnandano51@gmail.com

1 Introduction

For acquiring higher-order derivatives of LF signals are used in sensor applications. To study the speed and acceleration (ace), the first- and second-order digital differentiators (DD) added to LF scientifically signals [1]. In the solid conduction, the variation of the data of temperature is only at low frequencies. The temperature is in degrees Celsius. Using this temperature information, to get the heat transition and after that high frequency of noise is intensified. Then, it gets unstable and the SR is additionally expanded [2]. In the computerized numbered managed binary process, the ace and moment, i.e., work of another source of ace is displayed. The moment signal is obtained by using a DD [1]. But there is a problem with using old method differentiators to get a derivative signal. For a fewer cost data system, i.e., low ADC binary conversion for digital to analog conversion is used. The large quantization errors get in results. The sampling technique is improved to solve this problem. BME signals are considered in the LF signals, the vibrator sensor signal is with bandwidth near to 500 Hz. The sensor signal is conditioned and it contains some bandwidth. The DD carries out two functions (fun). The first fun is to re-construct the spectral line between two points of the quantization noise (QN) and next it went to high frequency (HF) ranges. The second fun is to filter the HFQN. Finally, the new un-error signal is acquired through the DS process. Hence, signal-to-noise ratio is obtained for the given derivative signal [1].

Most of the sensor applications are used in voltage signals. The working of this type of sensors is also used in many types of applications. They are solid and fluid mechanics, aerospace, heat transfer, and some security applications. In that, heat transfer is the main application [2]. The noise is amplified by the process of differentiation, particularly, at HFE. The data of the biological nature contains up to 60 dB, i.e., 10–12 bits [3].

2 Literature Review

Exact values and information are required in the heat treatment applications, but it is an integral relationship and it requires the time derivative. Hence, the data or the information are more accurate. Theoretically, heat is getting by using the TD (dT/dt). This theory is used to develop the voltage sensor for LF applications. The voltage rate working process is involved in the temperature sensor. The sensors are also used in many engineering applications. It is the simulation process type experiments like MATLAB and PSpice. Predicting heat transition from transient temperature information is not able to be poorly presented. Noise and errors predominant in all estimations will expand comparative with the sign, the SNR significantly in the wake of smoothing and shifting. By converting the amplitude of the signal spectra regulation, as recommended, the SNR of the subordinates will stay reasonable to warrant solid resulting process. The principle is to study about the all-inclusive

arrangement with an attempting to change over the voltage output from sensors to voltage rate. The SNR is developed in experimental cases. But the SNR is less than the simulations because of external noise and the op-amp limitations. All these sensor applications are mainly used in aerospace stations, engineering sciences and defense purposes [2].

Multi-rate signal processing is a technique that is used for acquiring higher-request derivatives of advanced signals. It is getting by an improved signal (SNR). It contains oversampling and is used to convert the analog to digital signals. Computer simulations and experimental data are processed using this method. In the multi-rate signal processing, FIR filters are involved and it has very small in size. In the digitalized numbered controlled machine process, the acc and jerk, i.e., subordinate of increasing speed is observed. The jerk signal is obtained by using a digital differentiator. The construction of a linear stage FIR derivative at higher sampling rate may have need of an enormous channel size. In [2], FIR DF length could arrive at more than 10,000 coefficients at the over sampler component of 128 to become its superior [1].

Computerized low-pass separation is involved with scientific and biomechanical information. The data is used in micro or mini computers. This information requirement for straight forward, low and quick separation methods. By using the first- and second-order differentiations simple algorithms are achieved, it has both practical and theoretical views. In the research on this topic, the main aim is to achieve good results accurately. The main is to achieve a simple algorithm by using low and high differential methods and it is almost optimum. The noise is improved by the procedure of separation, particularly, at higher-order frequencies. The information of the biological nature contains up to 60 dB, i. e., 10–12 bits and along these lines, basic integer number calculations can be utilized as floating-point. Since BMD are typically tested at a less rate, the number of sampling points is relatively little. Along these lines, we cannot use in their handling higher-order calculations, which require huge ranges of information. The low-pass first- and second-order separation channels are both theoretical and practical aspects. The biological and biomechanical applications are used in the micro and minicomputers. Some filters are proposed but it is not used as a practical purpose. We represented the tables and graphs for choosing the difference between low-order and low-pass derivative algorithms, these algorithms are used in particular applications. These algorithms are used in data processing in the fields [3].

By using multi-rate analysis, a time-frequency response is obtained and at the next stage, the decomposed parameters are cascaded. Due to multi-rate signal decomposition with high frequency processed accurately and simulated. Hence, we proposed an approach to the real-time simulation. The frequency bands are calculated based on wavelet analysis by using multi-rate filters. The multi-rate technique provides the best implementation of signal processors [4].

Low and high rate estimations are derived by using a filter bank. The dimensions are the same as for high and low rate measurements. Then, the better results are obtained. The filter bank is measured by a downsampling approach [5].

For representing the binary sequence, high values are taken in the new quantization scheme. The quantization errors are going to higher frequencies. These frequency

procedures are applied to drive the FIR channel to get the accurate value in the binary bits. The FIR filter's magnitude is more accurate over low frequencies. The FIR filters have more efficient designs. Narrowband FIR channels are utilized to take out the multi-bit multipliers. The scheme is executed by utilizing sigma and delta regulation. The finite impulse response is only based on the cascade technique. A designed algorithm is developed to give a number of examples. In all the cases, FIR filter approach is necessary to get low-frequency equivalent multipliers. The scale factor is fixed to the filter by using this approach and also fixes the lengths of the filter. Sigma to delta conversion may provide benefits to the narrow band-pass filters [6].

The signal processing is especially used in engineering applications and some computer-based experiments. The signal processing technique is developed in a laboratory by a hands-on approach. By the motivation of cliche, the laboratory is developed and he is one of the approaches to teach the signal processing in the laboratory. The research laboratory has been named as the "signal computing and real time". In the research laboratory, practical training is giving around 150 last year undergraduates each year. These are the highlights of the research laboratory, and the strategies in the laboratory depend on test improvement. The primary objective of this research laboratory is to give adequate training to the last year students to actualize signal handling in present-day gadgets. The laboratory provides computer-based technology experiments [7].

The high-speed radar data communication is by using a signal processing technique. In the data processing, the sampling frequency is at 100 MHz. The high-speed radar data processor possesses a block shape. Along with these signals, the quantity of its signal processing can be expanded or diminished by the necessities of calculations. Various types of methods in the computer hardware and the software are developed by this radar technique. A high and low-speed conversion technique is developed by this sampling algorithm technique. The high-speed radar data is a hardware system and it is planned for the radar target recognition in the field. In this field, data processing is processed and satisfies the radar data. It tends to be utilized in the processing of radar signals just as radar targets [8].

For reconfiguring the system signal processing is used and for the high-performance, DSP chips are used. It is by dataflow process it can do only parallel data processing. Without changing any system software different signal processing techniques are implemented. A pipelined way with information stream driven is executed in data processing. Every handling node has its memory space and neighborhood sharing memory. It very well may be parallel signal processing. This design technique is not just for parallel signal preparing of all radar purposes yet besides for different regions required large scale and constant parallel signal handling [9].

In the signal processing technique, Fourier transformation is used and it needs sampled data. It is used in the development of multi-rate FIR filters. There are three data sets of processing and these are handled for both insertions utilizing FIR channels and cubic lines and suited for parallel programming execution. The present methods for the calculation of uniformly and unevenly spaced arrangements were thought about, cubic lines and multi-rate FIR techniques. The examination of the

simulated beat arrangement demonstrated that the FIR channel is an increasingly exact frequency, this might be because the idea of the strategy, prompts an increasing sampling rate of the time arrangement, on the opposite to the cubic lines technique may give adjusting error to the main frequency. We develop parallel and matrix implementations for preparing uniformly with multi-rate channels. The parallel created application shows good improvement [10].

We are facing the issue of direction finding for non-stationary signals. By the utilization of frequency of time representation of the information, we can get the non-stationary signal from the source. We execute the morphological picture processing to approximation of time-frequency signature sections of every source and it is also a type of sensor [11].

Signal processing is used in daily time appliances. In that, smartphones are involved in signal processing. Smartphone plays a very vital role in society by implementing this processing technique. Some of the laboratory application is used in the smartphone by this technique. Smartphone is the goal of teaching signal processing in real-time experience. It is the source to employ signal processing algorithms in cell phones using the C programmable language. Now we are using smartphones as real-time experience [12].

3 Block Diagram

For example, let us assume $L = 2^M$, a series connection of M(multi)-stage signal processing method comprises an LPF with decimation of 2, followed by the k th-order FIR at sampling rate $2f_s$. At last, the processed signal will be decimated by a factor of 2 to accomplish the ideal k th-order derivative signal $y_k(m)$ at the Nyquist rate of f_s Hz (Fig. 1).

The derivative of k th-order ideal frequency response is $H_k(z)$, here $z = e^{j\Omega}$

$$H_k(e^{j\Omega}) = \begin{cases} 0 & \text{for } -\pi \leq \Omega < -\Omega_{\max} \\ (j\Omega/\Omega_{\max})^k & \text{for } -\Omega_{\max} \leq \Omega \leq \Omega_{\max} \\ 0 & \text{for } \Omega_{\max} < \Omega \leq \pi \end{cases}$$

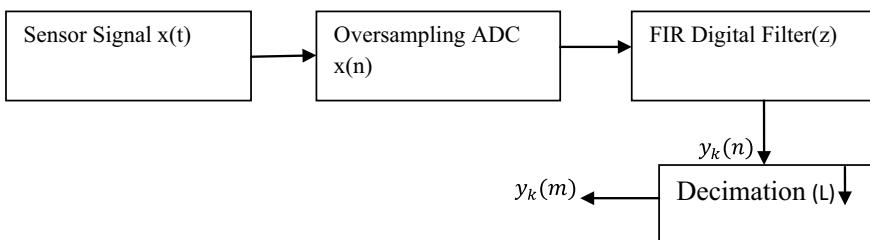


Fig. 1 Single-stage signal processing with oversampling technique

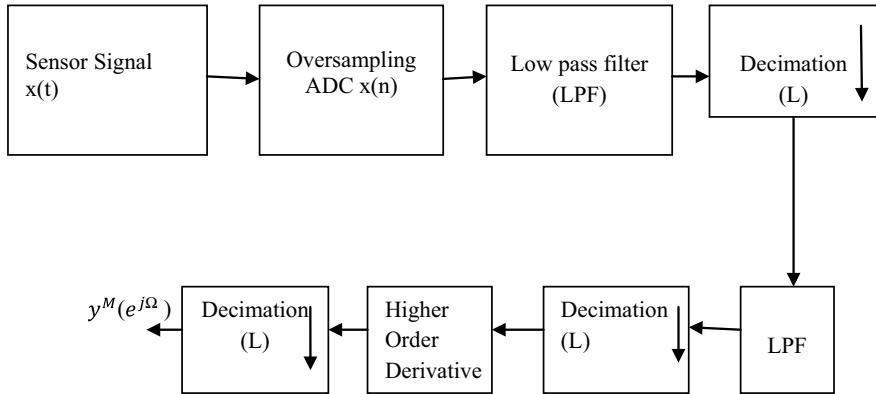
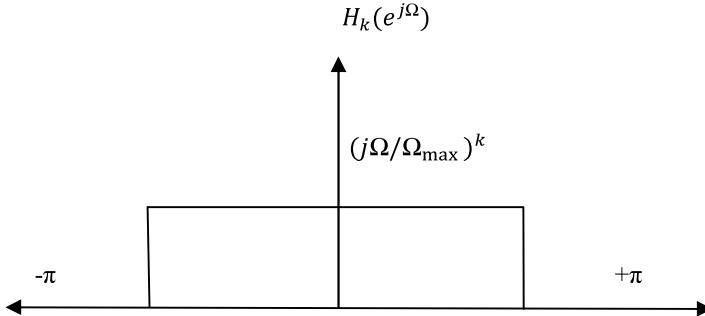


Fig. 2 Higher-order derivatives with multi-stage signal processing technique

where $\Omega_{\max} = 2\pi(f_s/2)/f_{sl} = \pi/L$ radian is the sensor signal of max frequency.

In a system of over sampling $\Omega_{\max} \ll \pi$, a channelize is needed to suit the performance. As appeared in Fig. 2, since the differentiator starts operation at $2f_s$, therefore, $\Omega_{\max} = \pi/2$. A smaller filter (channel) size is fundamentally required.

4 Graphical Representation



5 Discussion

In this paper, we discussed that the sensor applications are used by using the higher-order derivatives of low-frequency signals. It contains the oversampling techniques to convert the signal from analog to digital by using multi-stage downsampling.

Biological signals are involved and it is needed to implementation of mini and micro-computer applications. Biomechanical signals are considered in the low-frequency signals, the vibrator sensor signal is with bandwidth near to 500 Hz. The data of the biological nature contains up to 60 dB, i.e., 10–12 bits. By using multi-rate analysis, a time-frequency response is obtained, and at the next stage, the decomposed parameters are cascaded. The high-speed radar data communication is by using a signal processing technique. In the data processing, the sampling frequency is at 100 MHz. The high-speed radar data processor possesses a block shape. For reconfiguring, the system signal processing is used and for the high performance, DSP chips are used. It is by dataflow process and it can do only parallel data processing.

6 Conclusion

In this paper, we concluded that obtaining HOD of LF signals with higher SNR techniques are used to process the signal. It contains the oversampling techniques to convert the signal from analog to digital by using multi-stage downsampling. Technique followed by good performance computerized differentiator cascaded by a decimation unit with a factor of two. In the signal processing, biological signals are involved and it is needed to implementation of mini and microcomputer applications. In most of the research works previously done in this field, it is not only for accuracy but also in the simplicity of the algorithm resulting in much less memory and computational complexity in comparison with the single-stage process. Final stage of filter is to eliminate noise further, resulting in the high SNR.

References

1. Tan L et al (2013) Obtaining higher-order derivatives of low-frequency signals using multi-rate signal processing. In: Conference of record—IEEE international instrumentation and measurement technology conference, pp 1277–1282. <https://doi.org/10.1109/I2MTC.2013.6555619>
2. Kruttiventi J et al (2010) Obtaining time derivative of low-frequency signals with improved signal-to-noise ratio. IEEE Trans Instrum Meas 59(3):596–603. <https://doi.org/10.1109/TIM.2009.2025069>
3. Usui S, Amidror I (1982) Digital low-pass differentiation for biological signal processing. IEEE Trans Biomed Eng BME-29, 10:686–693. <https://doi.org/10.1109/TBME.1982.324861>
4. Schoenle M et al (1993) Parametric approximation of room impulse responses by multirate systems. In: Proceedings of ICASSP, IEEE international conference on acoustics, speech, and signal processing, vol 1, pp 153–156. <https://doi.org/10.1109/icassp.1993.319078>
5. Tian T, Sun S (2015) Optimal filtering for multi-rate systems with one-step auto-correlated noises. In: Proceedings of 2015 international conference on estimation, detection and information fusion, ICEDIF 2015, ICEDIF, pp 13–17. <https://doi.org/10.1109/ICEDIF.2015.7280149>

6. Powell SR, Chau PM (1994) Efficient narrowband FIR and IFIR filters based on powers-of-two sigma-delta coefficient truncation. In: IEEE transactions on circuits and systems II: analog and digital signal processing, vol 41(8), pp 497–505. <https://doi.org/10.1109/82.318938>
7. Chandran V et al (1994) The design and development processing of an undergraduate signal Vi-41 Vi-42, pp 41–44
8. Yulan L et al (1996) High speed radar data acquisition and processing system. In: International conference on signal processing proceedings, ICSP, vol 1, pp 449–452. <https://doi.org/10.1109/icsig.1996.567299>
9. Unser M, Zerubia J (1998) A generalized sampling theory without band-limiting constraints. In: IEEE transactions on circuits and systems II: analog and digital signal processing, vol 45(8), pp 959–969. <https://doi.org/10.1109/82.718806>
10. Risk MR et al (2007) Time series calculation of heart rate using multi rate FIR filters. Comput Cardiol 34:541–544. <https://doi.org/10.1109/CIC.2007.4745542>
11. Atkins PR et al (2007) Transmit-signal design and processing strategies for sonar target phase measurement. IEEE J Sel Top Signal Process 1(1):91–104. <https://doi.org/10.1109/JSTSP.2007.897051>
12. Heidenreich P, Cirillo LA, Zoubir AM (2007) Time-frequency distributions and morphological image processing. In: Signal processing group Darmstadt University of Technology, image process, pp 1137–1140

Factors that Determine Advertising Evasion in Social Networks



Jesús Silva, Yisel Pinillos-Patiño, Harold Sukier, Jesús Vargas,
Patricio Corrales, Omar Bonerge Pineda Lezama, and Benjamín Quintero

Abstract The present work is framed within the study of advertising evasion online and particularly in social networks. Social networks are a growing phenomenon, where users spend most of their time online and where companies are moving part of their advertising investment, as they are considered an ideal place for commercial campaigns. In order to deepen in the variables that precede advertising evasion in social networks, a relationship model was developed based on the theoretical framework of advertising evasion on the Internet, which was contrasted at an empirical level through a panel of users. For this purpose, a structural equation model was designed, which highlighted the relationships between the main antecedent variables of evasion, such as perceived control, advertising intrusion, and psychological reaction.

J. Silva (✉) · P. Corrales
Universidad Peruana de Ciencias Aplicadas, Lima, Peru
e-mail: jesussilvaupc@gmail.com

P. Corrales
e-mail: patriciocorralesd@gmail.com

Y. Pinillos-Patiño
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: [y whole](mailto:ypinilos@unisimonbolivar.edu.co)

H. Sukier · J. Vargas
Universidad de la Costa, Barranquilla, Colombia
e-mail: hsukier@cuc.edu.co

J. Vargas
e-mail: jvargas41@cuc.edu.co

O. B. P. Lezama
Universidad Tecnológica, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

B. Quintero
Corporación Universitaria minuto de Dios, UNIMINUTO, Barranquilla, Colombia
e-mail: benjamin.quintero@uniminuto.edu.co

Keywords Perceived control · Intrusion · Reactance · Advertising evasion · Social networks

1 Introduction

Nowadays, any individual frequents several social media in which he or she is subject to advertising messages, presented through various forms. In the Internet, advertising can be considered even more annoying than in other traditional media as is the case, for example, of mass and unauthorized advertising through spam [1]. Recent studies start from the idea that advertisements are generally intrusive [2, 3].

In addition, social media are becoming increasingly popular among Internet users and consequently, advertisers are recognizing social networks as ideal platforms to carry out their advertising campaigns [4], as evidenced by the increase in advertising spending on these platforms [5, 6].

In particular, Internet Social Networks (ISN) are used to promote brands, products, or services through marketing campaigns, but the true effectiveness of the campaigns carried out on ISN is still unknown [7]. One of the main particularities in the advertising communications by ISN is that the information shared in them has a strong social component, coming from the different members of the network. Among the motivations for using ISN and spend more time in them are: keeping in touch with friends, maintaining relationships with people living far away, and finding out what the friends are doing [8].

Therefore, in the study of the effectiveness of advertising in social networks, the composition of the social environment of an individual in an ISN is a variable to take into account, since it affects the results of the campaign. Specifically, in [9], they conducted an empirical study on the acceptance of advertising in social networks, concluding that social norms are an essential indicator of advertising acceptance, as they influence attitudes towards the campaign, the brand, and behavioral intentions.

Based on the above, it is reasonable to think that ISN in general, and the advertising inserted in them influence the attitudes and behavior of users. However, few studies have considered the causal relationships that precede negative advertising behavior in an empirical way [10]. This paper contributes to the existing literature by focusing on those background variables of advertising avoidance that stand out in the online environment such as the lack of perceived control [11], the perceived advertising intrusion [12] and the levels of psychological reactance developed by the individual [13–16].

2 Research Model and Hypothesis

The results of the scientific literature review support the idea that consumers have developed negative attitudes towards intrusive digital advertising on the Internet and

mobile platforms, from a push or pull perspective [2, 8, 9, 16–18]. Based on the above theoretical propositions, the following causal relationships are proposed:

H1: There is a direct and negative relationship between perceived control and perceived intrusion.

H2: There is a direct and positive relationship between the perceived intrusion and the psychological reaction of the social network user.

The literature on psychology indicates that attitudes can influence behaviors and intentions [19, 20]. As a result, the advertising shown on social networks is perceived as intrusive because it occupies a place that does not correspond to it. This situation will lead to a loss of control, where levels of psychological reactance will increase as the above is confirmed [13]. This affects the level of intrusion into attitudes towards the platform, advertisement, and brand, as well as negative behaviors such as avoidance [17]. Therefore, the following causal relationships are proposed:

H3: There is a direct and negative relationship between perceived control and advertising avoidance.

H4: There is a direct and positive relationship between perceived intrusion and advertising evasion.

As mentioned above, in the case of an intrusive situation, the individual will increase his or her levels of experienced psychological reaction, with the intention of awakening mechanisms of evasion of publicity and of getting out of the situation that causes discomfort [14, 18]. So, the following relationship is proposed:

H5: There is a direct and positive relationship between psychological reaction and advertising avoidance.

Figure 1 shows the relationships with the above hypotheses.

3 Study Method

3.1 Measuring Instrument

The interview method used in the study was a self-administered web questionnaire, structured as follows: the first part consisted of a series of items related to the variables of the proposed model, perceived control (CONTROL), perceived intrusion (INTRU), psychological reactance (REACT) and advertising evasion (EVA). The scales of measurement were adapted from previous studies, so the adapted scale of [19] was applied for the perceived control. For perceived intrusion, the scale of [4] was used. For psychological reactance, the scale adapted from [6], and for advertising avoidance in its behavioral dimension, the scale developed by [18] was used. These four constructs were measured from their respective items under a seven-point Likert scale ranging from 1 “strongly disagree” to 7 “strongly agree”.

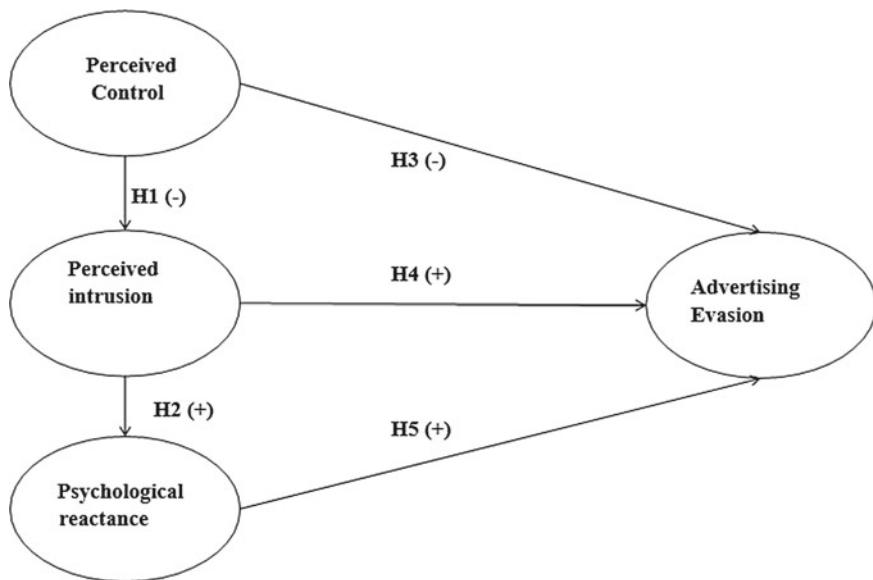


Fig. 1 The proposed relationship model

3.2 Data Collection

The population under study was 30.5 million Colombian Internet users with accounts in social networks and a profile on Facebook. The contracted company contacted the panel users via email. Once users agreed to participate in the study, the hired company would filter the panelists to see if they were Facebook users. After filtering, users accessed the home page where the purpose of the research and the task to be performed on the network (search for a promotional code) were explained. The fieldwork was developed between the months of July and August 2019. Finally, a number of 3542 valid questionnaires were reached with a sampling error of 3.33% for the estimation of a proportion.

4 Analysis and Results

The study sample was composed of 51.3% men and 48.7% women. Sixty percent were between the ages of 25 and 45, and 95% had secondary or university education. The results of the study were analyzed in three stages: (1) previous analysis to examine the unidimensionality of the constructs (exploratory factorial analysis), (2) analysis of the reliability of the scales (internal consistency) as well as the validity of the proposed variables (composite reliability and extracted variance) through a confirmatory factorial analysis with the intention of discriminating between the correct

inclusion or elimination of items in the proposed scales. In addition to the previous verification of the psychometric properties of the constructions that are part of the model, (3) the causal analysis contrasts the proposed structural relations. For data analysis, SPSS 20 and AMOS 18 software was used.

First, exploratory factor analysis was performed using the varimax rotation principal component method. The first analysis showed that all the items proposed should be maintained, as most of them showed values above the acceptance values ($ij > 0.4$) [22]. Thus, the KMO (Kaiser–Mayer–Olkin) index was 0.800, suggesting that the data are sufficiently related to each other, with factorial analysis being feasible. By using the four constructs of the model, the variance is explained by 73.8% (see Table 1). Finally, the composite reliability and extracted variance data are higher than recommended in all cases (see Table 2).

In order to evaluate the psychometric properties of the proposed scales, Confirmatory Factor Analysis (CFA) was used with the use of structural equations, using AMOS software. Maximum likelihood with bootstrapping [21] was used as the estimation method under 500 iterations. The selection of this procedure allows for a

Table 1 Main component analysis

Items	Main components			
	Factor 1—Perceived intrusion	Factor 2—Perceived control	Factor 3—Psychological reactance	<i>m</i>
INTRU1	0.512	**	**	**
INTRU2	0.747	**	**	**
INTRU3	0.858	**	**	**
INTRU4	0.914	**	**	**
INTRU5	0.924	**	**	**
INTRU6	0.872	**	**	**
CONT1	**	0.935	**	**
CONT2	**	0.821	**	**
CONT3	**	0.974	**	**
REACT1	**	**	0.8478	**
REACT2	**	**	0.796	**
REACT3	**	**	0.847	**
EVA1	**	**	**	0.958
EVA2	**	**	**	0.934
EVA3	**	**	**	0.882
α Cronbach	0.795	0.878	0.771	0.947
Cumulative variance (%)	24.775	44.447	60.257	74.13

Extraction method: a main component analysis. Rotation method: Varimax Normalization with Kaiser. **: weights less than 0.4

Table 2 Correlation and reliability matrix

Variables	Composite reliability	Variance extracted	Correlation matrix			
			REACT	CONT	INTRU	EVA
REACT	0.771	0.557	0.778	–	–	–
CONT	0.858	0.702	0.030	0.888	–	–
INTRU	0.872	0.601	0.314	–0.088	0.757	–
EVA	0.899	0.814	0.257	–0.130	0.201	0.878

Recommended Reliability values: FC > 0.7; Convergent validity: FC > VE; Discriminant validity: VE > 0.5 [21]

Table 3 Model fitting indices

Model fitting index	Recommended values	Results in the study
GFI	≥0.92	0.947
AGFI	≥0.79	0.910
CFI	≥0.92	0.942
NFI	≥0.92	0.955
RMSEA	≤0.07	0.071

global adjustment of the proposed model, analyzing various statistics were corrected by assuming the non-normality of the data. The analysis was carried out in successive stages: initially, the validity of the proposed scales was checked, then the validity of the model was checked on the basis of the adjustment of the data to the proposed model, a situation which was essential in order to verify the initial hypotheses proposed. The goodness-of-fit indices can be seen in Table 3, showing that they are within the acceptable range of values. This allows us to indicate that the data fits well with the proposed model.

The results of the structural model are presented graphically in Fig. 2. The values of R2 for perceived intrusion, psychological reaction, and advertising evasion are shown as 7.0%, 15%, and 9%, respectively.

It should be noted that the structural relations proposed are significant, except for hypothesis H1, which linked the perceived loss of control with the intrusion generated, which did not find empirical support to be verified. Therefore, the perceived control had a negative and direct influence on avoidance (H3: $\beta = -0.012^*$), the intrusion had a direct and positive effect on the user's social network reactance levels (H2: $\beta = 0.33^*$), the intrusion had a direct positive effect on avoidance (H4: $\beta = 0.12^*$) and the psychological reactance had a direct and positive effect on avoidance (H5: $\beta = 0.22^*$). The standardized coefficients (β) between relationships can be seen in Fig. 2, all of which are significant (* $p < 0.01$).

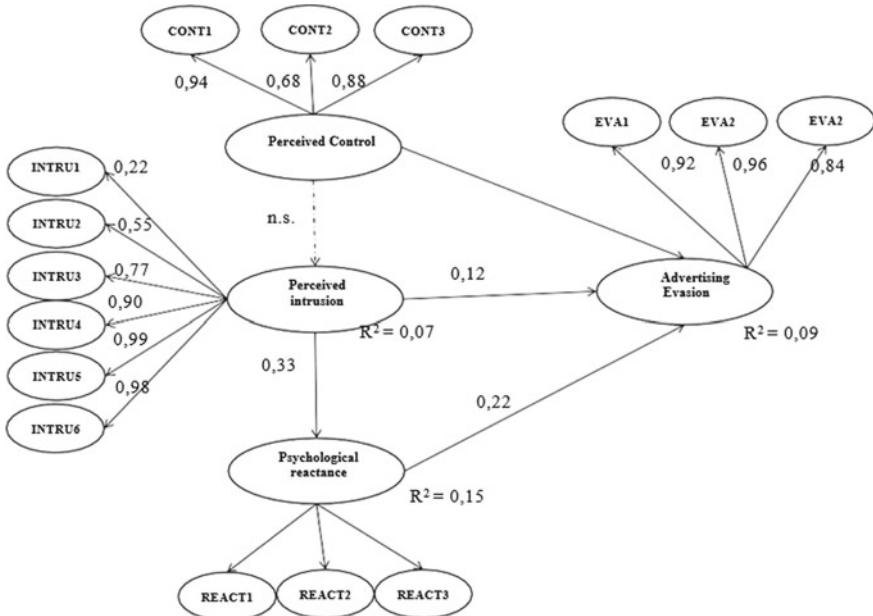


Fig. 2 Results of the structural model

5 Conclusions

The analysis of the literature shows the interest that supposes to know the antecedent factors of the advertising evasion in social networks. This is due, in part, to the increasing advertising investment being made in these media. Thus, it is understood that as the user tends to avoid advertising, the objectives set out in it will no longer be valid or will lose effectiveness. Therefore, the literature on evasion has studied this concept fundamentally, from an advertising effectiveness approach.

In order for online users not to initiate avoidance mechanisms, advertising should not compromise freedom of navigation or prevent the completion of tasks. Otherwise, if the perceived control during navigation is reduced by the presence of advertising, based on the theory of psychological reactance of [14], the consumer will try to restore independence or control of the situation by advertising evasion. These types of defensive responses are not good for the advertised brands. They have associated negative consumer attitudes towards the brand and the advertisement that ultimately harm the objectives of the advertising campaigns.

For the present study, a model of relations precedent of the advertising evasion in social networks was proposed, structured according to the relations proposed by the literature, and evaluated for the first time in the field of social networks. The variables and relationships proposed as a background to advertising evasion considered the particular way in which social network users interact, as well as the amount of

time spent on social networks which, together with the type of information shared on them, means that advertising on social networks can be perceived as annoying. If this happens, negative attitudes related to advertising emerge; in particular, the literature emphasizes the perception that advertising in social networks is intrusive, as it occupies a place that does not belong to it. So, when advertising is considered as intrusive, it will increase the levels of the psychological reactance of the social network user, trying to restore his freedom in the social network. In addition, the literature proposes the loss of control perceived as an antecedent of the advertising intrusion, however, this relation was rejected in the study.

References

1. Fayyad U, Piatetsky-Shapiro G, Smyth P (1996) From data mining to knowledge discovery in databases. *AI Magazine* 17(3):37–54
2. Witten I, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann Publishers
3. WEKA 3 (2016) Data mining software in Java homepage. <https://www.cs.waikato.ac.nz/ml/weka/>
4. Singh Y, Chanuhan A (2009) Neural networks in data mining. *J Theor Appl Inf Technol* 5(1):37–42
5. Orallo J, Ramírez M, Ferri C (2008) Introducción a la Minería de Datos. Pearson Education
6. Aladag C, Hocaoglu G (2007) A tabu search algorithm to solve a course timetabling problem. *Hacettepe J Math Stat* 53–64
7. Moscato P (1989) On evolution, search, optimization, genetic algorithms and martial arts: towards Memetic algorithms. Caltech Concurrent Computation Program (report 826) (1989)
8. Frausto-Solís J, Alonso-Pecina F, Mora-Vargas J (2008) An efficient simulated annealing algorithm for feasible solutions of course timetabling. Springer, pp 675–685
9. Joudaki M, Imani M, Mazhari N (2010) Using improved Memetic algorithm and local search to solve University Course Timetabling Problem (UCTTP). Islamic Azad University, Doroud, Iran
10. Coopers PWH (2014) IAB internet advertising revenue report. http://www.iab.net/insights_research/industry_data_and_landscape/adrevenuerreport
11. Tuzhilin A (2006) The lane's gifts v. Google report. Official Google Blog: Findings on invalid clicks, pp 1–47
12. Ponce H, Ponce P, Molina A (2014) Artificial organic networks: artificial intelligence based on carbon networks. In: Studies in computational intelligence, vol 521. Springer
13. Ponce H, Ponce P, Molina A (2013) A new training algorithm for artificial hydrocarbon networks using an energy model of covalent bonds. In: 7th IFAC conference on manufacturing modelling, management, and control, vol 7, issue 1, pp 602–608
14. Viloria A, Lis-Gutiérrez JP, Gaitán-Angulo M, Godoy ARM, Moreno GC, Kamatkar SJ (2018) Methodology for the design of a student pattern recognition tool to facilitate the teaching—learning process through knowledge data discovery (big data). In: Tan Y, Shi Y, Tang Q (eds) Data mining and big data. DMBD 2018. Lecture notes in computer science, vol 10943. Springer, Cham
15. Moe WW (2013) Targeting display advertising. Advanced database marketing: Innovative methodologies & applications for managing customer relationships. Gower Publishing, Londres
16. Granitto PM, Furlanello C, Biasioli F, Gasperi F (2006) Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products. *Chemometr Intell Lab Syst* 83(2):83–90

17. Kuhn W, Wing J, Weston S, Williams A, Keefer C et al (2012) Caret: classification and regression training. R package, v515
18. Miller B, Pearce P, Grier C, Kreibich C, Paxson V (2011) What's clicking what? Techniques and innovations of today's clickbots. In: Detection of intrusions and malware, and vulnerability assessment. Springer, pp 164–183
19. Kamatkar SJ, Tayade A, Viloria A, Hernández-Chacín A (2018) Application of classification technique of data mining for employee management system. In: International conference on data mining and big data. Springer, Cham, pp 434–444
20. Kamatkar SJ, Kamble A, Viloria A, Hernández-Fernandez L, Cali EG (2018) Database performance tuning and query optimization. In: International conference on data mining and big data. Springer, Cham, pp 3–11
21. Ellison NB, Steinfield C, Lampe C (2007) The benefits of Facebook "Friends:" Social capital and college students' use of online social network sites. *J Comput Med Commun* 12(4):1143–1168
22. Silva J, Hernández-Fernández L, Cuadrado ET, Mercado-Caruso N, Espinosa CR, Ortega FA, Hernández H, Delgado GJ (2019) Factors affecting the big data adoption as a marketing tool in SMEs. In: International conference on data mining and big data. Springer, Singapore, pp 34–43

Classification of Academic Events from Their Textual Description



Jesús Silva, Nicolas Elias María Santodomingo, Ligia Romero, Marisol Jorge, Maritza Herrera, Omar Bonerge Pineda Lezama, and Francisco Javier Echeverry

Abstract The aim of this paper is to compile dictionaries of slang words, abbreviations, contractions and emoticons to help the preprocessing of texts published in social networks. The use of these dictionaries is intended to improve the results of the tasks related to data obtained from these platforms. Therefore, a hypothesis was evaluated in the task of identifying author profiles (author profiling).

Keywords Lexicon · Social networks · Author profiling · Text classification

1 Introduction

The use of social networks is steadily increasing worldwide. Hundreds of users register daily in the different existing platforms, therefore, the content extracted from

J. Silva (✉) · N. E. M. Santodomingo · L. Romero · M. Jorge · M. Herrera
Universidad de la Costa, Barranquilla, Colombia
e-mail: jesussilvaUPC@gmail.com

N. E. M. Santodomingo
e-mail: nmaria1@cuc.edu.co

L. Romero
e-mail: lromero11@cuc.edu.co

M. Jorge
e-mail: marisol.jorge@upc.pe

M. Herrera
e-mail: luz.herrera@upc.pe

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

F. J. Echeverry
Corporación Universitaria Minuto de Dios—Uniminuto, Barranquilla, Colombia
e-mail: francisco.echeverry@uniminuto.edu

social networks is fundamental for tasks such as sentiment analysis [1], author profile detection [2], author identification [3], opinion mining [4], plagiarism detection [5], calculation of similarity between texts [6] and to develop robust systems to help decision making in related areas such as politics, education, economy, among others.

The processing of messages posted on social networks is not an easy task to solve [7]. Messages published on these platforms are generally short (hundreds of words) and do not follow the conventional rules of language, for example, slang words, abbreviations and emoticons are often used to compose texts [8].

The objective of this work is to obtain information about the author of a text, specifically his age and gender, by analyzing messages published by the author on Twitter.

2 Related Studies

This section presents some of the main studies that demonstrate the importance of the data preprocessing phase in several automatic text processing tasks. Proper preprocessing leads to proper analysis and helps to increase the accuracy and efficiency of text analysis processes. Some of the challenges faced when preprocessing social network texts are presented in detail in the study of [9].

In the study developed by [10], problems related to the processing of messages obtained from social networks are discussed. The reported results indicate that the system is capable of reducing the average error per message from 15 to 5%.

The study conducted by [11] presents some of the text preprocessing steps that should be undertaken to improve the quality of the messages obtained through Twitter. Among the techniques mentioned are: removing URLs, special characters, repeated letters of a word and question words (what, when, how, etc.). This study showed that the result of the sentiment analysis task improves considerably by performing the steps mentioned above.

In the research conducted by [12], a combination of different preprocessing techniques was used such as HTML tag cleaning, abbreviation extension, negation word handling, stop word removal and use of methods to reduce a word to its root. The aim of this paper is to analyze the feelings about opinions related to movies. The authors reported that appropriate text preprocessing can improve the performance of the classifier and considerably increase the results of the sentiment analysis task.

In [13], the authors propose the spelling correction of messages found in social networks. This includes repeated letters, omitted vowels, substitution of letters with numbers (typically syllables), use of phonetic spelling, use of abbreviations and acronyms. In a data-driven approach [14], a URL filter is applied in combination with standard text preprocessing techniques.

As can be observed, there are various investigations related to the preprocessing of texts published on social networks. In this work, a lexical resource is presented and its importance for the task of identifying author profiles is demonstrated. The

following section describes the procedure used for the compilation of the dictionaries and shows examples of their content.

3 Creation of the Social Network Lexicon

This research includes the analysis and compilation of shortened vocabulary (used in social networks) for the creation of dictionaries in various languages such as English, Spanish, Dutch and Italian. The dictionaries were compiled for these four languages since they are necessary for the preprocessing of tweets for the task of identifying the author of Etica Editorial (EE) 2018 [15]. The EE is an evaluation laboratory on plagiarism discovery, authorship and misuse of social software.

The type of shortened vocabulary generally used in social networks can be divided into three categories: slang words, abbreviations and contractions. Each category is briefly described below [16, 17].

Slang words: structured vocabulary in a given language, usually used among people in the same social group. It is a metalanguage used to enrich expressions, and the words have an intact phonological representation. Some examples of slang words found in the Spanish language are bb (bebé), xq (porque), dnd (donde), tb (tambien), tqm (te quiero mucho) and xfa (por favor).

Abbreviations are orthographic representations of a word or phrase. Also included in this category are acronyms, which are formed from the initial letters of a name or parts of words or phrases. Within this category, the following examples can be found: Arq. (Arquitecto), Sr. (Señor), NY (New York), kg. (kilogram), Av. (Avenue), among others.

Contractions occur when two words are reduced to one and an apostrophe takes the place of the missing letter. There are many rules between languages and create contractions. However, this research will not take into account any of them. Examples of contractions are: al (a el) and del (de el).

Another type of element that frequently appears in social network messages is the emoticon. Emoticons are typographic visualizations that allow to represent the facial expressions of emotions, that is to say, it is a way to give an emotive load to a text. Two styles of emoticons were included, known as Western and Eastern. The western style is commonly used in the United States and Europe. Emoticons in this style are written from left to right, as if a face is turned 90° to the right. The emoticons shown below belong to this style: :-) (smiling face), :-/ (doubtful face) and :-o (surprised face). On the other hand, there are the eastern type emoticons that are popular in East Asia and unlike the western style, the eastern emoticons are not rotated. In this style, the eyes are often seen as an important characteristic of the expression. Some examples of this style are (^v^) (smiley face), ((+ +)) (doubtful face) and (o.o.) surprised face.

This paper is a compilation of abbreviated vocabulary and emoticons that are generally used in social networks. The following describes the compiling process of the dictionaries [18, 19]:

Table 1 # Entries in dictionary

Dictionary type	Dutch	Italian	English	Spanish
Abbreviations	2352	114	2147	624
Slangs	300	452	1478	1047
Contractions	20	55	201	23
Emoticons	–	–	521	854
Totals	2672	621	4347	2548

1. Search and identification of Web sites that are used as sources for the extraction of slang word lists, abbreviations and contractions in the four languages (English, Spanish, Italian and Dutch).
2. Manual or semi-automatic extraction of all slang words, abbreviations and contractions along with their respective meanings from each Web site in the different languages.
3. Identification and merging of all files in the same category. Cleaning, formatting and standardization of each file, eliminating duplicates. Manual verification of the meanings of each dictionary entry.

Through the process described above, twelve dictionaries were created, divided into four languages, one for each category (slang words, abbreviations and contractions). The dictionaries are freely available on the Website2, which also includes a brief description of the dictionaries, a list of Web sites used to collect the three vocabulary categories for the four languages, and a list of Web sites used to obtain the emoticons. In the case of the dictionary of slang words in Spanish, entries were also included from the study on [20], in which a manual extraction of slang words from a collection of Twitter messages was performed.

Each dictionary was stored in a different file, the elements are ordered alphabetically and the information is coded using two columns separated by a tab. The first column corresponds to an entry of word slang, abbreviation or contraction, depending on the nature of the dictionary, and the second column corresponds to the meaning of the corresponding entry.

Table 1 presents the statistics of each dictionary, where it can be seen that there is a significant number of slang words available for English and Spanish, while for Dutch and Italian the number of entries is lower. On the other hand, it can be seen that there is a large number of abbreviations in the Dutch language. The total number of entries in the social network lexicon is 10,188.

4 Identification of Author Profiles

The task of identifying author profiles consists of identifying some aspects of a person such as their age, sex, or some behavioral features based on the analysis of text samples. The profile of an author can be used in many areas, for example, in

forensic sciences to obtain the description of a suspect by analyzing the messages published in social networks [21].

In recent years, different methods have been proposed to address the task of identifying author profiles, most of them using automatic learning techniques, data mining and natural language processing. From a self-learning point of view, the author profile identification task can be considered as a multi-class and multi-label classification problem, where each S_i element of a set of text samples $S = \{S_1, S_2, \dots, S_n\}$ multiple tags are assigned (l_1, l_2, \dots, l_k), each one representing one aspect of the author (gender, age, behavioral features) and the value assigned in each tag represents a category within the corresponding aspect. The problem is translated into the construction of an M classifier that assigns several labels to the unlabeled texts [22].

In the training stage, a vectoral representation of each of the example texts in each category is obtained, i.e., $v_i = \{v_{i1}, v_{i2}, \dots\}$ where v_{ij} is the vectoral representation of the example text S_i [5]. A classifier is then trained to use the vector representation of the labeled samples. In this paper, a Support Vector Machine (SVM) is used and different classification models are generated for each aspect of an author's profile, i.e., a model is learnt to determine the age and another model to determine the gender of an author.

The characteristics used in this paper are based on a vectoral representation of the frequency of occurrence of words using the standard Bag Of Words (BOW) model, which has proven to be effective in tasks related to the characterization of authors in previous studies [22]. In this paper, only the frequency of words that occur in the training text set is used to construct the model of representation.

In the test or evaluation phase, the vectoral representation of the unlabeled texts is obtained using the same characteristics extracted in the training stage. Then, the classifier is used to assign values to the labels of each aspect of the author profile for each user of the test set.

In order to evaluate the usefulness of these dictionaries, the corpus designed for the task of identifying the author profiles of the EE 2018 is used. The corpus is composed of tweets in four different languages: English, Spanish, Italian and Dutch. Each language has a set of tagged tweets corresponding to the age and gender of the author of that tweet. The values of the gender class tags can be: male or female. The values of the age class tags can be: 18–24, 25–34, 35–49, 50–xx.

The EE-2018 author profile identification corpus is partially available. Due to the organizers' policy, only the training corpus has been released. In this sense, the experiments were performed using the training corpus and a 10-layer cross-validation was performed to evaluate the proposal.

Tables 2 and 3 present the accuracy obtained for the gender and age classes respectively, with and without corpus preprocessing. It can be concluded that the best results were obtained for each language when preprocessing is done using these dictionaries.

The preprocessing stage basically consists of identifying within the corpus of words found in the dictionaries and replacing them with their respective meanings. It is worth mentioning that this study does not involve any process of disambiguation

Table 2 Results obtained for the classification of gender

Language	Liblinear SVM	
	Without preprocessing	With preprocessing
English	75.23	77.35
Spanish	81.45	82.45

Table 3 Results obtained for the classification of age

Language	Liblinear SVM	
	Without preprocessing	With preprocessing
English	75.02	77.24
Spanish	69.47	70.02

of the meaning of words and therefore, only the first available meaning is selected for each term.

5 Conclusions

This paper presents a social network lexicon containing dictionaries of slang words, abbreviations, contractions and emoticons that are most popular on social networks. The resource contains dictionaries in English, Spanish, Dutch, and Italian. Additionally, the methodology of data collection is described, the URLs used as sources for the creation of each dictionary are listed, and the process of standardization of the dictionaries is explained. Later, a description of the structure of the dictionaries and a description of the length of each dictionary are provided.

When using the dictionaries for preprocessing texts, it was noticed that there are some terms commonly used in social networks that are not present in the Web sources, especially for the English, Italian and Dutch languages.

References

- Schler J, Koppel M, Argamon S, Pennebaker JW (2006) Effects of age and gender on blogging. In: Computational approaches to analyzing weblogs, Papers from the 2006 AAAI spring symposium, Technical Report SS-06-03, Stanford, California, USA, 27–29 March 2006, pp 199–205
- Viloria A, Lis-Gutiérrez JP, Gaitán-Angulo M, Godoy ARM, Moreno GC, Kamatkar SJ (2018) Methodology for the design of a student pattern recognition tool to facilitate the teaching—learning process through knowledge data discovery (Big Data). In: Tan Y, Shi Y, Tang Q (eds) Data mining and big data. DMBD 2018. Lecture notes in computer science, vol 10943. Springer, Cham
- Tang J (2016) AMiner: mining deep knowledge from big scholar data. In: Proceedings of the 25th international conference ComEEion on world wide web. International world wide web

- conferences steering committee, Republic and Canton of Geneva, Switzerland, pp 373–373
4. Obit JH, Ouelhadj D, Landa-Silva D, Vun TK, Alfred R (2011) Designing a multi-agent approach system for distributed course timetabling, pp 103–108. <https://doi.org/10.1109/his.2011.6122088>
 5. Lewis MRR (2006) Metaheuristics for university course timetabling. Ph.D. Thesis, Napier University
 6. Deng X, Zhang Y, Kang B, Wu J, Sun X, Deng Y (2011) An application of genetic algorithm for university course timetabling problem, pp 2119–2122. <https://doi.org/10.1109/ccdc.2011.5968555>
 7. Mahiba AA, Durai CAD (2012) Genetic algorithm with search bank strategies for university course timetabling problem. Procedia Eng 38:253–263
 8. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Knowl Data Eng 17:734–749
 9. Camacho-Vázquez V, Sidorov G, Galicia-Haro SN (2016) Machine learning applied to a balanced and emotional corpus of tweets with many varieties of Spanish. Submitted
 10. Nguyen K, Lu T, Le T, Tran N (2011) Memetic algorithm for a university course timetabling problem, pp 67–71. https://doi.org/10.1007/978-3-642-25899-2_10
 11. Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. Procedia Comput Sci 17, 26–32. In: First international conference on information technology and quantitative management
 12. Hemalatha I, Varma DGPS, Govardhan DA (2012) Preprocessing the informal text for efficient sentiment analysis. Int J Emerg Trends Technol Comput Sci (IJETTCS) 1(2):58–61
 13. Pinto D, Vilarinho-Ayala D, Alemán Y, Gómez-Adorno H, Loya N, Jiménez-Salazar H (2012) The soundex phonetic algorithm revisited for sms-based information retrieval. In: II Spanish conference on information retrieval CERI 2012
 14. Torres-Samuel M, Vásquez C, Viloria A, Lis-Gutiérrez JP, Borrero TC, Varela N (2018) Web visibility profiles of Top100 Latin American universities. In: International conference on data mining and big data. Springer, Cham, pp 254–262
 15. Henao-Rodríguez C, Lis-Gutiérrez JP, Bouza C, Gaitán-Angulo M, Viloria A (2019) Citescore of publications indexed in Scopus: an implementation of panel data. In: International conference on data mining and big data. Springer, Singapore, pp 53–60
 16. Peersman C, Daelemans W, Van Vaerenbergh L (2011) Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on search and mining user-generated contents. New York, NY, USA, ACM, pp 37–44
 17. Nguyen D, Gravel R, Trieschnigg D, Meder T (2013) How old do you think i am?: a study of language and age in twitter. In: Proceedings of the seventh international AAAI conference on weblogs and social media. ICWSM 2013
 18. Rangel F, Rosso P (2013) Use of language and author profiling: Identification of gender and age. In: Proceedings of the 10th workshop on natural language processing and cognitive science (NLP-CS-2013)
 19. Bedford D (2013) Evaluating classification schema and classification decisions. Bull Am Soc Inf Sci Technol 39:13–21
 20. Toutanova K, Klein D, Manning C, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human language technology conference (HLT-NAACL 2003)
 21. McGrail MR, Rickard CM, Jones R (2006) Publish or perish: a systematic review of interventions to increase academic publication rates. High Educ Res Dev 25:19–35
 22. Costas R, van Leeuwen TN, Bordons M (2010) A bibliometric classificatory approach for the study and assessment of research performance at the individual level: the effects of age on productivity and impact. J Am Soc Inf Sci 61:1564–1581

Geosimulation as a Tool for the Prevention of Traffic Accidents



Amelec Viloria, Noel Varela, Luis Ortiz-Ospino,
and Omar Bonerge Pineda Lezama

Abstract Traffic accidents represent a never-ending tragedy, and according to the World Health Organization (2018), 1.33 million people die in the world every year [1]. Most efforts in modeling phenomena of a dynamic nature have focused on working with static snapshots that reduce the natural depth of the world's dynamics to simplify it, abstracting perspectives that are fixed or static in some way. In the case of traffic accidents, most models used are those based on the principle of cause and effect, where the appearance of one or several variables gives rise to the event, like a domino effect. In this research, the problem of traffic accident avoidance was addressed through the use of a dynamic type model, based on the technique called geosimulation, where all the elements involved are interrelated.

Keywords Traffic accidents · Geosimulation · Agent-based modeling · Geographic information systems · Dynamic models · Traffix

1 Introduction

Currently, geographic information sciences focus on capturing the dynamism of geographic environments and describing the semantics of their entities (objects and events), as well as the spatial relationships between them over time. There is a large

A. Viloria (✉) · N. Varela

Universidad de La Costa, St. 58 #66, Barranquilla, Atlántico, Colombia
e-mail: aviloria7@cuc.edu.co

N. Varela
e-mail: nvarela2@cuc.edu.co

L. Ortiz-Ospino
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: lortiz27@unisimonbolivar.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

number of these that, by their nature, are not static, that is, their behavior is dynamic within a geographic environment [2]. If a traffic accident is defined as a collision in a road environment, resulting from a combination of factors related to the system components, the best way to model and project it is through a dynamic model [3].

In this study, the modeling and prevention of traffic accidents is approached through the use of a dynamic model, using the technique called geosimulation. The objective was to demonstrate that dynamic models provide greater benefits in terms of their representation and information, which is generated for decision making in their understanding and prevention. The methodology used was based on the solution of a case study and the comparison of results between the different types of models.

2 Methods and Data

In order to show the advantages and disadvantages between the different models, a representation of Caracas Avenue in Bogotá, Colombia, was made in the section that includes the junction of 26th to 36th streets.

The first model used was based on a geographic information system (GIS). With the increasing availability of road accident data and the popularity of GIS, accident analysis with this tool allowed, at a macroscopic level, the identification of areas with high accident incidence, the frequency of accidents related to each crossing, accident rate analysis and spatial queries, which allow the user to analyze and manipulate data quickly and identify possible problem areas and zones [4, 5]. The analysis of traffic accidents using a GIS was as follows:

- (a) First, the geographical location of the accidents was chosen, which was obtained from [6], where the information on time, geographical coordinates and type of accident is obtained.
- (b) The first analysis with the GIS was the elaboration of thematic maps, which show, based on the size of the points, the places where the greatest number of accidents occur.
- (c) A geostatistical function was applied to the number of geocoded accidents to project them over time, using the technique known as Kriging3, which calculates an unbiased linear estimator of a characteristic studied, which projects the number of accidents in the same unit of time as the input data [4].
- (d) Another function that was used was the elaboration of a heat map, where the areas where the greatest number of accidents are concentrated are emphasized in red.

The second type of model was based on neural networks, which are the preferred tool for many data mining applications, since it is predictive because of its power, flexibility and ease of use. One of the most accurate definitions is that of neural networks [7]. The use of neural networks allowed to project the future traffic accidents and the place where they happen, through an equation where the advantage is that it imitates the pattern of behavior of independent or predictive variables, which in this

case, were the number of historical accidents and vehicle flow that occurred at the intersections of the avenue under study [8]. It was obtained using the SPSS software, and the type of neural network used was a multi-layer perceptual type, which allowed the simulation of non-linear behaviors (i.e., it allows the reproduction of defined patterns that show the data) [9], generating a predictive model for a dependent variable for traffic accidents.

The third model was a dynamic, agent-based model. The stages that were followed for its development were the following:

- (a) Virtualization of the real situation, where the elements to be included in the model, their characteristics and behaviors are defined. In order to achieve this, a review was made of the state of the art of road traffic accidents and the theories on the way they are generated, mainly those published by the World Health Organization (WHO) in its World Report on Road Traffic Injury Prevention [10].
- (b) The second stage was the implementation of the model in a computer platform. For the selection of the tool, a review of several studies was made around the comparison of several agent-based modeling environments, some of them were [11–13].

The Traffix platform was adequate for the development of the model, since it has the following characteristics

- (1) It is based on Repast and works on the Eclipse development environment. It is written in Java language, which makes it possible to add new elements and behaviors to the model [14].
- (2) The road models, comply with the classical theory of the basic measures of vehicle flow and its relationship between them, which makes the platform able to simulate real environments. The main variables are: speed (V), which is the average speed of vehicles, density (K), which is the number of vehicles that are in a lane at speed V in a linear unit of distance, and flow (Q), which is the number of vehicles that pass in a lane in a unit of time [15], also complying with the following equations [16]:

$$V = V_{\max} \left(1 - \frac{K}{K_d} \right) \quad (1)$$

$$Q = V_{\max} \left(K - \frac{K^2}{K_d} \right) \quad (2)$$

- (c) The description of the functioning of each of the elements that form of the model, as well as the interaction between them, is carried out using the methodology called Unified Software Development Process [17].

The algorithm of the model is explained through the description of its main classes [18, 19]:

- (1) insertNewStatCar class: Creates and allows a vehicle to enter the road network, which was previously created according to a number of lanes and their measurements, in proportion to the areas; when the vehicle enters the network it is assigned a color and a vehicle type according to a random number function called Random.uniform.nextIntFromTo (0, 4) and based on the number that was generated, four types of vehicles are incorporated: 0-Jeep, 1-PickUp, 2-Truck, 3-SlowTruck and 4-Standard-Car.
 - (2) assignDriver class: Assigns a driver to each vehicle, which will count a percentage to pass a stop, pass a vehicle when it has another in front, take its distance to the next vehicle and stop in a cruise; these values were assigned at random; the correct assignment of values is out of the scope of this study.
 - (3) getContainerConfigs class: When the vehicle enters the system, its position and speeds are determined. These are configured and obtained through a study of vehicle gauges in person and the formulas for determining the service speed in the Road Capacity Manual, the number of observations were determined according to a first sample and considering a normal distribution.
 - (4) desiredMove class: When the vehicle is moving to the next position, the user must make the decision whether to go through a traffic light, to pass because there are cars in front of him, to go through the intersection, to observe the color of the traffic light, to pass or to yield to another vehicle.
 - (5) moveCar class: Based on the above decision, the vehicle is moved.
 - (6) setCrashed class: When two vehicles touch the same point, it is considered as an accident and is counted; this is part of the contribution in programming and in the functionalities of the platform. This is also a window that shows the number of accidents that are registered.
 - (7) Specific behaviors such as blocking avenues or public transportation that are in third rows are placed in the model directly through the modification of the configuration files.
 - (8) Another of the contributions made in functionalities through programming, was the creation of a statistics window, in which the number of vehicles that exist in the system is graphically displayed, as well as the average time of stay.
- (d) Testing the model with real data, the model was fed with the following information, which was taken in the field [20].
- (1) Circulation conditions were taken in the time range of 6:00 am to 9 am (peak hours).
 - (2) Vehicle types: 4 and their maximum speeds were: 45, 35, 50 and 25 km/h, respectively; their decelerations: 3, 2, 2 and 2 km/h, respectively.
 - (3) Drivers' characteristics: passing when there is a 5% chance, taking the shortest route, 5% probability of passing a traffic light, distance of proximity to the vehicle in front 0.5 m and distraction factor of 5%.
 - (4) Running time 50,000 ticks, which is equivalent in time to 34 days.

- (5) With a contour map, the change in slope along the avenue was determined, which was reflected in the model as an increase or decrease in vehicle speed by 1%. The determination of this value was done randomly without any support, and is outside the scope of the project.

3 Discussion of Results

With the use of geographic information systems, the areas where accidents occur most frequently were determined, and a projection was made over time. This tool considers the current behavior of accidents; i.e., this pattern will be repeated where they occur most often. Figure 1 shows a thematic map, the size of the circle is proportional to the number of accidents. Figure 2 shows a heat map, the dark areas indicate greater danger.

With respect to the results of the neural network model, the values of the constants of the equation are shown in Table 1 and the result of the forecast in Table 2. Reviewing these projections, it can be noted that the behavior pattern is similar, that is, where there are currently the greatest number of accidents, they will maintain the same behavior over time and will be directly proportional to vehicle flow. It also provides the number of accidents projected over time at the cruise level, but does not provide an understanding of the process, nor the specific actions that should be taken to reduce them.

The platform includes the main elements of a road environment, and the fact that no secondary factors are integrated does not invalidate the results, since the Traffic platform is a program based on open source objects, which allows the inclusion of additional factors. Furthermore, if the existence of an under-record in the count



Fig. 1 Thematic map of road accidents

Fig. 2 Heat map, dark color indicates increased presence of accidents

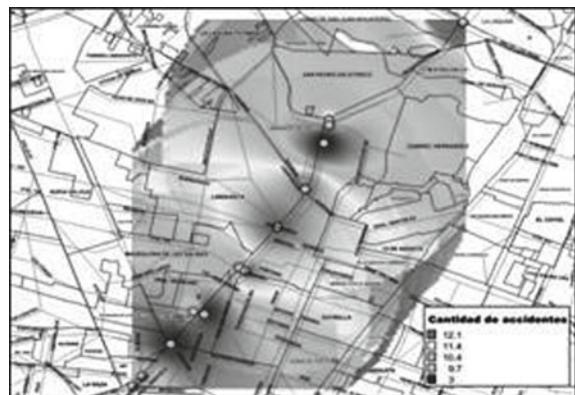


Table 1 Values of the constants of the predictive variable from traffic accidents and vehicle capacity

Predictor	Predicted			
	Hidden layer 1			Output layer
	H(1:1)	H(1:2)	H(1:3)	
Input layer	(Bias)	0.752	0.458	0.247
	VAR00003	0.235	-0.090	-0.589
Hidden layer 1	(Bias)			-0.825
	H(1:1)			-0.633
	H(1:2)			-1.9520
	H(1:3)			-0.3520

Table 2 Projected values of traffic accidents using neural networks

Intersection	Street 1	Street 2	2018	Vehicle capacity	Predicted values
1	Caracas Ave.	26	33	60	40
2	Caracas Ave.	27	30	50	33
3	Caracas Ave.	28	32	41	35
4	Caracas Ave.	29	31	35	28
5	Caracas Ave.	30	20	30	24
6	Caracas Ave.	35	20	30	23
7	Caracas Ave.	36	22	30	18
Total of accidents			258	300	267

of these is considered, the number provided is better, in addition to understanding the phenomenon and providing information for decision making and reducing them through the change of some main variables as mentioned in the following paragraphs.

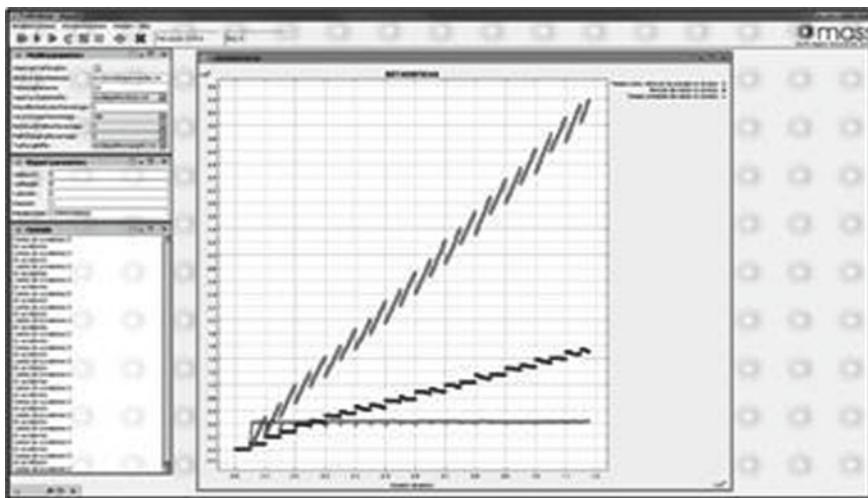


Fig. 3 Graphical interface showing the behavior of the vehicles in the study area

In the graphic interface that was developed, one can see the effect of the blocking of the three lanes that make up the public transport, where the number of vehicles (dark color) grows exponentially, as well as the time that they are on the road under study (Fig. 3).

If the lanes in the area are unblocked and the model is run again, the number of accidents is reduced to 10, which represents a decrease of 30% with respect to those registered during the year 2018. This allows to know the effect that the change of the different scenarios has according to the traffic accidents. Figure 4 shows the interface with this effect.

4 Conclusions

This study explored the use of geosimulation, and specifically, agent-based modeling to analyze and prevent road accidents, using an advanced simulation platform called Traffix. This platform was adapted by creating a virtual environment representative of traffic accidents, where the elements that make it up interact and provide a unique behavioral result.

According to the results, it is concluded that Traffix is the best tool for the projection, understanding and generation of information for decision making. Likewise, it achieves the reduction with the following advantages with respect to the static ones.

A higher level of aggregation in the information generated allows the phenomenon to be understood graphically.

It reports unforeseen behavior, since its elements have individual interaction.

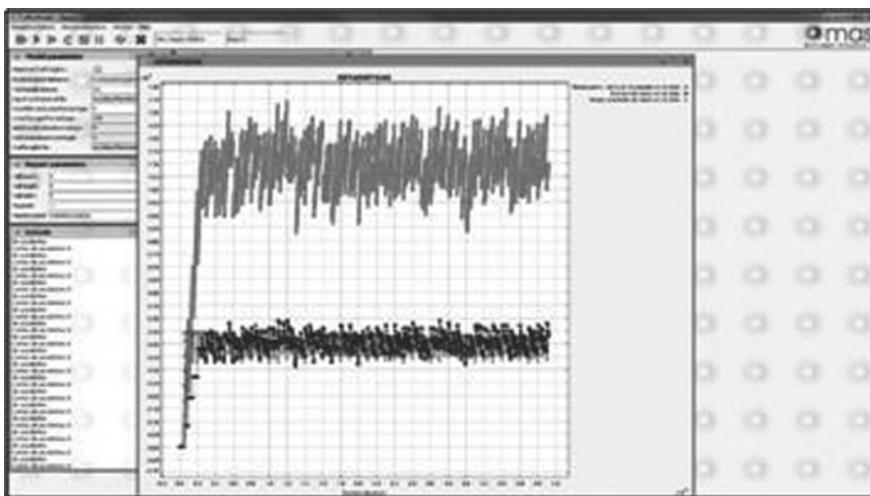


Fig. 4 Graphical interfaces of vehicle behavior with the unlocking of the lanes by public transport

It contains intelligent organisms that exhibit complex behavior, even though it inevitably has many weaknesses.

References

1. Marston S, Li Z, Bandyopadhyay S, Zhang J, Ghalsas A (2011) Cloud computing—the business perspective. *Dec Supp Syst* 51(1):176–189
2. Bifet A, De Francisci Morales G (2014) Big data stream learning with Samoa. Recuperado de https://www.researchgate.net/publication/282303881_Big_data_stream_learning_with_SAMOA
3. Lomax T, Schrank D, Turner S, Margiotta R (2003) Report for selecting travel reliability measures. Federal Highway Administration, Washington, DC
4. Anderson JA (2007) In: S.A. de C.V. (ed) Redes neuronales, 1a ed. Alfa Omega Gru-po Editor, México, pp 120–125. ISBN 9789701512654
5. Pardillo J, Sánchez V (2015) Apuntes de Ingeniería de Tránsito. ETS Ingenieros de Caminos, Canales y Puertos, Madrid, España
6. Skabardonis A, Varaiya P, Petty K (2003) Measuring recurrent and non-recurrent traffic congestion. *Transp Res Rec J Transp Res Board* 1856:60–68
7. U.S. Department of Transportation (2004) Archived data management systems—a cross-cutting study. Publication FHWA-JPO-05-044. FHWA, U.S. Department of Transportation
8. Yong-chuan Z, Xiao-qing Z, li-ting Z, Zhen-ting C (2011) Traffic congestion detection based on GPS floating-car data. *Proc Eng* 15:5541–5546
9. Thame L, Schaefer D (2016) Software defined cloud manufacturing for industry 4.0. *Procedia CIRP* 52:12–17
10. Viloria A, Neira-Rodado D, Lezama OBP (2019) Recovery of scientific data using intelligent distributed data warehouse. In: ANT/EDI40 2019, pp 1249–1254
11. Viloria A, Lezama OBP (2019) Improvements for determining the number of clusters in k-Means for innovation databases in SMEs. In: ANT/EDI40 2019, pp 1201–1206

12. Alcalá R, Alcalá-Fdez J, Herrera F (2007) A proposal for the genetic lateral tuning of linguistic fuzzy systems and its interaction with rule selection. *IEEE Trans Fuzzy Syst* 15(4):616–635
13. Alpaydin E (2004) Introduction to machine learning. The MIT Press, Massachusetts
14. Álvarez P, Hadi M, Zhan C (2010) Using Intelligent transportation systems data archives for traffic simulation applications. *Transp Res Rec J Transp Res Board* 2161:29–39
15. Bizama J (2012) Modelación y simulación mediante un microsimulador de la zona de influencia del Puente Llacolén. Memoria de Título, Universidad del Bío Bío
16. Levinson H, Rakha H (2010) Analytical procedures for determining the impacts of reliability mitigation strategies. Cambridge Systematics, Texas A&M University, Dowling Associates, Street Smarts
17. Cortés CE, Gibson J, Gschwender A, Munizaga M, Zúñiga M (2011) Commercial bus speed diagnosis based on GPS-monitored data. *Transp Res Part C* 19(4):695–707
18. Diker AC (2012) Estimation of traffic congestion level via FN-DBSCAN algorithm by using GPA data. In: 2012 IV international conference problems of cybernetics and informatics (PCI), Baku, Azerbaijan
19. Amelec V (2015) Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. *Adv Sci Lett* 21(5):1406–1408
20. Viloria A, Robayo PV (2016) Inventory reduction in the supply chain of finished products for multinational companies. *Indian J Sci Technol* 8(1)

Identification of Author Profiles Through Social Networks



Jesús Silva, Nicolas Elias María Santodomingo, Ligia Romero, Marisol Jorge, Maritza Herrera, Omar Bonerge Pineda Lezama, and Francisco Javier Echeverry

Abstract The aim of this paper is to compile dictionaries of slang words, abbreviations, contractions, and emoticons to help the pre-processing of texts published in social networks. The use of these dictionaries is intended to improve the results of the tasks related to data obtained from these platforms. Therefore, a hypothesis was evaluated in the task of identifying author profiles (author profiling).

Keywords Lexicon · Social networks · Author profiling · Text classification

J. Silva (✉) · M. Jorge · M. Herrera
Universidad de la Costa, Barranquilla, Colombia
e-mail: jesussilvaUPC@gmail.com

M. Jorge
e-mail: marisol.jorge@upc.pe

M. Herrera
e-mail: luz.herrera@upc.pe

N. E. M. Santodomingo · L. Romero
Universidad de la Costa, Barranquilla, Colombia
e-mail: imaria1@cuc.edu.co

L. Romero
e-mail: lromero11@cuc.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

F. J. Echeverry
Corporación Universitaria Minuto de Dios—Uniminuto, Barranquilla, Colombia
e-mail: francisco.echeverry@uniminuto.edu

1 Introduction

The use of social networks is steadily increasing worldwide. Hundreds of users register daily in the different existing platforms, therefore, the content extracted from social networks is fundamental for tasks such as sentiment analysis [1], author profile detection [2], author identification [3], opinion mining [4], plagiarism detection [5], calculation of similarity between texts [6], and to develop robust systems to help decision making in related areas such as politics, education, economy, among others.

The processing of messages posted on social networks is not an easy task to solve [7] Messages published on these platforms are generally short (hundreds of words) and do not follow the conventional rules of language, for example, slang words, abbreviations and emoticons are often used to compose texts [8].

The objective of this work is to obtain information about the author of a text, specifically his age, and gender, by analyzing messages published by the author on Twitter.

2 Related Studies

This section presents some of the main studies that demonstrate the importance of the data pre-processing phase in several automatic text processing tasks. Proper pre-processing leads to proper analysis and helps to increase the accuracy and efficiency of text analysis processes. Some of the challenges faced when preprocessing social network texts are presented in detail in the study of [9].

In the study developed by [10], problems related to the processing of messages obtained from social networks are discussed. The reported results indicate that the system is capable of reducing the average error per message from 15 to 5%.

The study conducted by [11] presents some of the text pre-processing steps that should be undertaken to improve the quality of the messages obtained through Twitter. Among the techniques mentioned are: removing URLs, special characters, repeated letters of a word and question words (what, when, how, etc.). This study showed that the result of the sentiment analysis task improves considerably by performing the steps mentioned above.

In the research conducted by [12], a combination of different pre-processing techniques was used such as HTML tag cleaning, abbreviation extension, negation word handling, stop word removal, and use of methods to reduce a word to its root. The aim of this paper is to analyze the feelings about opinions related to movies. The authors reported that appropriate text pre-processing can improve the performance of the classifier and considerably increase the results of the sentiment analysis task.

In [13], the authors propose the spelling correction of messages found in social networks. This includes repeated letters, omitted vowels, the substitution of letters with numbers (typically syllables), use of phonetic spelling, use of abbreviations,

and acronyms. In a data-driven approach [14], a URL filter is applied in combination with standard text pre-processing techniques.

As can be observed, there are various investigations related to the pre-processing of texts published on social networks. In this work, a lexical resource is presented and its importance for the task of identifying author profiles is demonstrated. The following section describes the procedure used for the compilation of the dictionaries and shows examples of their content.

3 Creation of the Social Network Lexicon

This research includes the analysis and compilation of shortened vocabulary (used in social networks) for the creation of dictionaries in various languages such as English, Spanish, Dutch, and Italian. The dictionaries were compiled for these four languages since they are necessary for the pre-processing of tweets for the task of identifying the author of Etica Editorial (EE) 2018 [15]. The EE is an evaluation laboratory on plagiarism discovery, authorship, and misuse of social software.

The type of shortened vocabulary generally used in social networks can be divided into three categories: slang words, abbreviations, and contractions. Each category is briefly described below [16, 17].

Slang words: structured vocabulary in a given language, usually used among people in the same social group. It is a metalanguage used to enrich expressions, and the words have an intact phonological representation. Some examples of slang words found in the Spanish language are bb (bebé), xq (porque), dnd (donde), tb (tambien), tqm (te quiero mucho), and xfa (por favor).

Abbreviations are orthographic representations of a word or phrase. Also included in this category are acronyms, which are formed from the initial letters of a name or parts of words or phrases. Within this category, the following examples can be found: Arq. (Arquitecto), Sr. (Señor), NY (New York), kg. (kilogram), Av. (Avenue), among others.

Contractions occur when two words are reduced to one and an apostrophe takes the place of the missing letter. There are many rules between languages to create contractions. However, this research will not take into account any of them. Examples of contractions are al (a el) and del (de el).

Another type of element that frequently appears in social network messages is the emoticon. Emoticons are typographic visualizations that allow us to represent the facial expressions of emotions, that is to say, it is a way to give an emotive load to a text. Two styles of emoticons were included, known as Western and Eastern. The western style is commonly used in the United States and Europe. Emoticons in this style are written from left to right as if a face is turned 90° to the right. The emoticons shown below belong to this style: :-) (smiling face), :-(/doubtful face), and :-o (surprised face). On the other hand, there are the eastern type emoticons that are popular in East Asia and unlike the western style, the eastern emoticons are not rotated. In this style, the eyes are often seen as an important characteristic of

the expression. Some examples of this style are (^v^) (smiley face), ((++)) (doubtful face) and (o.o.) surprised face.

This paper is a compilation of abbreviated vocabulary and emoticons that are generally used in social networks. The following describes the compiling process of the dictionaries [18, 19]:

1. Search and identification of web sites that are used as sources for the extraction of slang word lists, abbreviations, and contractions in the four languages (English, Spanish, Italian and Dutch).
2. Manual or semi-automatic extraction of all slang words, abbreviations and contractions along with their respective meanings from each website in the different languages.
3. Identification and merging of all files in the same category. Cleaning, formatting, and standardization of each file, eliminating duplicates. Manual verification of the meanings of each dictionary entry.

Through the process described above, twelve dictionaries were created, divided into four languages, one for each category (slang words, abbreviations, and contractions). The dictionaries are freely available on the website2, which also includes a brief description of the dictionaries, a list of websites used to collect the three vocabulary categories for the four languages, and a list of websites used to obtain the emoticons. In the case of the dictionary of slang words in Spanish, entries were also included from the study on [20], in which a manual extraction of slang words from a collection of Twitter messages was performed.

Each dictionary was stored in a different file, the elements are ordered alphabetically and the information is coded using two columns separated by a tab. The first column corresponds to an entry of word slang, abbreviation, or contraction, depending on the nature of the dictionary, and the second column corresponds to the meaning of the corresponding entry.

Table 1 presents the statistics of each dictionary, where it can be seen that there is a significant number of slang words available for English and Spanish, while for Dutch and Italian the number of entries is lower. On the other hand, it can be seen that there is a large number of abbreviations in the Dutch language. The total number of entries in the social network lexicon is 10,188.

Table 1 # Entries in dictionary

Dictionary Type	Dutch	Italian	English	English
Abbreviations	2352	114	2147	624
Slangs	300	452	1478	1047
Contractions	20	55	201	23
Emoticons	–	–	521	854
Totals	2672	621	4347	2548

4 Identification of Author Profiles

The task of identifying author profiles consists of identifying some aspects of a person such as their age, sex, or some behavioral features based on the analysis of text samples. The profile of an author can be used in many areas, for example, in forensic sciences to obtain the description of a suspect by analyzing the messages published in social networks [21].

In recent years, different methods have been proposed to address the task of identifying author profiles, most of them using automatic learning techniques, data mining, and natural language processing. From a self-learning point of view, the author profile identification task can be considered as a multi-class and multi-label classification problem, where each S_i element of a set of text samples $S = \{S_1, S_2, \dots\}$ If multiple tags are assigned (l_1, l_2, \dots, l_k), each one representing one aspect of the author (gender, age, behavioral features) and the value assigned in each tag represents a category within the corresponding aspect. The problem is translated into the construction of an M classifier that assigns several labels to the unlabeled texts [22].

In the training stage, a vectoral representation of each of the example texts in each category is obtained, i.e., $v_i = \{v_{i1}, v_{i2}, \dots\}$ where v_i is the vectoral representation of the example text S_i [5]. A classifier is then trained to use the vector representation of the labeled samples. In this paper, a Support Vector Machine (SVM) is used and different classification models are generated for each aspect of an author's profile, i.e., a model is learnt to determine the age and another model to determine the gender of an author.

The characteristics used in this paper are based on a vectoral representation of the frequency of occurrence of words using the standard Bag Of Words (BOW) model, which has proven to be effective in tasks related to the characterization of authors in previous studies [22]. In this paper, only the frequency of words that occur in the training text set is used to construct the model of representation.

In the test or evaluation phase, the vectoral representation of the unlabeled texts is obtained using the same characteristics extracted in the training stage. Then, the classifier is used to assign values to the labels of each aspect of the author profile for each user of the test set.

In order to evaluate the usefulness of these dictionaries, the corpus designed for the task of identifying the author profiles of the EE 2018 is used. The corpus is composed of tweets in four different languages: English, Spanish, Italian, and Dutch. Each language has a set of tagged tweets corresponding to the age and gender of the author of that tweet. The values of the gender class tags can be male or female. The values of the age class tags can be 18–24, 25–34, 35–49, 50–xx.

The EE-2018 author profile identification corpus is partially available. Due to the organizers' policy, only the training corpus has been released. In this sense, the experiments were performed using the training corpus and 10-layer cross-validation was performed to evaluate the proposal.

Tables 2 and 3 present the accuracy obtained for the gender and age classes,

Table 2 Results obtained for the classification of gender

Language	Liblinear SVM	
	Without pre-processing	With pre-processing
English	75.23	77.35
Spanish	81.45	82.45

Table 3 Results obtained for the classification of age

Language	Liblinear SVM	
	Without pre-processing	With pre-processing
English	75.02	77.24
Spanish	69.47	70.02

respectively, with and without corpus pre-processing. It can be concluded that the best results were obtained for each language when preprocessing is done using these dictionaries.

The pre-processing stage basically consists of identifying within the corpus of words found in the dictionaries and replacing them with their respective meanings. It is worth mentioning that this study does not involve any process of disambiguation of the meaning of words and therefore, only the first available meaning is selected for each term.

5 Conclusions

This paper presents a social network lexicon containing dictionaries of slang words, abbreviations, contractions, and emoticons that are most popular on social networks. The resource contains dictionaries in English, Spanish, Dutch, and Italian. Additionally, the methodology of data collection is described, the URLs used as sources for the creation of each dictionary are listed, and the process of standardization of the dictionaries is explained. Later, a description of the structure of the dictionaries and a description of the length of each dictionary is provided.

When using the dictionaries for pre-processing texts, it was noticed that there are some terms commonly used in social networks that are not present in the web sources, especially for the English, Italian, and Dutch languages.

References

1. Schler J, Koppel M, Argamon S, Pennebaker JW (2006) Effects of age and gender on blogging. In: Computational approaches to analyzing weblogs, papers from the 2006 AAAI spring symposium, technical report SS-06-03, Stanford, California, USA, 27–29 Mar 2006, pp 199–205
2. Viloria A, Lis-Gutiérrez JP, Gaitán-Angulo M, Godoy ARM, Moreno GC, Kamatkar SJ (2018) Methodology for the design of a student pattern recognition tool to facilitate the teaching—learning process through knowledge data discovery (Big Data). In: Tan Y, Shi Y, Tang Q (eds) Data mining and big data. DMBD 2018. Lecture notes in computer science, vol 10943. Springer, Cham
3. Tang J (2016) AMiner: mining deep knowledge from big scholar data. In: Proceedings of the 25th international conference companion on world wide web, international world wide web conferences steering committee, republic and canton of Geneva, Switzerland, pp 373–373
4. Obit JH, Ouelhadj D, Landa-Silva D, Vun TK, Alfred R (2011) Designing a multi-agent approach system for distributed course timetabling. pp 103–108 <https://doi.org/10.1109/his.2011.6122088>
5. Lewis MRR (2006) Metaheuristics for university course timetabling. Ph.D. Thesis, Napier University
6. Deng X, Zhang Y, Kang B, Wu J, Sun X, Deng Y (2011) An application of genetic algorithm for university course timetabling problem. pp 2119–2122 <https://doi.org/10.1109/ccdc.2011.5968555>
7. Mahiba AA, Durai CAD (2012) Genetic algorithm with search bank strategies for university course timetabling problem. Procedia Eng 38:253–263
8. Adomavicius G, Tuzhilin A (2005) Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans Know Data Eng 17:734–749
9. Camacho-Vázquez V, Sidorov G, Galicia-Haro SN (2016) Machine learning applied to a balanced and emotional corpus of tweets with many varieties of Spanish. Submitted
10. Nguyen K, Lu T, Le T, Tran N (2011) Memetic algorithm for a university course timetabling problem. pp 67–71. https://doi.org/10.1007/978-3-642-25899-2_10
11. Haddi E, Liu X, Shi Y (2013) The role of text pre-processing in sentiment analysis. In: Procedia computer science first international conference on information technology and quantitative management vol 17, pp 26–32
12. Hemalatha I, Varma DGPS, Govardhan DA (2012) Preprocessing the informal text for efficient sentiment analysis. Int J Emerg Trends Technol Comput Sci (IJETTCS) 1(2):58–61
13. Pinto D, Vilarinó-Ayala D, Alemán Y, Gómez-Adorno H, Loya N, Jiménez-Salazar H (2012) The soundex phonetic algorithm revisited for sms-based information retrieval. In: II Spanish conference on information retrieval CERI 2012
14. Torres-Samuel M, Vásquez C, Viloria A, Lis-Gutiérrez JP, Borrero TC, Varela N (2018) Web visibility profiles of Top100 Latin American universities. International conference on data mining and big data. Springer, Cham, pp 254–262
15. Henao-Rodríguez C, Lis-Gutiérrez JP, Bouza C, Gaitán-Angulo M, Viloria A (2019) Citescore of publications indexed in scopus: an implementation of panel data. International conference on data mining and big data. Springer, Singapore, pp 53–60
16. Peersman C, Daelemans W, Van Vaerenbergh L (2011) Predicting age and gender in online social networks. In: Proceedings of the 3rd international workshop on search and mining user-generated contents. New York, NY, USA, ACM, pp 37–44
17. Nguyen D, Gravel R, Trieschnigg D, Meder T (2013) how old do you think i am?: a study of language and age in twitter. In: Proceedings of the seventh international AAAI conference on weblogs and social media. ICWSM
18. Rangel F, Rosso P (2013) Use of language and author profiling: Identification of gender and age. In: Proceedings of the 10th workshop on natural language processing and cognitive science (NLP-CS-2013)

19. Bedford D (2013) Evaluating classification schema and classification decisions. *Bull Am Soc Inform Sci Technol* 39:13–21
20. Toutanova K, Klein D, Manning C, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human language technology conference (HLT-NAACL 2003)
21. McGrail MR, Rickard CM, Jones R (2006) Publish or perish: a systematic review of interventions to increase academic publication rates. *Higher Educ Res Dev* 25:19–35
22. Costas R, van Leeuwen TN, Bordons M (2010) A bibliometric classificatory approach for the study and assessment of research performance at the individual level: the effects of age on productivity and impact. *J Am Soc Inf Sci* 61:1564–1581

Real Road Networks on Digital Maps with Applications in the Search for Optimal Routes



Amelec Viloria, Noel Varela, David Ovallos-Gazabon,
Omar Bonerge Pineda Lezama, Alberto Roncallo,
and Jairo Martinez Ventura

Abstract Google Maps web mapping service allows, through its extensive API development tool, to extract, process and store updated and real-time road information such as the aerial view of a road network, the travel time and distance between two points and the geographic coordinates of intersections (Di Natale et al. in understanding and using the controller area network communication protocol. Springer, New York, NY, 2012 [1]). However, trivial data required in the construction of the digraph, such as the relationship of the streets associated to those intersections and the type of direction that corresponds to each street, do not exist as an attribute in the API since they are not freely accessible or an excessive cost must be paid for the database. Therefore, a practical way to obtain this specific information is through the development of an application that allows the visual selection of the characteristic elements of a network and the extraction of the necessary data in the construction of related digraphs as a tool in the solution of road problems (Rutty et al. in Transp Res Part Transp Environ 24:44–51, 2013 [2]). This research proposes a method to

A. Viloria (✉) · N. Varela

Universidad de la Costa, St. 58 #66, Barranquilla, Atlántico, Colombia

e-mail: aviloria7@cuc.edu.co

N. Varela

e-mail: nvarela2@cuc.edu.co

D. Ovallos-Gazabon

Universidad Simon Bolívar, Barranquilla, Colombia

e-mail: david.ovallos@unisimonbolivar.edu.co

O. B. P. Lezama

Universidad Libre, San Pedro Sula, Honduras

e-mail: omarpineda@unitec.edu

A. Roncallo

Corporación Universitaria minuto de Dios, UNIMINUTO, Barranquilla, Colombia

e-mail: alberto.roncallo@uniminuto.edu.co

J. M. Ventura

Corporación Universitaria Latinoamericana, CUL, Barranquilla, Colombia

e-mail: academico@ul.edu.co

build digraphs with an application in the Google Maps API in the visual extraction of elements such as vertices (intersections), edges (streets) and direction arrows (road direction), allowing the application of Dijkstra's algorithm in search of alternative routes.

Keywords Digraph · Road network · Google Maps API · Dijkstra's algorithm · Alternative route

1 Introduction

Currently, graph theory is used in the construction of networks as an interconnection system between elements of a set that share a similarity, so the graph is the means by which many methods and techniques have been applied in the form of algorithms aimed at solving optimization problems [3]. Digraphs are used in the representation of road networks as one of the most used applications in areas of study such as intelligent cities and GIS systems, being useful in logistics and planning issues in the search for shorter routes between two points, the shortest route in stopping points, etc.

Most of the studies on road networks found in the literature deal with the analysis of various algorithms in the solution of routes on hypothetical or randomly created networks [4], using expressions such as: "Consider a network formed by a set of N vertices connected to each other ...", "In this hypothetical idea, the road network is illustrated as a network ...", "For example, a route map is modeled as a network ...", etc. [5].

To a lesser extent, there are studies where algorithms are applied in the search for optimal routes using limited GIS databases of road networks, so the results are only dedicated to evaluate the computational efficiency and complexity of the algorithms due to the limitation or lack of properties in the network [6]. In other projects, web mapping service applications are used to show only the routes of a digit network with connection elements, but it does not establish the indications of the routes to follow such as intersections, streets and traffic direction, leaving the task of rendering to the application to show the connection between points without considering the shortest route [7].

The objective of this study is to develop a mapping application that allows the extraction of the geographical characteristics of a road network in the construction of either digraphs or equivalent data structures, allowing the application of optimization algorithms as a solution to road and urbanization problems. In this sense, the proposal presented here works with real data, updated and, in real time, having the ability to choose any road area of study, unlike other works where road maps are loaded only as background images [8].

2 Methods and Data

In the methodology of development, the Google Maps API in JavaScript language is initially used to create an application to extract information about the characteristics of a road network from a digital map in the form of a digraph in a red.txt text file [9].

With the information of the digraph obtained from a road network and stored in the red.txt file, this file is read with the MATLAB mathematical software to create a corresponding incidence list, which contains a series of vectors with the vertices and edges of the digraph without losing relation with the properties of the network. Likewise, MATLAB R2016B has a set of graphic functions that allow to show the incidence lists or matrices as digraphs with labels in edges and vertices [10].

With the built digraph, a route optimization algorithm is applied to validate the data obtained by the application, in this case, to obtain the shortest alternative route among other solution routes. One of these algorithms is Dijkstra's, capable of finding the shortest route from a source point to a destination point within a weighted digit [11].

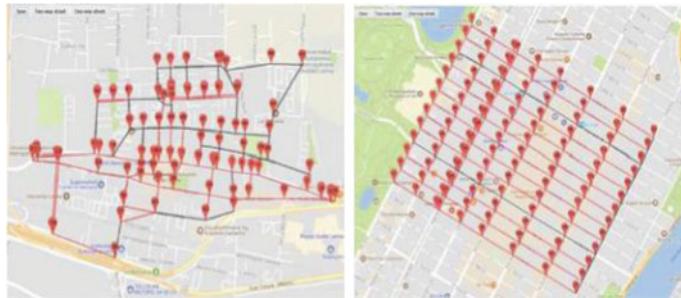
Finally, with the routes obtained in the form of a sequence of vertices applying Dijkstra's algorithm programmed in MATLAB [12], the output file called route.txt is created and is read again with the same Google Maps application to show the alternative route solution on the road network map, with the help of the polyline function and textual indications (address and streets) from the point of origin to the destination.

2.1 *Construction of a Network as a Digraph on a Digital Map with Road Network*

The development in the construction of the digraph as a road network starts with the creation of the network on the digital road map. The application developed with the Google Maps Java Script API contains two buttons that specify the type of signaling of the streets, either one-way with the label “one-way street” in the presence of arrows on the streets or two-way, with the label “two-way street” in the absence of arrows on the streets [13].

With the previous selection of the type of direction per street, the network of relations between road intersections is carried out by placing a first marker with the “click” event on an intersection (initial vertex) [14], and then placing a second marker again with the “click” event on an adjacent intersection (final vertex), thus generating a connection line between vertices as a road section (edge) with direction from start to end. The color of the connecting lines indicates the type of direction, where red represents a single road direction and black two directions.

The selection of intersections already identified with a marker is made with the “right click” event, and the placement of the second marker in a new intersection adjacent to the “click” event, or of an existing marker with the “right click” event,



a) Network in map of Medellin. b) Network in map of N.Y.

Fig. 1 Networks as digraphs in road networks

thus relating each of the intersections through the interconnection paths. The network data are saved in a red.txt text file with the action of a button labeled “Save” [15].

Figure 1 shows the application developed with the Google Maps API in the construction of an irregular grid over the characteristics of the road network of the city of Medellin, Colombia and a regular grid in New York City in New York, USA, as examples of types of road networks.

The application starts with the specification of the geographic area where the digraph network will be built, followed by the button corresponding to the type of direction of the street and the “click” or “right click” events to place a new marker or take an existing one as source and destination vertexes, respectively [16].

2.2 Creation of Digraphs with Real Data from Google Maps

MATLAB’s mathematical development software, version R2016B, has an extensive variety of digraph design and construction functions. The function $G = \text{digraph}(U, V)$, creates a directed network (digraph) with three vectors U , V and D , as source, destination and weight labels as distance for each of the edges formed respectively [17].

In order to create the vectors U , V and D , necessary for the construction of the digraph with information from the Google Maps API, an algorithm was made in MATLAB to convert numerical characters to rational numbers, without losing their position by the row y (,). Figure 2 shows the isomorphic weighted digraphs of the constructed nets, where the width of the edges is proportional to their distance in meters.

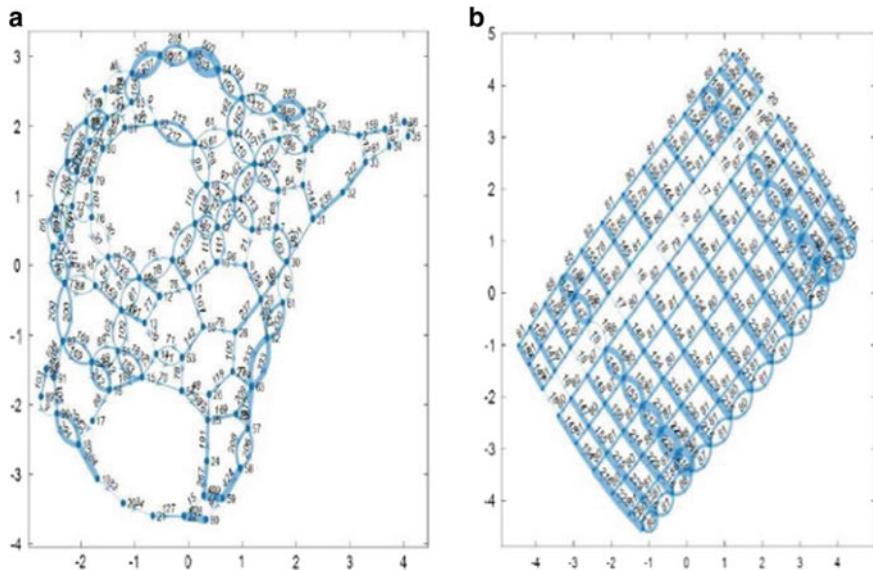


Fig. 2 Digraph from a road network. **a** Medellin. **b** NY

2.3 Validation of Geographical Data in Road Networks as Elements of a Network

With the representation of geographic data as a digraph of a real road network, a solution to the problem of road blockages is proposed with the search for the shortest alternative route [18], as a way of validating the geographic data obtained through the corresponding Google Maps application with the elements of a digraph. As a first step, the vertices corresponding to the intersections that form the route or area to be blocked on the built network are chosen. Figure 3 shows an example of a route or blocked area (red color) in the road network of the map in Fig. 1 as well as in the corresponding vertices and edges with the digits in Fig. 2. Additionally, a random route (green) is included as a vehicle route specifying the origin vertex (orange) and destination vertex (magenta), which intersects with the route or blocked area marked to arrive from a source point to a destination point.

The function `highlight(h, path, "Node Color", "color", "Edge Color", "color")` highlights the nodes and edges specified in the “path” vector within a digit with a specific color. One solution that finds the shortest route S_p is the application of optimization algorithms. Dijkstra's algorithm [19] is a voracious algorithm that finds the shortest route from a source vertex to $\in V$, through all the remaining vertices $[x_2, x_3, \dots, x_{n-1}] \in V$ to the destination vertex in the digit, exploring all possible routes.

The search algorithm for the shortest alternative path in MATLAB is shown in Program 1 (Fig. 4), starting with the elimination of the subset of vertices $G' = (U', V', D')$ that form the blocking zone within the vectors U , V and D .

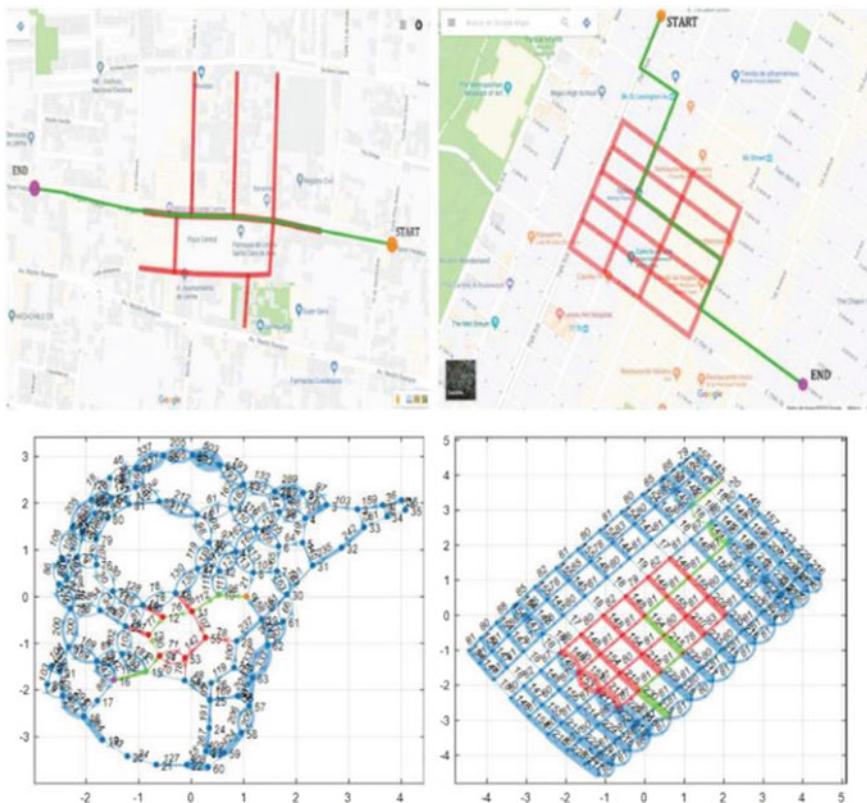


Fig. 3 Route or block zone on a vehicle's route on a road network and digraph

Subsequently, if the source and/or destination vertices are within the blocking zone, one or two new vertices are defined outside the blocking zone. With the digit without the blocking vertices, Dijkstra's algorithm is applied from the source vertex $a = x_1$ to the target vertex $b = x_n$, with edges $e(x_i, x_{i+1})$ and weights $p(x_i, x_{i+1})$ for $i = 1, 2, 3, \dots, n - 1$. The route is obtained by adding the weights on edges by Formula (1). Dijkstra's algorithm found 15,147 routes for the City of Medellin's built network, while for the New York network, 190,458 routes were found, where R1 is the shortest alternative route.

$$S_p = \min \left(\sum_{i=1}^{n-1} p(x_i, x_{i+1}) \right) \quad (1)$$

```

1: Function Dijkstra () = (Vinicio, Vfin, distant, time)
2:   Norigen = vertice;
3:   Ndestino = vertice;
4:   path = [o, o, Norigen, Norigen];
5:   rutaOptima = [o, o, o, o];
6:   ColumnaD = 1;
7:   ColT = 1;
8:   filaR = 1;
9:   Recorrido (1) = o;
10:  while ColumnaT >= 1
11:    noNodo = false;
12:    while noNodo ~= true
13:      nodos = [o, o,o, o], j = o;
14:      For i = 1: NoAristas
15:        if Vinicio(i) == path (4)
16:          NodoNo = false;
17:          for j = 1: lgt (recorrido)
18:            if Vfin(i) == recorrido(j)
19:              NodoNo = true;
20:            end, end
21:            If NodoNo == false
22:              j = j+1; nodos (j) = [distancia(i), Vinicial(i), Vfinal(i)]
23:              end, end, end
24:              if nodos ~= o
25:                path (filaR (1) +1, 1:4) = nodos (1,:);
26:                path(filaR (1) +1,2)=path(filaR(1),2)+nodos (1, 2);
27:                fila=fila+1;
28:                recorrido (fila (1),1) = nodos (1,3);
29:                for i = 2: j
30:                  ruta(1:fila(1)-1,(ColT+1)*4-3: (ColT+1)*4) = ruta(1:fila(1)-1,:);
31:                  recorrido (1: fila (1)-1, ColT+1) = recorrido (1: fila (1)-1,1);
32:                  ruta(fila(1), (ColT+1)*4-3:(ColT+1)*4)=nodos(i,:);
33:                  ruta(fila(1),(ColT+1)*4-2=nodos(i,2)+ruta(fila(1)-1,(ColT+1)*4-2
34:                  recorrido (fila (1), ColT+1) = nodos(i,3);
35:                  ColT = ColT+1;
36:                  fila(ColT) = fila (1);
37:                  end, end, end
38:  end function

```

Fig. 4 Dijkstra's algorithm in the search for the shortest alternative route between two points

3 Results Analysis

With the alternative routes found in the form of a list of vertices using the digraphs as a representation of road networks against road blockage problems, it is possible to represent such solutions on a road map with the Google Maps API. With the developed application, a new button is added to read a text file called route.txt created by MATLAB with the list of geographic coordinates related to the vertices of the solution routes found by Dijkstra's algorithm. Subsequently, with the obtained geographical points, the route service tool is executed by indications with the list of points (waypoints) over the intersections, drawing a sequence of polylines on the map as sections of straight interconnected lines from the origin point through all the remaining intersections in the list, to destination point B. Figure 5 shows the alternative routes R1 and R2 printed on polyline maps formed next to the nodes of each of the intersections where a vehicle turns within the determined routes. These routes are printed on the maps corresponding to the road areas of Medellin City and New York City in the USA.

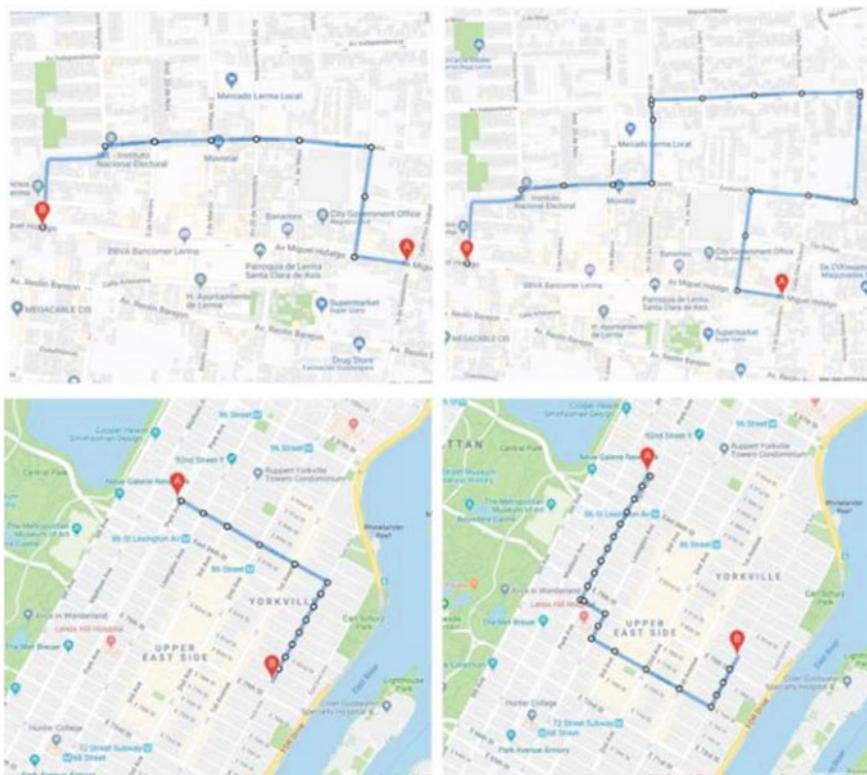


Fig. 5 Alternative routes printed on polyline maps

In addition, the Google Maps API has the route service tool by directions, which allows established routes to be indicated as road instructions through a series of textual descriptions such as: the distance and time of travel, direction of rotation indications on named streets, and the number and street of the origin and destination position of the route.

4 Conclusions

The application of Dijkstra's algorithm in digraphs built with data extracted from the developed application allowed to find the shortest alternative route from an origin to a destination, among all possible solutions. This allowed the validation of the information obtained by the Google Maps API for its application and study with other route optimization algorithms with real data. It is important to consider the use of other algorithms of route searches such as Expansion Tree, Floyd–Warshall, Bellman–Ford, A* search, Kruskal, which can find the same route in a shorter time. However, Dijkstra's algorithm, by the nature of its development in digraphs, allows to find more than one additional alternative route to the optimal shortest route between those two points.

References

1. Di Natale M, Zeng H, Giusto P, Ghosal (2012) Understanding and using the controller area network communication protocol. Springer, New York, NY
2. Rutty M, Matthews L, Andrey J, Matto TD (2013) Eco-driver training within the City of Calgary's municipal fleet: monitoring the impact. *Transp Res Part Transp Environ* 24:44–51
3. Zarkadoula M, Zoidis G, Tritopoulou E (2007) Training urban bus drivers to promote smart driving: a note on a Greek eco-driving pilot program. *Transp Res Part Transp Environ* 12(6):449–451
4. Strömberg HK, Karlsson ICM (2013) Comparative effects of eco-driving initiatives aimed at urban bus drivers—results from a field trial. *Transp Res Part Transp Environ* 22:28–33
5. Vagg C, Brace CJ, Hari D, Akehurst S, Poxon J, Ash L (2013) Development and field trial of a driver assistance system to encourage eco-driving in light commercial vehicle fleets. *IEEE Trans Intell Transp Syst* 14(2):796–805
6. Ferreira JC, de Almeida J, da Silva AR (2015) The impact of driving styles on fuel consumption: a data-warehouse-and-data-mining-based discovery process. *IEEE Trans Intell Transp Syst* 16(5):2653–2662
7. Rionda A et al (2014) Blended learning system for efficient professional driving. *Comput Educ* 78:124–139
8. Restrepo J, Sánchez J (2004) Aplicación de la teoría de grafos y el algoritmo de Dijkstra para determinar las distancias y las rutas más cortas en una ciudad. *Scientia et technica* 10(26):121–126
9. Nathaniel O, Nsikan A (2017) Anapplication of Dijkstra's Algorithm to shortest route problem. *IOSR J Math* 13(3):20–32
10. Saboohi Y, Farzaneh H (2009) Model for developing an eco-driving strategy of a passenger vehicle based on the least fuel consumption. *Appl Energy* 86(10):1925–1932

11. Hellström E, Åslund J, Nielsen L (2010) Design of an efficient algorithm for fuel-optimal look-ahead control. *Control Eng Pract* 18(11):1318–1327
12. Saerens B, Vandersteen J, Persoons T, Swevers J, Diehl M, Van den Bulck E (2009) Minimization of the fuel consumption of a gasoline engine using dynamic optimization. *Appl Energy* 86(9):1582–1588
13. Mensing F, Trigui R, Bideaux E (2011) Vehicle trajectory optimization for application in ECO-driving. In: 2011 IEEE vehicle power and propulsion conference, pp 1–6
14. Rionda Rodriguez A, Martinez Alvarez D, Paneda XG, Arbesu Carbalj D, Jimenez JE, Fernandez Linera F (2013) Tutoring system for the efficient driving of combustion vehicles. *IEEE Rev Iberoam Tecnol Aprendiz RITA* 8(2):82–89
15. Pañeda G et al (2016) An architecture for a learning analytics system applied to efficient driving. *IEEE Rev Iberoam Tecnol Aprendiz RITA* 11(3):137–145
16. Mokhtar K, Shah MZ (2006) A regression model for vessel turnaround time. Tokyo academic industry & culture integration tour, pp 10–19
17. Viloria A, Lezama OBP (2019) Improvements for determining the number of clusters in k-means for innovation databases in SMEs. *ANT/EDI40 2019*, pp 1201–1206
18. Amelec V (2015) Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. *Adv Sci Lett* 21(5):1406–1408
19. Viloria A, Robayo PV (2016) Inventory reduction in the supply chain of finished products for multinational companies. *Indian J Sci Technol* 8(1)

Design of a Network with Sensor-Cloud Technology Applied to Traffic Accident Prevention



Amelec Viloria, Noel Varela, Yaneth Herazo-Beltran,
and Omar Bonerge Pineda Lezama

Abstract The main goal of this research is to facilitate the connection of sensors, people and objects to build a centralized community with parameter measurement applications, where people can share and analyze sensor data in real-time. For example, public institutions that plan, regulate and control land transport, traffic and road safety, and that aim to increase the level of road safety, would be able to prevent traffic accidents by monitoring public transport vehicles in real-time. To obtain real-time data, transport units would be installed with biosensors and a video camera, which would send a series of data automatically for storage and processing in the sensor cloud. The data will allow the recognition of facial expressions and the analysis of physiological variables to determine the driver's mood, according to which immediate actions will be taken to control this variable; and through an audible alarm or a led screen, the actions to be taken by the driver or the passenger would be explained, and even an automatic setting of the interior of the vehicle can be carried out and the emergency units informed according to the case.

Keywords WSN · Cloud computing · 6LowPAN · Sensor cloud

A. Viloria (✉) · N. Varela
Universidad de la Costa, Street 58 #66, Barranquilla, Atlántico, Colombia
e-mail: aviloria7@cuc.edu.co

N. Varela
e-mail: nvarela2@cuc.edu.co

Y. Herazo-Beltran
Universidad Simona Bolívar, Barranquilla, Colombia
e-mail: aherazo4@unisimonbolivar.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

1 Introduction

In recent years, wireless sensor networks (WSN) have become a technology that allows the physical environment to be adapted to the digital world. The sensor nodes cooperatively monitor conditions in different locations, such as temperature, humidity, vehicle movement, light conditions, pressure, noise levels, presence or absence of certain types of objects, as well as characteristics such as speed, direction, size of objects [1]. The sensor nodes are small, low power consumption, low cost and provide multiple functionalities such as detection capability, power processing, memory, bandwidth for communications and battery consumption, which are used in various applications such as measurement of environmental parameters, health care, education, defense, manufacturing, smart homes or home automation [2].

The integration of Cloud Computing to the wireless sensor network infrastructure provides reliable resources, software and data on demand, allowing observation and long-term data exchange, as well as innovation in application services. The sensor cloud infrastructure, or Sensor Cloud, is the extended form of cloud computing, with functionality to manage sensors that are dispersed across wireless sensor networks (WSN). In general, the Sensor-Cloud model allows the exchange of sensor data in real-time through Cloud Computing [3].

According to data from the Spanish Sleep Society, “1 out of every 5 traffic accidents is caused by sleep and fatigue, and this is among the top five causes of accidents with victims” (European Commission Foundation for the Automobile [4]). Therefore, the implementation of a network with Cloud Sensor Technology in vehicles could be an option for the prevention of traffic accidents applied in the road monitoring environment.

2 Theoretical Background

2.1 Wireless Sensor Network (WSN)

A sensor network is a network of tiny sensor-equipped devices that collaborate on a common task. Sensor networks are made up of a group of sensors with certain sensing and wireless communication capabilities that allow for the formation of ad-hoc networks without pre-established physical infrastructure or central administration [5].

Wireless Sensor Networks (WSN) are based on low-cost and low-power devices (nodes), which are able to obtain information from their environment, process it locally, and communicate it through wireless links to a central coordination node. The nodes act as elements of the communication infrastructure by forwarding messages transmitted by more distant nodes to the coordination center [6]. The architecture of a wireless sensor network consists of wireless nodes, gateway and base station.

Wireless Node: These are electronic devices capable of capturing information from the environment in which they are located, processing it, and transmitting it wirelessly to another recipient. The hardware of these devices consists of processor, power supply, wireless communication (RF radio transceiver), sensor, and memory [7].

Gateway: Elements for the interconnection between the sensor network and a data network (TCP/IP). It is a special node without a sensor element, whose objective is to act as a bridge between two networks of different types. Devices that perform the function of interconnecting two networks of different natures are called gateway devices, but the best-known term in the network environment is Gateway [8].

Base Station: It is in charge of collecting the data based on a common computer or embedded system. In a normal structure, all the data goes to a server machine inside a database, from where users can remotely access, observe, and study the data.

2.2 *Types of Sensors Applied in Transport and Mobility*

According to some researches conducted by [9–15], the types of sensors applied in transport and mobility are summarized in Table 1.

Table 1 Types of sensors applied in transport and mobility

Types of sensors applied in transport and mobility

GPS	Facial emotion sensor
Help buttons	Automotive
Satellite connection module	Q sensor
Sensor on the steering wheel (Detects the intensity of pressure and the position of the hands)	Light sensor (Located in the headrest warning of possible driving head tilts)
Drowsiness detector	Eye-tracking
Video camera	Galvanic skin response (GSR)
Driver's temperature sensor	Electroencephalography (EEG)
Heart rate sensor	Skin sensor
Blood pressure/blood pressure sensor	Information led display and overspeed speaker
Biomedical sensors	Infrared video cameras

2.3 Sensor Cloud

According to MicroStrain, cited by [16], Sensor-Cloud infrastructure is defined as a single sensor data storage, visualization, and remote management platform that leverages powerful cloud computing technologies to provide excellent data scalability, rapid visualization, and user-programmable analysis. Roberto Fernández Martínez [17] explains that a Sensor-Cloud Network consists of wireless sensor networks (WSN) and cloud resources such as computers, servers, disk arrays for processing and storing of sensor data.

Figure 1 shows the relationships between the actors and the infrastructure with regard to the Cloud Sensor, according to [18].

3 Proposal: Design of the Sensor-Cloud Network Applied to Traffic Accident Prevention

This item presents the design of a network with Sensor Cloud technology applied to the prevention of traffic accidents. Next, its components are described and the design of the network infrastructure, connections, and explanation of its protocols and operation and the respective security mechanisms are proposed [19–22].

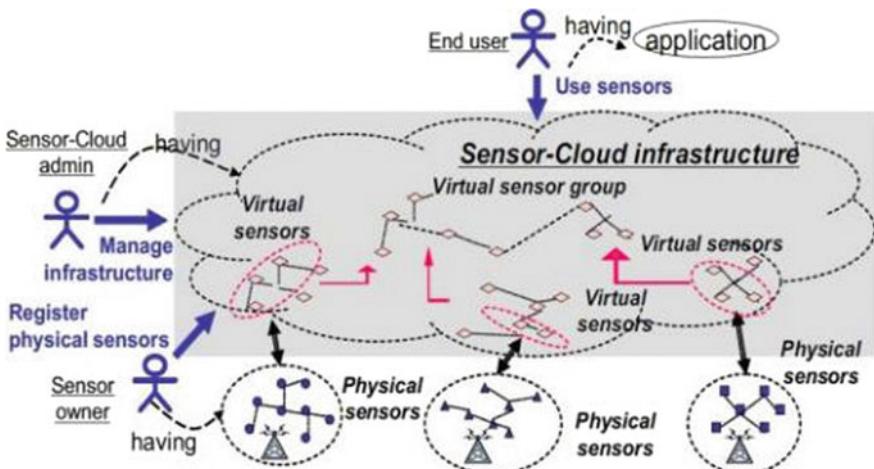


Figure 3. Relationship among Actors and Sensor Cloud Infrastructure

Fig. 1 Relationship among actors and Sensor-Cloud infrastructure

3.1 Main Components of the Sensor-Cloud Network Architecture Applied to Traffic Accident Prevention

The proposed architecture comprises the following components:

3.1.1 WSN Sensor Platform (Wireless Sensor Infrastructure)

The WSN can be a sensor architecture with low power radios, which requires a gateway to connect to an IP network, or the sensor architecture can incorporate technology with direct IP connectivity, WiFi style or 2G, 3G, 4G, LTE cellular network. Depending on the architecture of the WSN network, the following technologies or communication protocols are used to send data from the base station or WSN coordinator to the gateway: ZIGBEE, 6LOWPAN, these technologies are based on the IEEE 802.15.4 protocol for communication.

In the proposed design, the WSN platform is composed of the following sensors: temperature sensor, balance sensor, and video camera, which are installed in the vehicle. In addition, some actuators are installed that will serve as alarm mechanisms for the driver and passengers of the vehicle, including an audible alarm and a LED display to show messages. The power supply for the sensors is provided by the vehicle's battery.

3.1.2 Publishing/Subscription Agent

This module is responsible for the monitoring, processing, and delivery of events to registered users through service applications (SaaS).

3.1.3 Application-Specific Interface

This project proposes SaaS as the specific application interface. SaaS is the Cloud Computing Software as a Service platform that combines data and applications together and runs on the Cloud server. This interface provides flexibility for the control and monitoring entity to access Sensor-Cloud services remotely hosted over the Internet.

3.1.4 System Administrator (SM)

This component within the proposed model is responsible for processing, archiving of sensor data and management of system resources, as well as storage of sensor data; by means of the computer cycles, the data issued by the sensors is processed. It manages the resources (hardware and software) of the computer or servers.

3.1.5 Monitoring and Measurement MaM

This module tracks the use of resources in the primary cloud, as well as the resources of collaborating CLPs (Cloud Providers) so that the resources used can be attributed to a given user. When sensor data arrives at the publish/subscription agent, the system administrator (System Manager) makes decisions for the storage processes.

3.1.6 Service Registry

In the proposed model, the service registry is responsible for discovering and storing information on hardware and software resources, in addition to the policies handled in the local domain.

3.1.7 Identity and Access Management Unit IAMU

When the user or road safety entity requires information from the Sensor-Cloud platform, it connects to the specific SaaS application through the IAMU, which is in charge of providing authentication between the client (monitoring and road safety entity) and the provider (SaaS application), in addition to providing cloud resources with policy-based access control as security mechanisms.

3.1.8 Architecture Web Servers

Operation of the proposed cloud sensor architecture model:

1. The wireless sensor network is connected through the base station or coordinator to the gateway via a common interface and communications protocol, based on IEEE 802.15.4, in different ways.
2. The gateway receives the raw data from the communication ports and converts it into a packet. The packet is kept in a buffer for further processing.
3. In the Virtualization Manager module, the Data Processor retrieves the packets from the buffer and processes them according to their type. The type of packet depends on the application running on the platform. In this case, the proposed application for road control and monitoring is SaaS. The data is processed in a storage format and then sent to the Data Repository (DR) [23].
4. The command interpreter provides the reverse communication channel from the gateway to the wireless sensor network (WSN), which processes and interprets various commands issued by different applications and generates the code, which is understood by the sensor nodes for the actions to be performed by the WSN actuator nodes [10].

5. The vehicle's sensor data then arrives at the publishing/subscription agent, which is responsible for providing the necessary capabilities for various applications to access the same sensor data (parallel execution).
6. The Stream Monitoring and Processing component receive the flow of sensors in many different ways. In some cases, it is raw data that must be captured, filtered, and analyzed in real-time, and in other cases, it is stored in cache memory. The style of calculation depends on the nature of the streams. Therefore, the SMPC running in the cloud monitors the event streams and invokes the correct method of analysis, depending on the data rates and the amount of processing required; in addition, SMP manages the parallel execution model in the cloud.
7. For each application, the Registration Component stores the user subscriptions for that application and the sensor data types (temperature, light, pressure, etc.) [7].
8. When sensor data or events arrive at the publishing/subscription agent, the Analyzer Component determines which applications they belong to and whether they require periodic or emergency delivery [8].
9. The Disseminator Component, using the event matching algorithm, finds the appropriate subscribers for each SaaS application and delivers the events. The Cloud parallel run model can be used for fast event delivery.
10. Computer cycles are provided internally by the SM (System Manager) as needed to process data from the sensors. The SR (Service Registry) manages subscriptions and user credentials.
11. MaM (Monitoring and Metering) calculates the price of the offered services [24, 25].

3.2 Security in the Proposed Cloud Sensor Model

To achieve security and privacy in the proposed design, the Secure Socket Layer (SSL) technique is applied.

In this design, the symmetric Advanced Encryption Standard (AES) algorithm is proposed. Data will be stored in the cloud using this format.

3.3 Advantages of the Proposed Design

Provide real-time data collection.

The manual elimination of the data collection process, which includes, at some point, data entry errors.

Make it possible to track a large number of drivers who depend on limited number of monitoring staff and who work all day without resting periods, which can affect their performance and lead to accidents. Constant monitoring allows the risks of accidents to be reduced by sending out alarms.

Ensure that there are no accidents that are caused by driver stress problems.

In a Sensor-Cloud infrastructure—unlike sensor networks—which are typically used for targeted applications—information obtained from different sensors can be shared and used by multiple applications, which is achieved by virtualizing the various nodes in a network into a cloud platform.

4 Conclusions

The design of the Sensor-Cloud infrastructure applied in the proposed prevention of traffic accidents allows the exchange of sensor data from vehicles in real-time through Cloud Computing. In order to solve the security problems in the cloud, IAMU (Identity and Access Management Unit) is proposed, an identity and access control module that will provide security to the system by managing the infrastructure through a policy repository defined by the user. The proposed scheme is one of publication/subscription, and the specific application interface is of the SaaS type, which permits the possibility of considering the web platform, under the SOAP or REST style.

In this research, GPRS and VANET technologies were analyzed as a potential for the development of vehicle applications. GPRS has a permanent connection to data networks, and there are several devices with this technology adapted for connection to a vehicle and monitoring of variables. VANET networks, on the other hand, have a wide communication capacity with different technologies such as GPRS, which shows the wide coverage that will be available in the future. The two technologies analyzed are complementary in vehicular applications, likewise, the Sensor-Cloud Technology can implement GPRS for communication from the gateway modules.

The Sensor-Cloud infrastructure design that was proposed can be used as a basis for other applications such as in the field of health, video surveillance, telemedicine, and any activity that requires monitoring of variables.

The implementation of the network with Sensor-Cloud Technology applied to the prevention of traffic accidents, will allow us to obtain large amounts of information that can be applied for the creation of models of vehicle behavior, as well as, for the estimation of traffic behavior, among others.

References

1. Ye Q, Zhuang W (2017) Distributed and adaptive medium access control for Internet-of-Things-enabled mobile networks. *IEEE Internet Things J* 4(2):446–460
2. Agencia Nacional de Regulación y Control del Transporte Terrestre (2015) Proyecto Seguridad Integral para el Transporte Público y Comercial. Ecuador
3. Delgado C et al (2018) Joint application admission control and network slicing in virtual sensor networks. *IEEE Internet Things J* 5(1):28–43

4. Alamri A, Shadab Ansari W, Mehedi Hassan M, Shamim Hossain M, Alelaiwi A, Anwar Hossain M (2013) A survey on sensor-cloud: architecture, applications, and approaches. *Int J Distrib Sens Netw* 9(2). Recuperado de <https://journals.sagepub.com/doi/pdf/10.1155/2013/917923>
5. Giraldo MA (Mayo de 2013) Estudio y Simulación de Redes Ad-Hoc Vehiculares VANETS. Universidad Católica de Pereira Ingeniería de Sistemas y Telecomunicaciones
6. Hernandez J (2014) AutoEmotive, bringing empathy to the driving experience to manage stress. MIT Media Lab, Cambridge
7. Hernández JV (2010) Redes inalámbricas de sensores: una nueva arquitectura eficiente y robusta basada en jerarquía dinámica de grupos. Editorial Universitat Politècnica de València, Valencia
8. Beng LH (2009) Sensor cloud: towards sensor-enabled cloud services. Intelligent Systems Center Nanyang Technological University
9. Consuegra MC (2014) Conectando los vehículos a Internet en el sistema de transporte inteligente estandarizado por el ETSI. Departamento de Ingeniería Telemática, Leganés
10. ESEC (2014) Introducción a las Redes de Sensores Inalámbricas. Plataforma Tecnológica Española, España
11. Misra S, Chatterjee S, Obaidat MS (2017) On theoretical modeling of sensor cloud: a paradigm shift from wireless sensor network. *IEEE Syst J* 11(2):1084–1093
12. Fundación Comisariado Europeo del Automóvil (2015) Informe sobre la influencia de la fatiga y el sueño en la conducción. Fundación CEA. Recuperado de <https://www.fundacioncea.es/np/pdf/estudio-somnolencia-al-volante.pdf>
13. Ibañez P (2011) Sistemas de detección en los coches para evitar accidentes. Xataka. Recuperado de <http://www.xataka.com/automovil/sistemas-de-deteccion-en-los-coches-para-evitar-accidentes>
14. Joyanes L (2012) Computación en la nube. Estrategias de Cloud Computing en las empresas. Alfaomega
15. Bharat KS, Priyanka AN (2014) Sensor information management using cloud computing. *Int J Comput Appl* 103(14)
16. Zanella A, Bui N, Castellani A, Vangelista L, Zorzi M (2014) Internet of things for smart cities. *IEEE Internet Things J.* 1(1):22–32
17. Roberto Fernández Martínez JO (2009) Redes inalámbricas de Sensores: Teoría y Aplicación Práctica. España: Grupo de Investigación EDMANS-Universidad de la Rioja
18. Scanaill MJ (2013) Sensor technologies, healthcare, wellness and environmental applications. Apress open. *Int J Comput Appl* 14(103):14–22
19. Kelion L (2014) Crea dispositivo que detecta las emociones humanas a partir de la piel. BBC News. Recuperado de 1–20. http://www.bbc.com/mundo/NOTICIAS/2014/06/140626_CIENCIA_MONITOR_PIEL_DE_GALLINA_MZ.SHTML
20. Mell P, Grance T (2011) The NIST definition of cloud computing. Recommendations of the National Institute of Standards and Technology. National Institute of Standards and Technology, 800–845. Recuperado de <https://nvlpubs.nist.gov/NISTPUBS/LEGACY/SP/NISTSPECIALPUBLICATION800-145.PDF>
21. Mohapatra S (2014) The scalable architecture for future generation computing. Department of Computer Science and Engineering. National Institute of Technology, Springer, India, pp 963–974
22. Dinh T, Kim Y (2017) An efficient sensor-cloud interactive model for on-demand latency requirement guarantee. In: 2017 IEEE International Conference on Communications (ICC). pp 1–6
23. Peñaherrera AF (2009) Pago electrónico a través de teléfonos móviles. ESPOL, Guayaquil, pp 85–96
24. Amelec V (2015) Increased efficiency in a company of development of technological solutions in the areas commercial and of consultancy. *Adv Sci Lett* 21(5):1406–1408
25. Viloria A, Robayo PV (2016) Inventory reduction in the supply chain of finished products for multinational companies. *Indian J Sci Technol* 8(1):47–55

Comparison of Bioinspired Algorithms Applied to Cancer Database



Jesús Silva, Reynaldo Villareal-González, Noel Varela, José Maco,
Martín Villón, Freddy Marín-González, and Omar Bonerge Pineda Lezama

Abstract Cancer is not just a disease; it is a set of diseases. Breast cancer is the second most common cancer worldwide after lung cancer, and it represents the most frequent cause of cancer death in women (Thurtle et al. in: PLoS Med 16(3):e1002758, 2019, 1]). If it is diagnosed at an early age, the chances of survival are greater. The objective of this research is to compare the performance of method predictions: (i) Logistic Regression, (ii) K-Nearest Neighbor, (iii) K-means, (iv) Random Forest, (v) Support Vector Machine, (vi) Linear Discriminant Analysis, (vii) Gaussian Naive Bayes, and (viii) Multilayer Perceptron within a cancer database.

Keywords Big data · Machine learning · Cancer prediction

J. Silva (✉) · J. Maco · M. Villón
Universidad de la Costa, Barranquilla, Colombia
e-mail: jesussilvaUPC@gmail.com

J. Maco
e-mail: jose.maco@upc.edu.pe

M. Villón
e-mail: pcafmvil@upc.edu.pe

R. Villareal-González
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: rvillareal2@unisimonbolivar.edu.co

N. Varela · F. Marín-González
Universidad de La Costa (CUC), Barranquilla, Colombia
e-mail: nvarela2@cuc.edu.co

F. Marín-González
e-mail: fmarin1@cuc.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

1 Introduction

Cancer is the leading cause of death in developed countries and the second leading cause of death in developing countries. Within the categories of cancer, breast cancer is the most frequent diagnosis and the leading cause of death in women [2]. The importance of supervised and unsupervised learning methods and models for identifying patterns and characteristics in cancer classification is evidenced by progress in the performance of classifications based on them. The aim of this paper is to show a comparison of the cancer classification performance of nine models, framed within supervised learning (Machine Learning) and unsupervised learning (Deep Learning) [3].

The recognition of characteristics of “serumproteomics” is widely used for the detection of ovarian cancer [4], breast cancer [5], among others. Multiple studies have used gene selection for cancer classification using support vector machines [5], random forest [6], Bayesian networks [7], Deep Learning [8] and comparing their performance between pairs [9] of methods; however, there is no evidence of a comprehensive comparison of this type, using nine models [10–12].

Machine Learning was defined by Arthur Samuel as the field of study that gives computers the ability to learn without being explicitly programmed [13]. Deep Learning originates from research into artificial neural networks [14]. Deep learning discovers a complex structure in large datasets by using the retro propagation algorithm to indicate how a machine should change its internal parameters that are used to calculate the representation in each layer, from the representation in the previous layer [15]. A confusion matrix was generated, which can be used to measure the performance of a machine learning algorithm, usually a supervised learning algorithm. Each row in the confusion matrix represents the instances of a real class and each column represents the instances of a predicted class. Within the classification tasks, the precision of a class is the number of true positives [16]. In that sense, the interpretation of the confusion matrix of exercise one is presented, which shows three sets of data, outside the main diagonal of the confusion matrix.

2 Methodology

2.1 Automatic Learning

There are several conceptual frameworks for automatic learning. Here is a summary of some aspects associated with the used architectures. The machine learning algorithms are briefly described in [17–21].

2.2 Data and Analysis

This paper is based on the database analysis of eleven breast cancer tumors from diagnostic data from the University of Wisconsin (BCW). The preparation process for processing includes adjustment of the data within the Anaconda Spyder (supervised learning) and H2O (unsupervised learning) platforms [22].

The performance comparison measures of the forecasts of each model include descriptors such as the mean and standard deviation of predictive outcomes within the supervised learning models. In addition, a measure of the efficiency of each of the models is established.

3 Results

This section is presented in two components: (i) performance analysis of supervised learning models based on five models, and (ii) unsupervised model performance.

3.1 Supervised Learning

Taking the document “Cancer Classification Based on Microarray Gene Expression Data Using Deep Learning” [23], the objective is to create a confusion matrix to summarize the performance of a classification algorithm, to the data supplied for this purpose, and identify as an output of this process [24].

The five models applied to this data are Logistic Regression, models.append ('LDA', Linear Discriminant Analysis, models.append 'KNN', K-Neighbors Classifier, models.append 'NB', Gaussian NB y el models.append [25].

In reference to supervised learning, it shows the result of the implementation of five models: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gaussian Naive Bayes (NB), and Vector Support Machines (SVM). These models are optimized based on the strategy of using 80% of the data for training and 20% of the data for validation. The models are evaluated using the “accuracy” evaluation technique by means of the mean and the standard deviation. Figure 1, box and mustache diagram, is presented in order to visualize the performance of the models.

Table 1 shows the comparison of performance results of the five supervised models. It is identified that the best performance is obtained based on logistic regression, which presents an average of 0.9 and a standard deviation of 0.079. The performance evaluated by the model shows efficiency of 100%, a result that ratifies the high relevance of the model for the analysis of the supplied data.

Calculating a confusion matrix can give a better idea of what the classification model is doing well and what kinds of mistakes it is making. The definition of

Fig. 1 Comparison of the performance of five models of supervised learning

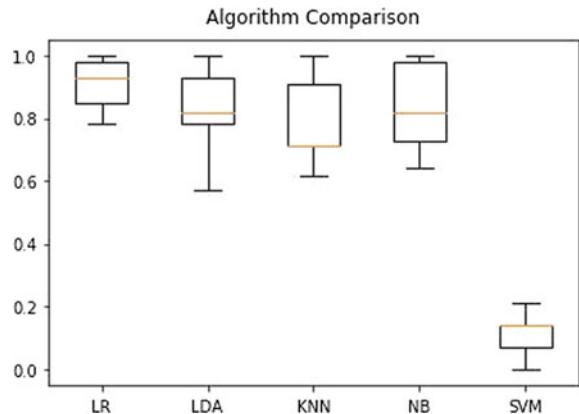


Table 1 Performance comparison of supervised models

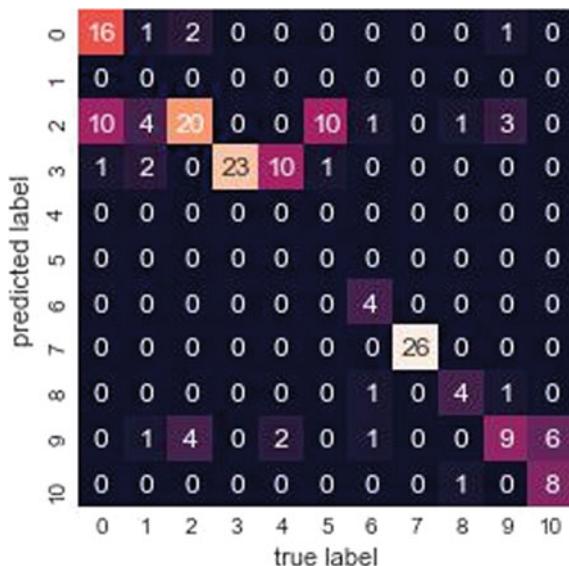
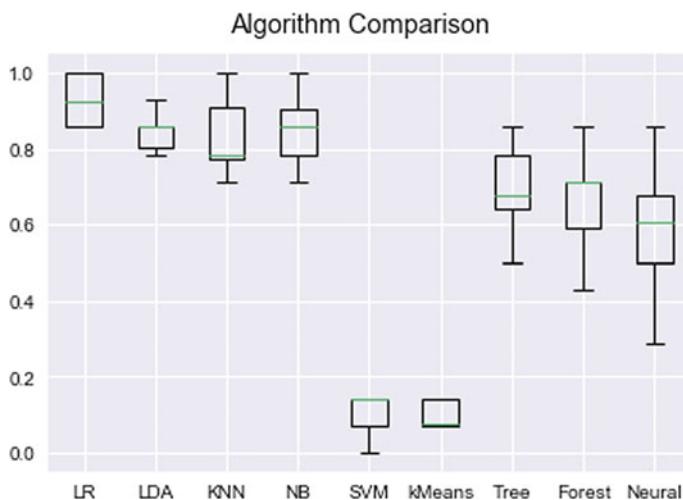
Model	Mean	Standard deviation
LR	0.921458	0.0821
LDA	0.845875	0.125478
KNN	0.801254	0.142587
NB	0.869875	0.135247
SVM	0.954875	0.089875

the confusion matrix is a fundamental step for understanding and sustaining the validity of machine-based learning analyses. Specifically, confusion matrices allow the identification of key parameters such as the accuracy of forecasts [26]. In this sense, the code is constructed for the reading of the variables and the construction of the matrix, Fig. 2.

3.2 Unsupervised Learning

As a measure to establish a relationship between the performance of supervised and unsupervised learning models, algorithm comparisons are made through the box-and-whisker graph shown in Fig. 3.

This comparison is implemented based on the standard conditions of each of the unsupervised methods, including the parameters. Based on this standard condition, low precisions are obtained based on the k -means method (0.15). The other algorithms are Random Forest (0.68), Neural Multilayer (0.6247) and Decision Tree (0.689). In order to improve the classification performance of each of the models, the individual work strategy of each algorithm is implemented, based on the identification of key parameters for each particular model.

**Fig. 2** Confusion matrix**Fig. 3** Comparison of the performance of five unsupervised learning models

In the case of the tumor classification optimization strategy, based on the K -means algorithm, additional parameters are included in the default parameters. The four parameters included and their values are: `n_clusters = 10`, `random_state = 170`, `n_jobs = 10`, `tol = 0.001`. $k = 10$ is addressed because there is information that there are 10 classes within the data. The parameter `n_jobs` is tested from $n = 4$,

obtaining an accuracy of 50% with tolerance: 0.001. Finally, the estimated accuracy is 0.735874522%. This strategy results in a significant improvement in performance from 0.15 to 0.74.

In the case of Random Forest, the choice of seed is decisive. It is contrasted with values of 4 and 170, evidencing changes in the order of 50% improvement in accuracy. The estimated value with seed = 170 was 0.64789. This improvement strategy does not generate better results than the original one, where the first value obtained of 0.70 is greater than the value of 0.64.

In the case of Neural Multilayer Perceptron, an arrangement based on a data partition of 80% for training, 20% test and a seed of 44 is implemented. Two layers are implemented with 50 neurons each. Parameters include alpha = 0.0001 and a constant learning rate. The result is an accuracy of 95.36. In addition, a recall = 0.93 and accuracy = 0.96 are obtained.

The comparative analysis of the algorithms shows better performance results within unsupervised methods. The best performance within the supervised learning algorithms is SVM with a value of 0.9412587; and the best performance for the case of unsupervised learning is Neural MP, with an accuracy value of 0.96398.

4 Conclusions

Five appropriate recognition methods were applied in Logistic Regression class, models.append ('LDA', Linear Discriminant Analysis, models.append 'KNN', K-Neighbors Classifier, models.append 'NB', GaussianNB and models.append), for data analysis identifies eleven classes, a result that demonstrates the structure and cancer classes of the data supplied.

The model that shows better results, measured based on the mean and standard deviation, is LR, which presents a mean equal to 0.91245 and standard deviation of 0.08966. The performance evaluated by the model shows efficiency of 100%, a result that ratifies the high relevance of the model for the analysis of the provided data.

Within the classification tasks, the precision of a class is the number of true positives. In that sense, the interpretation of the confusion matrix of exercise one is presented, which shows three sets of data, outside the main diagonal of the confusion matrix. In addition, two predicted classes are identified with high-pressure values, between seven and eight.

References

1. Thurtle DR, Greenberg DC, Lee LS, Huang HH, Pharoah PD, Gnanapragasam VJ (2019) Individual prognosis at diagnosis in nonmetastatic prostate cancer: development and external validation of the PREDICT Prostate multivariable model. *PLoS Med* 16(3):e1002758. <https://doi.org/10.1371/journal.pmed.1002758>
2. Nima T, Shin JY, Gurudu SR, Hurst RT, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging* 35(5):1299–1312
3. Desautels T, Das R, Calvert J, Trivedi M, Summers C, Wales DJ et al (2017) Prediction of early unplanned intensive care unit readmission in a UK tertiary care hospital: a cross-sectional machine learning approach. *BMJ Open* 7:e017199
4. Hahsler M, Karpienko R (2017) Visualizing association rules in hierarchical groups. *J Bus Econ* 87:317–335
5. Velikova M, Lucas PJF, Samulski M, Karssemeijer N (2013) On the interplay of machine learning and background knowledge in image interpretation by Bayesian networks. *Artif Intell Med* 57(1):73–86. <https://doi.org/10.1016/J.ARTMED.2012.12.004>
6. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinform* 9:1–10. <https://doi.org/10.1186/1471-2105-9-319>
7. Olivera AR, Roesler V, Iochpe C, Schmidt MI, Vigo Á, Barreto SM, Duncan BB (2017) Comparison of machine-learning algorithms to build a predictive model for detecting undiagnosed diabetes - ELSA-Brasil: accuracy study. *Sao Paulo Med J* 135(3):234–246. <https://doi.org/10.1590/1516-3180.2016.0309010217>
8. Viloria A, Lezama OBP (2019) Improvements for determining the number of clusters in k-means for innovation databases in SMEs. *Proc Comput Sci* 151:1201–1206
9. Kamatkar SJ, Kamble A, Viloria A, Hernández-Fernandez L, Cali EG (2018) Database performance tuning and query optimization. In: International conference on data mining and big data, June 21018. Springer, Cham, pp 3–11
10. Chen T, Chefd'hotel C (2014) Deep learning based automatic immune cell detection for immunohistochemistry images. In: Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pp 17–24
11. Viloria, Amelec, et al. Integration of Data Mining Techniques to PostgreSQL Database Manager System. *Procedia Computer Science*, 2019, vol. 155, p. 575–580
12. Clougherty E, Clougherty J, Liu X, Brown D (2015) Spatial and temporal analysis of sex crimes in Charlottesville, Virginia. In: Proceedings of IEEE systems and information engineering design symposium. IEEE, pp 69–74
13. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436
14. Dudoit S, Fridlyand J, Speed TP (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Asoc* 97(457):77–86. <https://doi.org/10.1198/016214502753479248>
15. D'Amico AC, Renshaw AA, Cote K, Hurwitz M, Beard C, Loffredo M et al (2004) Impact of the percentage of positive prostate cores on prostate cancer-specific mortality for patients with low or favorable intermediate-risk disease. *J Clin Oncol* 22(18):3726–3732 (pmid: 15365069)
16. Ontario HQ (2017) Prolaris cell cycle progression test for localized prostate cancer: a health technology assessment. *Ont Health Technol Assess Ser* 17(6):1–75 (pmid: 28572867)
17. Kleemann N, Roder MA, Helgstrand JT, Brasso K, Toft BG, Vainer B et al (2017) Risk of prostate cancer diagnosis and mortality in men with a benign initial transrectal ultrasound-guided biopsy set: a population-based study. *Lancet Oncol* 18(2):221–229 (pmid: 28094199)
18. Turner EL, Metcalfe C, Donovan JL, Noble S, Sterne JA, Lane JA et al (2016) Contemporary accuracy of death certificates for coding prostate cancer as a cause of death: is reliance on death certification good enough? A comparison with blinded review by an independent cause of death evaluation committee. *Br J Cancer* 115(1):90–94 (pmid: 27253172)

19. Celi LA, Mark RG, Stone DJ, Montgomery RA (2013) "Big Data" in the intensive care unit. Closing the data loop. *Am J Respir Crit Care Med* 187:1157–1160
20. Andrea DM, Marco G, Michele G (2016) A formal definition of Big Data based on its essential features. *Libr Rev* 65:122–135
21. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L (2008) Detecting influenza epidemics using search engine query data. *Nature* 457:1012
22. Feng M, McSparron JI, Kien DT, Stone DJ, Roberts DH, Schwartzstein RM et al (2018) Transthoracic echocardiography and mortality in sepsis: analysis of the MIMIC-III database. *Intensive Care Med* 44:884–892
23. Liu WY, Lin SG, Zhu GQ, Poucke SV, Braddock M, Zhang Z et al (2016) Establishment and validation of GV-SAPS II scoring system for non-diabetic critically ill patients. *PLoS ONE* 11:e0166085
24. Calvert J, Mao Q, Hoffman JL, Jay M, Desautels T, Mohamadlou H et al (2016) Using electronic health record collected clinical variables to predict medical intensive care unit mortality. *Ann Med Surg (Lond)* 11:52–57
25. Desautels T, Calvert J, Hoffman J, Jay M, Kerem Y, Shieh L et al (2016) Prediction of sepsis in the intensive care unit with minimal electronic health record data: a machine learning approach. *JMIR Med Inform* 4:e28
26. Sandfort V, Johnson AEW, Kunz LM, Vargas JD, Rosing DR (2018) Prolonged elevated heart rate and 90-day survival in acutely ill patients: data from the MIMIC-III database. *J Intensive Care Med*. <https://doi.org/10.1177/0885066618756828> 885066618756828

Indicators for Smart Cities: Tax Illicit Analysis Through Data Mining



Jesús Silva, Darwin Solano, Claudia Fernández, Lainet Nieto Ramos, Rosella Urdanegui, Jeannette Herz, Alberto Mercado, and David Ovallos-Gazabon

Abstract The anomalies in the data coexist in the databases and in the non-traditional data that can be accessed and produced by a tax administration, whether these data are of internal or external origin. The analysis of certain anomalies in the data could lead to the discovery of patterns that respond to different causes, being able to evidence these causes certain illicit by taxpayers or acts of corruption when there is the connivance of the taxpayer with the public employee or public official. The purpose of this research is the theoretical development of the causal analysis of certain anomalies of tax data, demonstrating that the data mining methodology

J. Silva (✉) · D. Solano · C. Fernández · L. N. Ramos · R. Urdanegui · J. Herz
Universidad de la Costa, Street 58 #66, Barranquilla, Atlántico, Colombia
e-mail: jesussilvaUPC@gmail.com

D. Solano
e-mail: dsolano1@cuc.edu.co

C. Fernández
e-mail: cfernand10@cuc.edu.co

L. N. Ramos
e-mail: lnieto2@cuc.edu.co

R. Urdanegui
e-mail: rosella.urdanegui@upc.pe

J. Herz
e-mail: Jeannette.Herz@upc.pe

A. Mercado
Corporación Universitaria Minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: alberto.mercado@uniminuto.edu.co

D. Ovallos-Gazabon
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: david.ovallos@unisimonbolivar.edu.co

contributes to evidence of illicit and corrupt acts, through the application of certain algorithms.

Keywords Data mining · Anomalous data · Algorithms · Automatic learning · Big data · Noise

1 Introduction

The methodological processes of data mining are used to extract proactive or analytical knowledge [1] from the data, optimizing the full potential of the non-trivial knowledge extraction process [2], knowledge that is implicitly found in the data, data that are contributed, to a large extent, by taxpayers.

Therefore, the following axiom can be inferred: The quality resulting from the non-trivial extracted knowledge will depend, to a large extent, on the quality of the data [3].

The data cleaning and transformation stage of the knowledge extraction process, according to or based on the CRISP-DM methodology, analyzes the influence and cause of data inconsistencies and/or anomalies [4].

The grouping algorithms are oriented to unsupervised learning, where the grouping of the data is related to the common characteristics that they have, in the present research, these data correspond to ex officio determinations of the inspectors that concluded with the process of determination, in the generation of Debt Liquidations (LD). These algorithms are widely used when someone wants to discover hidden knowledge, behavior patterns and extreme values of data [5]. When analyzing the distance between data in a dataset, the general criterion of analysis is that the greater the distance between a data in a database and the rest of the data, the greater the possibility of considering the data as anomalous, inconsistent or noisy.

The reasons why anomalous data may exist are [6]:

1. Incorrect data loading.
2. Errors in the programs used (software) or incompatibilities between different programs.
3. The data is from a different population.
4. Some kind of illicit, like tax evasion.
5. Some possible act of corruption.

Therefore, when not working with a standard data distribution, the identification of these inconsistencies, anomalies or noise detection in the data is very difficult. Search for anomalous data by performing manual consultations or formalize a sequential analysis on all the data of a tax administration, even if it is in one of its processes, such as the generation of Debt Liquidations (LD) requires prior knowledge of the inconsistencies and/or anomalies of the data that could appear [7].

2 Method

For the development of this research, RapidMiner Studio®2 version 8.2.1 was used, an easy-to-use tool with a wide range of algorithms and varied visualization options [8]. The chosen software can be integrated with other programs and languages such as Python and R. The data correspond to the Office of National Taxes and Customs (DIAN—Dirección de Impuestos y Aduanas Nacionales) in Colombia for the year 2017.

The Support Vector Clustering (SVC) algorithm is selected in contrast to other clustering algorithms because the rest of the algorithms have no mechanism for dealing with data noise, or outliers [9].

Vector Support Grouping deals with outliers and data noise by using a soft margin constant that allows the sphere in the characteristics space not to enclose all points [10].

Therefore, those points that are not in any “cluster” are considered noise, being noise for the present study everything that is not of interest or is irrelevant, which degrades or distorts the data, contaminates them and/or prevents or limits the study or use of the information in the analysis of the causes in the anomalies of the data. That is to say, noise in the data under analysis will be that which is outside the limits of the objectives that are proposed in Data Mining [11].

3 Results and Discussions

The analysis of anomalies in tax data will be applied to data resulting from tax inspections carried out on taxpayers, determining ex officio the tax they had to pay in a certain interval of time, limited by statute of limitations, which affects its enforceability.

When a taxpayer gives his consent in a carried-out inspection, that is to say, gives his consent and chooses not to appeal the determination ex officio, the regularization of that determined tax adjustment remains determined. This regularization is performed with the generation of a Debt Settlement (LD). This procedure for generating the Debt Settlement (LD) concludes with the issuance of a Debt Bond that allows the taxpayer to pay, imputing such adjustments and payments to the taxpayer's current account in the periods determined by the inspection [12].

The Debt Liquidations (LD) are generated by the operators of the system of the tax administration, a system that registers different fiscal transactional operations (payment plans, affidavits, etc.) and this generated Debt Liquidation (LD) is validated by the taxpayer upon receiving from the system operator, the Debt Ticket to proceed to the cancellation of his tax obligation [13, 14].

The number of records corresponding to Liquidations of Debts (LD) generated in fiscal year 2017 is approximately 250,000 (of different taxpayers and activities). From the total of records, a subset of data is made up of liquidations that have

characteristics in the data that show modifications of the taxpayers' affidavits and that the values entered in the Tax and Withholding fields are coincident. The sample selected at random to analyze the causes of imperfections is that corresponding to one of those contributors, and that sample has 70 records [15].

3.1 *Spreadsheet Data*

The data are contained in a spreadsheet, do not have referential integrity. The names of the columns are as follows:

SPO_CUIT: Unique Tax Identification Key of the taxpayer.
SPO_DENOMINATION: Name or corporate name of the taxpayer.
NRO_INSCRIPTION: Tax registration number.
ROL: Tax Identification.
TYPE_LD: Code that identifies the type of Debt Settlement.
NUMBER_LD: Number assigned to the Debt Settlement.
FINANCIAL YEAR: Fiscal Year.
ANTICIPAL: Month corresponding to the Fiscal Year.
TAX: Tax declared by the taxpayer.
ALDO CUENTA: payment of the previous month in favor of the taxpayer.
RETENTION: Withholding of the tax made to the taxpayer.
BULLET: Number of the payment ticket.
ACTIVITY: Taxpayer's Activity Code, Tax Law nomenclator.
IMPORT_MULTA: Fine generated for the taxpayer.
MULTA_TERMINO: Fine with compensatory interests.
MONTO_INSPECTOR: Percentage of the fine for the inspector (incentive).
FEC_ALTA: Registration date of the Debt Settlement.

The analysis is concentrated in two columns of the spreadsheet: the TAX column and the RETENTION column, where it is observed that there are coincident values, which would imply, without performing a comprehensive analysis of the causes of this anomaly in the data, that in certain tax periods corresponding to monthly affidavits according to the tax to which they refer, the taxpayer would not have to pay any tax since the balance of the tax to be paid will be zero, because that value has been withheld [16].

In all the matches under analysis, the value of the withholding never exceeds the value of the tax, which means that there is no negative balance, starting the interval at zero, toward positive values.

What is known, by virtue of the previous analysis in which the relationships between the data are identified, is that the values that have been recorded in the RETENTION column that are equal to the TAX column are data that do not fit either the data model or the established procedures resulting from its quality system [17].

3.2 Data Analysis

The Vector Support Grouping algorithm is applied to an attribute of the example set (ExampleSet), which is called IMP-RET (which means Tax minus Withholding) and that was created to understand the behavior of the data under analysis. That attribute will then have the values resulting from subtracting from the TAX column of the spreadsheet, the values from the RETENTION column. Values that are equal to zero (0) will be representative of those records in which the determined value of the Gross Income Tax is equal to the withholding charged manually by the operator and validated by the taxpayer in the generation of the Debt Bond [18].

3.3 Execution of the Algorithm

Having adjusted the default parameters of the Segmentation algorithm by changing the convergence value that specifies the precision of cluster conditions, the Support Vector Clustering (SVC) algorithm results in ten (10) clusters that group the following items, which can be viewed in Fig. 1—Dispersion of clusters in three dimensions—and in Fig. 2—Dispersion of clusters in two dimensions:

In Cluster 0, 20 items, set of different values.

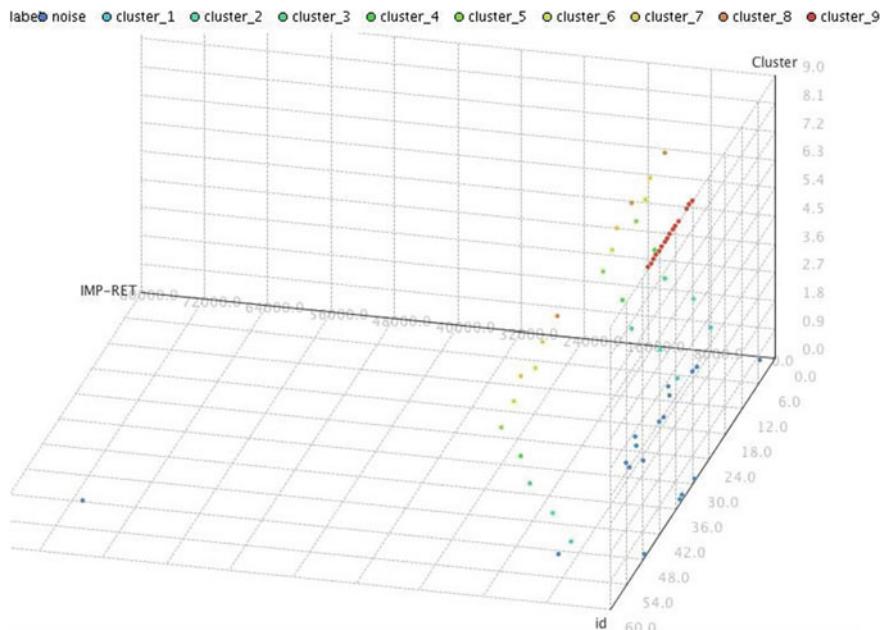


Fig. 1 Dispersion of clusters in three dimensions

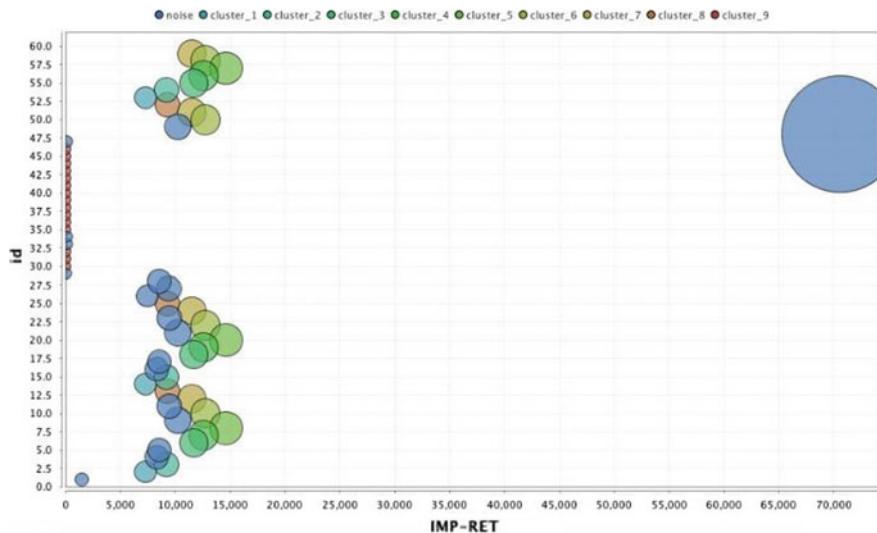


Fig. 2 Dispersion of clusters in two dimensions

- In Cluster 1, 04 items, with the value 6895.25.
- In Cluster 2, 04 items, with the value 10,254.14.
- In Cluster 3, 04 items, with the value 12,458.24.
- In Cluster 4, 04 items, with the value 14,254.32.
- In Cluster 5, 04 items, with the value 16,248.21.
- In Cluster 6, 05 items, with value 11,478.25.
- In Cluster 7, 05 items, with the value 10,258.32.
- In Cluster 8, 04 items, with the value 8587.21.
- In Cluster 9, 18 items, with the value 0.

3.4 First Deductions from the Analysis

The data grouped in the different clusters can receive information to arrive at certain deductions, almost as if it were the generation of a model of supervised learning:

1. The values imputed by the operator of the tax system, which are found in the RETENTION column, cannot be verified with documentary backing that supports them, being these values, in short, data that do not conform to the data model or to the procedures established by the quality system of the tax administration.
2. The system operator that loaded the data can be identified.
3. And from the analysis of the year of each one of the registers, a logic of the inconsistency of the data can be inferred.

Of the total sample of fifty-nine records, twenty-three (23) of them, representing 33% of the sample, would not demand an extra analysis to the objectives set, and are the data grouped in Cluster 0, defined by the algorithm selected as noise.

Twenty-eight (28) records were grouped in different clusters, because the value of the tax payable is the same in different years, which can be seen in the three-dimensional graph, this grouping representing 40% of the data set.

And nineteen (19) records, where the value of the declared tax is equal to that of the withholding, this cluster represents 27% of the sample.

Since the objective is to discover patterns in the items of the sample that explain the cause of the anomaly in the data and that show possible illicit acts on the part of the taxpayer, and/or acts of corruption, and the items were prepared for this purpose, cluster 0 is defined as noise, because these data do not clearly show possible illicit acts and/or acts of corruption, but this does not mean that they cannot exist.

The items from cluster 1 to cluster 8 represent 40% of the sample. These values reveal certain patterns and it can be established that:

1. The values from cluster 1 to cluster 8 may indicate a possible illicit, such as, for example, tax evasion, as long as in the analysis of the causes of data imperfection it cannot be asserted with certainty that these data correspond, for example, to errors in the loading of Debt Liquidations by the operator or to other causes.
2. The values of cluster 9, verifying that the imputation of the values to the RETENTION field was done manually by the tax system operator could evidence a possible act of corruption because there could be collusion between the taxpayer and the operator, since the load is made in the presence of the taxpayer, and it is the taxpayer that validates the load and generation of the Debt Settlement (LD), receiving the Debt Ticket, with which it cancels the adjustments resulting from the tax determination.

4 Conclusions

If the cause of the anomaly has a logical or documented explanation such as lost data, data errors, inconsistencies in the data, or missing or erroneous metadata, there is no indication of illicit and/or corrupt act. Applying the indicated algorithm presupposes not only knowing how the algorithm behaves with the set of examples (ExampleSet), but also understanding the data being evaluated, and their interrelations.

So far, there is no specific application with algorithms for Tax Data Mining [19]. There are also no Tax Automatic Learning algorithms, corresponding to predictive or descriptive models.

The analysis of tax data with segmentation algorithms contributes to the detection of the behavior of a taxpayer, or of a group of taxpayers, since the possibilities of parameterization and the creation of models respond to different alternatives by virtue of the objectives set opportunely and of the preparation of the data, according to the CRISP-DM method or some methodology based on it [20].

The observations of the groupings of tax data and their common characteristics may reveal anomalies, which should be warned because they may be exteriorizing illicit and corrupt acts, by virtue of their causes. Therefore, the detection of atypical tax data, or anomalous tax data, leads to the discovery of small sets of data that will be significantly different from the rest of the tax data under analysis, and it is precisely the analysis of these inconsistent data and their causes that will be more valuable than the general analysis of all the data in the sample, based on the fact that the objectives of data analysis are precisely to determine the causes of the anomaly, without losing sight of the fact that the premise is that there is quality in the data of the tax databases, so there would not be space for the existence of anomalies of this type, an even more striking fact, when these inconsistencies respond to a pattern of conduct of the same taxpayer, of a group of taxpayers, of a specific activity, of a specific exercise or of an operator of the tax system, without an assertive cause.

The correct analysis of anomalies in tax data will make it possible to explain the causes of these anomalies, and if it can be determined that they are not illicit on the part of taxpayers, nor acts of corruption, then it will make it possible to segregate and clean these anomalies from the tax bases, to purify them, correcting taxpayers' current accounts, optimizing the quality of the tax data, in order to contribute to automatic learning processes, and a correct and valuable extraction of non-trivial knowledge.

References

1. Newton K, Norris P (2000) Confidence in public institutions, faith, culture and performance? In: Pharr SJ, Putnam RD (eds) *Disaffected democracies*. Princeton. Princeton University Press, New Jersey, pp 52–73
2. León Medina FJ (2014) Mecanismos generadores de la confianza en la institución policial. *Indret: Revista para el Análisis del Derecho* 2:15–30
3. Liu B, Xu G, Xu Q, Zhang N (2012) Outlier detection data mining of tax based on cluster. In: 2012 international conference on medical physics and biomedical engineering (ICMPBE2012) 33 (Supplement C), pp 1689–1694. <https://doi.org/10.1016/j.phpro.2012.05.272>
4. Malone MFT (2010) The verdict is in: the impact of crime on public trust in Central American Justice Systems. *J Politics Lat Am* 2:99–128
5. Bedoya Velasco ÁY, Rojas Cruz AE, Sandoval Rozo O (2019) El derecho a la defensa técnica en el proceso jurisdiccional de cobro coactivo adelantado por la dirección de impuestos y aduanas nacionales de colombia
6. Bottia Rengifo, RE (2019) Apoyo en las actividades de los programas posconsumo en la coordinación de inventarios y almacén de la Unidad Administrativa Especial Dirección de Impuestos y Aduanas Nacionales (DIAN)
7. Lis-Gutiérrez JP, Reyna-Niño HE, Gaitán-Angulo M, Viloria A, Abril JES (2018) Hierarchical ascending classification: an application to contraband apprehensions in Colombia (2015–2016). In: International conference on data mining and big data, June 2018. Springer, Cham, pp 168–178
8. Amelec V, Carmen V (2015) Relationship between variables of performance social and financial of microfinance institutions. *Adv Sci Lett* 21(6):1931–1934
9. Romero R, Milena S, Torres González B (2019) Núcleo de apoyo contable y fiscal, NAF convenio DIAN Universidad Cooperativa de Colombia-campus Neiva

10. Godoy Godoy DL, González Gómez LA (2019) Big data para la priorización de zonas de atención a emergencias causadas por inundaciones en Bogotá Colombia: uso de las redes sociales
11. Becerra G, Alurralde JPL (2017) Big data y Data mining. Un análisis crítico acerca de su significación para las ciencias psicosociales a partir de un estudio de caso. *{PSOCIAL}* 3(2):66–85
12. Magnani E (2017) Big data y política: El poder de los algoritmos. Nueva sociedad, (269)
13. Segovia C, Haye A, González R, Manzi J (2008) Confianza en instituciones políticas en Chile: un modelo de los componentes centrales de juicios de confianza. *Revista de Ciencia Política* 28(2):39–60
14. Tankebe J (2008) Police effectiveness and police trustworthiness in Ghana: an empirical appraisal. *Criminol Crim Justice* 8:185–202
15. DANE (2019) Proyecciones de Población [database]. DANE, Bogotá
16. Rojas Nonzoque JD, Ramírez Barbosa N (2019) El Impuesto a la Renta y Complementarios en Colombia Desde el Punto de Vista del Contribuyente Persona Natural, Ley 1819 De 2016
17. Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B (2013) Orange: data mining toolbox in python. *J Mach Learn Res* 14(Aug):2349–2353
18. Viloria A, Neira-Rodado D, Pineda Lezama OB (2019) Recovery of scientific data using intelligent distributed data warehouse. *ANT/EDI40* 2019:1249–1254
19. Tarazona LTA, Gómez YH, Granados CM (2019) Caracterización y creación del manual de los procesos y procedimientos de importación de gráneles sólidos en el puerto marítimo de Buenaventura Colombia
20. Vargas D, Lineht L, Santis Criado AC (2019) Aplicación de las disposiciones tributarias actuales en las entidades sin ánimo de lucro en Colombia respecto al impuesto de renta y complementarios

Classification, Identification, and Analysis of Events on Twitter Through Data Mining



Jesús Silva, Pedro Berdejo, Yuki Higa, Juan Manuel Cera Visbal,
Danelys Cabrera, Alexa Senior Naveda, Yasmin Flores,
and Omar Bonerge Pineda Lezama

Abstract Due to its popularity, Twitter is currently one of the major players in the global network, which has established a new form of communication: the microblogging. Twitter has become an essential media network for the follow-up, diffusion and coordination of events of diverse nature and importance (Gonzalez-Agirre et al. in Multilingual central repository version 3.0. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, 2012, [1]), such as a presidential campaign, a disaster situation, a war or the repercussion of information. In such scenario, it is considered a relevant source of information to know the opinions that are emitted about different issues or people. This research proposes the evaluation of several supervised classification algorithms to address the problem of opinion mining on Twitter.

Keywords Machine learning · Twitter · Opinion mining · Classification

J. Silva (✉) · P. Berdejo · Y. Higa · J. M. C. Visbal · D. Cabrera · A. S. Naveda
Universidad de la Costa, Street 58#66, Barranquilla, Colombia
e-mail: aviloria7@cuc.edu.co

P. Berdejo
e-mail: pberdejo@upc.pe

J. M. C. Visbal
e-mail: jcerca7@cuc.edu.co

D. Cabrera
e-mail: dcabrera4@cuc.edu.co

A. S. Naveda
e-mail: asenior@cuc.edu.co

Y. Flores
Corporación Universitaria Minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: yasmin.flores@uniminuto.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

1 Introduction

The flow of information in this social network is such that it is not practical to process it directly without the aid of computer systems. To face this task, systems can perform analyses automatically, and many of these analyses are aimed at extracting global information from the network to assess various issues. In this sense, several studies report the use of opinion mining techniques [2]. There are various opinion mining approaches to address research on Twitter, and one of the most widely used is the supervised classification [3]. However, based on the referred studies, there is still no universally recognized solution as the best approach to address this field of study [4].

In this respect, this paper proposes an evaluation of supervised classification algorithms for the solution of the opinion mining problem on Twitter. For this purpose, representative algorithms of most of the existing approaches in the supervised classification were selected [5], prioritizing in each case the classical variants. The evaluation of this selection will allow to assess the possibilities of the supervised classification in order to face the problem of opinion mining in this particular domain without using any semantic information. In addition, as a result of the evaluation, evidence will be obtained to allow the formulation of new proposals that will integrate the supervised classification together with other techniques.

2 Opinion Mining on Twitter

Several research works have focused on carrying out opinion mining studies using tweets as a source of data. Usually, Twitter users are induced to give their opinion about products, services and politics [6]. In this way, tweets become an interesting source for the analysis of sentiments. Since the messages are short, about one sentence long, it can be assumed that they express a single idea [7]. Therefore, each message is assigned a single opinion, as a simplification of the issue.

One approach consists of recognizing subjective words and hashtags with subjective meaning. In addition, it is proposed to apply rules for the treatment of comparative judgments, negation and other expressions that change the orientation of the sentence [8]. Similarly, in the analysis of sentiment on tweets, techniques based on ontologies have also been used [9] (Table 1).

Another study on the classification of Twitter messages to determine their polarity leads to the use of supervised classification methods [12]. Some proposals suggest the use of n -grams, combined with some learning rhythms like Naive Bayes and the use of POS-TAGS like characteristics from the tweet [20]. A study on hybrid classification methods suggests the existence of two paradigms, one based on the use of lexical resources and the other one based on automatic learning techniques [5].

Table 1 Types of machine learning for opinion mining

Problem	At the level of	Type of learning	References
Determining subjectivity	Document	Supervised	[10]
	Sentence	Supervised	[11]
	Word	Supervised	[12]
Determining polarity	Document	Supervised	[13]
		No supervised	[14]
		Supervised	[15]
	Sentence	No supervised	[16]
		Semi-supervised	[17]
	Word	Semi-supervised	[18]
Determining the grade	Word	Supervised	[19]

3 Twitter Opinion Mining Methodology

It can be noted that one of the most used techniques for opinion mining of messages with similar length to tweets is the supervised classification [3]. However, none of the approaches to the problem of opinion mining is shown to be superior to the rest. Then, the evaluation of the algorithms that represent the monitored classification in the opinion mining on Twitter would be a result to consider in future solutions to this problem. So, it is essential to design an evaluation method that takes into account both the dynamics of Twitter and the process of opinion mining. Thus, for the evaluation of supervised classification techniques, the proposed opinion mining process on Twitter is divided into three stages [12]:

1. Pre-processing or normalization of data.
2. Reduction of dimensions: decomposition and grouping of terms.
3. Supervised classification.

In each stage, different algorithms can be used depending on the characteristics of the problem to be solved in each of them.

3.1 Standardization of Data

On Twitter, messages are not subject to the strict syntactic rules of the language [14], which makes it difficult to apply opinion mining algorithms. Therefore, the normalization of tweets into an appropriate text can affect their further processing.

The proposed standardization model consists of 7 independent and optional phases [7]:

Remove Twitter tags (hashtags) and URLs.

Separate the emoticons from the text.

Translate the lingo.

Delete the repeated letters. Remove the stop words.

Apply spelling correction. Apply stemming.

Once the data set is standardized, it is represented using a word bag model. Thus, a matrix is obtained where the columns represent the tweets and the rows represent all the terms or words present in the universe of the analyzed tweets [6].

3.2 Reduction of Dimensions

In the matrix representation of the data, where the documents are represented as vectors of terms, these documents are expressed in the canonical basis of the terms. In this way, this space has an enormous dimension and, since in this case the documents (tweets) are only written in 140 characters, it is very scattered.

A high data dimension affects the operation of several classifiers [18]. Then it is necessary to evaluate the use of algorithms for the process of reducing dimensions. Two main approaches were analyzed: matrix decomposition algorithms, discarding the less important components, and word space clustering algorithms, which group similar terms.

3.3 Classification

The classification is the final process where the existence or not of an opinion (subjectivity or objectivity) and its possible validity (positive, negative or neutral) are determined in each message (tweet). The classification is then divided into two stages [20]:

Determine whether the message is objective or subjective.

Classify subjective messages into positive, negative or neutral.

The same supervised classifiers can be used at each stage, as both face the same theoretical problem. In order to better explore the search space, linear and non-linear algorithms were analyzed.

4 Evaluation and Results

The evaluation of the classification algorithms in the opinion mining problem on Twitter, according to the proposed method, requires the implementation of the different classification algorithms as well as a corpus of tweets with an a priori assigned classification.

The sklearn tool was used in most of the cases for applying the decomposition and clustering algorithms as well as the classification algorithms [5]. Sklearn was used in the cases of Growing Neural Gas implementations and the Neural Network (linear, 1 layer) for which particular implementations were made.

Twitter does not allow the dissemination of corpus created from the data of that network. It was therefore necessary to develop a corpus to allow the evaluation of the classification algorithms. The tweet collection was built up from 300,000 messages from users on Twitter in the period February–April 2019. Subsequently, the tweets were filtered by language and, from these, the final set of messages that made up the corpus was randomly selected. Finally, these tweets were hand sorted to establish, a priori, their categories (objective-subjective, positive-negative-neutral). In this way, the corpus is composed of 4587 messages in Spanish language, 1475 of which are subjective.

According to the defined opinion mining method and the different classification algorithms to be evaluated, there is a total of 100,258 combinations to be evaluated (given by the 10 pre-processing, 14-dimension reduction algorithms and the 17 classification algorithms to be used in the 2 classification stages). This high number of combinations requires a great deal of experimentation. In order to reduce the processing volume, all combinations that would allow the classification of tweets into objective-subjective were initially evaluated. This first evaluation will allow to determine the most promising algorithms, at all stages of opinion mining, to complete the entire process.

In the first evaluation (objective-subjective classification) 45,258 combinations were analyzed. For each of these combinations the accuracy of the classification obtained was evaluated using cross validation with 60% of the data for training and 40% for validation. Each possible pre-processing or reduction or classification algorithm intervenes only in some of the combinations. So, a pre-processing algorithm intervenes in 1245 combinations, a dimension reduction algorithm in 170 and a classification algorithm in 140. Then, assuming independence between each stage up to the objective-subjective classification, it was obtained as an evaluation measure of each element the average of the classification accuracy of each combination in which they intervene (Fig. 1).

With the results of this first evaluation, it was observed that the pre-processing algorithms offer similar results, with the exception of the spelling correction that shows worse results. This is because the spell checker often chooses the wrong words and changes the meaning of the message.

In the decomposition algorithms, the best results are obtained with PCA and ICA. Clustering algorithms are much more susceptible to parameterization than decomposition algorithms. In the classification, the most notable algorithms were: Rocchio, Linear SVM, Decision Tree and in second place: Ridge, Perceptron Linear and Passive-Aggressive. Linear classifiers, due to the scattered nature of the data, obtain very good results (Fig. 1), as they are simpler and less sensitive to parameter adjustment.

Then in a second evaluation, the combinations including the algorithms that showed the best results in the first experiment were evaluated. As in the previous

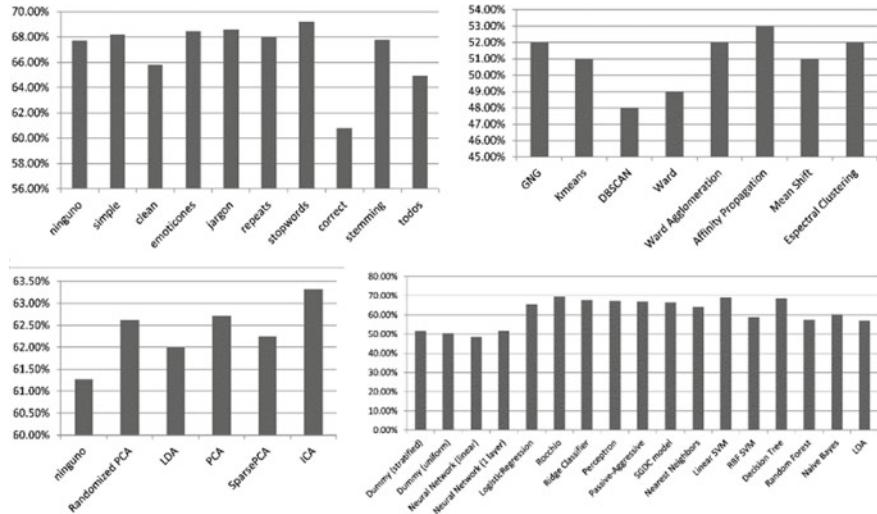


Fig. 1 Accuracy of elements in the first evaluation: upper-left pre-processing algorithms, upper-right clustering algorithms, lower-left decomposition algorithms, lower-right classification algorithms

experiment, the quality was determined by the accuracy of the classification using cross validation.

Table 2 shows the results in the target-subject classification while Table 3 shows the results in the positive-negative-neutral classification. The pre-processing applied was simple and clean, as it is more efficient and shows similar results to the rest.

Table 2 Precision: Standard deviation of each classifier, according to the pre-processing and decomposition applied, classifying in objective or subjective

	Ridge		Perceptron		Rocchio		SMVL		Pas-Agg		DTree	
Simple-PCA	72	1.41	68	4.25	73	1.55	72	1.14	62	1.02	61	3.25
Simple-ICA	70	2.32	65	5.36	73	1.55	65	1.75	66	1.02	62	3.25
Clean-PCA	73	1.14	65	1.58	74	1.66	72	1.58	60	1.02	69	3.01
Clean-ICA	72	1.75	63	3.25	74	1.66	65	1.69	69	1.02	65	3.22

Table 3 Accuracy: Standard deviation of each classifier, according to the pre-processing and decomposition applied, classifying in Positive, Negative or Neutral

	Ridge		Perceptron		Rocchio		SMVL		Pas-Agg		DTree	
Simple-PCA	52	1.75	47	1.03	49	2.36	49	1.35	43	2.58	42	1.78
Simple-ICA	52	1.36	42	1.47	49	2.58	49	1.75	49	2.58	46	1.57
Clean-PCA	54	1.02	41	1.48	50	3.25	47	1.02	41	2.47	41	1.25
Clean-ICA	52	2.35	41	1.49	42	1.25	49	1.58	43	1.47	42	1.01

Table 4 Precision: Standard deviation of each combination of classifiers, for general classification

OS\PN	Rocchio		Ridge		SV ML	
Rocchio	48	3.2	53	1.2	48	3.2 Ridge
Rocchio	48	3.2	53	1.2	48	3.2 Ridge

As a result of this evaluation, it was observed that for the objective-subjective classification, the Ridge and Rocchio algorithms offer the best results. On the other hand, for the positive-negative-neutral classification the algorithms that offer better accuracy are Ridge, Rocchio and SV ML. In the case of the reduction of dimensions, PCA stands out over ICA that affects the operation of some classifiers.

Finally, with these algorithms that offered the best accuracy in the previous experiments (Tables 2 and 3), the whole Twitter opinion mining process was evaluated. In this case the clean processing, which was the one with the best accuracy was used. The results are shown in Table 4.

In this final evaluation, the combination that reports the best results in the classification is Ridge-Ridge with 52% accuracy, classifying in 4 classes. This result doubles the effectiveness of the random classification.

5 Conclusions

The different experiments carried out with the proposed pre-processing show that the process of standardization of the text does not significantly influence the classification. Although it is necessary to perform a basic process that at least eliminates punctuation marks and separates words correctly. The reduction of dimensions with decomposition algorithms is generally superior to that performed with clustering algorithms. This is because the identification of clusters is very dependent on parameters, and requires large volumes of information and notions of distance using semantic information.

The results obtained with linear classifiers are more effective than the rest of the analyzed algorithms. Since classification occurs in a large space where data is widely spread, it is relatively easy to find a linear classifier with acceptable results. Meanwhile, non-linear classifiers, being more complex, depend more on the configuration of their parameters, often causing them to be overtaken by linear classifiers with the same set of training. Finally, the results achieved show the need to use supervised classification algorithms complemented with semantic information or in a mixed approach with other opinion mining techniques to achieve greater effectiveness in opinion mining on Twitter.

References

1. Gonzalez-Agirre A, Laparra E, Laparra G (2012) Multilingual central repository version 3.0. In: Proceedings of the eight international conference on language resources and evaluation (LREC'12), May 2012. European Language Resources Association (ELRA), Istanbul, Turkey
2. Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20(1):53–65 [Online]. Disponible: [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7)
3. Wilcoxon F (1945) Individual comparisons by ranking methods. *Biometrics Bull* 1(6):80–83
4. Riloff E, Janyce W (2003) Learning extraction patterns for subjective expressions. In: Proceedings of the 2003 conference on empirical methods in natural language processing, EMNLP'03. Association for Computational LinguisticsStroudsburg, PA, USA, pp 105–112
5. Lis-Gutiérrez JP, Gaitán-Angulo M, Henao LC, Viloria A, Aguilera-Hernández D, Portillo-Medina R (2018) Measures of concentration and stability: two pedagogical tools for industrial organization courses. In: Tan Y, Shi Y, Tang Q (eds) Advances in swarm intelligence. ICSI 2018. Lecture notes in computer science, vol 10942. Springer, Cham
6. Zhao WX, Weng J, He J, Lim EP, Yan H (2011) Comparing twitter and traditional media using topic models. In: 33rd European conference on advances in information retrieval (ECIR11). Springer-Verlag, Berlin, Heidelberg, pp 338–349
7. Viloria A, Gaitan-Angulo M (2016) Statistical adjustment module advanced optimizer planner and SAP generated the case of a food production company. *Indian J Sci Technol* 9(47). <https://doi.org/10.17485/ijst/2016/v9i47/107371>
8. Villada F, Muñoz N, García E (2012) Aplicación de las Redes Neuronales al Pronóstico de Precios en Mercado de Valores, *Información tecnológica* 23(4):11–20
9. Sapankevych N, Sankar R (2009) Time series prediction using support vector machines: a survey. *IEEE Comput Intell Mag* 4(2):24–38
10. Viloria A, Lezama OBP (2019) Improvements for determining the number of clusters in k-means for innovation databases in SMEs. *Procedia Comput Sci* 151:1201–1206
11. Toro EM, Mejía DA, Salazar H (2004) Pronóstico de ventas usando redes neuronales. *Scientia et technica* 10(26):12–25
12. Hernández JA, Burlak G, Muñoz Arteaga J, Ochoa A (2006) Propuesta para la evaluación de objetos de aprendizaje desde una perspectiva integral usando minería de datos. En A. Hernández y J. Zechinelli (eds.) *Avances en la ciencia de la computación*. Universidad Autónoma de México, México, pp 382–387
13. Romero C, Ventura S (2007) Educational data mining: a survey from 1995 to 2005. *Expert Syst Appl* 33(1):135–146
14. Romero C, Ventura S (2010) Educational data mining: a review of the state of the art. *Syst Man Cybern Part C Appl Rev IEEE Trans* 40(6):601–618
15. Choudhury A, Jones J (2014) Crop yield prediction using time series models. *J Econ Econ Educ Res* 15:53–68
16. Scheffer T (2004) Finding association rules that trade support optimally against confidence. *Intell Data Anal* 9(4):381–395
17. Ruß G (2009) Data mining of agricultural yield data: a comparison of regression models. In: Perner P (eds) *Advances in data mining. Applications and theoretical aspects, ICDM 2009*. Lecture notes in computer science, vol 5633
18. Viloria A, Lis-Gutiérrez JP, Gaitán-Angulo M, Godoy ARM, Moreno GC, Kamatkar SJ (2018) Methodology for the design of a student pattern recognition tool to facilitate the teaching - learning process through knowledge data discovery (Big Data). In: Tan Y, Shi Y, Tang Q (eds) *Data mining and big data. DMBD 2018*. Lecture notes in computer science, vol 10943. Springer, Cham

19. Berrocal JLA, Figuerola CG, Rodriguez AZ (2013) Reina at RepLab2013 topic detection task: community detection. In: Proceedings of the fourth international conference of the CLEF initiative
20. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. SIGKDD Explor Newsl 11(1):10–18

Algorithm for Detecting Polarity of Opinions in University Students Comments on Their Teachers Performance



Jesús Silva, Edgardo Rafael Sanchez Montero, Danelys Cabrera, Ramon Chacon, Martin Vargas, Omar Bonerge Pineda Lezama, and Nataly Orellano

Abstract Sentiment analysis is a text classification task within the area of natural language processing whose objective is to detect the polarity (positive, negative or neutral) of an opinion given by a certain user. Knowing the opinion that a person has toward a product or service is of great help for decision making, since it allows, among other things, potential consumers to verify the quality of the product or service before using it. This paper presents the results obtained from the automatic identification of the polarity of comments emitted by university students in a survey corresponding to the performance of their professors. In order to carry out the identification of the polarity of comments, a technique based on automatic learning is used, which initially makes a manual labeling of the comments and then these results allow to feed different learning algorithms in order to create the classification models that will be used to automatically label new comments, and thus determine their polarity as positive or negative.

Keywords Analysis of polarity · Opinion mining · Supervised classification

J. Silva (✉) · E. R. S. Montero · D. Cabrera · R. Chacon · M. Vargas
Universidad de la Costa, Street 58#66, Barranquilla, Colombia
e-mail: aviloria7@cuc.edu.co

E. R. S. Montero
e-mail: esanchez2@cuc.edu.co

D. Cabrera
e-mail: dcabrera4@cuc.edu.co

R. Chacon
e-mail: ramon.chacon@upc.pe

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

N. Orellano
Corporación Universitaria Minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: nataly.orellano@uniminuto.edu.co

1 Introduction

One of the studies focused on the classification of opinions, positive or negative, is the one presented in [1], one of the first researches on the analysis of sentiments in which data from movie reviews found in the Web are used. These data are employed in three classification algorithms, surpassing the baselines produced manually by a human. They published their work on the classification of documents based on the sentiment expressed in them and analyzed movie reviews, and found that the Automatic Learning techniques improved the performance of the baselines generated by the human experts.

They employed three Machine Learning algorithms: Naïve Bayes (NB) [2], Maximum Entropy (ME) [3] and Support Vector Machines (SVM) [4].

The generation of word lexicons, which are annotated with their corresponding polarity, is another approach that has been oriented by different researchers and that has helped to monitor opinions in the comments. In the case of Spanish language, examples of these approaches are those presented in [5, 6].

Making use of the data generated by people's comments is a great opportunity to gain time in decision making, because they provide information that can be used in different settings [7–9]. From the data acquired, an automatic analysis can be performed and statistics generated on the collective opinion (positive or negative) of a product, service or person. This analysis is very useful for media analysts due to the reduction of time and costs [10, 11] in contrast to manual studies in which an immense consumption of time and costs is observed [12].

Therefore, this paper presents an analysis of different automatic classification models that determine the performance of the process of automatic identification of the polarity of comments emitted by university students.

2 Methods and Data

This section presents the classification methods analyzed for the automatic detection of polarity in comments from university students, as well as the data set used in the experiment.

2.1 Aspects Related to Data

The instrument that serves as a basis for the analysis carried out in this paper was designed based on [13–16], and is used to obtain information on the teaching-learning processes, in order to enable intervention strategies to improve the teaching function. The instrument evaluates the following dimensions [9, 17, 18]:

- Mediation. Understood as the series of actions that the teacher proposes to the student, based on what the latter already knows and does, in order to transform and expand his or her level of knowledge about the issues addressed and their immediate reality.
- Strategies and resources. An action plan that is carried out to achieve a certain long-term goal; they specify the resources that the teacher uses in a reflective and flexible way to achieve significant learning.
- Learning outcomes. The teacher's work is focused on achieving tangible results for the student in terms of the development of their skills, knowledge and attitudes.
- Planning. The teaching activity requires planning, that is, the articulation of objectives, course contents, methodologies, educational strategies and resources, establishing a sequence of activities that allows the desired learning and the use of time.
- Learning evaluation. Learning assessment can be declarative, procedural and attitudinal-valuable, and must consider, at least, the following principles: make known in a timely manner the criteria under which the subject will be accredited; be fair (the same for all); favor timely feedback that allows for cognitive and process readjustments, and be objective (focused on what is expected to be learned according to the program).
- Transverse axes. These are the set of characteristics that define an academic educational model and that, in this case, correspond to the so-called Minerva University Model (MUM). Those considered are: Human and Social Training (HST), Development of Complex Thinking Skills (DCTS), Development of skills in the use of information and communication technologies (DSICT), the use of a second language, education for research and entrepreneurial culture.
- Relational. The teaching exercise requires a set of social skills to interact respectfully with students, namely: motivation, recognition and respect for difference and dialogue, including assertive communication and conflict resolution.
- Institutional Compliance. Compliance with 4 indicators: schedule, attendance, program coverage and mastery of the subject.

The questionnaire is answered on a scale of 1–4 and, according to the answers, a Weighted Satisfaction Index (WSI) is calculated which is used, along with other criteria, to evaluate the teacher's performance in the classroom. At the end of the questionnaire there is a section for comments open to the student's decision.

Since the objective of this study is the analysis of the polarity of comments from university students, it is necessary to have a manual label indicating whether the comment is positive or negative. For this purpose, a classification typology was constructed, which is manual and starts from the analysis of contents, containing the following categories [16]:

- Scope. Related to the teacher's academic, instrumental or relational environment.
- About what? About the subject, about the teacher, about another teacher, about the institution or about the instrument.
- Type. Positive, negative, denunciation, suggestion and substitution.
- Attention. Director, academy, teacher training school or the teacher himself.

2.2 Description of the Classifiers

Automatic, or supervised, learning techniques are capable of learning the human process to classify, among other things, the polarity in comments from college students. This process requires the extraction of characteristics from a corpus that is regularly manually annotated (supervised or training corpus). In this way, a classification model is generated, and can later be used to classify new samples whose final class is unknown. In this case, the training samples and the new samples are considered to be automatically classified as only one of the two following classes: positive or negative.

In the experiments conducted, it was considered to use the manually labeled corpus to form the training corpus and the test corpus in percentages of 80 and 20%, respectively. A technique known as v -fold cross-validation [19] was applied, making use of the total corpus and subdividing it, as mentioned above, into training and test corpus at times ($v = 10$ was used for the experiments presented in this paper). The values obtained from the 10 runs were averaged and presented as final results, results that will be shown in Sect. 3. In order to have a perspective of the type of classifier that can better treat the problem of polarity classification, the following four learning algorithms have been selected (each one belonging to a different type of classifier: Bayes, Lazy, Functions and Trees):

- Naïve Bayes: is a probabilistic classifier based on Bayes' theorem and some additional simplifying hypotheses [2].
- K-Star: This is the classifier of the nearest k neighbors with a generalized distance function [20].
- SMO: This is a sequential minimal optimization algorithm for the classification of supporting vectors [12].
- J48: This is an algorithm used to generate a decision tree [14].

All texts were represented by a frequency vector of n -grams, with values for = 1, 2 and 3. Frequencies higher than two, for the n -grams, are considered for the vector of characteristics.

3 Results

This section presents the results obtained after using each of the three evaluation corpora (C1000, C5000, and C10000). Table 1 shows the number of correctly and incorrectly classified instances for the C1000 corpus, for each of the four supervised classifiers (Naïve Bayes, K-Star, SMO and J48).

Table 2 shows the number of correctly and incorrectly classified instances for the C5000 corpus, for each of the four supervised classifiers (Naïve Bayes, K-Star, SMO and J48). Again, it is the support vector machine (SMO) based classification model that wins with a classification percentage of 92.55%, that is, 4559 samples out of a

Table 1 Results of the classification using the C1000 corpus

Classifier	Type	Correct instances (%)	Incorrect instances (%)
Naïve Bayes	Bayes	78.30	24.70
K-Star	Lazy	75.00	22.30
SMO	Functions	85.60	12.00
J48	Trees	81	22

Table 2 Classification results using the C5000 corpus

Classifier	Type	Correct instances (%)	Incorrect instances (%)
Naïve Bayes	Bayes	79.05	22.58
K-Star	Lazy	78.01	22.01
SMO	Functions	92.55	8.11
J48	Trees	83.89	16.04

Table 3 Results of the classification using the C10000 corpus

Classifier	Type	Correct instances (%)	Incorrect instances (%)
Naïve Bayes	Bayes	78.04	21.57
K-Star	Lazy	82.38	20.61
SMO	Functions	90.99	8.14
J48	Trees	86.40	13.57

total of 5000 were correctly classified. The error rate of this classifier is 8.11%, that is, 427 errors out of a total of 5000 samples. In particular, 176 samples that were positive were misclassified as negative, and 250 samples that were negative were misclassified as positive.

Table 3 shows the number of correctly and incorrectly classified instances for the C10000 corpus, for each of the four supervised classifiers (Naïve Bayes, K-Star, SMO and J48). Again, the classification model based on support vector machines (SMO) is the winner with a classification percentage of 90.99%, that is, 9147 samples out of 10,000 were correctly classified. The error rate of this classifier is 8.14%, i.e., 814 errors out of 10,000 samples. In particular, 333 positive samples were incorrectly classified as negative, and 474 samples that were negative were incorrectly classified as positive.

4 Discussion

Figure 1 shows the performance of each and every one of the classifiers evaluated on three different corpora (C1000, C5000 y C10000).

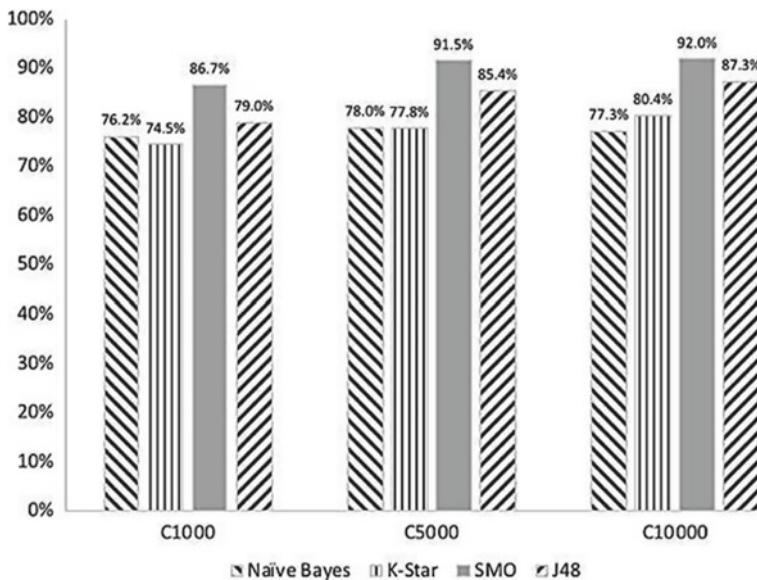


Fig. 1 Percentage of correctly classified instances

The K-Star classifier, of the Lazy type, is the one that mostly presents the lowest performance, mainly when the data set is small (1000 and 5000), although when the data set is larger (10,000) the performance reached 80.1%.

The Naïve Bayes classifier also exhibits poor performance compared to SMO, as its classification percentages are between 75 and 77%.

The J48 classifier, based on decision trees, is the second-best performer, reaching a maximum classification percentage of 86.6% when the data set is 10,000 samples. In general, it is observed that SMO is the best supervised classifier, reaching a percentage close to 91.8%, which is considered very good in the state of the art. Such behavior should be analyzed when the data set is larger, for example, by doubling the positive and negative samples. However, this analysis will be considered in future research.

Figure 2 shows the percentage of incorrectly classified instances by each of the supervised classifiers (Naïve Bayes, K-Star, SMO and J48) in the three different corpora evaluated (C1000, C5000 and C10000). It is relevant to mention that there is a significant difference in the error rate obtained by the SMO classifier with respect to the other three supervised classifiers (Naïve Bayes, K-Star and J48), which reaches up to 13.9 points when using the C10000 corpus.

From this perspective, the results are very good given the amount of data. It is clear that the best classifier is SMO, an implementation of the supporting vector machines.

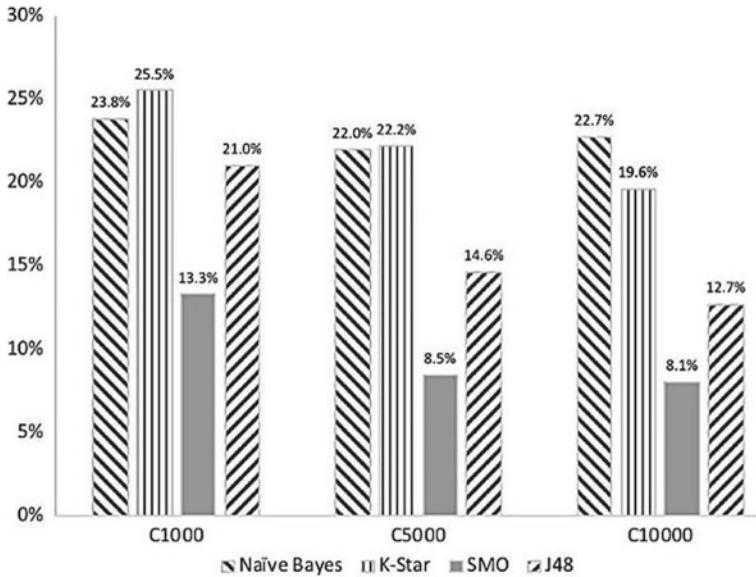


Fig. 2 Percentage of incorrectly classified instances

5 Conclusions

In this paper, an analysis of the polarity of comments from university students was presented. Four different methods of supervised classification were analyzed to determine which one exhibits the best behavior on the task at hand. Three different corpora were used to determine, also, the impact on the size of the samples used for the training process of the classification model.

Given the results obtained in the experiments, it is concluded that the support vector machine-based classifier has the best behavior on the polarity classification of comments from university students.

The main contributions of this study are:

- Balanced corpora to carry out experiments related to the calculation of the polarity of comments from university students.
- Analysis of four supervised classifiers for the task of classifying positive and negative comments.

As future research, it is considered important to experiment with a larger data set and with other classifiers such as conditional random fields (CRF), to cite one example. Also, and given the results obtained, it is believed that it is possible to generate a computer module that allows the process of polarity classification (positive/negative) of university students' comments to be carried out.

References

1. Sáiz J (2015) Sentiué: target and aspect-based sentiment analysis in semeval-2015 task 12. In: Proceedings of the 9th international workshop on semantic evaluation, Association for Computational Linguistics, Denver, Colorado, pp 767–771
2. Brun C, Pérez J, Roux C (2018) Xrce at semeval-2018 task 5: Feedbacked ensemble modeling on syntactico-semantic knowledge for aspect-based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation, Association for Computational Linguistics, San Diego, California, pp 282–286
3. Hercig T, Brychcín T, Svoboda L, Konkol M (2018) Uwb at semeval-2018 task 5: aspect based sentiment analysis. In: Proceedings of the 10th international workshop on semantic evaluation, Association for Computational Linguistics, San Diego, California, pp 354–361
4. Deng ZH, Luo KH, Yu HL (2014) A study of supervised term weighting scheme for sentiment analysis. *Expert Syst Appl* 41:3506–3513
5. Peñalver I, García F, Valencia R, Rodríguez MA, Moreno V, Fraga A, Sánchez JL (2014) Feature-based opinion mining through ontologies. *Expert Syst Appl* 41:5995–6008
6. Balaguer EV, Rosso P, Locoro A, Mascardi V (2010) Análisis de opiniones con ontologías. *Polibits* 41:29–36
7. Sanzón YM, Vilariño D, Somodevilla MJ, Zepeda C, Tovar M (2015) Modelos para detectar la polaridad de los mensajes en redes sociales. *Res Comput Sci* 99:29–42
8. Wilson T, Wiebe J, Hoffmann P (2005) Recognizing contextual polarity in phrase level sentiment analysis. In: HLT/EMNLP 2005, human language technology conference and conference on empirical methods in natural language processing, Proceedings of the Conference, Vancouver, British Columbia, Canada
9. Araújo M, Pereira A, Benevenuto F (2020) A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Inf Sci* 512:1078–1102
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
11. Buitinck L, Louppe G, Blondel M, Pedregosa F, Mueller A, Grisel O, Niculae V, Prettenhofer P, Gramfort A, Grobler J, Layton R, VanderPlas J, Joly A, Holt B, Varoquaux G (2013) API design for machine learning software: experiences from the scikit- learn project. In: ECML PKDD workshop: languages for data mining and machine learning, pp 108–122
12. Peng DL, Gu LZ, Sun B (2019) Sentiment analysis of Chinese product reviews based on models of SVM and LSTM. *Comput Eng Softw* 1:10
13. Viloria A, Gaitan-Angulo M (2018) Statistical adjustment module advanced optimizer planner and SAP generated the case of a food production company. *Indian J Sci Technol* 9(47). <https://doi.org/10.17485/ijst/2018/v9i47/107371>
14. Rousseeuw P (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20(1):53–65 [Online]. Disponible: [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7)
15. Toutanova K, Klein D, Manning CD, Singer Y (2003) Feature-rich part-of-speech tagging with a cyclic dependency network. In: Proceedings of the 2003 conference of the North American chapter of the association for computational linguistics on human language technology, vol 1, ser. NAACL'03. Association for Computational Linguistics, Stroudsburg, PA, USA, pp 173–180
16. Viloria A, Lezama OBP (2019) Improvements for determining the number of clusters in k-means for innovation databases in SMEs. *Procedia Comput Sci* 151:1201–1206
17. Viloria A, Acuña GC, Franco DJA, Hernández-Palma H, Fuentes JP, Rambal EP (2019) Integration of data mining techniques to PostgreSQL database manager system. *Procedia Comput Sci* 155:575–580
18. He Q, Yang J, Lu G, Chen Z, Wang Y, Sato M, Qie X (2019) Analysis of the first positive polarity gigantic jet recorded near the Yellow Sea in mainland China. *J Atmos Solar Terr Phys* 190:6–15

19. Funahashi Y, Watanabe T, Kaibuchi K (2020) Advances in defining signaling networks for the establishment of neuronal polarity. *Curr Opin Cell Biol* 63:76–87
20. Das S, Das D, Kolya AK (2020) An approach for sentiment analysis of GST tweets using words popularity versus polarity generation. In: Computational intelligence in pattern recognition, Springer, Singapore, pp 69–80

Prediction of the Efficiency for Decision Making in the Agricultural Sector Through Artificial Intelligence



Amelec Viloria, Alex Ruiz-Lazaro, Ana Maria Echeverría González, Omar Bonerge Pineda Lezama, Juan Lamby, and Nadia Leon Castro

Abstract Agriculture plays an important role in Latin American countries where the demand for provisions to reduce hunger and poverty represents a significant priority in order to improve the development and quality of life in the region. In this research, linear data analysis techniques and soil classification are reviewed through neural networks for decision making in agriculture. The results permit to conclude that precision agriculture, observation and control technologies are gaining ground, making it possible to determine the production demand in these countries.

Keywords Neural networks · Agricultural activity · Precision agriculture · Decision making · Prediction analysis

A. Viloria (✉) · A. M. E. González
Universidad de La Costa, Street 58 #66, Barranquilla, Colombia
e-mail: aviloria7@cuc.edu.co

A. M. E. González
e-mail: aecheverria@cuc.edu.co

A. Ruiz-Lazaro
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: aruiz25@unisimonbolivar.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

J. Lamby
Corporación Universitaria Minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: juan.lamby@uniminuto.edu.co

N. L. Castro
Corporación Universitaria Latinoamericana, Barranquilla, Colombia
e-mail: nleon@ul.edu.co

1 Introduction

Agricultural activities and their non-linear behaviors require dynamic technologies based on techniques that provide accuracy and a better understanding of decision making. Increasing productivity, reducing time, and avoiding errors when changing production methods are the main objectives for the design of state-of-the-art manufacturing and production systems [1].

The use of techniques that simulate human intelligence in computers takes advantage of detection technologies and cutting-edge actuators: artificial neural networks (ANN), which have proven to be an effective tool for characterizing, modeling and predicting a large number of non-linear processes with accurate results in decision making required in complex agricultural problems such as: prioritizing and classifying products, pattern recognition, crop prediction and product physical changes [2].

2 Application of Neural Networks

This article shows an exploration of the different neural network techniques applied to the agricultural sector and the research carried out in this area. Understanding the factors that influence productivity is essential for the analysis and synthesis of variables that correlate with each other, especially in modern and large-scale production systems [3], since these systems require the adoption of a different management technique known as precision agriculture, which intends to optimize the use of inputs, increase productivity and obtain more benefits from a better production process.

Precision agriculture is defined as “a set of techniques that allows localized management, and its success depends on three elements: information, technology and management” [4]. With the increasing adoption of different soil management techniques, such as variable rate fertilizers and the possibility of adopting specific machines in precision agriculture, it is necessary to establish specialized management sites and different management areas [5].

ANN have shown high performance due to factors such as their distributed or parallel and robust structure known as layers; their efficiency in learning and generalization, which allows them to solve complex problems; they are tolerant of typical or “outlier” values; they can model different variables and their non-linear relationships; and finally, they allow modeling with categorical variables [6].

A commonly used method to evaluate the degree of relationship between the variables involved in the modeling process is Pearson’s correlation analysis. Sequentially, models of different categories of configuration of neural networks and multiple regression can be adjusted to represent these models [7]. The ordinary least-squares multiple linear regression method can be used to estimate productivity:

$$Y = \beta_0 + \beta_{1x} MO + \beta_{2x} CTC + \beta_{3x} V(%) + \beta_{4x} TA \quad (1)$$

where Y is the average yield of the crop (kg) in the period to be evaluated; MO solids content, i.e., organic soil (mg); CTC Cation Exchange Capacity (mmol); V (%) is the base saturation; RT is defined as the resistance of the clay (mg); β_i = Estimators of parameters to be adjusted by $i = 0, 1, 2, 3$ and 4 [8].

Neural networks consider the same variables; however, ANNs use artificial intelligence to solve adjustment problems caused by simple processing elements, which are activated by a function (activation function), to achieve a unique response [9]. In these artificial neurons, the processing unit information consists of “ n ” inputs x_1, x_2, \dots, X_n (Dendrites) and an output (axon). The inputs are associated with the weights W_1, W_2, \dots, W_n representing the synapses. This model can be represented as follows [10]:

$$Y_k = \varphi(V_k) \quad (2)$$

where: Y_k = Output of artificial neuron k ; function φ = activation; V_k = Combiner result.

For using the network, the multilayer perceptron technique was chosen, where initially, the weights of all the networks are generated at random. Sequentially, this individual value update evolves during the error-based learning process.

Estimates of crop yields are simulated with the possible combinations of inputs, for a total of four combinations to the response variable. The activation function used in this method is sigmoid, since it is the most common method in the development of artificial neural networks [11].

The learning of the network is of a supervised type, and the most convenient way to train it is by providing sets of values: the set of input values and a set of output values. Therefore, the training consists in a problem of optimizing the network parameters (its synaptic weights) so that they could respond to inputs as expected until the error among the output patterns generated by the network reaches the minimum desired value [12]. For this purpose, the total number of cycles or times equal to 3000 or less than 1% of the mean square error was used, as suggested.

In order to demonstrate this method of productivity improvement, the following values, taken by the Fertilab laboratories in 2018, are shown in Table 1 [13]:

Table 1 Statistical data

	Mean	Maximum	Minimum	Standard deviation	Cv (%)
Pg (kg)	5852.24	8963.7	2125.81	1233.77	22.74
MO (mg)	25.3	25.58	20.67	1.36	7.25
CTC (cmol)	7.58	8.74	6.63	0.5	6.69
V (%)	41.07	47.36	24.3	6.24	15.4
TA (gr)	296.14	380.07	170.14	53.25	21.74
<i>Correlation between productivity (dependent variable) and predictive</i>					
Productivity	MO = 0.41	CTC = 0.55	V (%) = 0.50	TA = 0.43	

3 Networks in Decision Making

According to the previous table, and the method for perceptron neural networks, the following table is obtained which is an adaptation of the formula for handling patterns by means of ANN [14].

where R^2 is the coefficient of determination; ESP (%) is the percentage of standard error in the evaluation; CAIK: Akaike coefficient; CBY: Bayesian criteria; Prod. R: Network adjusted to estimate the productivity of the crop.

The behavior of the equations estimating productivity in line with the residual errors is shown in Fig. 1: ANN can predict productivity [15], the use of SOM, CTC, V (%) attributes and clay content in an acceptable way. Although these same variables, when subjected to technical regression lose their predictive capabilities [16], the lower return statistics show the resolution of ANN (Table 2).

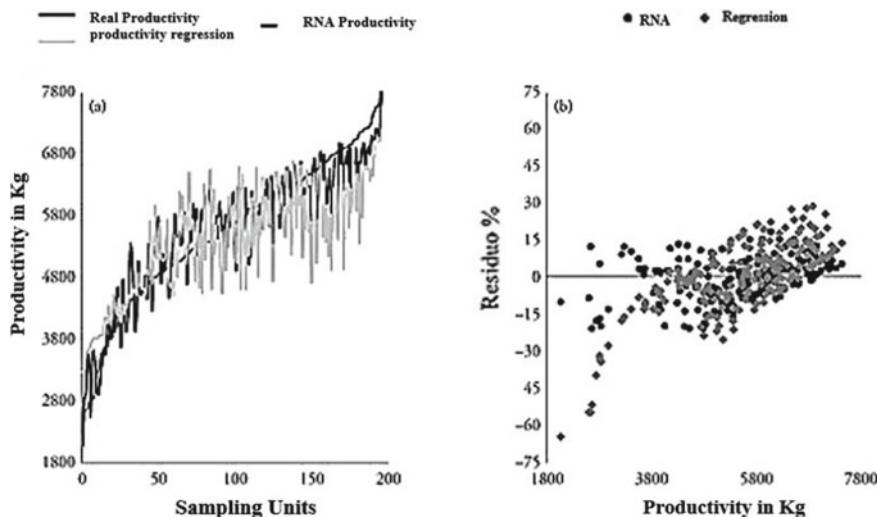


Fig. 1 Data modeling

Table 2 Neural network

Adjusted coefficient regression model								
Variable	β_0	β_1	β_2	β_3	β_4	Statistics		
	-833.14	-102.33	1052.14	17.11	4.32	R^2	ESP	CAIK
		Neural networks				0.51	18.14	1652
		Neurons per layer						1652
Network		Inputs	Hidden	Outputs		Statistics		
	Variables	3	3		2	0.83	12.47	1125
Prod. R	MO	CTC	V (%)	TA				1025

Table 3 Network statistics

Variables production	Medium	Minimum	Maximum	ESP	DA	Errm	χ^2	χ^2 tab (64 GI)
Real	5754	2647	7962				47.7	
RNAs	5251	3145	7742	14.14	11.14	196		19.32
Regression	5547	3236	7652	23.36	22.25	302		40.35

The comparison between regression and ANN by chi-square (χ^2) test; the standard error of the percentage estimate (ESP%); aggregate difference and mean error in absolute terms are shown in Table 3. Values of c^2 calculated against tabulated values were not significant at 95% of probability for the ANN [17].

Behavior of regression models/neural networks to estimate productivity performance through soil properties, compared to actual values (a) and residual distribution graph (b).

Where: ESP (%) Standard error of the estimate as a percentage; DA Aggregate difference in percentage; Errm (ABS), the error of the mean; c^2 calculated chi-square; χ^2 tab chi-square; GI the degree of freedom of the sample; 95% probability. As can be seen, the recession requires more studies in the field, in addition to its low availability, which represents an alternative for the future. In the meantime, ANN has to be feasible, since they are highly available in areas with high production [17]. In addition to regional studies, there is evidence of benefits from their adoption in cultivation [18].

The incorporation of statistics and artificial neural networks can produce highly satisfactory forecasts of wheat yields, as well as of the consequences of soil erosion in harvesting areas [19].

Erosion assessment is often a time-consuming task due to the direct field work required in the affected areas. A single rain event can drastically change the landscape; therefore, a method that allows results to be obtained in a short time is required [20].

The coverage of ground by images has different reflectances and the pixels could be classified as pure and mixed. Images of eroded areas were used to measure the performance of the soil image classification technique by delimiting the photograph with the field data. This allows a full knowledge of photography in accordance with reality. The network for the classification of the image is known as backward propagation and is recommended for the classification of patterns by means of the transfer function, at its output from the network by means of the sigmoid function, which takes the output values between zero and one.

The image was digitized for analysis, so each pixel is analyzed and classified by the network. If it was rejected it is assigned the value of zero (black area in the image) (Fig. 2).

Residuals in rows indicate types of actual coverage that were not included in the classification, while residuals in columns imply coverage (pixels) that do not match reality: in short, they represent errors of omission and commission, respectively [17]. Table 4 shows the result of soil classification and soil erosion.

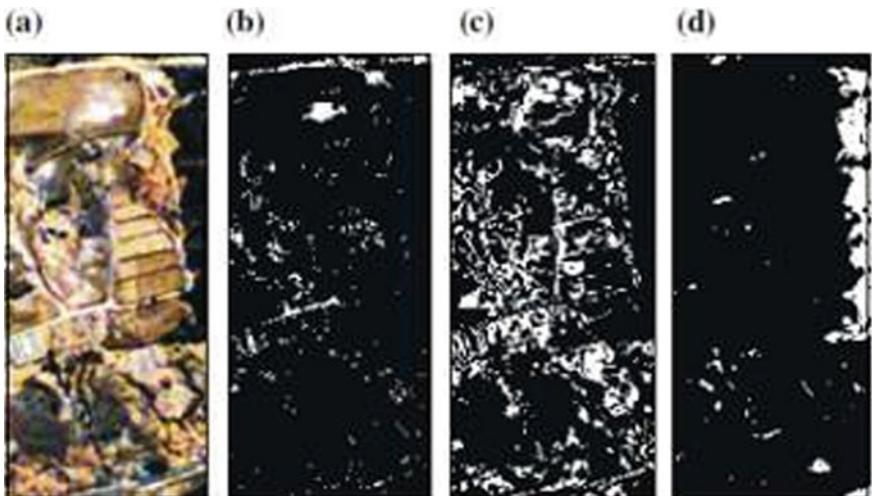


Fig. 2 Binary images

Table 4 Values obtained by network training

Categories	R	V	A	Pixel types	CME	Network performance
White tepetate	241	241	241	Pure	2.3×10^{-12}	1 1 1
Yellow tepetate	241	212	70	Mixed	3.1×10^{-12}	1 1 1
Trees	0	64	3	Mixed	2.2×10^{-14}	1 1 1
dVegetation	69	142	140	Mixed	3.1×10^{-11}	1 1 1
Gullies <20 cm	221	155	175	Mixed	6.8×10^{-12}	1 1 1
dGullies >20 cm	42	210	70	Mixed	6.8×10^{-12}	1 1 1
Pedestals	160	120	60	Mixed	1.7×10^{-10}	1 1 1

The neural network applied acceptably classifies the pure pixels of an aerial image, considering the high percentage of classified pixels representing white tepetates, trees and vegetation. The coding was done in C#, through which the images were captured with a webcam and in communication with MATLAB, developing in the latter the perceptron algorithm.

4 Conclusions

From the previously studied cases of perceptron networks and backpropagation for soil classification and grain productivity activities, it can be clearly observed that, in recent years, artificial neural networks acquired a powerful capacity and efficiency of

non-linear mapping in crop research; in particular, those based on material prediction resources for decision making.

By comparing the results through a reference model applied to the field and other traditional models against those obtained by network-based methods and other non-traditional models, they allow to conclude that precision agriculture, observation and control technologies are gaining ground. In this way, statistical analysis and the inclusion of technology are increasingly cost-effective options for agricultural field management that make it possible to offset the demand for production that is required to meet the needs of the growing population in various countries.

References

1. Abraira V (2014) El Índice Kappa. Unidad de Bioestadística Clínica. 2014. 89, Montreal: sf, 2014, SEMERGEN, vol 12, pp 128–130
2. Apraéz BE (2015) La responsabilidad por producto defectuoso en la Ley 1480 de 2011. Exploración a partir de una obligación de seguridad de origen legal y constitucional. Revista de Derecho Privado (28):367–399
3. FAO (2017) Organización de las Naciones Unidas para la Agricultura y Alimentación. Datos estadísticos. Recuperado el 09 de enero de 2018. <http://www.fao.org/faostat/>
4. García MI (2003) Análisis Y Predicción De La Serie De Tiempo Del Precio Externo Del Café Colombiano Utilizando Redes Neuronales Artificiales. Universitas Scientiarum, vol 8, pp 45–50
5. Matich DJ (2001) “Redes Neuronales: Conceptos básicos y aplicaciones”, Cátedra de Informática Aplicada a la Ingeniería de Procesos–Orientación I
6. Mercado D, Pedraza L, Martínez E (2015) Comparación de Redes Neuronales aplicadas a la predicción de Series de Tiempo. Prospectiva 13(2):88–95
7. Wu Q, Yan HS, Yang HB (2008) A forecasting model based support vector machine and particle swarm optimization. In: 2008 Workshop on power electronics and intelligent transportation system, pp 218–222
8. Clements CF, Ozgul A (2016) Rate of forcing and the forecastability of critical transitions. Ecol Evol 6:7787–7793
9. Comisión Económica para América Latina y el Caribe -CEPAL- (2013) Visión agrícola del TLC entre Colombia y Estados Unidos: preparación, negociación, implementación y aprovechamiento. Serie Estudios y Perspectivas, 25, 87
10. Henao-Rodríguez C, Lis-Gutiérrez JP, Gaitán-Angulo M, Malagón LE, Viloria A (2018) Econometric analysis of the industrial growth determinants in Colombia. In: Australasian database conference, Springer, Cham, pp 316–321
11. Viloria A, Gaitan-Angulo M (2016) Statistical adjustment module advanced optimizer planner and SAP generated the case of a food production company. Indian J Sci Technol 9(47). <https://doi.org/10.17485/ijst/2016/v9i47/107371>
12. Song YY, Ying LU (2015) Decision tree methods: applications for classification and prediction. Shanghai Arch 27:130
13. Mehdiyev N, Lahann J, Emrich A, Enke D, Fettke P, Loos P (2017) Time series classification using deep learning for process planning: a case from the process industry. Proc Comput Sci 114:242–249
14. Wang S, Liu P, Zhang Z, Zhang Y, Song C et al (2016) Development of management methods for “bohai sea granary” data. J Chinese Agric Mechanization 37(3):270–275
15. Liu B, Shao D, Shen X (2013) Reference crop evapotranspiration forecasting model for BP neural networks based on wavelet transform. Eng J Wuhan 34:69–73 [7-5g, Guangzhou: IEEE, 2013, 5102–2575]

16. Silveira CT (2013) Soil prediction using artificial neural networks and topographic attributes. *Geoderma*. 2013, IEEE, pp 192–197
17. Valiente Ó (2013) Education: current practice, international comparative research evidence and policy implications. OCDE, Chicago, pp 44–52 [133-133234-33]
18. Andrecut MK, Ali MA (2012) Quantum neural network model. 2012. *Int J Mod Phys* 12:75–88 [1573-1332]
19. Srinivas A (2013) Handbook of precision agriculture: principles and applications. CRC, New York, 683p
20. Rodrigues MS, Corá JE, Fernandes C (2014) Spatial relationships between soil attributes and corn yield in no-tillage system. *J Soil Sci Plant Nutr* 1:367–379 [1806-9657]

Model for Predicting Academic Performance in Virtual Courses Through Supervised Learning



Jesús Silva, Evereldys García Cervantes, Danelys Cabrera, Silvia García, María Alejandra Binda, Omar Bonerge Pineda Lezama, Juan Lamby, and Carlos Vargas Mercado

Abstract Since virtual courses are asynchronous and non-presential environments, the following of student tasks can be a hard work. Virtual Education and Learning Environments (VELE) often provide tools for this purpose (Zaharia et al. in Commun ACM 59(11):56–65, 2016, [1]). In Moodle, some plugins take information about students' activities, providing statistics to the teacher. This information may not be accurate with respect to leadership ability or risk of abandonment. The use of artificial neural networks (ANNs) can help predict student behavior and draw conclusions at early stages of the learning process in a VELE. This paper proposes a plugin for Moodle that analyzes social metrics through graph theory. This article outlines the advantages of integrating an ANN into this development that complements the use of

J. Silva (✉) · E. G. Cervantes · D. Cabrera · S. García · M. A. Binda
Universidad de la Costa (CUC), Barranquilla, Colombia
e-mail: aviloria7@cuc.edu.co

E. G. Cervantes
e-mail: egarcia12@cuc.edu.co

D. Cabrera
e-mail: dcabrera4@cuc.edu.co

S. García
e-mail: silvia.garcia@upc.pe

M. A. Binda
e-mail: maria.binda@upc.pe

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

J. Lamby
Corporación Universitaria Minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: juan.lamby@uniminuto.edu.co

C. V. Mercado
Corporación Universitaria Latinoamericana, Barranquilla, Colombia
e-mail: cvargas@ul.edu.co

the graph to provide rich conclusions about student performance in a Moodle virtual course.

Keywords Virtual education environments · Supervised learning · Moodle · Neural networks

1 Introduction

Virtual education is a model that tries to be flexible. It is based on the philosophy of asynchrony and non-concurrency, that is, it is not necessary for the agents participating in the learning process to coincide in time or space [2].

One of the problems that Virtual Education and Distance Education present is the little personal contact with the student and, therefore, the little knowledge of the personal situations that he is going through with respect to the learning in the course. In addition, there are also problems that it shares with the model of Presential Education in general, such as the fear that many people drop out the course [3]. For this reason, the development of tools that allow the measurement of relationships within the virtual classroom and their implications for student performance become valuable insofar as they allow to anticipate actions [4]. Previous work has developed a prototype plugin for a Moodle virtual course [5] to measure interactions in a course forum, applying the criteria of centrality metrics. In this way, the teacher has a tool that helps him/her to recognize, within that course, those students who may have greater participation than their classmates.

The growing use of platforms such as VELE requires more tools for teachers, such as behavioral prediction mechanisms that replace the perception they may have in face-to-face classes [6].

This study seeks to deepen, from the implementation of the plugin, in the support tools for teachers extending the prototype in such a way that includes the application of machine learning techniques that allow the model to give support to decisions of the teacher.

2 Student Activities in a VELE: The Moodle Case

The work was developed taking the VELE Moodle, which is used as a teaching support in different universities around the world [7]. A plugin was designed for using a test environment simulating forums and participants, and then applied to a real exercise in a forum, in a current chair of Information Systems Engineering career at the University of Mumbai in India.

Among the advantages of Moodle, its extensibility and widespread use stand out, reaching 9 million users in 229 countries, allowing the existence of a strong community that develops a large number of plugins.

Moodle is based on concepts of social constructivism, so the activities given in a virtual course of the platform make possible a collaborative teaching scheme, in which students contribute to their own training. The teacher, in addition to offering materials, must create an environment that allows students to build their own knowledge [8]. In each virtual course, the student can contribute by carrying out a varied set of activities. The most relevant activities are [9]:

Homework: Students can upload their work so that it can be graded and commented on by the teacher.

Choice/Consultation: The teacher can ask a multiple-choice question for students to answer. It is similar to the survey, but contains only one question.

Feedback: it is a type of survey that the teacher can create with his own questions, unlike predefined surveys.

Forum: May contain asynchronous discussions among all course participants.

Lesson: Allows to create a series of content pages that students navigate flexibly.

Exam/Questionnaire: Allows to build evaluations with multiple features.

Predefined survey: Survey with static questions about the course.

Workshop: It is a space in which students evaluate the work of other students, guided by the teacher.

In the case of predefined queries, feedback and surveys, the analysis of behavior lies in knowing whether the student answered them or not, just as it is useful for lessons to know whether he or she took them. The tasks and exams have a higher level of complexity because, in addition to the need to know if they were carried out, it is necessary to know the grade obtained. In terms of forums and workshops, they provide interactions between course participants, allowing students to detect characteristics such as centrality, leadership capacity or isolation.

In addition, there are other activities that will not be useful for the proposed development, such as: chat, database, access to external learning tools, glossary, SCORM (Shared Content Object Reference Model) and Wiki [10].

3 Plugin: Course Analysis Using Social Metrics and Graphs

In order to help the teacher apply improvements in the quality of learning in a Moodle virtual course, a plugin has been developed to use graph theory to apply social metrics in order to determine student behavior and identify problems. The plugin creates a graph with students as nodes and the interactions between them as arcs and calculates the metric “centrality” to know which is the most central student in the graph. As a first development, it was allowed to create a graph for each discussion within a forum. The latest enhancements include the realization of the entire forum’s graph.

The final objective of the plugin is the analysis of the entire course, taking into account forums, personal activities and group activities of students to draw conclusions that the teacher in charge of the course can interpret, to propose improvements [11].

The graph is created completely with JavaScript code in the view, using the D3.js library [12]. PHP scripts are limited to getting information from the Moodle database and presenting it in an orderly way.

4 Analysis of Student Performance Using Neural Networks

Artificial neural networks (ANNs) are Artificial Intelligence models that emulate the functioning of the brain in terms of the connection of its cells (neurons). These types of structures have the properties of performing distributed computing activities, tolerating noisy inputs, and learning. ANNs have the ability to understand the significance of complicated, inaccurate or unstructured data and can be used to extract patterns, detect trends or make predictions, tasks that may be too complex to otherwise perform [13, 14].

Each ANN is a set of connected perceptrons. A perceptron is a unit with many input channels and one output channel. Each connection is assigned a weight [15]. During the learning phase, the network learns by adjusting these weights in order to predict the correct output for the input tuples. Connectivity between the nodes of a neural network is related to the way in which the outputs of neurons are channeled to become inputs of other neurons. The output signal of a node can be an input from another process element, or even be an input from itself (self-recurring connection). Connections may be forward or backward with respect to levels and an ANN may have one or multiple of them [16].

During the learning period, it is iterated among many examples, leaving the weight when the result is correct compared to the desired one and changing it if it is not. If the output is binary, the problem can be solved with a single perceptron [17].

As will be explained later, in this case the result will be discrete, limiting the possible outputs to a vector of positive integers. Having a number of possible outputs greater than two makes the model necessarily multilayer.

4.1 Supervised Learning

Supervised learning is a sub-category of machine learning. The choice of this type of learning lies in the nature of the data that since a set of tagged trainings is used, that is, the origin of the data is known. The main objective of this learning style is to build a model that makes predictions from examples [18].

Specifically, a supervised learning algorithm takes a set of examples (input) and their response (output), and enters the model to make acceptable predictions for the response to new data.

Supervised learning is divided into 2 categories:

Classification: The objective is to assign a class or label of a series of classes to the prediction.

Regression: The objective is to assign a continuous value to the prediction [19].

4.2 ANNs Applied in Education

Using ANN for problems related to human behavior, more specifically in educational settings, may be appropriate for a number of reasons [20]:

They can find patterns in unstructured datasets.

They can be coupled with e-learning platforms, the use of which is constantly growing.

Its use is suitable for forecasts or predictions of categorization problems.

Studies conducted at the National Technical University of Athens, Greece, suggest that accurate predictions can be obtained in 10-week virtual courses from the third week onwards. The predictions made by ANN were compared with other statistical predictions of linear regression. The comparison showed favorable results for ANNs, concluding that they are more efficient at any stage of forecasting [21].

Regarding the accuracy of ANNs in the analysis of student behavior in educational environments, a study conducted at the Department of Engineering, University of Ibadan, Nigeria, shows that an ANN based on a multilayer perceptron model is able to correctly predict the performance of more than 70% of the analyzed students. The objective of the study was to predict the future performance of a student being considered for admission to the University. Personal data of each future student were analyzed, such as age, type and location of their secondary school, grades in subjects, among others. The network used was trained with data from University alumni [22].

5 Integration of a Neural Network to the Plugin

The main objective of this study is to expand the plugin with the ability to predict the behavior of virtual course learners within Moodle by taking an input data set composed of personal information and information related to each learner's activities within the VELE.

More specifically, the result will be obtained in numerical form, representing the student's final grade. This, in addition to the information provided by the graph on the interactions in the course, may allow the teacher to early identify students who will have problems facing the activities, as well as those who present leadership characteristics. To achieve this, supervised learning techniques will be used.

5.1 Data Set

To generate the model, already completed courses were used as a training set. As input data, the following was used for each student:

Logging activity: Number of connections per week.

Forum activity (metrics): The plugin is currently ready to calculate indegree and outdegree metrics, which refer to student input and output interactions. In addition, another input data would be the number of discussions generated by each student.

Performance in consultations and surveys: Number of consultations, feedback and predefined surveys on the total of those carried out in the course.

Performance in homework and exams: Number of homework assignments out of the total in the course and average grade. It will be represented with only one entry, since the note also reflects the absence or not of the student.

Personal data: Age, sex, number of courses attended.

And as output data the final grade of the course. In supervised learning, this translates into 9 tags for prediction.

In addition to predicting the final grade of the course, the objective is trying to determine the student's risk of dropping out of the course. To make this prediction, a tag was added.

As the output data are discrete, the problem to be solved is categorization.

5.2 Structure of the Neural Network

Figure 1 shows the connections of the upper neuron of each layer as an example.

The neural network that will be used to make the prediction is completely connected and has 3 layers.

The input layer, which has as many neurons as input data.

The output layer, with as many neurons as there are response tags. Label 0 indicates that the student dropped out of the course and label 1–10 indicates the grade.

The hidden layer. Initially, it has 9 neurons (the average between the input and output neurons). The number of neurons in this layer can be modified in order to improve the performance of the network.

5.3 Training

The objective of neural network training is to adjust the weights that each neuron has with respect to its input.

In order to carry out the training, a double validation is carried out, which consists of dividing the data set into two parts. The first part, which consists of 70% of the

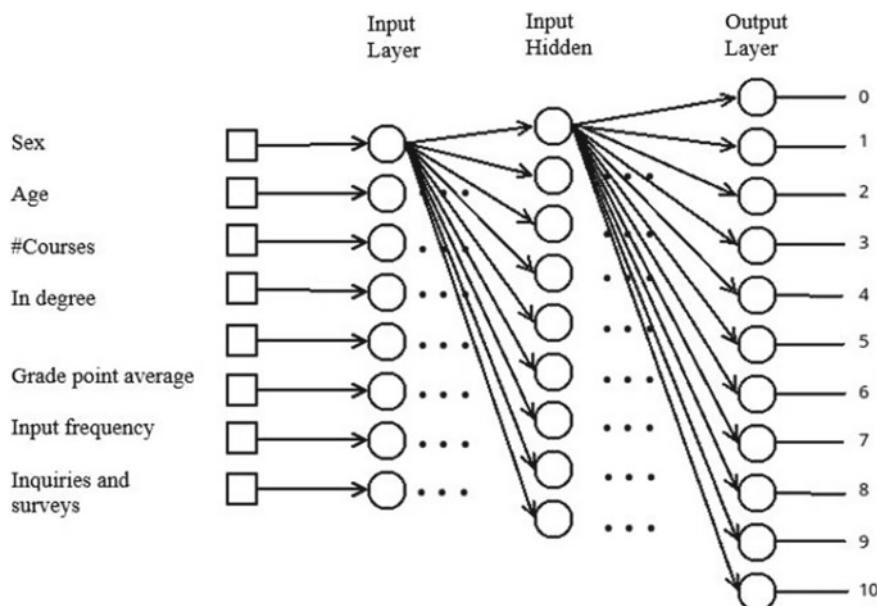


Fig. 1 Structure of ANN each neuron connects to all the neurons in the next layer

training set, is used to train the neural network. The second part, the remaining 30%, is used to validate the trained network.

If the error in the response is less than the acceptable error, the network is considered as trained. Otherwise, the values of the neural network must be adjusted (the initial weights, the pitch function of each network, the number of hidden layers and the number of neurons in each of these layers).

6 Conclusions

The use of artificial intelligence techniques in virtual teaching environments, specifically neural networks, allows the design of teaching strategies on the approach to different problems within the classroom. It also allows to improve the elaboration of contents and to orient the class according to interest and particular problems of each course. This results in improved performance of teachers and students in the learning processes.

The most visible advantage is the ability of prediction to have a better vision of the particularity of each student. Since these are virtual courses, the teacher does not have the individual contact that is so necessary in face-to-face courses. Based on the scores resulting from the analysis, palliative actions can be taken for students who would drop out or not pass, in addition to identifying the students with more

incidence in the course taking into account the high marks and complementing this analysis with the use of the graph that is also part of the plugin.

As a future research, the possibility of extending the architecture of the plugin to other educational platforms besides Moodle is proposed, as well as to extend it to social networks in which groups of students interact. The possibility of sharing anonymous data between different courses is also explored in order to evolve the neural network and generate better results.

References

1. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, Ghodsi A, Gonzalez J, Shenker S, Stoica I (2016) Apache spark: a unified engine for big data processing. *Commun ACM* 59(11):56–65
2. Agrawal R, Srikant R (1994) Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th international conference on very large data bases, VLDB, pp 487–499
3. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
4. Hahsler M, Karpienko R (2017) Visualizing association rules in hierarchical groups. *J Bus Econ* 87:317–335
5. Yuan M, Ouyang Y, Xiong Z, Sheng H (2013) Sentiment classification of web review using association rules. In: Ozok AA, Zaphiris P (eds) Online communities and social computing. OCSC 2013. Lecture notes in computer science, vol 8029. Springer, Berlin
6. Silverstein C, Brin S, Motwani R, Ullman J (2000) Scalable techniques for mining causal structures. *Data Min Knowl Disc* 4(2–3):163–192
7. Amelec V, Carmen V (2015) Relationship between variables of performance social and financial of microfinance institutions. *Adv Sci Lett* 21(6):1931–1934
8. Amelec V, Lezama OBP (2019) Improvements for determining the number of clusters in k-means for innovation databases in SMEs. *Procedia Comput Sci* 151:1201–1206
9. Kamatkar SJ, Kamble A, Viloria A, Hernández-Fernandez L, Cali EG (2018) Database performance tuning and query optimization. In: International conference on data mining and big data, Springer, Cham, pp 3–11
10. Cagliero L, Fiori A (2012) Analyzing Twitter user behaviors and topic trends by exploiting dynamic rules. In: Behavior computing: modeling, analysis, mining and decision. Springer, Berlin, pp 267–287
11. Erlandsson F, Bródka P, Borg A, Johnson H (2016) Finding influential users in social media using association rule learning. *Entropy* 18:164
12. Meduru M, Mahimkar A, Subramanian K, Padiya PY, Gunjgur PN (2017) Opinion mining using twitter feeds for political analysis. *Int J Comput (IJC)* 25(1):116–123
13. Abascal-Mena R, López-Ornelas E, Zepeda-Hernández JS (2013) User generated content: an analysis of user behavior by mining political tweets. In: Ozok AA, Zaphiris P (eds) Online communities and social computing. OCSC 2013. Lecture notes in computer science, vol 8029. Springer, Berlin
14. Dehkharhghani R, Mercan H, Javeed A, Saygin Y (2014) Sentimental causal rule discovery from Twitter. *Expert Syst Appl* 41(10):4950–4958
15. Oladokun VO, Adebanjo AT, Charles-Owaba OE (2008) Predicting students' academic performance using artificial neural network: a case study of an engineering course. *Pac J Sci Technol* 9(1):72–79
16. Finkel JR, Grenager T, Manning C (2005) Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL 2005) pp 363–370

17. Halachev P (2012) Prediction of e-learning efficiency by neural networks. *Cybern Inf Technol* 12(2):98–108
18. Collins M (2002) Discriminative training methods for hidden Markov models: theory and experiments with perceptron algorithms. In: Proceedings of the 2002 conference on empirical methods in natural language processing (EMNLP 2002), pp 1–8
19. Amelec V et al (2019) Integration of data mining techniques to PostgreSQL database manager system. *Procedia Comput Sci* 155:575–580
20. Torres Samuel M, Vásquez C, Viloria A, Hernández Fernandez L, Portillo Medina R (2018) Analysis of patterns in the university Word Rankings Webometrics, Shangai, QS and SIRScimago: case Latin American. Lecture notes in computer science (Including subseries Lecture Notes in Artificial Intelligent and Lecture Notes in Bioinformatics)
21. Jacznik R, Tassara M, D'Uva I, Baldino G (2016) Herramienta de software pedagógica para identificar relaciones y comportamientos en entornos de educación virtual, CyTal
22. Lykourentzou I, Giannoukos I, Mpardis G, Nikolopoulos V, Loumos V (2009) Early and dynamic student achievement prediction in e-learning courses using neural networks. *J Am Soc Inf Sci Technol* 60(2):372–380

Genetic System for Project Support with the Sequencing Problem



Amelec Viloria, Noel Varela, Carlos Herazo-Beltran,
Omar Bonerge Pineda Lezama, Alberto Mercado, Jairo Martinez Ventura,
and Hugo Hernandez Palma

Abstract One of the main problems faced by manufacturing companies in the production sequencing, also called scheduling, which consists of identifying the best way to order the production program on the machines for improving efficiency. This paper presents the integration of a simulation model with an optimization method to solve the problem of dynamic programming with stochastic demand.

Keywords Simulation · Programming · Dynamic sequencing · Job shop · Stochastic demand

A. Viloria (✉) · N. Varela
Universidad de la Costa, Street 58 #66, Barranquilla, Atlántico, Colombia
e-mail: aviloria7@cuc.edu.co

N. Varela
e-mail: nvarela2@cuc.edu.co

C. Herazo-Beltran
Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: carlos.herazo@unisimonbolivar.edu.co

O. B. P. Lezama
Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

A. Mercado
Corporación Universitaria minuto de Dios. UNIMINUTO, Barranquilla, Colombia
e-mail: alberto.mercado@uniminuto.edu.co

J. M. Ventura · H. H. Palma
Corporación Universitaria Latinoamericana, CUL, Barranquilla, Colombia
e-mail: Academico@ul.edu.co

H. H. Palma
e-mail: hernandez@ul.edu.co

1 Introduction

Production management involves the entire planning process from the long term (strategic planning), through tactical or medium-term programming, and reaching operational or short-term programming, in what is known as the hierarchical approach to production [1]. The planning processes of production act in the medium and long term of the company and the decisions are taken at this level of planning directly affect the programming processes of production [2]. The programming process of production aims to assign jobs to the machines in the corresponding stages and define the Processing sequence on each machine, in order to minimize the maximum completion time [3].

Likewise, the distribution of the machines in the production plant or layout defines both the sequence of the processes productive as the methodology for programming production. The flow shop is a special case of the well-known job shop, where the works (i) follow the same sequence of processes and is linear through the stages k present in the factory. In addition, the hybrid flow shop turns out to be an extension of the flow shop, in which there are two or more machines j in one or more stages k in the process [4].

Many research results show the great application that genetic algorithms have had as an approach to solve the flow shop programming problem and the hybrid flow shop, and which also define as a function target makespan [5]. For the programming problem in a flow shop configuration, a hybrid genetic algorithm was built demonstrating greater efficiency in the results regarding the conventional genetic algorithm. [6] Propose a genetic algorithm for programming two parallel machines not related to enlistment times dependent on the sequence [7]. Another publication shows how a genetic algorithm was applied to program a hybrid flow shop in the industry sector ceramics [8].

Since, generally, each specific integration approach uses its own representation of the system, the exchange of data becomes difficult and requires the development of dedicated interfaces [4]. In this sense, this study aims to find, through the integration between the simulation model and the optimization method, one or more satisfactory solutions to the problem of dynamic and stochastic sequencing of production in a Job-Shop environment [9, 10]. Therefore, the contribution of this study lies in the search for continuous improvement in the production process, specifically in production sequencing (scheduling), with a proposal of simple use approach that can be used in production planning, since it can offer more efficient responses in relation to the real environment of constant change and, at the same time, can strategically help decision-makers in the daily process of a manufacturing company.

2 Thematic Approach

2.1 Problem Actors

Now, we will proceed with the definition of the different elements of the Flow-Shop problem and how they were reflected in the application. Thus, it should be remembered that the elements considered in the Flow-Shop problem are:

Machine: The machine is responsible for processing a task at a certain time.

Work: is the one that is composed of different tasks.

Task: It is composed of a time and this must be processed by a machine in the time before mentioned.

Planning: It is the allocation of resources where it is defined which machine should process which task in a certain instant of time.

In this way, the similarity that exists between the elements described above and those of the company in which the planning process will be as follows:

Technician: The technician corresponds to the machine that is responsible for processing a task in a while determined.

Work: is the one that is composed of different tasks, a job corresponds to the repair.

Task: It is composed of a time and a technician must process it in the aforementioned time, a task has an order of precedence defined and must be respected.

Planning: It is the allocation of resources where it is defined which machine should process which task in a certain instant of time.

2.2 Simulation Module

The simulation model was built from a dynamic and stochastic production scenario composed of 8 machines and 10 types of jobs with predefined routes [11], and estimated production times determined according to Table 1.

Table 1 Dynamic and stochastic production scenery: jobs with predefined routes and estimated production times

	Route	Time (min)	Route	Time (min)
1	1, 2, 4, 5	17.3	1, 3, 3, 2, 1, 2, 4, 5, 8	35
2	1, 2, 7, 6, 8	15.2	2, 7, 4, 6, 5	21.3
3	1, 3, 7	10.2	1, 3, 4, 6, 7	18.4
4	1, 3, 4, 6, 7, 8	22.2	2, 1, 5, 6, 8	22.1
5	3, 5, 6, 8	15.7	1, 2, 3, 5, 7	20.6

Table 2 Parameters of the proposed simulation model (std = standard deviation)

Parameters	Expression	Description
Time between arrivals	EXPO (mean)	Exponential description: 8, 10, 12, 14, 16, and 18 min average
Processing time	Normal (5, 1, 5)	A normal distribution of average 4 min and 0.4 min std
Confidence factor	$k = 0, 4$	Confidence factor in relation to expiration date
Estimated production time		Estimated by <i>processing time x number of operation of tasks</i>
Expiration date	NORMAL ((1 + k) * 0.1 * (1 + k)*	Normal distribution of $(1 + k) * \text{average}$ and $0.1 * (1 + k) * \text{std}$

Other parameters of the proposed simulation model are presented in Table 2. The details of the development and assessment are presented in [12], carried out in accordance with [13].

2.3 Job Selection

Once a machine has been selected, it must select a job from among those assigned to modify its assignment. The purely random selection, despite the good results obtained, does not seem a very smart approach, so we proceeded to look for a more targeted way of choosing the job to modify its allocation among those assigned to the previously selected machine. After several attempts at work selection forms, the most promising form came after looking for which jobs among those assigned have shorter process times on other machines. That is, it is checked for each of the jobs assigned to a preselected machine, which of those jobs could be performed faster on another machine [14, 15].

2.4 Optimization Module

Genetic Algorithm (GA) is a meta-heuristic technique—stochastic, non-deterministic method of search and optimization. It imitates the evolutionary process that occurs with biological organisms in nature, based on the process of natural selection. However, GA has some peculiarities in relation to other optimization methods. Mitchell highlights these aspects in [2, 16, 17].

Table 3 summarizes the GA parameters that were adopted to execute the integration proposal.

Table 3 Parameters used in the experiments

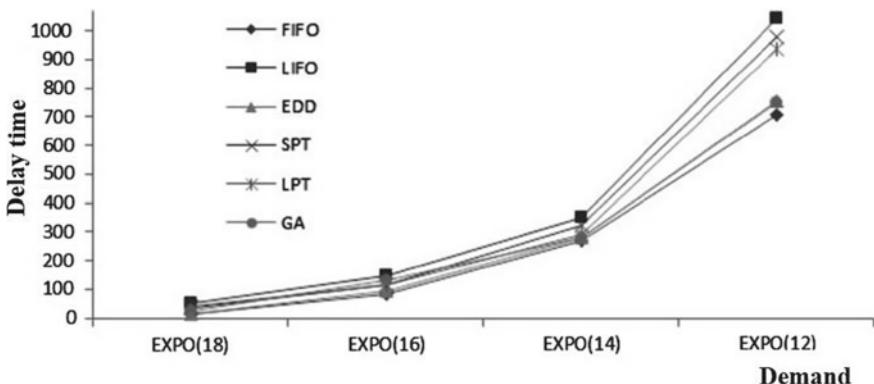
Parameters	Adopted value
Chromosome representation	Ordered operation-based
Selection	Steady-state
Replacement rate	79%
Population size	52
Crossing	Partial match (98%)
Mutation	Swap (3%)
Total number of generations	33
Stop criterion	—
Evaluation function	Simulation module

3 Numerical Results

In this case, the sequencing rule data has a minimum variation of 73% (EDD rule) when the demand changes EXPO(8)-EXPO(6), while the AG variation is less than 10% (increase from 77.1–84.7). This shows that integration can produce equally good results in situations of different demands. These results are shown in Fig. 2. AG-generated solutions are less sensitive to variations in demand.

In addition to promoting a significant reduction in the number of delays, particularly for higher demands, as shown in Figs. 1 and 2, AG-generated solutions are less sensitive to variations in demand. It is noted that the numbers obtained in the AG-generated solutions to EXPO(6) and EXPO(8) claims are almost equal and around 77 units, compared to over 115 units, in the best case, for the dispatch rules (LPT).

In addition to optimization, integration with simulation allows to evaluate other system performance measures and to infer/see important correlations between these measures. The maximum line size and use rate on each machine can be observed

**Fig. 1** Results for the time delay

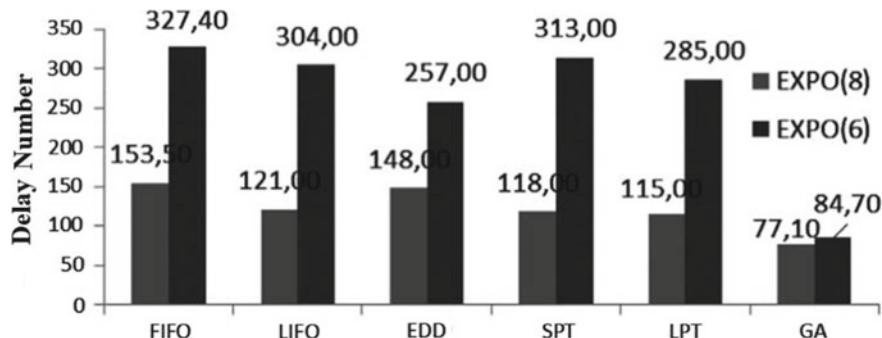


Fig. 2 Numbers of delays for the different approaches

allowing to identify the jobs that limit production (bottleneck, example machine 2) and to see the expected increase in lines and use rates due to increased demand.

4 Conclusions

This paper proposes an integration between the simulation model and the optimization method in order to solve the problem of dynamic and stochastic sequencing of production in a Job-Shop environment [18]. In addition, the solutions generated by the GA are less sensitive to variations in demand, which is very significant in job-shop environments with variable stochastic demand.

References

- Iassinovski S, Artiba A, Bachelet V (2003) Integration of simulation and optimization for solving complex decision-making problems. *Int J Prod Econ* 85(1): 3–10. [https://doi.org/10.1016/S0925-5273\(03\)00082-3](https://doi.org/10.1016/S0925-5273(03)00082-3), <http://www.issn.org/0925-5273>
- Hamid M, Hamid M, Musavi M, Azadeh A (2019) Scheduling elective patients based on sequence-dependent setup times in an open-heart surgical department using an optimization and simulation approach. *Simulation* 95(12):1141–1164
- Zhang B, Yi L-X, Xiao S (2005) Study of stochastic job shop dynamic scheduling. In: Proceedings of the fourth international conference on machine learning and cybernetics, Guangzhou, China, pp 18–21
- Zhang B, Xu L, Zhang J (2020) A multi-objective cellular genetic algorithm for energy-oriented balancing and sequencing problem of mixed-model assembly line. *J Clean Prod* 244:118845
- Banks J (2000) Introduction to simulation. In: Proceedings of the winter simulation conference, Orlando, FL, USA
- Mohammadi A, Asadi H, Mohamed S, Nelson K, Nahavandi S (2018) Optimizing model predictive control horizons using genetic algorithm for motion cueing algorithm. *Expert Syst Appl* 92:73–81

7. Keshanchi B, Souri A, Navimipour NJ (2017) An improved genetic algorithm for task scheduling in the cloud environments using the priority queues: formal verification, simulation, and statistical testing. *J Syst Softw* 124:1–21
8. Silva EB, Costa MG, Silva MFS (2014) Simulation study of dispatching rules in stochastic job shop dynamic scheduling. *World J Model Simul* 10(3):231–240. <http://www.issn.org/1746-7233>
9. Mosadegh H, Ghomi SF, Süer GA (2020) Stochastic mixed-model assembly line sequencing problem: Mathematical modeling and Q-learning based simulated annealing hyper-heuristics. *Eur J Oper Res* 282(2):530–544
10. Leal F, Costa RFS, Montevechi JAB (2011) A practical guide for operational validation of discrete simulation models. *Pesquisa Operacional* 31(1):57–77. <https://doi.org/10.1590/S0101-74382011000100005>, <http://www.issn.org/0101-7438>
11. Kelton WD, Sadowski RP, Sadowski DA (2000) Simulation with ARENA, 2nd edn. McGraw Hill, Boston, USA, pp 385–396. ISBN: 978-0071122399
12. Mitchell TM (1997) Machine learning, 1st edn. McGraw-Hill, New York, USA, pp 249–273. <http://www.issn.org/978-0070428072>
13. Wall M (1996) GALIB: A C++ library of genetic algorithm components. Mechanical Engineering Departament, Massachusetts Institute of Technology. <http://lancet.mit.edu/ga/dist/>
14. Seghir F, Khababa A (2018) A hybrid approach using genetic and fruit fly optimization algorithms for QoS-aware cloud service composition. *J Intell Manuf* 29(8):1773–1792
15. Rauf M, Guan Z, Sarfraz S, Mumtaz J, Shehab E, Jahanzaib M, Hanif M (2020) A smart algorithm for multi-criteria optimization of model sequencing problem in assembly lines. *Robot Comput Integr Manuf* 61:101844
16. Kumar M, Khatak P (2020) Development of a discretization methodology for 2.5 D milling tool-path optimization using genetic algorithm. In: Advances in computing and intelligent systems. Springer, Singapore, pp 93–104
17. Rajagopalan A, Modale DR, Senthilkumar R (2020) Optimal scheduling of tasks in cloud computing using hybrid firefly-genetic algorithm. In: Advances in decision sciences, image processing, security and computer vision. Springer, Cham, pp 678–687
18. Rekha PM, Dakshayini M (2019) Efficient task allocation approach using genetic algorithm for cloud environment. *Cluster Comput* 22(4):1241–1251

Method for the Recovery of Images in Databases of Skin Cancer



Amelec Viloria, Noel Varela, Narledys Nuñez-Bravo,
and Omar Bonerge Pineda Lezama

Abstract Deep learning is widely used for the classification of images since the ImageNet competition in 2012 (Zaharia et al. in Common ACM 59(11):56–65, 2016, [1]; Tajbakhsh et al. in IEEE Trans Med Imaging 35(5):1299–1312, 2016, [2]). This image classification is very useful in the field of medicine, in which there is a growing interest in the use of data mining techniques in recent years. In this paper, a deep learning network was selected and trained for the analysis of a set of skin cancer data, obtaining very satisfactory results, as the model surpassed the classification results of trained dermatologists using a dermatoscope, other automatic learning techniques, and other deep learning techniques.

Keywords Deep learning · Medical images · Clinical data analysis

1 Introduction

Deep learning has been used in the field of computer vision for decades [3, 4]. However, its true value was not discovered until the ImageNet competition in 2012 [5], a success that caused a revolution through its efficient use in Graphics Processing Units (GPUs). The main power of deep learning lies in its architecture [6, 7], which allows discrimination at multiple levels of abstraction for a set of characteristics.

A. Viloria (✉) · N. Varela

Universidad de La Costa, Street 58 #66, Barranquilla, Atlántico, Colombia
e-mail: aviloria7@cuc.edu.co

N. Varela

e-mail: nvarela2@cuc.edu.co

N. Nuñez-Bravo

Universidad Simón Bolívar, Barranquilla, Colombia
e-mail: mnunez3@unisimonbolivar.edu.co

O. B. P. Lezama

Universidad Libre, San Pedro Sula, Honduras
e-mail: omarpineda@unitec.edu

Deep learning techniques have been used successfully in fields such as medicine, in which deep learning comes to solve problems presented by automatic learning algorithms with some widely used data structures in medical images.

The clinical data mining is the application of data mining techniques to clinical data, with the aim of interpreting the available data. It allows the creation of knowledge models and provides assistance for making clinical decisions. In the last 10 years, there has been a growing interest in the application of data mining techniques to clinical data [8, 9]. MEDLINE has seen a strong increase of factor 10 in the number of papers with the term data mining in its title [10]. To do a complete training, deep learning requires a large amount of tagged training data, a requirement that can be difficult to meet in the field of medicine, where the expert annotation is expensive and diseases (injuries) do not present large datasets. In addition, it requires a large amount of computer resources so that training does not become excessively slow [11].

This study intends to apply deep learning techniques to analyze a set of skin cancer image data, obtaining very satisfactory results. In the following sections, the methods and the obtained results are shown.

2 Method

This section describes the methods used for the development of the study.

2.1 *Tensorflow*

Tensorflow [12] is one of the best deep learning frameworks. It has been adopted by a lot of big companies like Airbus, Twitter, IBM and others because of its flexibility and versatility. Tensorflow is developed by Google, which uses it in all its automatic learning and deep learning projects. Tensorflow is not a deep learning framework itself, it is a framework that allows us to work very quickly with matrices thanks to its parallelization in GPUs.

Since almost all calculations made to train and predict with a neural network are matrix calculations, this tool is ideal for using in the construction of neural networks. Tensorflow has an internal package that comes with the functionality needed to run a convolutional neural network: Convolutional layers, optimizers, optimization functions, etc. A disadvantage of Tensorflow is that it requires writing a great quantity of code to make it function. Writing all that code gives total control over all the elements of the neural network architecture, but it also causes to make a lot of mistakes [13].

2.2 Keras

Keras [14] is a specific deep learning framework. It is not a competitor of Tensorflow because Keras runs “on top” of Tensorflow. Keras provides an extremely easy syntax for the creation of neural networks, and then converts this syntax to Tensorflow models using its power to run all the learning machinery. Due to the fact that this research consists of a case of neural network use, and the manipulation at a very low level of the deep learning models is not needed, the researchers opted for the use of Keras for the development of the present project.

3 Results

The data set selected to evaluate the proposal was the set of skin cancer images from The International Skin Imaging Collaboration (ISIC) (<https://isic-archive.com/>). It is a platform that aims to bring dermatological professionals together with the aim of fighting skin cancer. The data set consists of 23,906 images. In addition to the images, some metadata are provided, including age, sex, anatomical location of the melanoma, type of diagnosis, class of melanoma, and thickness of the melanoma [15]. First, the data sets were analyzed with respect to the benign/malignant class. As shown in Fig. 1, it is very unbalanced.

For this reason, a resampling technique was applied to the minority class that is in charge of choosing random elements with replacement and adding them to the minority class. The advantage of this technique is that it does not eliminate

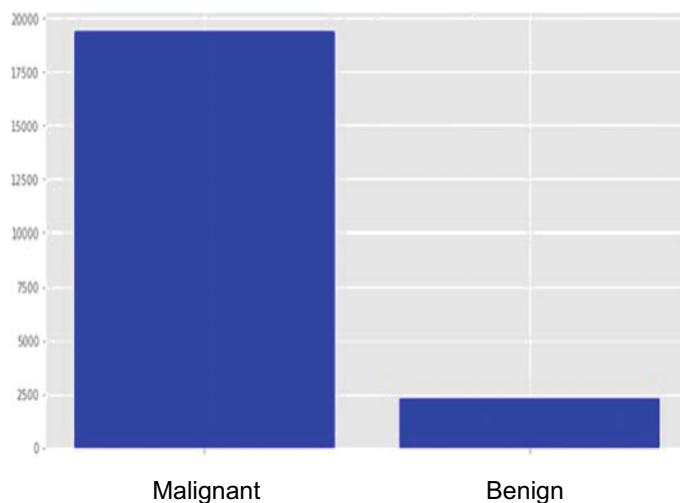


Fig. Fig. 1 Unbalanced skin cancer dataset (absolute frequency)

information, simply reinforces it about the minority class. Second, the data sets were split into training (80%) and test (20%), making sure to keep the classes balanced in each of the sets [16].

Third, a data augmentation mechanism was applied, consisting of carrying out different random transformations on each of the elements of the training set. By performing random transformations on the data set, a double effect was achieved: increasing the size of the data and making the system generalize much better. In particular, the method included the application of techniques known as automatic rotation, vertical and horizontal translation of pixels, shearing, zooming, turning, and salt and pepper filter [17].

3.1 Selection and Application of the Deep Learning Model

At the moment of selecting the model, the researchers faced a moment of maximum uncertainty because there are many parameters that had to be chosen, between these parameters is the neural network architecture: Own, InceptionV3, VGG16, VGG19, Xception, etc.; type of training: Knowledge transfer, from scratch, etc.; Parameters of knowledge transfer: layers to train, complete training, etc.; Learning algorithm: Gradient descent, RMSProp, Adagrad, etc.; parameters of the learning algorithm: learning rate, activation functions, times, lots, etc. To select the appropriate parameters, researchers had to use the grid search. This technique is based on the definition of a set of possible values that parameters can take, and generate the Cartesian product of all the values for parameters between them, using each of these generated parameter sets to build a model and evaluate goodness [18].

Three parameters were fixed in the search grid: times, batches, and activation functions. The epochs were established since it is evident that the more epochs the better the models will work, so two epochs were established for all of them to run through the data set twice [19]. The batches go along with the capacity of the Graphics Processing Unit, although they can influence the training. In this case, calculations were made so that the set of batches is the maximum that the Graphics Processing Unit can support, lowering the search time per grid. The activation function is fixed to the ReLu activation function because all architectures use ReLu since learning is done faster with this type of function.

The model selected as optimal is the Inception V3 model, created by Google. As described in [20], it was expected that knowledge transfer would be selected instead of training from scratch, as it may have a higher performance. It is also recommended that, although starting from a few defined weights (for the transfer of knowledge), it is necessary to allow the readjustment of some weights in the convolutional layers, in order to adapt it to the problem. Adam is used as the learning algorithm, with a rather small rate. This small rate is due to the fact that thanks to the knowledge transfer it is very close to the optimum, taking very small steps to get closer to the optimum. Once the model and parameters are defined, the researchers proceed to increase the number of epochs in order to obtain a better model and be able to evaluate in the next

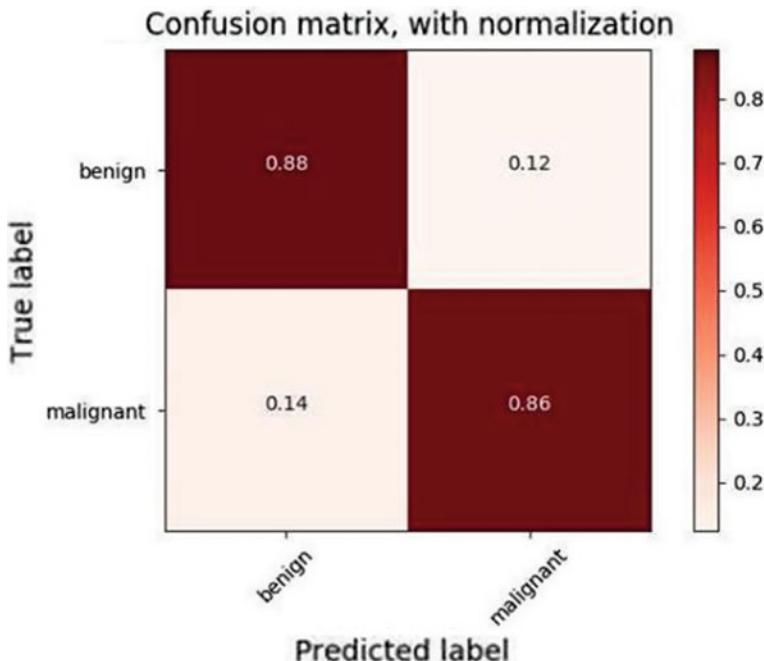


Fig. Fig. 2 Standard confusion matrix

section. Once the training of the model was completed, the classifier was applied to the set of tests, obtaining the results shown in Fig. 2.

3.2 Model Evaluation

The evaluation of the model is shown below. Accuracy is the default metric used by Keras to represent the goodness of the model throughout the different training seasons. Figure 3 shows how the classifier behaves with respect to the accuracy throughout the different periods for the validation and training sets.

As can be observed in the graph that they are always in growth, except in epoch 6 that the set of tests presents something hard-to-interpret, but in the following epoch it recovers. The classifier was training for 10 seasons, but an option that saves resources was activated. This option consists in that if the classifier during the stage of classification begins to remain for some time without changing its accuracy, the learning stage stops when the epoch is over. For this reason, it can be observed that, although it was training for 10 seasons, only 8 are observed in the graph.

Also provided are the measures that can be seen in Table 1 regarding the performance of the deep learning model. Figure 4 represents the area under the ROC curve.

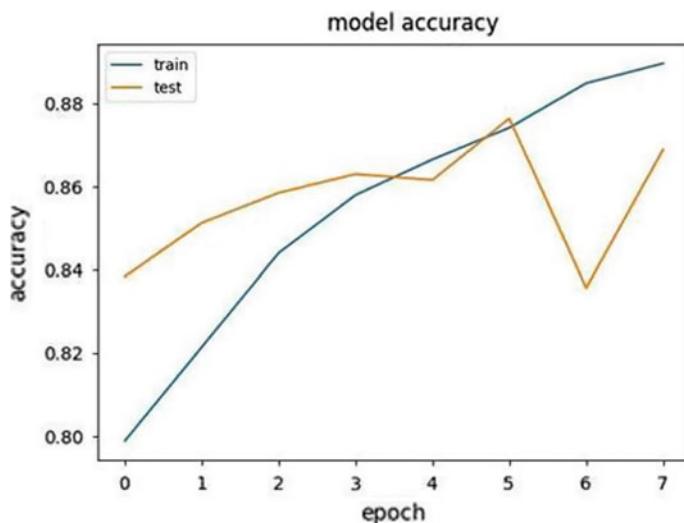


Fig. Fig. 3 Accuracy of the model throughout the different epochs

Table Table 1 Results of the deep learning model

Measure	Value
Accuracy	88.20%
Precision	86.37%
Sensitivity	85.14%
Specificity	88.14%
ROC area	0.91

Visual inspections without the help of dermatological experts obtain an average of 60% accuracy [21], due to the complexity of observation on differentiating characteristics with the naked eye. With the help of a trained expert along with a dermatoscope, the accuracy can be increased up to 76–85%. Starting from the base of the previously exposed assumptions, and the metrics obtained by the classifier, it can be observed that this classifier surpassed in success a dermatologic expert trained for doing use of a dermatoscope [22].

This looks like a good starting point. Knowing that the classifier is capable of surpassing an expert dermatologist, researchers proceed to compare the classifier with other techniques. The first intuition when trying to apply artificial intelligence techniques in the analysis of medical images is to use automatic learning or deep learning. So, these two approaches will be used in order to compare the classifier to study its performance with respect to other classifiers. First, the focus will be on the comparison of the classifier with the automatic learning techniques.

The automatic learning on the ISIC data set obtained the metrics shown in Table 2 [4]. As can be observed from the table, the classifier outperforms the automatic

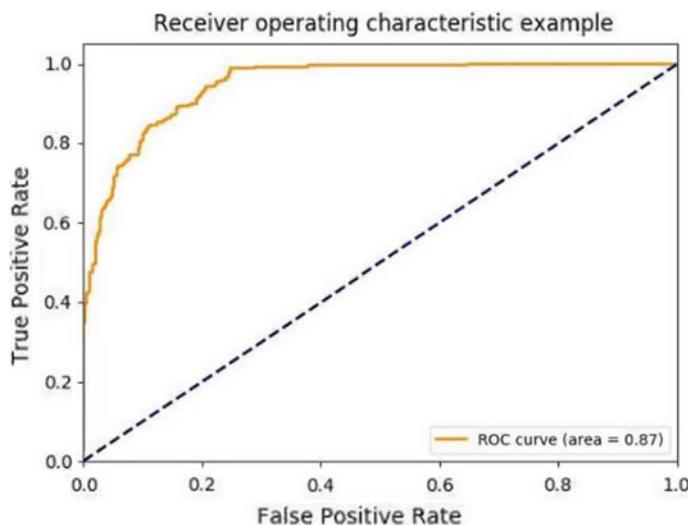


Fig. Fig. 4 Area under the ROC curve

Table Table 2 Comparison of results between deep and automatic learning

Measure	Proposed classifier (%)	Machine learning (%)
Accuracy	88.20	74.21
Precision	86.37	70.14
Sensitivity	85.14	69.54

learning classifiers focused on the analysis of medical images for the ISIC data set. The result is logical because the automatic learning algorithms are not designed to extract characteristics from images, so a lower performance is always assumed.

Second, the classifier was compared with another deep learning classifier [6]. The classifier with which this classifier was compared, is used by a team participating in the ISIC competition. The results can be seen in Table 3. It can be observed that, although the performance in accuracy in both classifiers is similar, this one behaves much better in most of the situations. They have managed to get a very high

Table Table 3 Comparison of results between proposed deep learning and other models

Measure	Proposed classifier	Another classifier
Accuracy	88.20%	84.2%
Precision	86.37%	62.5%
Sensitivity	85.14%	49.8%
Specificity	88.14%	92.5%
ROC area	0.91	0.7

specificity, which is what allows them to compensate for the sensitivity to get a quite acceptable accuracy and AUC ROC.

4 Conclusions

This paper presented the results of applying deep learning techniques to a set of skin cancer image data. It described the phase of data preprocessing, selection of parameters for the selected model and different metrics were calculated for the obtained results. The metrics show that the classifier is quite homogeneous, and there are no hard-to-interpret results. The classifier presents a fairly stable level of success in both classes, which indicates that it has correctly generalized the detection of cancer in the images.

This model has surpassed the results of classification by trained dermatologists using a dermatoscope, other automatic learning techniques, and other deep learning techniques. Definitively, the classifier performed very well in front of this set of images with respect to other classifiers and techniques. Even knowing that medical images are hard to treat, the classifier was able to generalize the details that separate an image of a patient with skin cancer from another who does not, thus accurately predicting most of the cases.

References

1. Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, Ghodsi A, Gonzalez J, Shenker S, Stoica I (2016) Apache spark: a unified engine for big data processing. *Commun ACM* 59(11):56–65
2. Tajbakhsh N, Shin JY, Gurudu SR, Todd Hurst R, Kendall CB, Gotway MB, Liang J (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging* 35(5):1299–1312
3. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: Proceedings of the 52nd annual meeting of the association for computational linguistics: system demonstrations, pp 55–60
4. Hahsler M, Karpienko R (2017) Visualizing association rules in hierarchical groups. *J Bus Econ* 87:317–335
5. Alves LGA, Ribeiro HV, Rodrigues FA (2018) Crime prediction through urban metrics and statistical learning. *Phys A Stat Mech Appl* 505:435–443
6. Silverstein C, Brin S, Motwani R, Ullman J (2000) Scalable techniques for mining causal structures. *Data Min Knowl Disc* 4(2–3):163–192
7. Amelec V, Carmen V (2015) Relationship between variables of performance social and financial of microfinance institutions. *Adv Sci Lett* 21(6):1931–1934
8. Amelec V, Lezama OBP (2019) Improvements for determining the number of clusters in k-Means for innovation databases in SMEs. *Procedia Comput Sci* 151:1201–1206
9. Kamatkar SJ, Kamble A, Viloria A, Hernández-Fernandez L, Cali EG (2018) Database performance tuning and query optimization. In: International conference on data mining and big data, Springer, Cham, pp 3–11

10. Erlandsson F, Brodka P, Borg A, Johnson H (2016) Finding influential users in social media using association rule learning. *Entropy* 18:164
11. Baculo MJC, Marzan CS (2017) Remedios de Dios Bulos, and Conrado Ruiz. Geospatial-temporal analysis and classification of criminal data in manila. In: Proceedings of 2nd IEEE international conference on computational intelligence and applications, IEEE, pp 6–11
12. Amelec V et al (2019) Integration of data mining techniques to PostgreSQL database manager system. *Procedia Comput Sci* 155:575–580
13. Clougherty E, Clougherty J, Liu X, Brown D (2015) Spatial and temporal analysis of sex crimes in Charlottesville, Virginia. In: Proceedings of IEEE systems and information engineering design symposium, IEEE, pp 69–74
14. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M et al (2016) Tensorflow: a system for large-scale machine learning. *OSDI* 16:265–283
15. Codella NCF, Gutman D, Celebi ME, Helba B, Marchetti MA, Dusza SW, Kalloo A, Liopyris K, Mishra N, Kittler H et al (2017) Skin lesion analysis toward melanoma detection: a challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018), IEEE, pp 168–172
16. Iavindrasana J, Cohen G, Depersinge A, Müller H, Meyer R, Geissbuhler A (2009) Clinical data mining: a review. *Yearb Med Informatics* 18(01):121–133
17. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J (2008) LIBLINEAR: a library for large linear classification. *J Mach Learn Res* 9:1871–1874
18. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The WEKA data mining software: an update. *SIGKDD Explorations* 11(1):10–18
19. Kang H-W, Kang H-B (2017) Prediction of crime occurrence from multimodal data using deep learning. *PLoS ONE* 12(4):e0176244
20. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521(7553):436
21. Leitão JC, Miotto JM, Gerlach M, Altmann EG (2016) Is this scaling nonlinear? *Roy Soc Open Sci* 3(7):25–36
22. Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A (2016) Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). Preprint <http://arXiv.org/1605.01397>

Author Index

A

- Abilash, R., 291
Acharya, Debopam, 323
Adiraju, Rama Vasantha, 697, 849, 857
Aghav, Shubham, 737
Aher, Jayshree, 737
Ahmed, Muqeem, 385
Ahmed, Syed Musthak, 279
Akshara, Revelly, 21
Angadi, Basavaraj M., 473
Arun Kumar Gudivada, A., 619
Aswathi, P., 425
Azman, Mohamed, 73, 93

B

- Babhlugaonkar, Arun, 217
Bade, Dattatray, 609
Baghel, Vimal Anand, 305
Bansal, Saumya, 355
Behera, Manas Chandan, 435
Berdejo, Pedro, 939
Bharne, Smita, 365
Bhat, Naveen N., 465
Bhavani, R. Durga, 315
Bhonekar, Amol P., 1, 785
Biju, Mable, 407
Binda, María Alejandra, 967
Bodkhe, Sanjay, 45
Bokka, Raveendranadh, 725
Bothra, Aditya, 355

C

- Cabrera, Danelys, 939, 949, 967

- Castro, Nadia Leon, 959
Cervantes, Evereldys Garcia, 967
Chacon, Ramon, 949
Chaitanya Kumar, D., 849
Chakradhar, A., 239
Chakravaram, Venkamaraju, 145
Challa, Madhavi Latha, 203
Chandrasekaran, K., 511
Charulatha, B. S., 291
Chatterjee, Santanu, 481
Cheggoju, Naveen, 671, 681
Chidambaram, M., 333
Chockaiah, Nikhila Shri, 415
Choudhry, Anurag, 375
Corrales, Patricio, 865

D

- Das, Bhaskarjyoti, 745
Das, Shreyas, 251
Deepthi, S., 715
Deepudev, S., 755
Devarapalli, Danny Joel, 225
Devi, M. Nirmala, 415
Dhathri, D., 635
Dhiman, Surender, 355
Dhondiram, Patil Manoj, 189
Dubey, Kumkum, 525
Dubey, Shivendu, 121

E

- Echeverry, Francisco Javier, 875, 893

F

- Fernández, Claudia, 929
 Fernandes, Warren Mark, 33
 Flores, Yasmin, 939

G

- Gandhi, Hetal, 579
 García, Silvia, 967
 Gautam, Minakshi, 553
 Ghallab, Abdullatif, 767
 Ghosh, Swastik, 251
 González, Ana María Echeverría, 959
 Gopi, Varun P., 491, 755
 Gunjan, Vinit Kumar, 279
 Gupta, Anil Kumar, 579
 Gupta, Ankur, 203
 Gupta, Ritik, 435
 Gupta, Vishal, 157
 Gurubelli, Yugeswararao, 171

H

- Halder, Biswajit, 441
 Haresh, R., 817
 Harikrishnan, P. M., 491
 Herazo-Beltran, Carlos, 977
 Herazo-Beltran, Yaneth, 911
 Herrera, Maritza, 875, 893
 Herz, Jeannette, 929
 Higa, Yuki, 939

J

- Jadav, Nilesh Kumar, 9
 Jagadeeswara Rao, E., 569, 635, 689
 Jagtap, Santosh, 609
 Jorge, Marisol, 875, 893

K

- Kakkasageri, Mahabaleshwar S., 473, 627, 663
 Karar, Vinod, 785
 Karedula, Rajkamal, 397
 Kaur, Arshpreet, 1
 Kaur, Manpreet, 157
 Kaur, Ravreet, 157
 Kaw, Rishab, 785
 Kayal, S. K. Swetha, 415
 Khandare, Hrishikesh, 217
 Khatoon, P. Salma, 385
 Khot, Amruta, 179
 Kiran, Dendukuri Ravi, 543

Kirithika, P.

- Kondur, Monica, 553
 Kori, Gururaj S., 627
 Kottath, Rahul, 785
 Kovela, B., 279
 Krishna, Aki Vamsi, 715
 Krishna, R. V. V., 533, 689, 697
 Kumar, Naresh, 795
 Kumar, Nikhil, 323
 Kumar, Sanjeev, 561, 647
 Kumar, S. Suresh, 553
 Kumar, V. S., 239
 Kunda, Parvateesam, 647
 Kurdukar, Atharv, 217

L

- Lakshmi Akhila, M., 689
 Lamby, Juan, 959, 967
 Lekha, Shekar, 425
 Lezama, Omar Bonerge Pineda, 865, 875, 883, 893, 901, 911, 921, 939, 949, 959, 967, 977, 985
 Lohani, Divya, 323

M

- Maco, José, 921
 Madhukar Rao, G., 589, 599
 Magdum, Anmol, 179
 Majumder, Koushik, 481
 Malar, J. Kavin, 415
 Malhotra, Ruchiika, 453
 Mandaviya, Himani, 135
 Manju, I., 817
 Marín-González, Freddy, 921
 Mathew, Monica Merin, 407
 Menon, Balu M., 425
 Mercado, Alberto, 929, 977
 Mercado, Carlos Vargas, 967
 Mishra, Ashish, 121
 Mishra, Bhabani Shankar Prasad, 251
 Mishra, Manoj Kumar, 251
 Mishra, Shivendu, 525
 Mohsen, Abdulqader, 767
 Montero, Edgardo Rafael Sanchez, 949
 Moulik, Sayantan, 441
 Mounika, N., 315
 Mudengudi, Shailaja S., 663
 Musale, Manali, 217

N

- Nandan, Durgesh, 533, 561, 569, 619, 635, 689, 697, 849, 857
Nandimandalam, Mohan Krishna Varma, 811
Nandu, Eslavath, 93
Narasimhulu, C. Venkata, 55
Narendran, Danussvar Jayanthi, 291
Naveda, Alexa Senior, 939
Nawandar, Neha K., 671, 681
Nelli, Manjunath K., 755
Nirmala Devi, M., 345, 715
Nisha, J. S., 491
Nuñez-Bravo, Narledys, 985

O

- Ojha, Muneendra, 305
Orellano, Nataly, 949
Ortiz-Ospino, Luis, 883
Ovallos-Gazabon, David, 901, 929

P

- Palanisamy, P., 491, 755
Palit, Achinta K., 67
Palma, Hugo Hernandez, 977
Pal, Nisha, 525
Panday, Suman Kumar, 533
Panicker, Jithu G., 73
Paramkusham, Spandana, 55
Pasupuleti, Swarnalatha, 849
Patel, Ankit, 269
Patel, Ashish, 501
Patel, Ashish Singh, 305
Patil, Chinmay, 737
Patil, Shubham, 737
Patil, Suryakant, 365
Pavan Avinash, G., 857
Pinillos-Patiño, Yisel, 865
Poddar, Shashi, 785
Ponnusamy, Palanisamy, 171
Popale, Shraddha, 579
Porika, Dhanrajnath, 827
Prasad, M. Shiva, 239
Premchand, Anshu, 375
Priya, B. Jyothi, 647

Q

- Quintero, Benjamín, 865

R

- Raj, Patnala Prudhvi, 745

Rajat, 795

- Rajesh Kumar Reddy, C., 811
Rajpoot, Prince, 525
Ramakrishnan, M., 707
Ramanathan, Malmathanraj, 171
Ramesh, Dharavath, 589, 599
Ramesh Kumar, P., 857
Ramos, Lainet Nieto, 929
Ramu, Vaishnavi S., 553
Rashid, Mohammad Saad, 33
Ratnakaram, Sunitha, 145
Ravi, A., 315
Ravi Shankar, V. Ch. S., 697
Reddy, G Pradeep, 543
Reddy, Kothur Dinesh, 543
Reddy, Pranay Kumar, 553
Reddy, Sandhi Kranthi, 21
Renjith, Shini, 407
Revanasiddappa, M., 465
Rishikeshan, C. A., 811
Rohith, Aki, 543
Rohith Kumar, E., 817
Romero, Ligia, 875, 893
Roncallo, Alberto, 901
Ruiz-Lazaro, Alex, 959

S

- Sadasivam, Tamilselvan, 725
Sadhukhan, Swarnali, 481
Sahu, Amit, 121
Saif, Mohammed H., 767
Saikiran, Gogineni, 827
Sai Kiran, K. V., 93
Sai Prakash, S. K. L. V., 93
Sangale, Sagar, 737
Santhoshi, M. Siva, 561
Santodomingo, Nicolas Elias Maria, 875, 893
Sardal, Nihar, 269
Sarkar, Subhanjan, 481
Sathwara, Snehal, 135
Satpute, Vishal R., 671, 681
Sawant, Vinaya, 269
Shah, Jigarkumar, 501
Shah, Preet, 745
Shanmugam, R., 333
Sharith Babu, K., 561
Sharma, Hemant, 203
Sharma, Manoj, 795
Shashvat, Kumar, 1
Shekar, Y., 239
Shirsath, Mahesh, 217
Shivalal Patro, B., 435

- Shukla, Anupam, 511
 Silva, Jesús, 865, 875, 893, 921, 929, 939, 949, 967
 Singh, Mithilesh Kumar, 33
 Singh, Pradeep, 397
 Sing, Mihiir, 481
 Sinha, Sachin, 553
 Sinha, Yash Kirti, 305
 Solano, Darwin, 929
 Sonwani, Bhupendra Kumar, 305
 Srinivasan, K. S., 817
 Srinivas, Suyoga, 465
 Srinivasulu, Kothuru, 259
 Subiksha, K. P., 707
 Sukier, Harold, 865
 Suresh, Pragnya, 745
 Surya Narayana, G., 827
 Surya, Sri Sai, 619
- T**
 Talhar, Archana, 45
 Tanaji, Khude Anupam, 189
 Tatikonda, Neelakantam, 145
 Tekale, Adwait, 217
 Thomas, Anju, 491
 Tiwari, Priya, 609
- U**
 Unnikrishnan, Aravind, 511
 Upadhyay, Ashish, 305
 Urdanegui, Rosella, 929
- V**
 Vadariya, Arpita, 9
 Vara Prasad, R. U. S. D., 697
 Varela, Noel, 883, 901, 911, 921, 977, 985
 Vargas, Jesús, 865
 Vargas, Martin, 949
 Venkata Ganesh, P., 569
 Ventura, Jairo Martinez, 901, 977
 Verma, Karan, 1
 Verma, Neetu, 525
 Vidushi, 453
 Vihari, Nitin Simha, 145
 Villareal-González, Reynaldo, 921
 Villón, Martín, 921
 Viloria, Amelec, 883, 901, 911, 959, 977, 985
 Vinit Kumar, Gunjan, 827
 Visbal, Juan Manuel Cera, 939
 Vishnupriya, R., 345
- Y**
 Yaduvanshi, Ritika, 525