

Yu-Chen Hu
Shailesh Tiwari
Krishn K. Mishra
Munesh C. Trivedi *Editors*

Ambient Communications and Computer Systems

RACCCS-2018

Advances in Intelligent Systems and Computing

Volume 904

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 Hagrass, Electronic Engineering, University of Essex, Colchester, UK

László T. Kóczy, Department of Automation, Széchenyi István University, Győr, 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>

Yu-Chen Hu · Shailesh Tiwari ·
Krishn K. Mishra · Munesh C. Trivedi
Editors

Ambient Communications and Computer Systems

RACCCS-2018

 Springer

Editors

Yu-Chen Hu
Department of Computer Science
and Information Management
Providence University
Taichung, Taiwan

Shailesh Tiwari
Department of Computer Science
and Engineering
ABES Engineering College
Ghaziabad, Uttar Pradesh, India

Krishn K. Mishra
Department of Computer Science
and Engineering
Motilal Nehru National Institute
of Technology
Allahabad, Uttar Pradesh, India

Munesh C. Trivedi
Department of Computer Science
and Engineering
ABES Engineering College
Ghaziabad, Uttar Pradesh, India

ISSN 2194-5357 ISSN 2194-5365 (electronic)
Advances in Intelligent Systems and Computing
ISBN 978-981-13-5933-0 ISBN 978-981-13-5934-7 (eBook)
<https://doi.org/10.1007/978-981-13-5934-7>

Library of Congress Control Number: 2018966446

© Springer Nature Singapore Pte Ltd. 2019

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

RACCCS-2018 is a major multidisciplinary conference organized to provide a forum for researchers, educators, engineers, and government officials involved in the general areas of communication, computational sciences, and technology to disseminate their latest research results and exchange views on the future research directions of these fields; to exchange computer science and integrate its practice and application of the academic ideas; to improve the academic depth of computer science and its application; to provide an international communication platform for educational technology and scientific research for the universities and engineering field experts and professionals.

Nowadays, globalization of academic and applied research is growing with speedy pace. Computer, communication, and computational sciences are the heating areas with a lot of thrusts. Keeping this ideology in preference, the third version of International Conference on Recent Advancement in Computer, Communication and Computational Sciences (RACCCS-2018) has been organized at **Aryabhata College of Engineering and Research Center, Ajmer, India, during August 10–11, 2018.**

Ajmer is situated in the heart of India, just over 130 km south-west of Jaipur, a burgeoning town on the shore of the Ana Sagar Lake, flanked by barren hills. Ajmer has historical strategic importance and was sacked by Mohammed Gauri on one of his periodic forays from Afghanistan. Later, it became a favorite residence of the mighty Mughals. The city was handed over to the British in 1818, becoming one of the few places in Rajasthan controlled directly by the British rather than being part of a princely state. The British chose Ajmer as the site for Mayo College, a prestigious school opened in 1875 exclusively for the Indian princes, but today open to all those who can afford the fees. Ajmer is a perfect place that can be symbolized for the demonstration of Indian culture and ethics and for the display of perfect blend of a wide plethora of diverse religion, community, culture, linguistics, etc., all coexisting and flourishing in peace and harmony. This city is known for the famous Dargah Sharif, Pushkar Lake, Brahma Temple, and many more evidences of history.

This is the third time Aryabhata College of Engineering and Research Center, Ajmer, India, is organizing international conference based on the theme of computer, communication, and computational sciences, with a foreseen objective of enhancing the research activities at a large scale. Technical Program Committee and Advisory Board of RACCCS include eminent academicians, researchers, and practitioners from abroad as well as from all over the nation.

RACCCS-2018 received around 235 submissions from 752 authors of 5 different countries such as USA, Saudi Arabia, China, and many more. Each submission has gone through the plagiarism check. On the basis of plagiarism report, each submission was rigorously reviewed by at least two reviewers with an average of 2.43 per reviewer. Even some submissions have more than two reviews. On the basis of these reviews, 46 high-quality papers were selected for the publication in this proceedings volume, with an acceptance rate of 19.5%.

We are thankful to the speakers, delegates, and the authors for their participation and their interest in RACCCS as a platform to share their ideas and innovation. We are also thankful to Prof. Dr. Janusz Kacprzyk, Series Editor, AISC, Springer, and Mr. Aninda Bose, Senior Editor, Hard Sciences, Springer Nature, for providing continuous guidance and support. Also, we extend our heartfelt gratitude to the reviewers and Technical Program Committee members for showing their concern and efforts in the review process. We are indeed thankful to everyone directly or indirectly associated with organizing team of the conference leading it toward the success.

Although utmost care has been taken in compilation and editing, a few errors may still occur. We request the participants to bear with such errors and lapses (if any). We wish you all the best.

Ajmer, India

Yu-Chen Hu
Shailesh Tiwari
Krishn K. Mishra
Munesh C. Trivedi

About This Book

The field of communication and computer sciences always deals with new arising problems and finding effective and efficient techniques, methods, and tools as solutions to these problems. This book captures the latest and rapid developments that happen in the immediate environmental surroundings of computer and communication sciences.

Nowadays, various cores of computer science and engineering field entered in a new era of technological innovations and development, which we are calling “*Smart Computing and smart communication.*” Smart computing and communication provides intelligent solutions to the problems which are more complex as compared to the general problems. We can say that the main objective of *Ambient Computing and Communication Sciences* is to make software, techniques, and computing and communication devices which can be used effectively and efficiently.

Computational and computer sciences have a wide scope of implementation in engineering sciences through networking and communication as a backbone. Keeping this ideology in preference, this book includes the insights that reflect the immediate surroundings developments in the field of communication and computer sciences from upcoming researchers and leading academicians across the globe. It contains the high-quality peer-reviewed papers of ‘*International Conference on Recent Advancements in Computer, Communication and Computational Sciences*’ (RACCCS-2018), held at Aryabhata College of Engineering and Research Center, Ajmer, India, during August 10–11, 2018. These papers are arranged in the form of chapters. The contents of this book cover five areas: *intelligent hardware and software design; advanced communications; intelligent computing technologies; web and informatics; and intelligent image processing.* This book helps the prospective readers from computer and communication industry and academia to derive the immediate surroundings developments in the field of communication and computer sciences and shape them into real-life applications.

Contents

Part I Advanced Communications

EERP: Energy-Efficient Relay Node Placement for k-Connected Wireless Sensor Networks Using Genetic Algorithm	3
Akhilesh Kumar Srivastava and Suneet Kumar Gupta	
Detection of Online Malicious Behavior: An Overview	11
D. S. Deshpande, S. P. Deshpande and V. M. Thakare	
Reliable Data Delivery with Extended IPV4 Using Low-Power Personal Area Network	25
Shambhavi Mishra, Pawan Singh and Anil Kumar Tiwari	
Design and Investigations of Multiband Microstrip Patch Antenna for Wireless Applications	37
Ajay Dadhich, Preeti Samdani, J. K. Deegwal and M. M. Sharma	
A Novel ZOR-Inspired Patch Antenna for Vehicle Mounting Application	47
Chetan Barde, Arvind Choubey, Rashmi Sinha, Santosh Kumar Mahto and Prakash Ranjan	
A Nascent Approach for Noise Reduction via EMD Thresholding	55
Rashi Kohli and Shubhi Gupta	
Congestion Control Network Coding Scheme in Delay-Tolerant Network	67
Uroosa Zaidi, Praneet Saurabh, Ritu Prasad and Pradeep Mewada	
IoT: Architecture, Technology, Applications, and Quality of Services	79
Vidhyotma and Jaiteg Singh	
High-Speed Optical Mode Division Multiplexing of Hermite–Gaussian Modes in Multimode Fiber	93
Saumya Srivastava, Kamal K. Upadhyay and Nar Singh	

Part II Intelligent Computing Techniques

A Novel Algorithm for Automatic Text Summarization System Using Lexical Chain	103
Ashima Tiwari and Deepak Dembla	
Interactive User Interface to Advent HCI Artefact	113
Anil Kumar Dubey	
A Comprehensive Survey on Artificial Bee Colony Algorithm as a Frontier in Swarm Intelligence	125
Shiv Kumar Agarwal and Surendra Yadav	
Modeling and Simulation of Al6082 MMC of Gravity Die Casting for Solidification Time	135
Harendra Pal, Dinesh Kumar Kasdekar and Sharad Agrawal	
A Survey on the Detection of Windows Desktops Malware	149
Sanjay K. Sahay and Ashu Sharma	
Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image	161
Chethan Dev, Kripa Kumar, Arjun Palathil, T. Anjali and Vinitha Panicker	
Extractive Summary: An Optimization Approach Using Bat Algorithm	175
Anshuman Pattanaik, Santwana Sagnika, Madhabananda Das and Bhabani Sankar Prasad Mishra	
Phylogenetics Algorithms and Applications	187
Geetika Munjal, Madasu Hanmandlu and Sangeet Srivastava	
Comparative Study of Forecasting Model for Price Prediction of Rice	195
Kesari Verma, N. K. Nagwani and Shrish Verma	
Wind Power Forecasting Using Hybrid ARIMA-ANN Technique	209
Pavan Kumar Singh, Nitin Singh and Richa Negi	
Part III Hardware and Software Design	
On the Development of Feature-Based Sprint in AGILE	223
Sarika Sharma and Deepak Kumar	
A Reliable Novel Framework of User-Oriented Software Engineering	237
Gurpreet Singh Saini, Sanjay Kumar Dubey and Sunil Kumar Bharti	

Android-Based Blind Learning Application 247
 Abhishek Ranjan and T. M. Navamani

An Approach for Test Case Prioritization Using Harmony Search for Aspect-Oriented Software Systems 257
 Abhishek Singhal, Abhay Bansal and Avadhesh Kumar

Onboard Data Acquisition System to Monitor the Vehicle 265
 Adesh Kumar Pandey and Sangeeta Arora

Part IV Web and Informatics

Toward Adapting Metamodeling Approach for Legacy to Cloud Migration 275
 Pooja Parnami, Aman Jain and Navneet Sharma

Application of Cloud Computing for Priority Job Scheduling by Multiple Robots Operating in a Co-operative Environment 285
 Amitava Kar, Ajoy K. Dutta and Subir K. Debnath

A Framework for Security Management in Cloud Based on Quantum Cryptography 295
 Priya Raina and Sakshi Kaushal

Analyzing Student Performance Using Data Mining 307
 Pankhurhi Mallik, Chandrima Roy, Ekansh Maheshwari, Manjusha Pandey and Siddharth Rautray

Prediction of Employee Turnover Using Ensemble Learning 319
 Shubham Karande and L. Shyamala

Automated Review Analyzing System Using Sentiment Analysis 329
 A. C. Jishag, Vishnu Rakhesh, Suraj Mohan, N. Vinayak Varma, Vaisakh Shabu, Lekshmi S. Nair and Maya Menon

A Comprehensive Survey on Extractive and Abstractive Techniques for Text Summarization 339
 Abhishek Mahajani, Vinay Pandya, Isaac Maria and Deepak Sharma

Cloud Storage–Optimization of Initial Phase for Privacy-Preserving Public Auditing 353
 Deepak Kumar Verma, Purnima Gupta and Rajesh Kumar Tyagi

A New Approach for Cloud Security Using Hybrid Querying System Over Cloud Scenario 367
 Priya Sen, Ritu Prasad and Praneet Saurabh

A Novel Approach for Meta-Search Engine Optimization 377
 S. Siji Rani and S. Goutham

Why Adopting Cloud Is Still a Challenge?—A Review on Issues and Challenges for Cloud Migration in Organizations	387
Mohammed Shuaib, Abdus Samad, Shadab Alam and Shams Tabrez Siddiqui	
Analysis of Block-Level Data Deduplication on Cloud Storage	401
L. Suresh and M. A. Bharathi	
Computing Narayana Prime Cordial Labeling of Web Graphs and Flower Graphs	411
B. J. Balamurugan, K. Thirusangu and B. J. Murali	
Reliability-Aware Green Scheduling Algorithm in Cloud Computing	421
Chesta Kathpal and Ritu Garg	
An Empirical Study to Classify Website Using Thresholds from Data Characteristics	433
Ruchika Malhotra and Anjali Sharma	
Part V Intelligent Image Processing	
Assessment of Spectral-KMOD Composite Kernel-Based Supervised Noise Clustering Approach in Handling Nonlinear Separation of Classes	449
Ishuita SenGupta, Anil Kumar and Rakesh Kumar Dwivedi	
Multilevel Steganography for Data Protection	461
Sourabh Tyagi, Priyanka Anurag and Smitha N. Pai	
Feature Extraction of Normalized Colorectal Cancer Histopathology Images	473
Alok Kumar Jain and Shyam Lal	
An Efficient License Plate Text Extraction Technique	487
Anuj Kumar, Anuj Sharma and R. K. Singla	
Toward Recognition and Classification of Hindi Handwritten Document Image	497
Shalini Puri and Satya Prakash Singh	
Efficient Image Deblurring Using Alpha Plane Blending on Images Recovered with Linearly Varied Point Spread Function (PSF)	509
Poonam Sharma, Ashwani Kumar Dubey and Ayush Goyal	
Biometric Authentication-Based Data Encryption Using ECG Analysis and Diffie–Hellman Algorithm	523
Archana Bhardwaj, Shikha Chaudhary and Vijay Kumar Sharma	
Author Index	533

About the Editors

Prof. Yu-Chen Hu received his PhD degree in computer science and information engineering from the Department of Computer Science and Information Engineering, National Chung Cheng University, Chiayi, Taiwan in 1999. Currently, Dr. Hu is a professor in the Department of Computer Science and Information Management, Providence University, Sha-Lu, Taiwan. He is a member of ACM and IEEE. He is also a member of Computer Vision, Graphics, and Image Processing (CVGIP), Chinese Cryptology and Information Security Association, and Phi Tau Phi Society of the Republic of China.

He also serves as the Editor-in-Chief of International Journal of Image Processing since 2009. In addition, he is the managing editor of Journal of Information Assurance & Security. He is a member of the editorial boards of several other journals.

His areas of interest are: image and signal processing, data compression, information hiding, and data engineering.

Dr. Shailesh Tiwari currently works as a Professor in Computer Science and Engineering Department, ABES Engineering College, Ghaziabad, India. He is an alumnus of Motilal Nehru National Institute of Technology Allahabad, India. His primary areas of research are software testing, implementation of optimization algorithms and machine learning techniques in various problems. He has published more than 50 publications in International Journals and in Proceedings of International Conferences of repute. He is edited Scopus, SCI and E-SCI-indexed journals. He has also edited several books published by Springer. He has organized several international conferences under the banner of IEEE and Springer. He is a Senior Member of IEEE, member of IEEE Computer Society, Fellow of Institution of Engineers (FIE).

Dr. Krishn K. Mishra is currently working as a Assistant Professor, Department of Computer Science and Engineering, Motilal Nehru National Institute of Technology Allahabad, India. His primary area of research includes evolutionary algorithms, optimization techniques and design, and analysis of algorithms. He has

also published more than 50 publications in International Journals and in Proceedings of Internal Conferences of repute. He is serving as a program committee member of several conferences and also edited Scopus and SCI-indexed journals. He is also member of reviewer board of Applied Intelligence Journal, Springer.

Dr. Munesh C. Trivedi currently works as an Associate Professor in Department of IT, REC, Azamgarh, UP, India. He has published 20 text books and 80 research publications in different International Journals and Proceedings of International Conferences of repute. He has received Young Scientist and numerous awards from different national as well international forum. He has organized several international conferences technically sponsored by IEEE, ACM and Springer. He is on the review panel of IEEE Computer Society, International Journal of Network Security, Pattern Recognition Letter and Computer & Education (Elsevier's Journal). He is Executive Committee Member of IEEE UP Section, IEEE India Council and also IEEE Asia Pacific Region 10.

Part I
Advanced Communications

EERP: Energy-Efficient Relay Node Placement for k -Connected Wireless Sensor Networks Using Genetic Algorithm



Akhilesh Kumar Srivastava and Suneet Kumar Gupta

Abstract In this article, we propose relay node placement for providing k -connectivity to randomly deployed sensor nodes with energy efficacy using Genetic Algorithm (GA). Here, we also explain the basic step of GA with suitable examples. Also, we carried out the extensive simulations to study proposed algorithm's performance with existing one in terms of number of deployed nodes and lifetime of the network.

Keywords Energy efficiency · k -connectivity · WSNs · Genetic algorithm

1 Introduction

In the past few years, the area of WSNs has been devoted enormous look because of its vast application in general scenario, e.g., monitoring of environment, army field deployment in various factories, etc and many more [1]. The layout of Wireless Sensor Networks (WSNs) is a tedious task with enormous effect on the standard, price, and efficiency of general life sensor needs [2]. The crucial goal of such network is to cover an area and pass the information to remotely placed Base Station (BS) known as sink. In WSN, energy is one of the most important constraints because every sensor is battery operated and once deployed, it is very critical to change the energy cell, so while designing such network lifetime must be considered [3]. The random deployment of sensor nodes leads to high cost and waste a lot of energy [4]. Therefore, the design of WSN which provide energy efficacy known as Deployment and Power Assignment Problem (DPAP) [5] connectivity is also a critical issue in WSNs because network is divided into disjoint subnetworks due to failure of

A. K. Srivastava (✉)

CSE Department, ABES Engineering College, Ghaziabad, India

e-mail: akhilesh.srivastava@abes.ac.in

S. K. Gupta

CSE Department, Benett University, Gautam Buddha Nagar, India

e-mail: suneet.banda@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,

Advances in Intelligent Systems and Computing 904,

https://doi.org/10.1007/978-981-13-5934-7_1

sensor(s) [6]. The generalized definition of k -connectivity or k -fault tolerance is that entire nodes are adjoined with minimum k other nodes, k -connectivity provides fault tolerance of faulty $k - 1$ sensor nodes [7].

To provide desired connectivity having help of minimum number of sensing points is NP hard problem [8], and to have solution of such problems, Genetic Algorithm (GA) is the emerging prominent method [6, 7]. To provide required connectivity with the help of minimum number of sensor node is not sufficient because there is no consideration of energy efficacy, due to which some relay node(s) die quickly and network is partitioned.

Younis and Akkaya [4] presented a nice outlook of different methods for coverage problems. Best known and famous strategy for retaining the network establishment in the span of unresponsive node is to place duplicate sensor units later to the deployment of a WSN. To handle problem of this nature, many heuristics are available to generate approximate results. Research publication on sensor disposition can be clubbed into two divisions. One attempt is to merely confirm connectivity among edge nodes, i.e., $k = 1$ [9]. The other one attempt is to gain maximum connectivity [9]. The second one is preferred for this paper given the context. k -connectivity ensures the network to bear the fault of maximum $k - 1$ consecutive node. In [6] Suneet et al., GA- and Greedy-based approaches are presented to deploy the relay nodes in target-based WSN with considering the k -connectivity ($k \geq 1$), but author did not consider the energy efficacy parameter.

In contrast to [6], we have taken two objectives, in first objective, we consider the minimization of relay node and other target is to improve the life span of relaying nodes. In [6], the paper represents initial population as a parameter as the number of goals and we represent initial population in terms of potential positions, which are predefined. Here, we try to establish a trade-off between number of relaying nodes and lifetime.

Rest of the research work is presented as follows: In Sect. 2, we discussed terminologies and network lifetime with energy model. The proposed work has been discussed in Sect. 3. The experimental setup and results are presented in Sect. 4. Conclusion is presented in Sect. 5.

2 Terminology and System Model

2.1 Terminologies

1. $S = \{S_1, S_2, S_3 \dots S_m\}$ represents the set of m sensor nodes.
2. $Z = \{Z_1, Z_2, Z_3 \dots Z_n\}$ denotes the n predefined potential positions.
3. $R = \{R_1, R_2, R_3 \dots R_k\}$ represents the k relay node. The relay node R_i may be placed at potential position Z_i . R_{k+1} represents the base station.
4. $\text{Com_Range}(R_k)$ represents the range of a relay node R_k and $R_k \in R$.

2.2 Network Span, Energy Model

The network span indicates that for how much time wireless sensor networks is in operational condition. The network span is totally dependent on the application. In [10], the various definitions for network lifetime has been discussed. Some of them are as follows:

- Failure of any sensor node,
- Failure of any cluster head (CH),
- Failure of definite fraction of total number of sensor nodes,
- Failure of definite fraction of total number of cluster head (CH), and
- If any sensor node is in working condition.

For the experimental result, it is assumed that the network is in existence till all the nodes have sufficient energy to forward the data to base station. In proposed work, we have used the same energy model as discussed in [11].

3 Proposed Work

In the past, several research items used the genetic algorithm for area of wireless sensor networks to solve the problems, i.e., routing [12–14] clustering [15, 16], deployment, etc. Now, the author of this paper proposes GA-based technique for ensuring k -connectivity to all objects. Procedure of GA, e.g., chromosome representation, selection of population, crossover, and mutations are explained as follows.

3.1 Representation of Chromosome and Initial Population

Authors propose the representation of the chromosome as a series of potential position. The size of chromosome is assumed to be constant and equal to number of points. For a chromosome, if i th gene is 1, meaning that i th potential position is selected for placing the relay node. Here, we randomly generate the initial population by putting the 0 or 1 for each gene value in chromosome. In Fig. 1, a subgraph of WSN is depicted with five sensor nodes and nine potential positions, which are as follows.

A chromosome for a part of graph (refer Fig. 1) is depicted in Fig. 2. Here, the gene value at potential position 1, 5, 7, 8 is 0, it means these potential positions are not in use and in rest positions relay nodes are placed.

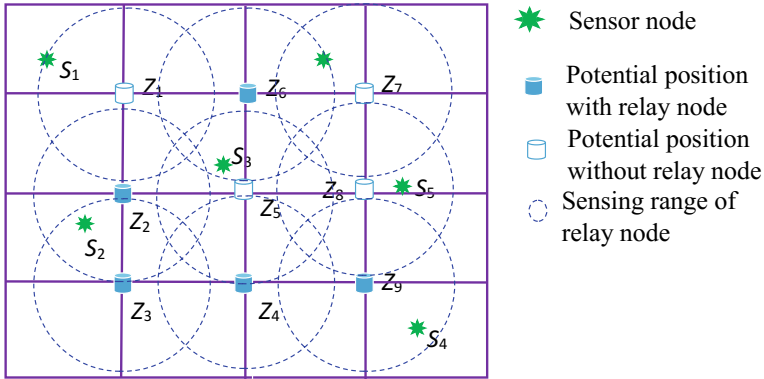


Fig. 1 A subgraph of a WSN with 5 sensor nodes and 9 potential positions

Fig. 2 Chromosome representation for subgraph represented in Fig. 5

Potential position	1	2	3	4	5	6	7	8	9
Gene value	0	1	1	1	0	1	0	0	1

3.2 Fitness Function

This value of a chromosome signifies its appropriateness with regard to facilitate k -connectivity with energy efficacy to every sensor nodes with the help of least number of relay points. In proposed work, there are two objectives, these are as follows:

Objective 1: Limiting number of relay nodes

Our first objective is to deploy a least number of relay nodes to ensure k -connectivity to all targets. Suppose we target m of k potential position, then

$$M = \left| \bigcup_{i=1}^m X_i \right|$$

So, *Objective 1 Maximize* $F_1 = 1/M$

Objective 2: Prolong the network lifetime

Here, we calculate the network lifetime, with previously discussed energy-efficient routing algorithm known as GAR [17]. In proposed work, the authors have used the same energy model proposed in [17]. So, another objective is

Objective 2: Maximize $F_2 = \text{Network lifetime}$

So, our fitness function is

$$\text{Fitness} = W_1 \times F_1 + W_2 \times F_2$$

$$W_1 + W_2 = 1 \text{ and } 0 \leq W_i \leq 1; \forall W_i \text{ and } i = 1, 2.$$

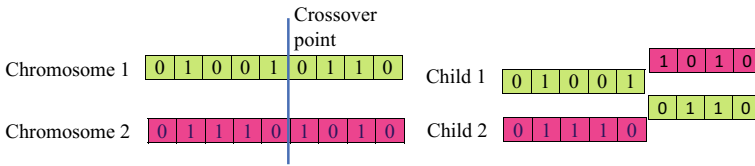


Fig. 3 Generation of child chromosome after one-point crossover operation

3.3 Selection

In the process of selection in GA, authors have picked 10% fittest population using Roulette wheel for attempting one-point crossover. In mutation operation, we randomly select a potential position with gene value 1 and replace it with 0.

3.4 Crossover and Mutation

For the crossover operation, one-point crossover method has been used. For the same, two chromosomes have been randomly selected and share the information with each other. The crossover operation is pictorially represented in Fig. 3, which is as follows.

After crossover operation, there are four chromosomes (two parents and two children). On the basis of fitness value, 2 out of 4 chromosomes are selected and placed in the initial population. For the mutation operation, randomly select a gene value and replace that value with 0.

4 Experimental Results

For applying Simulation, authors have assumed two nonidentical network scenarios (*WSN #1* and *WSN #2*). These two have the sensing blocks of 100 m × 100 m and 300 m × 300 m, respectively. In each of these scenarios, authors have considered a grid-sized 10 m and location of relay nodes at the intersection of these grids except boundary positions. For the execution of this GA-based method, authors have prepared an initial population of some 60 chromosomes. Authors compare the efficiency of proposed algorithms with GA-based algorithm discussed in [7]. Authors have executed the proposed algorithm for $w_1 = 0.5$ and $w_2 = 0.5$.

The comparison of required number of relay nodes is represented in Fig. 4 for the desired connectivity (i.e., k). From the figure, it was noticed that the efficiency of proposed method lies between Greedy-based and GA-based approaches discussed in [7]. The efficiency of proposed method is not the optimum because in proposed

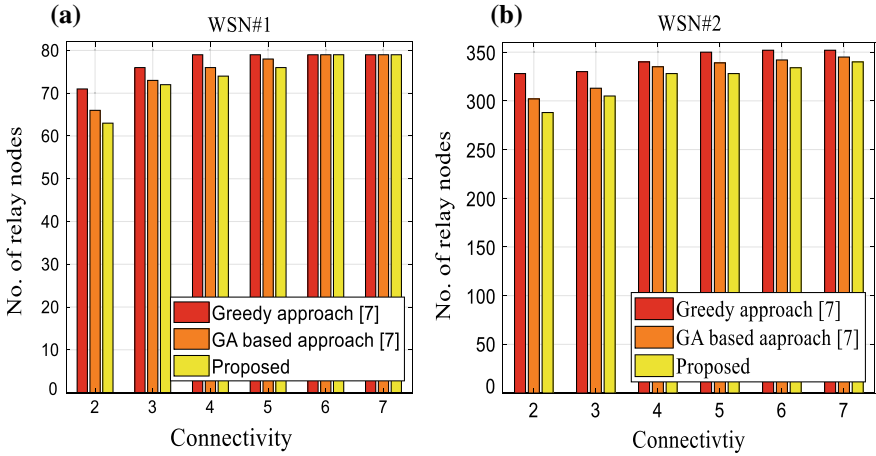


Fig. 4 Number of relay nodes required in scenario **a** WSN #1 and **b** WSN #2

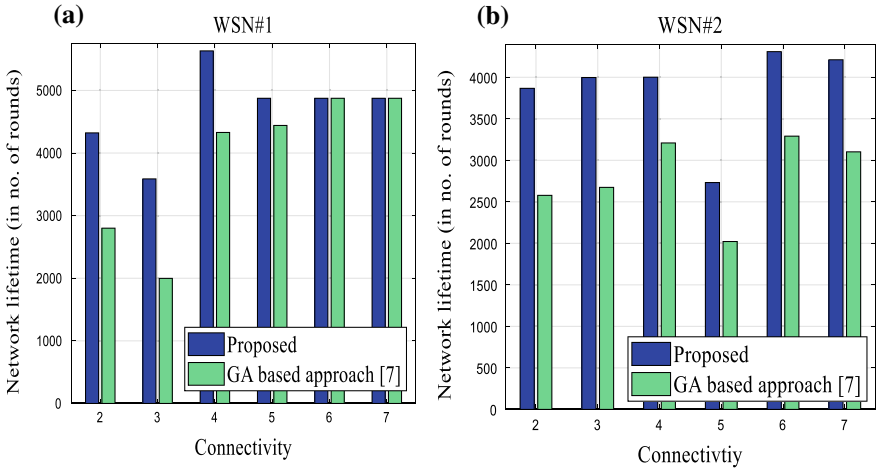


Fig. 5 Number of relay nodes required in scenario **a** WSN #1 and **b** WSN #2

algorithm two objectives have been considered and 50% weight is assigned to both the parameters.

The efficiency of proposed method with respect to network span has been represented in Fig. 5 and it is noticed that the efficiency of proposed method is superior to the GA-based method discussed in [7]. Moreover, the performance is better because energy efficacy is also an important parameter in the proposed algorithm. From the figure, it is also observed that for the smaller region the performance of both algorithms is very similar with higher degree of connectivity. However, as the size of region is large, then the proposed algorithm performs far better.

5 Conclusion

Through this research paper, authors have presented the algorithm for node deployment using genetic algorithm. Methodology of GA, i.e., chromosome representation, selection, crossover, and mutation has been explained with suitable examples. From the experimental setup, we conclude that the number of deployed nodes are more than the GA-based methods published in [7]. However, the proposed algorithm enhances the network lifetime by 130% (approximately) for the network scenario having lower connectivity as compare to GA-based approach discussed in [7]. Moreover, for the network scenario having higher connectivity, the performance of proposed method (with regard to network span) is very close to algorithm discussed in [7]. In proposed work, we have only considered the connectivity of the nodes with energy efficacy and in future we will try to develop the algorithms with desired coverage and connectivity simultaneously.

References

1. Akyildiz, I.F., Vuran, M.C.: *Wireless Sensor Networks*, vol. 4. Wiley, Chichester (2010)
2. Akyildiz, I.F., Su, W., Sankarasubramanian, Y., Cayirci, E.: *Wireless sensor networks: a survey*. *Comput. Netw.* **38**(4), 393–422 (2002)
3. Akyildiz, I.F., Su, W., Sankarasubramanian, Y., Cayirci, E.: *A survey on sensor networks*. *IEEE Commun. Mag.* **40**(8), 102–114 (2002)
4. Younis, M., Akkaya, K.: *Strategies and techniques for node placement in wireless sensor networks: a survey*. *Ad Hoc Netw.* **6**(4), 621–655 (2008)
5. Konstantinidis, A., Yang, K., Zhang, Q., Zeinalipour-Yazti, D.: *A multi-objective evolutionary algorithm for the deployment and power assignment problem in wireless sensor networks*. *Comput. Netw.* **54**(6), 960–976 (2010)
6. Gupta, S.K., Kuila, P., Jana, P.K.: *Genetic algorithm approach for k-coverage and m-connected node placement in target based wireless sensor networks*. *Comput. Electr. Eng.* **56**, 544–556 (2015)
7. Gupta, S.K., Kuila, P., Jana, P.K.: *Genetic algorithm for k-connected relay node placement in wireless sensor networks*. In: *Proceedings of the Second International Conference on Computer and Communication Technologies*, pp. 721–729. Springer, Berlin (2016)
8. Cardei, M., Thai, M.T., Li, Y., Wu, W.: *Energy-efficient target coverage in wireless sensor networks*. In: *INFOCOM 2005. 24th Annual Joint Conference of the IEEE Computer and Communications Societies*. *Proceedings IEEE*, vol. 3, pp. 1976–1984. IEEE (2005)
9. Han, X., Cao, X., Lloyd, E.L., Shen, C.-C.: *Fault-tolerant relay node placement in heterogeneous wireless sensor networks*. *IEEE Trans. Mob. Comput.* **9**(5), 643–656 (2010)
10. Bari, A., Wazed, S., Jaekel, A., Bandyopadhyay, S.: *A genetic algorithm based approach for energy efficient routing in two-tiered sensor networks*. *Ad Hoc Netw.* **7**(4), 665–676 (2009)
11. Heinzelman, W.B., Chandrakasan, A.P., Balakrishnan, H.: *An application-specific protocol architecture for wireless microsensor networks*. *IEEE Trans. Wirel. Commun.* **1**(4), 660–670 (2002)
12. Gupta, S.K., Kuila, P., Jana, P.K.: *Energy efficient multipath routing for wireless sensor networks: a genetic algorithm approach*. In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 1735–1740. IEEE (2016)
13. Shukla, R.N., Chandel, A.S., Gupta, S.K., Jain, J., Bhansali, A.: *GAE³BR: genetic algorithm based energy efficient and energy balanced routing algorithm for wireless sensor networks*. In:

- 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 942–947. IEEE (2015)
14. Gupta, S.K., Kuila, P., Jana, P.K.: GA based energy efficient and balanced routing in k -connected wireless sensor networks. In: Proceedings of the First International Conference on Intelligent Computing and Communication, pp. 679–686. Springer, Berlin (2017)
 15. Kuila, P., Gupta, S.K., Jana, P.K.: A novel evolutionary approach for load balanced clustering problem for wireless sensor networks. *Swarm Evol. Comput.* **12**, 48–56 (2013)
 16. Gupta, S.K., Jana, P.K.: Energy efficient clustering and routing algorithms for wireless sensor networks: GA based approach. *Wirel. Pers. Commun.* **83**(3), 2403–2423 (2015)
 17. Gupta, S.K., Kuila, P., Jana, P.K.: GAR: An energy efficient GA-based routing for wireless sensor networks. In: ICDCIT, pp. 267–277. Springer, Berlin (2013)

Detection of Online Malicious Behavior: An Overview



D. S. Deshpande, S. P. Deshpande and V. M. Thakare

Abstract Online malicious behavior is performed in a certain kind of pressure, availability of opportunity, and through rationalized way. Web systems are accessed through browser and integrated with database so they usually face many types of vulnerabilities and online threats. The survey is focused on categorization of online malicious behavior on certain web platforms such as education, information technology, finance, and government. The characteristics of malicious behavior are explained. The research purpose is to gather, observe, compare, and study different malicious behavior, detection systems, tools and technologies used, results, and their drawbacks. The numerical observations of malicious behavior are given in order to understand severity of this behavior and its impact. The systems are observed comparatively to point out the challenges. The possible suggestions are mentioned about current requirements in online malicious behavior detection systems. The mind condition behind all malicious behavior is dishonesty and it is contagious by nature. The hybrid detection model is required which will detect malicious behavior in real time, will be flexible enough to configure newly arrived malicious behavior with good accuracy, and will work on multiple domains.

Keywords Malicious behavior · Suspicious behavior · Fraud detection · Intrusion detection model · Online banking · Attacks · Web security · Online vulnerabilities · Online threats

D. S. Deshpande (✉)
SGBAU, Amravati, India
e-mail: ghanashree.rani@gmail.com

S. P. Deshpande
Department of Computer Science & Technology, DCPE, Amravati, India
e-mail: shrinishvadeshpande68@gmail.com

V. M. Thakare
Department of Computer Science & Technology, SGBAU, Amravati, Maharashtra, India
e-mail: vilthakare@yahoo.co.in

1 Introduction

1.1 Overview

Online malicious behavior is an illegal activity performed in order to vandalize the online system with the knowledge of computer technology. The destructive mindset of user behavior does not remain same for the period of time, it is always growing and creates more threat to the web system. These behaviors create more security challenges. The malicious behaviors can occur in different domains such as educational institutes, IT companies, financial services, government entities, Wikipedia, e-commerce applications, and online social networks. Online anonymity increases malicious behavior because it decreases social restrictions and gives facility of many-to-many communication [1]. The existence of malicious users is ubiquitous and they are much faster in their actions as compared to legitimate users [2]. So, it is necessary to study, understand, and detect malicious behavior patterns on all web platforms. There are few motivations to perform this research study. In research, 73% of web users observe online harassment and 40% faced it [3]. The severe behaviors include 10% physical threats, 7% stalking, 6% sexual harassment, and 7% harassment over a sustained period of time. Less severe behaviors include 27% online offensive name-calling and 22% intentional efforts to embarrass someone [4]. The crime complaints are approximately 280,000 per year and with \$1.33 billion loss in year 2016 [5]. India was on second rank and got 2188 total number of online crime reports in year 2016. The kinds of malicious behaviors are credit card fraud, virus, terrorism, extortion, identity theft, gambling, re-shipping, nonpayment, lottery/sweepstakes, scareware, copyright, tech support, data breach, overpayment/nondelivery, threat of violence/harassment, government impersonation, business e-mail compromise, confidence fraud/romance, and criminal forums. Mostly, OSN tool is used to facilitate the crime [5]. Malicious users pretend as benign users by copying their attributes, interest, relations, and connections. It has become challenging to differentiate between malicious and benign users. Existing malicious behaviors encourage other groups for this activity [6]. People think that a little bit of malicious behavior will give them benefit without damaging their good image [7]. Aim of this paper is to study and analyze characteristics of malicious user behavior and to classify them across various web domains.

This paper is organized in seven sections. The relevant work is explained in Sect. 2. Section 3 describes methods, behavior types, used technology, results, and limits of malicious behavior on various web platforms. Section 4 describes categorization of malicious behavior on various web platforms. Research observations are given in Sect. 5. The discussion, challenges, and suggestions are summarized in Sect. 6 and paper is concluded in Sect. 7.

2 Related Work

2.1 Malicious Behavior Detection Systems

Malicious behavior occurred in web applications due to loopholes and lack of testing. Machine learning methods are used for classification of different malicious behavior patterns [8]. In 2011, more than 20% of web users posted malicious comments in Korea. In 2013, a 12-year-old girl committed suicide after getting targeted for cyberbullying in US. In 2013, 28.5% users were targeted for unpleasant online comments in Singapore. It provokes violence in cyberspace [9]. The Finite State Machine (FSM) traces the behavior node and distinguishes between malicious and legitimate behavior by checking element, running order, and cause and effect. The web browser attacks are different due to new functionality of HTML5 [10]. The framework identifies malicious advertisements at the publisher's end. It contains behavioral analysis of advertisement in a secure sandbox environment to detect malicious behavior [11]. Network Security Analyzer (NSA) system detects most of the attacks using signatures [1]. CAPTCHA and MMOGs techniques differentiate between human and malicious behaviors. MMOGs require monitoring inputs of users which includes distributions keystroke durations and the efficiency of mouse movements [12]. Feature selection method has shown increased performance in feature size and classification accuracy. The temporal classification method has shown good performance as compared to other methods. The automatic scanner tool is produced to detect the SQLIAs. It captures various injection attacks that affect database access [13]. The detected malicious activities depend upon the sequence of network events. The results have shown that the cybersecurity knowledge accurately detects malicious behavior and has minimized the false classification of legal behavior as malicious activity. The experts performed consistently better than novices [14]. Network traffic behavior is analyzed and botnet activities are identified through classification of traffic behavior using machine learning [15]. Early Bird learning method is optimized to identify malicious activity in javascript code that extends SVMs learning method. The detection time and accuracy are optimized in learning. The analysis of javascript run time code is effective in malicious behavior identification [16]. The implementation of online teacher answers efficiently and traces about malicious behavior. Defense rules, JS replay mechanism, and data dependency analysis are used instead of using a huge number of training sets in offline learning. JS* performs better in existing commercial solutions to detect malware, their types, and also new attacks [17]. The developed method has removed the SQL query attribute values in web pages at the time of parameters submission and compares it with previous values. The static and dynamic analysis is used together in this method. It also detects stored procedure attacks efficiently [18]. Total 61% of attacks have occurred during the daytime and evening. The statistics of attacks which is performed against web applications in 2017

are represented as below. The aim of half of the total attacks is to access sensitive information. The most common attacks were SQLIA and cross-site scripting attack. These are represented as one-third of total number of attacks [19]. If we categorize the attacks in sectors, then it represents the scenario given in Figs. 1 and 2. In 2013, SQLI was rated as number one attack on Open Web Application Security Project (OWASP). The malicious query remains vulnerable to SQL injection attack in the absence of input purification [20]. The model could defend all kinds of SQL injection attack by comparing strings for equality, if not equal then warning is generated [21]. The SQLIA contains unintended context transition. The number of context transitions occurred at run time allows detection of injection attempts [22]. Anomaly-based system was trained by the normal database access profiles retrieved by using different models. This solution is improvement in previous approach because it minimizes the chances of executing SQLIA based on mimicry. Anomaly score was measured and calculated for each query. If exceeds then the query is considered anomalous by generating an alert [23]. The challenges in predicting human behaviors through ML tools are noisy data, disambiguating the data on “rare” individuals from innocent users, human behavior is dynamically changing [24]. Transaction Risk Score Generation Method (TRSGM) is used for malicious transactions using Bayesian theorem to calculate risk score and spending behavior is identified using K-mean clustering algorithm. Spending behavior, geographical location, and time since last transaction are considered as the parameters. The fraud is detected much faster than existing system [25]. The input of SQLI occurred through user input, cookies, server variables, second-order injection, and physical means of user input [26, 27].

The intention behind this activity is to extract and destroy sensitive information, performs DoS, evade detection, determine database schema, bypass authentication, execute remote commands, and perform privilege escalation [28–31]. On OSN, more negative interactions include more negative links [32].

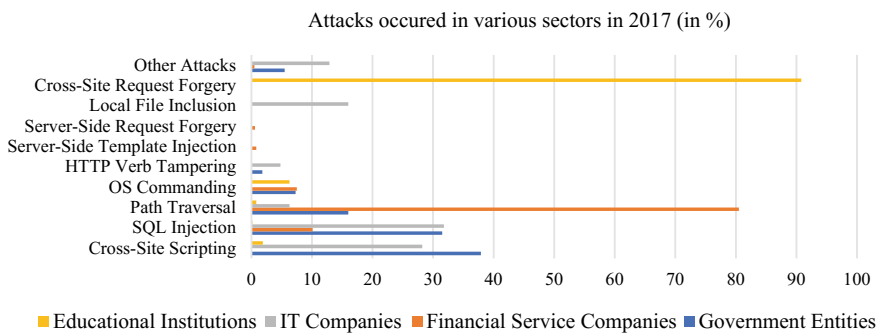


Fig. 1 Attack occurred in various sectors in 2017

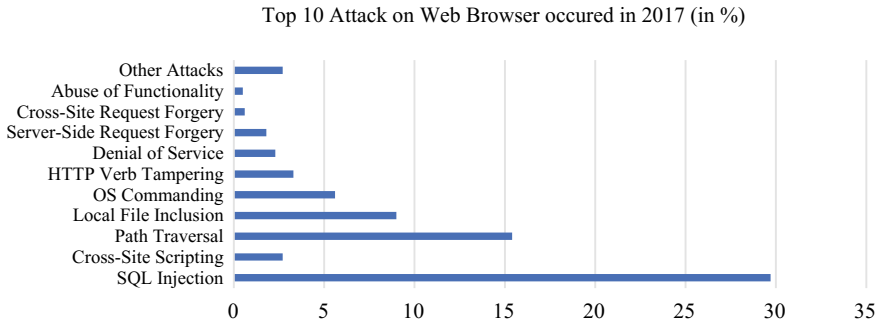


Fig. 2 Top 10 attacks on web applications recorded in 2017

3 Classification of Malicious Behavior Detection Systems

See Table 1.

Table 1 Malicious behavior detection system

Sr.	Name of method and algorithm	Type of user behavior	Tools/technology used	Result	Limits
1	FSM-script-based attack detection method. It works on behavior of web browser [10]	JS API, DOM, and EVENT	XML, HTML5, Webkit to implement virtual browser	90% malicious samples are detected except errors	It is hard for virtual browser to run additional web elements
2	Sandbox environment is used to check dynamic properties of the advertisements, SVM classifier [11]	Registry changes, network traffic trace, memory dumps of malware process, files created, deleted, and downloaded by the advertisements. during its execution	Cuckoo sandbox environment to visit the pages using browsers (cuckoo open-source malware analysis framework)	Suspicious redirects—69%, registry changes—3%, memory dumps—4%, file changes—6%, drive by download—18%. Very low false positives and false negatives to detect mal-advertisements	Various more features should be extracted from advertisement and can compare using more efficient supervised classifier

(continued)

Table 1 (continued)

Sr.	Name of method and algorithm	Type of user behavior	Tools/technology used	Result	Limits
3	Cognitive modeling and natural interaction [12]	Bots carried out scripted actions designed to look like human behavior	CAPTHA, MMOGs, HIP, HOP. HSP	Typing speed is correlated with the familiarity of the text being typed, which can be used	Distracted users make mistakes compared to focused users. Experts are more familiar
4	Web intrusion detection system, hybrid feature relevance algorithm, temporal classification method [13]	SQLIA executes on database system and runs harmful code that reveals confidential information	Fuzzy cognitive map technique, automatic scanner tool	Accuracy > 0.5	Many issues in outdated web IDS includes continuous updating, low detection capability to unknown attacks
5	Intrusion Detection System (IDS), linear regression model [14]	Threat in a network, cyber attacks, vulnerabilities in httpd and ftpd services, sniffer (sniff out passwords), DoS (increases traffic load within the network)	IDS tool	7 network events are malicious out of 20. Experts detected—67% attack scenarios, 12% false detections, novices detected—68, 14% false detections, cyber prof. detected—55, 15% false detection, novices detected—44, 18% false detection	Cybersecurity analysts should take continuous efforts to acquire latest knowledge needed to successfully defend a network, IDS can generate false alerts

(continued)

Table 1 (continued)

Sr.	Name of method and algorithm	Type of user behavior	Tools/technology used	Result	Limits
6	Detection framework is decision tree using Reduced Error Pruning algorithm (REPTree), 10-fold cross-validation technique [15]	Botnet	Java, Weka machine learning framework, TcpReplay tool	TPR > 90%, FPR < 5%, identifies new threats without training of existing malicious datasets and testing sample novel botnets. It identifies old and new botnet activities with high accuracy and with very small time windows	Unacceptable performance impacts if the detectors are distributed over individual devices on the network
7	Early Bird method extends SVM algorithm [16]	Malicious Javascript behavior	Cujo detector	Attacks detected—93.2%, false alarms <5	Only SVM is user in Early Bird method
8	Deterministic Finite Automaton (DFA), JS* learning framework with online learning algorithm L* [17]	Malicious Javascript behavior	Code clone detection tool CloneMiner, JS* compared with JSAND, VIRUSTOTAL, etc.	JS* accurately detects 149 from 156, malicious samples detected—95.51%, 9957 from 10,000, benign samples detected—99.57%	JS* is evaluated only using 8 popular types of attacks. It should be evaluated on more general attacks
9	Static and dynamic novel method [18]	Malicious SQL query	Attack tool—Paros 3.2.13	Efficiently detect stored procedure type of attacks also	All other attacks should be covered
10	Learning-based intrusion detection system [23]	Malicious SQL query	Libmysqlclient library, C++, Yacc-based parser, LibAnomaly	Novel attacks are detected with few false positives and little overhead	More complex database features such as stored procedure should be used
11	Community detection algorithm [33]	Sybil attack [33, 34]	–	Identifying a local community that surrounds the trusted node	Ranking of nodes used to reach local commu. is not strongly correlated

(continued)

Table 1 (continued)

Sr.	Name of method and algorithm	Type of user behavior	Tools/technology used	Result	Limits
12	SVM, DT-J48, and PART information gain feature selection method [8]	Scanning and attacking	High-interaction honeypot, PHPMySQL, IIS, SSH servers	J48 detects 75% new and unseen attacks and is best performing machine learning method here	SVM requires more time for execution
13	TRSGM, K-mean, and Bayesian algorithm [25]	Sudden change in location and time	–	Model adapts changing behavior	Very less parameters are used

4 Classification of Malicious Behavior on Various Web Platforms

4.1 Online Education System

Online distance courses became vulnerable to student aid fraud. Their identity can be easily replaced and falsified. The schools and fraud ring colleges offer online programs in very low price. They obtain federal financial aid by enrolling straw students. Then, the leaders in the ring acquire share of the students' loan amount which school needs to disburse to students after allowable costs are paid. Fraud rings get success and benefit although they are not eligible for this financial aid. Within last 6 years, members of 42 different fraud rings have been punished. More than \$7.5 million fines have been collected [35]. In 2001, the Shah brothers hijacked 8 million e-mail addresses of students in 2000 colleges in U.S. and used those for marketing spam campaign and earned more than \$4.1 million. In 2006, Halton got grant of \$538,932 from distance learning school in illegal way by showing fake diploma documents. In 2005, Dixie and Randock sold 1000s of fake online degrees from many real universities to students from 131 countries and earned \$6 million in profits and prosecuted in 2008. In 2007–2009, 23 separate financial aid applications had been filed to Webster University using the names of other inmates and earned \$467,500. In 2005, Trump University made lofty promises to receive industry secrets from Donald Trump and students paid over \$60,000 [36]. The plagiarism is detected at Nigerian institutions of higher education in doctoral research. In India, few of the higher education deemed universities violated existing rules and regulations of UGC and their distance learning programs are not granted by UGC [37]. American Public University System (APUS) has found multiple students with same e-mail address, IP address, and they have changed their address prior to disbursement [38].

4.2 IT Sectors

The fraud risk in every IT organization depends upon factors such as industry size, structure, geographical location, internal control environment, etc. IT works in few layers. In these layers, increased level of risk is to database layer in the form of online fraud followed by application risk, host, and then infrastructure. Deficiency in application layer has direct impact on database [39]. The knowledge of information technology is found as the cause of malicious activities and also solution to these activities. Cybercriminals combine their IT skill with social engineering and access confidential company information and personal users' data [40]. In Gurgaon, an employee of BPO sold a CD which contains confidential data of British bank customer. One employee in Hyderabad-based company could withdraw salary even after resigning the company. In India, one employee sold the source code of developed software to the competitors. It has found that senior or middle management employees who have responsibility and trust of everyone are more likely to commit fraud [41]. Malicious behaviors in organization are increased due to speed and refinement provided by technology. In between 50% and 90% of companies waste their finance in malicious activities [42].

4.3 Financial Sectors

As soon as bank implements any technology to become secure from online fraud, cybercriminals find loopholes to vandalize. The Account Takeover (ATO) attack is the theft of credentials to perform illegal activity, and it may include identity theft, ACH, and wire fraud. Malwares through phishing attacks capture customer's keystrokes including bank login credentials. They take benefit of user's same login credentials at more than one websites. A case involving in the loss of \$900,000 due to Denial of Service (DoS) attack [43]. Online banks face risks by implementing new technology. The risks are phishing, vishing, spyware, payment card fraud, terrorists financing, cyberstalking, fake accounts, identity theft, viruses, adware, website cloning, and card skimming. In 2013, India has got third rank in black money outflows and got fifth rank in cybercrime incidents, i.e., identity theft, phishing attacks, and ransomware, total 4356 number of cases of cybercrimes registered. 65% of total frauds were committed through internet banking [44]. Since 2011, 497 crore cyber fraud cases are reported (Figs. 3 and 4).

Credit-related frauds have the maximum impact in all banking frauds in India [45]. Online payments, trading in stock markets, and procurement of materials areas were identified as vulnerable. E-commerce payment fraud is on rise. In India, the most banking frauds occurred in the retail banking sector [46]. In 2016, 60% of people used online banking within 3 months compared to 35% in 2008 [47]. Most banking frauds occur within the retail banking sector [48]. Malicious users attack on banks repeatedly for a period of time and then reacts to deter the fraudsters [48].

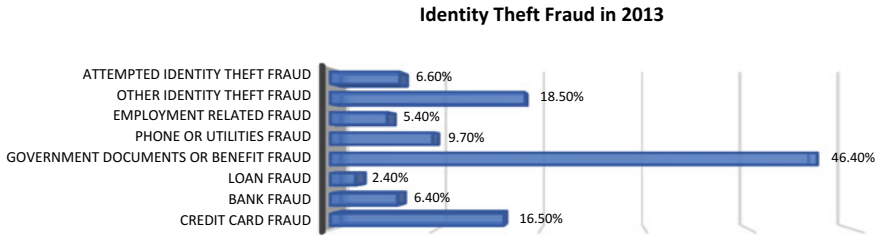


Fig. 3 Identity theft frauds in 2013 [44]

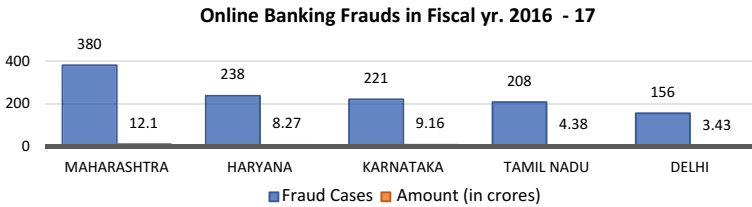


Fig. 4 Online banking frauds in fiscal year 2016–17 [49]

In December 2017, RBI reported 10,220 cases of frauds related to ATM/credit/debit cards and net banking. The total amount involved was 111.85 crores and over 25,800 online banking fraud cases were reported. The Maharashtra state topped in online banking fraud cases in fiscal year 2016–17 with 380 cases and amount was 12.10 crores [49]. Brazil has got maximum number of banking malware attacks followed by Russia and Germany. Younger people reported more loss of money in online [50]. Dating and romance website scams acquire gifts, personal details, and money through an emotional trigger [51].

4.4 Government Sectors

Bogus government websites cheat candidates by offering jobs. In October 2014, Uttar Pradesh police unveiled railway recruitment board scam. They have cheated around 10,000 candidates by charging them registration and processing fees. A person from Kolkata arrested in the case of cheating people through fake website name as www.pmay-gov.in of the Pradhan Mantri Adarsh Gram Yojana (a village development scheme). In 2006, government implemented the Common Services Centers scheme in order to promote e-governance. Then, in 2015, this scheme was integral part of PM Narendra Modi’s Digital India campaign. The fake website of the same content was created [52].

5 Observations

Many of the attacks are common on these all web platforms. Money and confidential information are the primary motives of these attacks, and only risk varies if platforms differ. SQLIA behavior is involved primarily followed by path traversal, DoS, cross-site scripting, phishing, etc. Banking sectors have faced mainly path traversal behavior of malicious users. Cross-site request forgery attack has occurred mainly in education sectors. Government sectors have faced cross-site scripting attack. Banking sectors have faced maximum malicious behavior and lost money in huge form. Maharashtra was in top rank in the cases of malicious behavior in banking sector in 2017.

6 Discussion, Challenges, and Suggestions

Suspicious behavior cannot be entirely concluded as malicious. After failure of three login attempts, the activity is called as suspicious but cannot be concluded as malicious attempt because legitimate user can forget the password. Illegal users perform some suspicious attempts to find out weaknesses in web application if any. These attempts may be more than one. The suspicious activity observation is necessary before actual fraud activity implementation takes place. This comes under preventive task. These observations can be for weak and insecure websites. Prevention of malicious activity is less expensive and more effective than detection. Identification of suspicious behavior is helpful for this task. Most of the research works mentioned here are performed to find out intrusion detection systems to detect malicious activity. However, there is a necessity of mechanism which will mitigate the risk and detect fraud at very early stage of an incident. Along with monetary loss, businesses also get loss of credibility and reputation which results into ban from doing business. Finance sectors have major risk and maximum possible attempts of attacks performed.

There are few challenges that are needed to be resolved to detect malicious behavior. It is difficult to trace frequently changing behavior of online user. Malicious activity does not have any consistent behavior patterns. Many of the detection systems detect legitimate user as malicious user by giving false alarms. Organizations are not getting favorable results in detection of fraud and mitigation of risk because continuous fraud monitoring is challenging. They have lack of skilled resources to manage data analytics tools on daily basis. The fraudsters are innovating new frauds faster than finding solutions by organizations. There are huge numbers of transactions taking place each day. The more efficient technology is required to acquire huge form of data. The data is noisy. Fraud is detected after it has already been executed and it worsens the resources.

Malicious behavior is playing long-term battle with the banks. Time is the essential thing in this whole detection process. Hybrid detection model can be developed to prohibit newly arriving malicious behavior which generates threat. The system must

be flexible to identify malicious behavior in real time and categorizes it into high-, medium, and low-risk users. Website administrator will have facility to immediately configure these newly arriving malicious behaviors' scenario and can be able to update the existing behavioral patterns. Few parameters can be used in combination to separate malicious user from legitimate user such as IP address, browser, operating system, time, location, accessed links, and frauds they are associated with. Particular IP address may involve into emerging vulnerabilities. New potential areas of fraud can be detected by observing current behavior and analysis of prediction. Previous and historical behavioral patterns of users from data can also be compared with current behavior in order to identify suspicious patterns. Web administrator can get the information about high-risk or low-risk users in real time and he can block those suspicious users. Domain-specific suspicious behavior detection systems are available but a system is required to work on all the abovementioned web platforms.

7 Conclusion

This study, in this paper, presents an overview of online malicious behavior patterns in various web platforms such as education, IT, finance, and government sectors. The malicious behavior detection systems have been classified within the few factors such as the methods they have used, type of malicious behavior they have performed, used tools, results, and limits of the system. The status of current malicious activities has been explained. The most dominating attacks, their severity, and impacts on society are going toward vandalization. There is a necessity to mitigate the identified risk and overcome challenges. Real-time suspicious user behavior hybrid early detection model is required which will be flexible enough to acquire and configure newly arising all types of malicious behavior with good accuracy.

References

1. Nilakshi, J., Shwetambari, P., Dhananjay K.: Network security analyzer: detection and prevention of web attacks. Springer International Publishing Switzerland. In: Satapathy, S.C., Das, S. (eds.) Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems, vol. 1. Smart Innovation, Systems and Technologies, vol. 50, pp. 497–505 (2016)
2. Srijan, K.: Characterization and Detection of Malicious Behavior on the Web. Ph.D. Thesis, pp. 1–225 (2017)
3. Maeve, D.: Online Harassment. Pew Research Center, pp. 1–65 (2014)
4. Maeve, D.: Online Harassment. Pew Research Center (2017)
5. FBI Internet Crime Complaint Center (IC3) annual reports. <https://www.ic3.gov/media/annualreports.aspx> (2017)
6. Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., Leskovec, J.: Anyone can become a troll: causes of trolling behavior in online discussions. In: Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing (2017)

7. Mazar, N., Amir, O., Ariely, D.: The dishonesty of honest people: a theory of self-concept maintenance. *J. Mark. Res.* 1–48 (2008)
8. Katerina, G., Goce, A., Ana, D., Risto, P., Brandon, M.: Characterization and classification of malicious web traffic. *Comput. Secur.* **42**, 92–115 (2014)
9. Lee, S., Kim, H.: Why people post benevolent and malicious comments online. *Commun. ACM* **58**, 74–79 (2015)
10. Soojin, Y., Hyun-lock, C., Hanchul, B., Hwankuk, K.: Behavior-based detection for malicious script-based attack. In: Park, J., et al. (eds.), *Advances in Computer Science and Ubiquitous Computing*, Lecture Notes in Electrical Engineering, pp. 97–103. Springer Nature Singapore (2017)
11. Poornachandran, P., Balagopal, N., Pal, S., Ashok, A., Sankar, P., Krishnan, M.R.: Demalvertising: a kernel approach for detecting malwares in advertising networks. In: Mandal, J.K., et al. (eds.) *Proceedings of the First International Conference on Intelligent Computing and Communication*, *Advances in Intelligent Systems and Computing*, vol. 458, pp. 215–224. Springer Science+Business Media Singapore (2017)
12. Amant, R.S., Robert, D.L.: Natural interaction for bot detection. *Natural Web Interfaces. IEEE Internet Comput.* **20**, 69–73 (2016)
13. Maheswari, K.G., Anita, R.: An intelligent detection system for SQL attacks on web IDS in a real-time application. In: *Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC—16’)*, Smart Innovation, Systems and Technologies, vol. 49, pp. 93–99. Springer International Publishing Switzerland (2016)
14. Noam, A., Cleotilde, G.: Effects of cyber security knowledge on attack detection. *Comput. Hum. Behav.* **48**, 51–61 (2015)
15. David, Z., Issa, T., Bassam, S., Wei, L., Sherif, S., Ali, G., Dan, G.: Botnet detection based on traffic behavior analysis and flow intervals. *Comput. Secur.* **39**, 2–16 (2013)
16. Kristof, S., Marius, K., Alexander, B., Rieck, K.: Early detection of malicious behavior in JavaScript code. In: *AISeC’12*, pp. 15–24. ACM (2012)
17. Yinxiang, X., Junjie, W., Yang, L., Hao, X., Jun, S., Mahinthan C., Detection and classification of malicious JavaScript via attack behavior modelling. In: *ISSTA’15*, pp. 48–59. ACM (2015)
18. Inyong, L., Soonki, J., Sangsoo, Y., Jongsub, M.: A novel method for SQL injection attack detection based on removing SQL query attribute values. *Math. Comput. Model.* **55**, 58–68 (2012)
19. *Web Application Attack Statistics*, pp. 1–11, Q1 (2017)
20. Thiyab, R., Ali, M., Basil, F.: The impact of SQLIAs on the security of databases. In: *Proceedings of 6th ICOCI*, no. 80, pp. 323–331 (2017)
21. Ashish, D., Sanjay, J.: Neutralizing SQL injection attack using server side code modification in web applications. *Hindawi, Secur. Commun. Netw. Research Article* 1–12 (2017)
22. Victor, P., Kim, C., Helen, A.: Context-oriented web application protection model. *Appl. Math. Comput.* **285**, 59–78 (2016)
23. Valeur, F., Mutz, D., Vigna, G.: A learning-based approach to the detection of SQL attacks
24. Subrahmanian, V.S., Kumar, S.: Predicting human behavior: the next frontiers. *Science* **355**(6324), 489 (2017)
25. Singh, P., Singh, M.: Fraud detection by monitoring customer behavior and activities. *IJCA* **111**(11), 23–32 (2015)
26. William, H., Jeremy, V., Alessandro, O.: A classification of SQL injection attacks and countermeasures, pp 1–11. IEEE (2006)
27. Roshni, C., Manoj, M., Santhi, T., Dipankar, S.: SQL injection attack mechanisms and prevention techniques. In: *ADCONS*, pp. 524–533. Springer, Berlin (2012)
28. Jai, S.: Analysis of SQL injection detection techniques. *Theor. Appl. Inform.* **28**(1 & 2), 37–55 (2016)
29. Atefeh, T., Maslin, M., Mohammad, H., Suhaimi, I.: SQL injection detection and prevention tools assessment, pp. 518–522. IEEE (2010)
30. *SQL Injection Attacks: Detection in a Web Application Environment*, DB Networks, www.dbnetworks.com, pp. 1–13 (2016)

31. Manisha, B., Vanita, M.: Protection of web application against SQL injection attack. *Int. J. Sci. Res. Publ.* **3**(10), 1–5 (2013)
32. Liu, Huan, Morstatter, Fred, Tang, Jiliang, Zafarani, Reza: The good, the bad, and the ugly: uncovering novel research opportunities in social media mining. *Int. J. Data Sci. Anal.* **1**(3–4), 137–143 (2016)
33. Viswanath, B., et al.: An analysis of social network-based sybil defenses. In: *Proceedings of SIGCOMM*, pp. 363–374. ACM (2010)
34. Long, J., Yang, C., Tianyi, W., Pan, H., Athanasios, V.: Understanding user behavior in online social networks: a survey. *IEEE Commun. Mag.* **51**, 144–150 (2013)
35. <https://www.nytimes.com/2011/10/06/opinion/fraud-and-online-learning.html>
36. <https://www.onlinecoursereport.com/education-scams/>
37. <https://acei-global.blog/2015/06/25/9-recent-episodes-of-cheating-fraud-and-scams-in-education-from-around-the-globe/>
38. David, B.: *Fraud and Distance Education*, Session 60
39. *Fraud and Role of Information Technology* (2008)
40. The impacts of fraud on Information Technology. Essay, <https://www.ukessays.com/essays/information-technology/the-impacts-of-fraud-on-information-technology-information-technology-essay.php> (2015)
41. Ever increasing fraud risks in the IT and ITeS sector. *Fraud Investigation and Dispute Services*, ERNST & young, pp. 1–12
42. da Cunha, J.V.A., Cornachione, E.B.: Frauds and information technology: analysis of the influence on accounting and company systems. In: *IFIP Springer Book Series*, vol. 105, pp. 179–193 (2003)
43. *Fighting Online Fraud: An Industry Perspective*, vol. 3, pp. 1–8. ACI Universal Payments
44. Current fraud trends in the financial sector. ASSOCHAM India, pp. 1–28 (2015)
45. Singh, C., Pattanayak, D., et al.: Frauds in Indian banking industry. IIMB-WP NO. 505, pp. 1–24 (2016)
46. Deloitte, *Indian Fraud Survey* (2014)
47. “Online Fraud”, National Audit Office, pp. 1–50 (2017)
48. “Online fraud: increased threats in a real-time world”, SAS, pp. 1–4
49. <https://timesofindia.indiatimes.com/business/india-business/over-25800-online-banking-fraud-cases-reported-in-2017-govt/articleshow/62296962.cms>
50. <https://www.consumer.ftc.gov/blog/2018/03/top-frauds-2017>
51. <https://www.acorn.gov.au/learn-about-cybercrime/online-scams-or-fraud>
52. “As fake website scams abound, a government registry for Indian internet domains may be in the works”, Scroll.in, Article, Nov. 2017

Reliable Data Delivery with Extended IPV4 Using Low-Power Personal Area Network



Shambhavi Mishra, Pawan Singh and Anil Kumar Tiwari

Abstract In most of the industries and research and developments, IoT is rapidly increasing technology because of its efficiency feature in the M2M communication. The industries and R&D tell that it is most cost-effective, efficient, optimization in communication, and till 2025 it will optimize every device. IoT is mainly the interconnection of embedded devices that can connect the physical world to virtual work with objects and things. It helps to access the data or information remotely at anytime, anywhere, etc. The IoT technology has a technical concept for communicating between M2M such as smart home, remote security control, remote devices monitoring, smart healthcare system, etc. There are various enabling technologies that help to communicate between the embedded devices through protocols such as RFID, addressing scheme, WSN, etc. Approximately, 4.3 billions of IP addresses (232) have been provided by IPv4. There is topical development of the Internet and its objects forced to find best solution for IPv4 extension. As a result, IPv6 is gaining lots of preference in research and it is expected that IPV6 will provide optimized performance by 2025. This paper proposes a current solution in the existing IP requirement with a new IP format which is compatible between IPV4 and IPV6. The proposed method is developed and analyzed using Contiki operating system platform with a Cooja simulator.

Keywords Internet of Things (IoT) · Enabling technologies · Cooja · IPv4 · IPv6

S. Mishra (✉) · P. Singh · A. K. Tiwari
Department of Computer Science and Engineering, Amity University, Noida, Uttar Pradesh, India
e-mail: shambhavimishra1000@gmail.com

P. Singh
e-mail: pawansingh51279@gmail.com

A. K. Tiwari
e-mail: aniltiwari19640@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_3

1 Introduction

“The internet of things explains as dynamic global network infrastructure which is self-configuring capabilities that depend upon standard and interoperable communication protocols, intelligent interfaces uses physical attributes, virtual personality, and are impeccably included into the information network where it is physical and virtual.” In 1999, the Internet of Things was first developed by Kevin Ashton, MIT, and Auto Labs in the field of supply chain management. It is mainly the interconnection of embedded devices that can connect the physical world to virtual work with objects and things, by Cisco in 2020 there will be over 50 billion connected objects against a population of 7 billion [1]. The main feature of IoT is the unique addressing methods that help to secure the message while communicating. In every communication, the main problem faced is security. Many different types of attacks such as application layer attack, data notification, and sniffer attack make an interruption in the communication between the various devices (M2M). The IPv6 availability and low-power gadgets are contained by IoT. The main aim of IoT may be to form a smart city, provide smart health care, offer smart transportation, etc. The modern setup of IoT in the WSN with a concept idea of things and object around us such as mobile phones, RFID tags, actuators, sensors, etc. that have a unique addressing schemes to reach their common goals to neighbors and also that are able to interact with each other [2] (Fig. 1).

The Contiki OS specially designed for memory-constrained environments such as wireless sensor network (WSN) uses sensor nodes to gather information and enables the sensor platform. A basic version always supports the emblematic sensor node platforms that are most important because users can add new devices easily. This is also for an event-driven kernel build up and supported full TCP/IP stack via the uIP library as well as the programming abstraction Protothreads. Contiki OS released in 2003, which can port more than 20 different platforms and implemented in the C language, can be easily portable for new platforms. The Contiki OS is

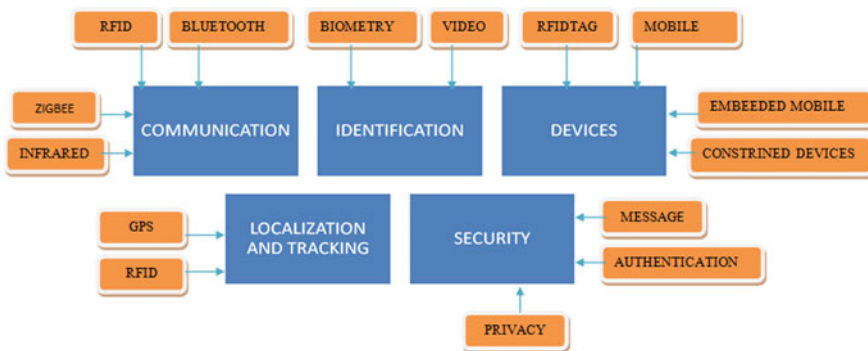


Fig. 1 IoT connectivity

Fig. 2 Contiki interface

User App			Built-In App
Uipv6			Contiki Operating System
Socket API			
UDP	TCP	ICMP	
IPV6 LoWPAN			
MAC			
Plate form			CPU
Hardware Drivers			

based on the Java virtual interface (JVI). The main design goal is extensibility which diverges greatly between the users and development phases. Main benefit of using the simulator would be larger scale testing for the various protocols while communicating the messages or information such as communication protocol which behaves as development and testing phases during coding and it is flexible, scalable, and usable simulator that fits into the surrounding development environment (Fig. 2).

2 Enabling Technologies

We have provided a complete survey of each technology. IoT is a concept where actuators or sensors that connect into the real world to the virtual world with a possible integrate to the several enabling technologies (Fig. 3).

2.1 Radio Frequency Identification (RFID)

The design of microchips for wireless data communications can be enabled by RFID technology. The unique identifier and electronic product code (EPC) have an automatic identification of memory units. A particular tag is recognized universally by electronic product code (EPC) function. RFID tags classification [3] (Fig. 4):

- **Active tag:**

Active RFID tags are the self-motorized and instantiate the announcement with battery supply. It provides the unique electronic product code interface remotely with its neighbor in a limited distance. The main applications for active RFID tags are remote monitoring and auto manufacturing.



Fig. 3 Technology associated with IoT

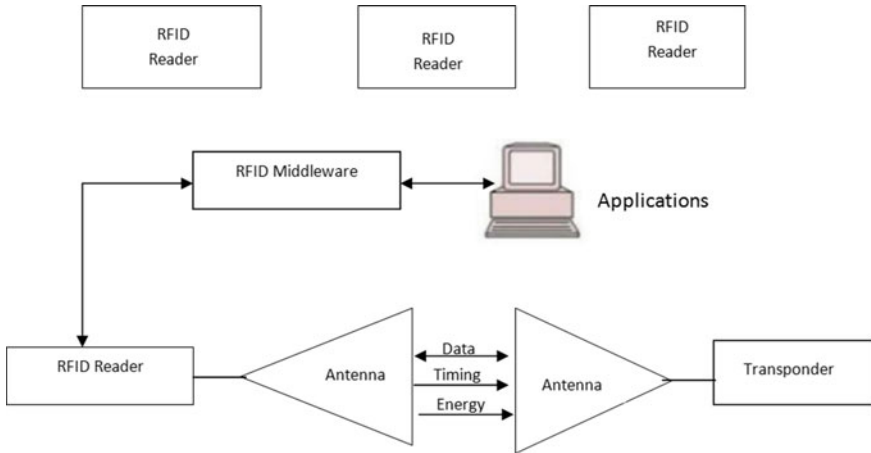


Fig. 4 RFID component

• **Passive tag:**

This is not a battery power driven tag, the power is only driven for reader's to examine the signals that has been communicated to the IDs so it pass on RFID reader. As a source of energy through inductive coupling, the tag readers are emitted with the utilization of electromagnetic signals which are substituted by the passive tag where the internal battery is less [4].

2.2 Wireless Sensor Network

Low energy consumption in remote sensing area, higher efficiency applications, nanotechnology, and tiny devices in WSN are the recent progresses to ensure that are available at lower costs, wireless low-power integrated circuits, and communications devices. The wireless communication takes more bandwidth with limited frequency [5]. It communicates with several parts as follows:

- Memory,
- Sensor,
- Radio transceiver,
- Microcontroller, and
- Battery (Fig. 5).

During implementation, the large amount of intelligent sensor node is deployed in a network to make it feasible for utilizing sensor network. An active RFID node limited processing capability and storage and it is nearly to same as the lower end of WSN. WSNs are the most multidisciplinary and substantial in nature for the scientific challenges and it must give the solution in order to comprehend the enormous potential [6]. WSN includes some components that monitor the networks which are given below:

- Wireless sensor network hardwares,
- Wireless sensor networks communication stack,
- Wireless sensor networks middleware, and
- Data aggregation security.

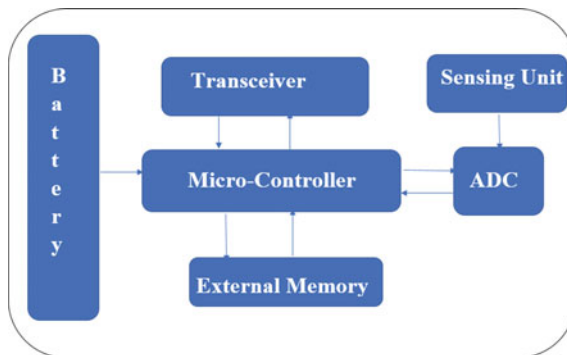


Fig. 5 WSN architecture, distributed node, gateway, and component

2.3 Addressing Scheme

The addressing schemes uniquely identify the objects deployed across the network. The unique identification is important to provide to all objects because it is a success of IoT. The IPv4 can identify uniquely address and group of sensor devices that are distributed in nature. Numerous issues are faced by the IPv4 such as mobility, which is easy to resolve by implementing the paradigm of IPv6. The conjugate sensor devices are supported by IPv4 with group of unique identification addresses. IPv6 provides the remote access to devices along with their unique identification and is capable of providing billions of devices with unique addresses. The unique identification of household goods can be assisted by IPv6 addressing scheme with a notable progress of lightweight foundations. This can also make the data confluence and exacerbates from the devices make the advance problem [7].

2.4 Data Storage and Analytics

IoT produces huge volumes of data and the outcome for this emerging technology is to create an extraordinary quantity of data fields. To ensure the efficiency and the reliability of energy they must be centralized and run on the harvested energy of data centers. The smart and intelligent way to monitor the stored data is with the help of artificial algorithm that may be centralized or distributed based on the needs of the data. Novel fusion algorithm develops and gathers the data from the evolutionary algorithm, genetic algorithm, artificial intelligence, and neural network. The storage and analytics are required in the middleware layer with the numerous challenges in it that is supported by the centralized infrastructure. Cloud-based solution of storage is becoming popular in the advanced year of visualization and analytics of platforms as per 2012.

2.5 Visualization

Visualization is a momentous aspect of an IoT application that allows interacting from the user with given environment. In the visualization screens, the movement of screen is from two dimension to three dimension and meaningful information can be provided in many ways for the users. The data can be converted into knowledge and fast decision-making by enabling this policy.

3 Related Work

3.1 Addressing in IoT

The most important issue in the IoT architecture is mapping and addressing while communicating machine-to-machine (M2M) nodes, that creates a standard policies for mapping and addressing technique in the IoT. In future, an increase in number of connected devices may not be applicable but presently IPv4 protocol identifies the nodes by 4-byte address. 6LoWPAN for low-power personal area network is proposed with a concept of IPv6 protocol. Mapping and addressing of IoT with various header formats for TCP/IP protocols are summarized and only few researches have been carried out [8, 9].

3.2 Operating System for IoT

IoT interconnects a wide range of sensor devices, micro-electronic devices that are equipped with a micro-controller, and external devices that can be accessed through Internet connection. This section briefs out some of the operating systems that are compatible with wide range of devices in IoT environment [10]. These wide ranges of devices need the operating system to establish M2M communication standards.

- **Contiki OS** [11]: It is an OS which is planned for simulating wireless sensor network protocols that enabled sensor platform. This simulator is equipped with a module of various sensor motes like Sky motes, Trexb motes, MicaZ motes, and Eth motes with its basic operation-based features. With this type of ranges of support including network feature that proposed addressing scheme, it is simulated using Contiki OS. In a Contiki OS, we can design and implement the sensors which carry the informations from one to another sensors. Contiki OS comes under a JVI (JAVA Virtual Interface).
- **mbed Operating System** [10]: It is designed for running over a resource constraints device with limited processing, storage, and memory capacity. This operating system is supported by a C++ framework for developing IoT applications. Security monitoring and management are also many options for device level.
- **TinyOS** [12]: This is widely used for wireless sensor networks with many features like network management and ubiquitous computing. An open-source noncommercial operating special system designed for running on resource compatible devices and constraints for IoT environment.

4 Experiment Result and Setup

4.1 Recommendation

In Fig. 6, we have proposed the IP structure where it is 40 bit addressing and we have added the 8 bit more in the IPv4 structure. In order to increase the IP number, we have increased the 8 bit in the global addressing network. Due to 32-bit address space we can only assign 4.3 billion of addresses in IPv4. The proposed model is compatible with IPv4 that will increase the bits in the same structure to make more global addresses assigning. In the upcoming solution with sufficient requirement of IP addresses in IPv4 structure, we have increased the bits and approximately it will increase the 1094 billion of global IP addresses.

4.2 Experimental Setup

In simulation, Contiki operating system is used for IP address proposed model. The Contiki OS work depends upon Cooja simulator and is built to configure and simulate the IoT algorithms. It is a very popular OS used for simulation as well as Cooja tool is also used for simulations. The Cooja platform is for creating more types of motes like sky, Cooja, Z1 mote, etc. which can be used for the simulation of the Internet algorithms on them. Contiki OS enables to simulate the low power memory devices.

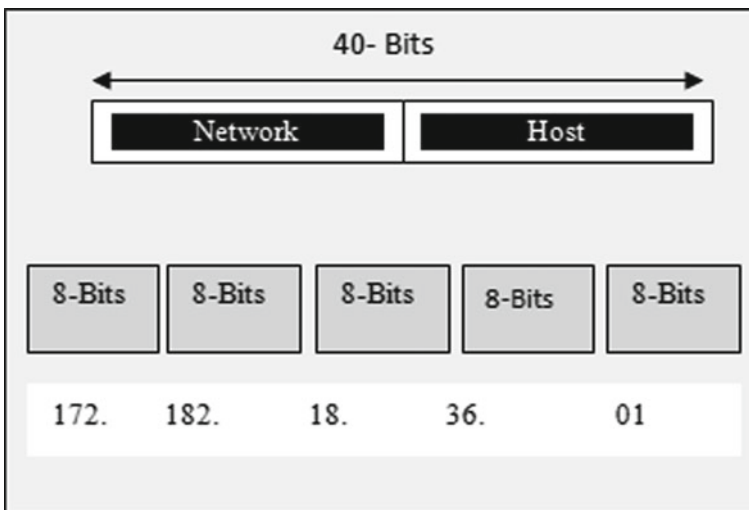


Fig. 6 Proposed IP structure

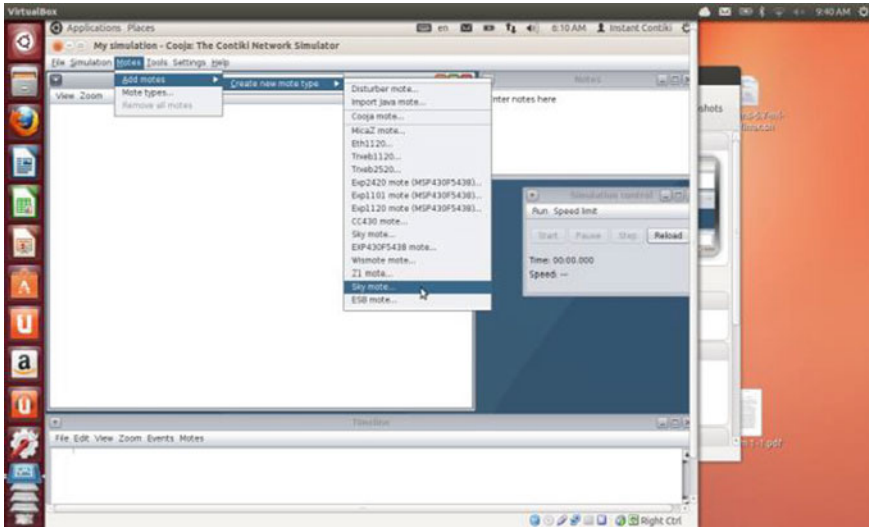


Fig. 7 Initialization modes

The process of new simulation has started now named as my simulation, and the motes will be added for the simulation. The Cooja provides the different types of motes like Sky motes, Trexb motes, MicaZ motes, Z1 mote, Cooja mote, and Eth motes with its basic operation-based features. For the project, the sky motes have been configured, and it is compatible with the designed algorithm for the IPs. For the selection of the motes, click on the motes and then add the motes and select the sky motes.

Figure 7 shows the addition of the motes to the simulator sky platform for the assignment of the IPs in this process and the project is programmed for the random generation of the motes. We have to select the number of the motes to be simulated and the number of motes to be added for the simulation. After the addition of the motes, select the program for the compilation, and after the compilation process the generated motes are assigned with the designed IP address and the out of the generated IP address are shown on the GUI and the same with features of the motes position, IP address, and the mote Id are stored in the text document (Fig. 8).

5 Application of IoT

See Table 1.

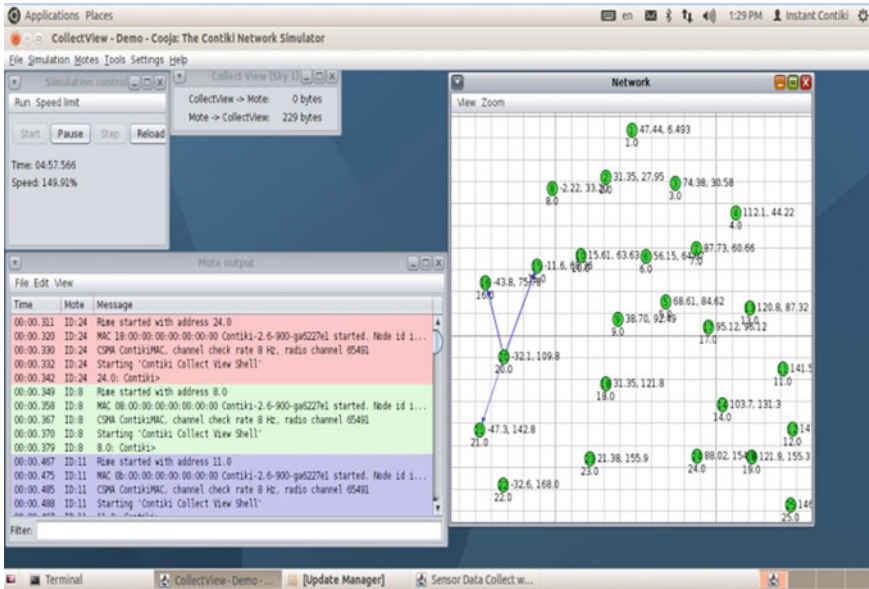


Fig. 8 Simulation

Table 1 Applications of Internet of Things

Application	Description
Human	Devices are ingestible and also it is wearable to monitor and maintain human health and wellness, and productivity, fitness, and disease management should be increased
E-health	Real-time patient health status should be monitored with personnel tracking, predictive expertise information, auctioneers’ and patient monitoring, doctor tracking
Home	Security systems and home controllers
Smart environment	Industrial plants and gym comfortable homes/offices
Retail environment	Stores, inventory optimization, banks, restaurants, arenas, self-checkout, in-store offers
Energy conservation	Smart grid and smart devices
Offices	Improved productivity and security into office buildings, and energy management
Environmental monitoring	Air pollution, waterways, industry monitoring, noise monitoring

6 Conclusion

In the existing network architecture, there is a solution to IP extinct with a new way of addressing objects and mapping objects in IoT. This would be carried out to be a research in future as proposed, whereas in the networking of IoT the scalability is also an issue. Simulated proposed work may be a solution of IPv4 extinct addressing and it designs a compatible IPv4 model that enlarge the addressing structure with 1096 billion of IP addresses in it. This can reduce the complexity of global IP addresses configuring assignment.

References

1. Evans, D.: Internet of Things. Cisco, https://www.cisco.com/.../IoT_IBSG_0411FINAL.pdf
2. Giusto, D., Iera, A., Morabito, G., Atzori, L. (eds): The Internet of Things. Springer J (2010)
3. Shen, G., Liu, B.: The visions, technologies, applications and security issues of Internet of Things. In: International Conference IEEE (2011)
4. Prodanoff, Z.G.: Optimal frame size analysis for framed Slotted ALOHA based RFID networks. <https://doi.org/10.1016/j.comcom.2009.11.007>
5. Akyildiz, I.F, Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE e-Commun. Mag. (2002) (Atlanta in Georgia Institute of Technology, USA)
6. Akyildiz, I.F, Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey: wireless sensor networks. Comput. Netw. (2002)
7. Zorzi, M., Gluhak, A., Lange, S., Bassi, A.: Future Internet of Things: a wireless- and mobility-related view. IEEE Wirel. Commun. **17** (2010)
8. Xu, B., Liu, Y., He, X., Tao, Y.: On the architecture and address mapping mechanism of IoT. In: The International Conference of IEEE at Intelligent Systems and Knowledge Engineering (2010)
9. Ma, Y.-W., Lai, C.-F., Huang, Y.-M., Chen, J.-L.: Mobile RFID with IPv6 for phone services. In: The IEEE Proceedings ISCE2009, pp. 169–170 (2009)
10. Malche, T., Maheshwary, P.: Harnessing the Internet of Things (IoT): a review. Adv. Res. Comput. Sci. Softw. Eng. J. **5**(8) (2015). ISSN: 2277 128X
11. “Contiki: The Open Source OS” the topic available at the link 2016: <http://www.contiki-os.org/>
12. “Contiki: The Tiny OS” the link available 2016: <http://www.tinyos.net/>

Design and Investigations of Multiband Microstrip Patch Antenna for Wireless Applications



Ajay Dadhich, Preeti Samdani, J. K. Deegwal and M. M. Sharma

Abstract Novel design for compact Multiband Microstrip Patch Antenna (MMSPA) is proposed. Antenna has rectangular-shaped patch and V and U slots are etched on patch and ring slot on ground plane which is covering patch from back along with line separation in ground plane. By connecting and disconnecting separation, it gives another design that also presents multiband operation along with frequency tuning capability. Physical dimension of proposed antenna is $0.1833 \lambda_0 \times 0.147 \lambda_0$ in mm where λ_0 is the wavelength (at free space) of the first resonant frequency 2.4 GHz. Inset feed microstrip line is used. MMSPA is simulated on CST software. Antenna resonates at 2.4, 4.1, and 7.98 GHz when separation cut is present at ground and antenna resonates at 2.5, 7.7, and 10.5 GHz when separation cut is connected at back of feed line (ground plane) S_{11} of -21.3 , -13.7 and -20 and -13.6 , -16.5 and -20.2 dB, respectively. Proposed antenna may be utilized for IEEE 802.15.1 (2.402–2.480 GHz band), TD-LTE 2300/2500 (2.2–2.5 GHz), ITU#10 band, and X-Band Satellite Communication Service (XSCS) applications.

Keywords ITU#10 · Multiband antenna · Microstrip antenna · WLAN · X-band

A. Dadhich (✉) · J. K. Deegwal
Government Engineering College Ajmer, Ajmer, India
e-mail: ajaydadhich13@ecajmer.ac.in

J. K. Deegwal
e-mail: jitendradeegwal@gmail.com

P. Samdani
NHITM, Mumbai, India
e-mail: preeti.samdani@gmail.com

M. M. Sharma
MNIT, Jaipur, India
e-mail: mmsjpr@gmail.com

1 Introduction

In modern wireless communication system antenna plays an important role. Many studies have been done to develop compact, small size, efficient, and robust design of antenna and the same has been reported in recent past four to five decades. Wide-band antenna with uniform radiation pattern, and high-gain antenna designs have been reported but their sizes are very big, and hence small size, low-weight antennas are required for WLAN, WIMAX, mobile phones, and other many wireless applications. Reduction of size may result in sacrificing some merits of patch antenna. Size can be reduced by cutting slots on patch and the use of high dielectric coefficient substrate material which results in severe polarization and surface-wave weakening and the etching of slots in the ground plane of antenna will wreck the back radiation. Size can also be reduced by the use of shorting post/wall, but may result in deformation of radiation pattern when small ground plane is used. These antenna miniaturization techniques also result in narrow bandwidth. Reduction of size of microstrip antenna is in demand but multiband operation from single antenna is quite necessary for many applications such as GSM900/DCS1800, Global Positioning System (GPS), Wireless Local Area Networks (WLAN-2.5/5.2/5.8 GHz), navigation system, TD-LTE 2300/2500 (2.2–2.5 GHz), ITU#10 band, and X-Band Satellite Communication Service (XSCS) applications [1–5]. So, small-sized MMSPA designs are highly required.

On the other hand, C-shaped and E-shaped patch antennas are well researched [6–8]. These antennas present wide bandwidth but their sizes are large [1–8]. In [9], E-shaped patch with the use of shorting walls between two arms gives reduction in size and wide bandwidth. However, most of these antennas are single band and are not suitable for multiband applications [1–13].

In this article, a compact MMSPA is investigated and designed based on rectangular-shaped patch by etching V and U slots on patch and ring slot on ground covering patch on back along with line separation in ground plane. By connecting separation from back of feed line, another antenna design is introduced that also presents multiband operation and broad frequency tuning ability. Proposed both antennas are compact, simple, and multiband characteristics are achieved in comparison with other reported works. Antennas may be utilized in IEEE 802.15.1 (2.402–2.480 GHz), TD-LTE 2300/2500 (2.2–2.5 GHz), ITU#10 band, and XSCS application.

2 Antenna Geometry

Physical dimensions of Antenna 1 and Antenna 2 are given in Fig. 1. Proposed structure is designed and fabricated on FR-4 lossy dielectric with $\epsilon_r = 4.4$, $\tan \delta = 0.025$, and substrate thickness (h) = 1.6 mm. Dimension of antenna is $40 \times 40 \text{ mm}^2$.

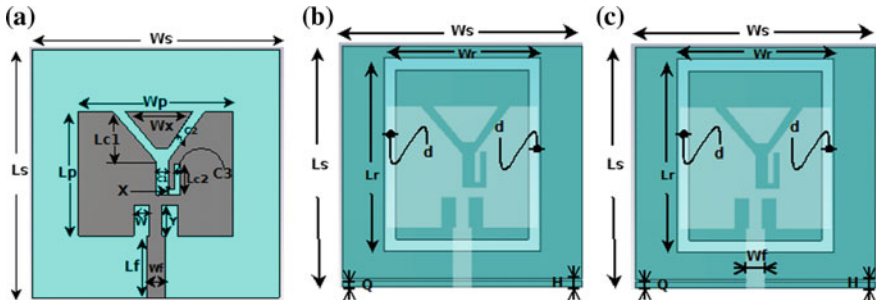


Fig. 1 Antenna Geometry **a** Patch of Antennas 1 and 2, **b** and **c** Ground of Antennas 1 and 2

Table 1 Dimensions of Antenna 1 and Antenna 2

Parameter	(mm)
Length of substrate (L_s)	40
Width of substrate (W_s)	40
Length of patch (L_p)	20
Width of patch (W_p)	25
Length of V of Y slot (L_{C1})	8
Width of V of Y slot (W_X)	11
Length of U of Y slot (L_{C2})	5
Width of U of Y slot (C_1)	2
Width of V of Y slot (C_2)	2
Width of right-side arm of U slot (C_3)	1
Width of bottom of U slot X	1.5
Length of inset slot in patch Y	5
Width of inset slot in patch W	2.5
Length of feed line (L_f)	10
Width of feed line (W_f)	05
Length of rectangular ring at ground (L_r)	26
Width of rectangular ring at ground (W_r)	32
Width of slot of rectangular ring d , Q and H at ground	2, 1.9, and 2

Dimensions of the above antennas are shown in Table 1, and it is shown that separation line is connected with dimension of w_f in Fig. 1c at ground of Antenna 2. Antenna 1 and Antenna 2 both show the multiband characteristics on different resonant frequencies.

3 Simulation Results and Discussion

Computer Simulation Tool (CST) software is used to design and analyze the proposed Antennas 1 and 2. Development of the proposed antenna from initial antenna to proposed Antenna 1 and Antenna 2 is given in Fig. 2. From Fig. 2, it can be seen that Antenna 1 resonates at 5.44, 9.0, and 9.9 GHz that is simple rectangular patch and by introducing inset feed in rectangular patch design of antenna without DGS is prepared which is resonating at 3.5, 5.5, 7.38, and 9.62 GHz. Further by etching V slot connected to U slot on patch, design of antenna with DGS is formed which resonates at 3.15, 5.62, 7.82, and 9.67 GHz and again Defective Ground Structure (DGS) at ground plane is introduced in antenna with DGS design of proposed antenna with separation which resonates in three desired frequency bands of 2.4, 4.1, and 7.98 GHz, respectively, and final design is created by connecting the separation line with w_f dimension from back of feed line at ground plane named as “antenna without separation” as shown in Fig. 2 which resonates at 2.5, 7.7 and 10.5 GHz.

CST is used to observe/analyze various parameters of proposed antennas like VSWR, return loss, and 2D far-field radiation pattern for frequency range of 1–12 GHz. This Antenna 1 resonates at three multiple bands 2.4, 4.1, and 7.98 GHz and Antenna 2 resonates at 2.4, 4.1, and 7.98 GHz. The analysis is illustrated below.

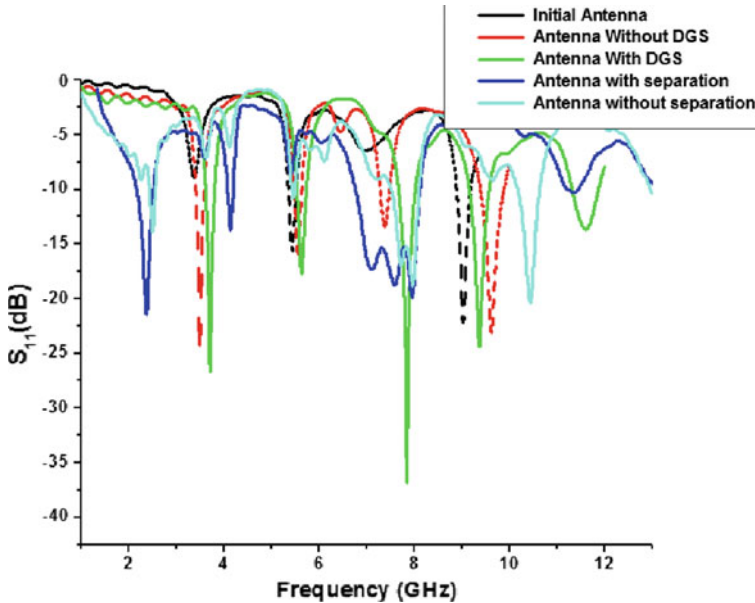


Fig. 2 Development of proposed Antennas 1 and 2

4 Return Loss

The simulated return loss of the proposed Antenna 1 and Antenna 2 is presented in Fig. 3. Antenna 1 resonates at 2.4, 4.1, and 7.98 GHz when separation cut is present at ground and Antenna 2 resonates at 2.5, 7.7, and 10.5 GHz when separation cut is connected at back of feed line (ground plane) having return loss of -21.3 , -13.7 , and -20 dB and -13.6 , -16.5 , and -20.2 dB, respectively.

5 VSWR

VSWR represents the radiation of antenna whenever the VSWR is less than 2 or near about to 2. It represents that the antenna is radiating. Figure 4 shows, at each resonating frequency of the proposed Antennas 1 and 2, values of the VSWR are below or near to 2.

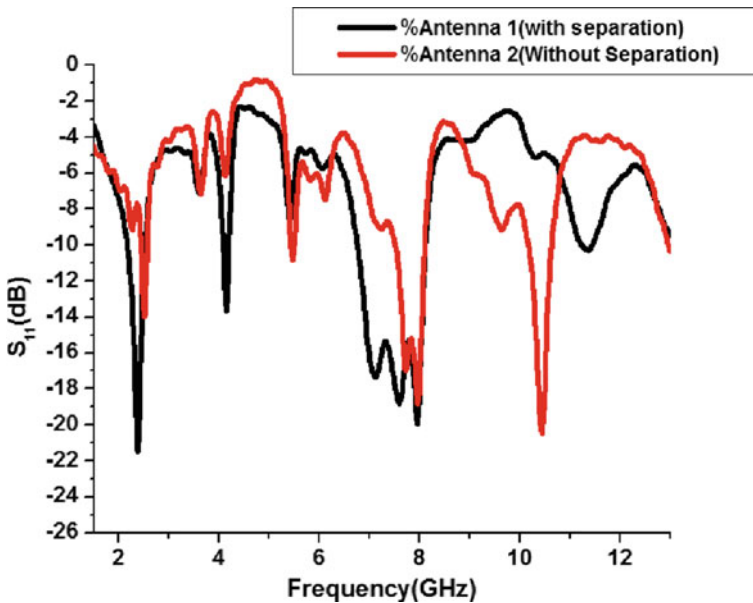


Fig. 3 Simulated return loss of proposed Antenna 1 and Antenna 2

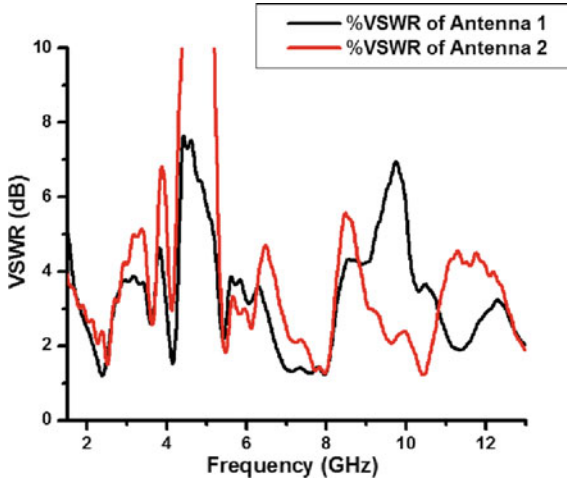


Fig. 4 VSWR of Antenna 1 and Antenna 2

6 Surface Current

Surface current at resonating frequency is shown in Fig. 5 which shows that at different resonating frequencies, the antenna is resonating at different parts of the patch.

7 Radiation Pattern

Far-field radiation pattern of Antenna 1 is presented in Fig. 6 at resonating frequency for phi variations ($\phi = 0^\circ$ and $\phi = 90^\circ$).

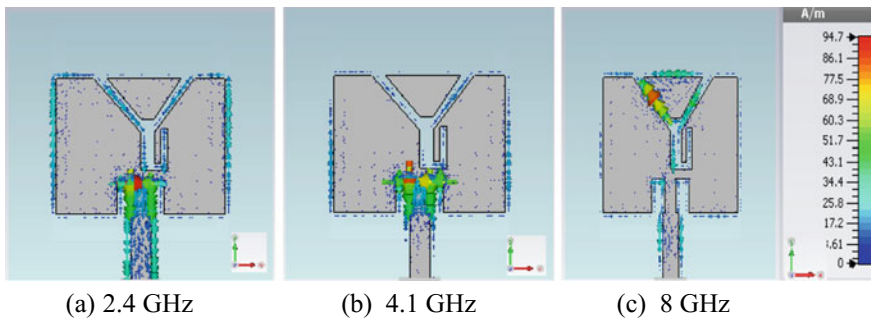


Fig. 5 Surface current at resonating frequencies for Antenna 1

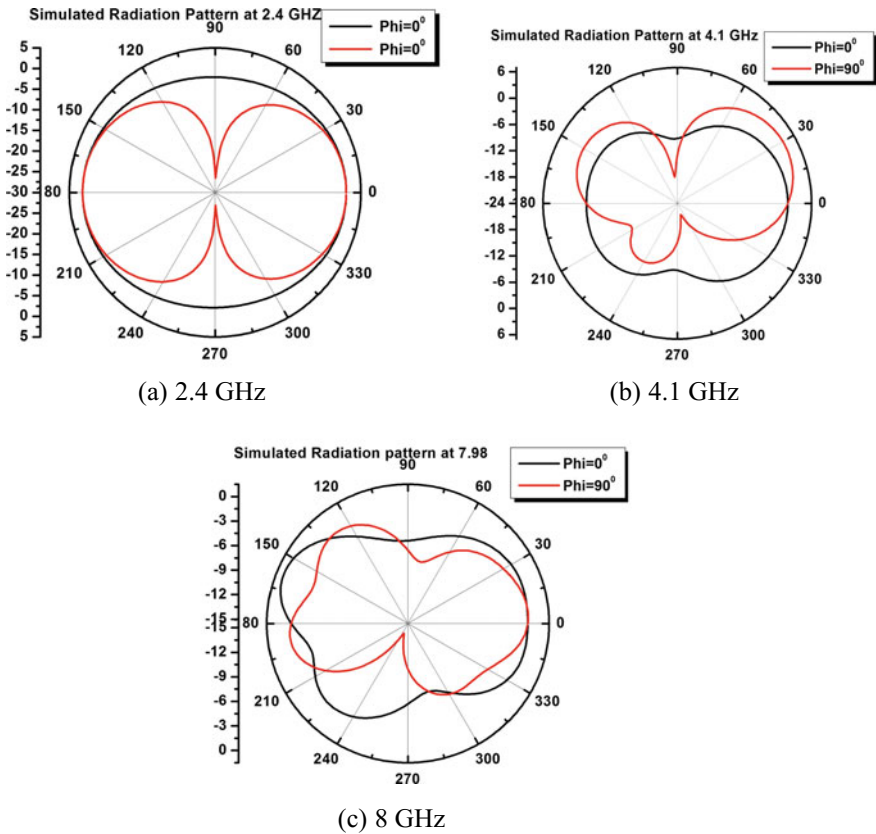


Fig. 6 Radiation pattern (simulated) of Antenna 1 at resonating frequencies

8 Fabricated Design Structure

Fabricated prototype design structure for Antenna 1 and Antenna 2 is shown in Fig. 7 and measured results are to be analyzed for above design.

9 Conclusion

Compact Multiband Microstrip Patch Antenna (MMSPA) is designed and investigated for different conditions of separation at ground plane both Antenna 1 and Antenna 2 show the multiband characteristics. Various parameters were simulated and analyzed like return loss, surface current, radiation pattern, and VSWR for both Antenna 1 and Antenna 2. Proposed antenna may be utilized for IEEE 802.15.1

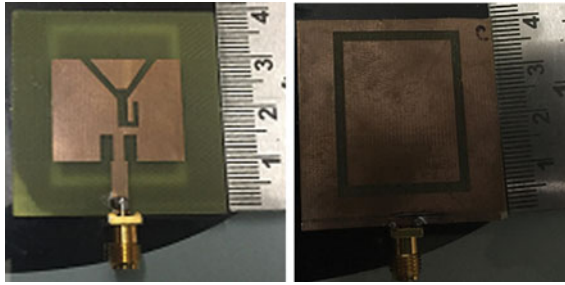


Fig. 7 Fabricated Antenna 1 on FR4 substrate

(2.402–2.480 GHz), TD-LTE 2300/2500 (2.2–2.5 GHz), ITU#10 band, and X-Band Satellite Communication Service (XSCS) applications.

References

1. Sheta, A.F., Alkanhal, M.A.: Compact dual-band tunable microstrip antenna for GSM/DCS-1800 applications. *Microw Antennas Propag.* **2**(3), 274–280 (2008)
2. Chang, F.S., Chao, K.C., Lu, C.H., Su, S.W.: Compact vertical patch antenna for dual-band WLAN operation. *Electron. Lett.* **44**(10), 612–613 (2008)
3. Huang, C.Y., Wu, J.Y.: Compact microstrip antenna loaded with very high permittivity superstrate. *Proc. IEEE AP-S Int. Symp.* **2**, 680–683 (1998)
4. Dadhich, A., Deegwal, J.K., Sharma, M.M.: Study and design of multiband antenna with V and U slot on patch for wireless application. In: *IEEE Sponsored Applied Electromagnetics Conference (AEMC-2017) at Hotel Rama International, MIT, Aurangabad, during 19th Dec. to 22nd Dec 2017*
5. Dadhich, A., Porwal, S., Yadav, S., Mewara, H.S., Sharma, M.M.: Dual band step shaped antenna array for WLAN and WiMax Application. In: *Sixth International Conference on Computer and Communication Technology 2015 (ICCCT'15)*, pp. 297–299, ISBN 978-1-4503-3552-2. <https://doi.org/10.1145/2818567.2818661>
6. Sanad, M.: Double C-patch antennas having different aperture shapes. *Int. Symp. AP-S Dig.* **4**, 2116–2119 (1995)
7. Yang, F., Zhang, X.X., Ye, X.N., Rahmat-Samii, Y.: Wide-band E-shaped patch antennas for wireless communications. *IEEE Trans. Antennas Propag.* **49**(7), 1094–1100 (2001)
8. Wong, K.L., Hsu, W.H.: A broad-band rectangular patch antenna with a pair of wide slits. *IEEE Trans. Antennas Propag.* **49**(9), 1345–1347 (2001)
9. Ge, Y., Esselle, K.P., Bird, T.S.: E-shaped patch antennas for high-speed wireless networks. *IEEE Trans. Antennas Propag.* **52**(12), 3213–3219 (2004)
10. Peng, L., Ruan, C.L., Zhang, Y.: A novel compact broadband microstrip antenna. In: *Proc. APMC*, pp. 1–4 (2007)
11. Dadhich, A., Deegwal, J.K., Sharma, M.M.: Design and investigation of wideband and multi-band microstrip patch antenna for bluetooth, TD-LTE, ITU and X-band applications. In: *International Conference on Emerging Trends in Engineering Innovations & Technology Management (ICET: EITM-2017)*, ISBN 978-93-86724-30-4, Vol. II, pp. 409–412 (2017)

12. Dadhich, A., Deegwal, J.K.: Multiband microstrip patch antenna with rectangular slots on patch for bluetooth and C-band applications. *Int. J. Eng. Technol.* ISSN: 2227-524X, **7**(1.2), 191–193 (2018) <https://doi.org/10.14419/ijet.v7i1.2.9064>
13. Deshmukh, A.A., Ray, K.P.: Compact broadband slotted rectangular microstrip antenna. *IEEE Antennas Wirel. Propag. Lett.* **8**, 410–413 (2009)

A Novel ZOR-Inspired Patch Antenna for Vehicle Mounting Application



Chetan Barde, Arvind Choubey, Rashmi Sinha,
Santosh Kumar Mahto and Prakash Ranjan

Abstract In this paper, a novel structure of hexagonal microstrip patch antennas is proposed which is based on the concept of zeroth-order resonator (ZOR). The simulated result shows that the proposed antenna achieves return loss of -19.23 dB and has bandwidth of 0.453 GHz centered at 6.92 GHz. It achieves gain of 4.374 dB with radiation pattern nearly half sphere, which is suitable for rooftop antenna of vehicles. The proposed antenna with dimension $16 \times 20 \times 1.6$ mm³ can be used for wireless and satellite communication. All the simulations have been done using commercially available finite element method (FEM) solver ANSYS-HFSS. Outmost precision has been taken by assigning $\lambda/20$ mm mesh size, and therefore the obtained results are much presized.

Keywords Patch antenna · Zeroth-order resonator · Printed antenna · Wide beam width

C. Barde (✉) · A. Choubey · R. Sinha · S. K. Mahto · P. Ranjan
Electronics and Communication Engineering,
National Institute of Technology Jamshedpur, Jamshedpur, India
e-mail: 2017rsec004@nitjsr.ac.in

A. Choubey
e-mail: achoubey.ece@nitjsr.ac.in

R. Sinha
e-mail: rsinha.ece@nitjsr.ac.in

S. K. Mahto
e-mail: santoshkumar6990@gmail.com

P. Ranjan
e-mail: 2014rsec001@nitjsr.ac.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_5

1 Introduction

Antenna is the main component of communication system generally used for transmitting and receiving signals from one end to another, in which signals travel through the medium in the form of radio waves directed from the transmitter antenna toward the receiving one. In the past few decades, microstrip antenna is widely used for the wireless communication but their size, bandwidth, and performance parameters such as radiation pattern are somewhat not meeting the requirements of present wireless communication system. To overcome these limitations, ZOR or left-handed material (LHMs) concept was utilized in [1].

ZOR is one of the novel applications of left-handed materials and widely used to achieve wider beam width and radiation pattern similar to dipole antenna from a directional patch antenna. The concept of LHM arises from metamaterial. Nowadays, researchers of electromagnetic community are working on metamaterial [2, 3]. These materials are artificially engineered material with unconventional electromagnetic properties such as negative permeability (μ) and negative permittivity (ϵ). The negative μ and ϵ cannot be found in naturally occurring material. Tremendous work is already done in the field of synthesizing metamaterial structures and is used for various applications such as cloaking devices [2], metamaterial-based filters [3], metamaterial absorbers [4], antennas [5–13], etc.

Metamaterial with negative permittivity and negative permeability, i.e., $\text{Re}(\epsilon) < 0$; $\text{Re}(\mu) < 0$ over a certain frequency region were studied by Veselago in 1968; these negative index material (NIM) attracted the interest because they have properties to reflect light in different directions as compared to normal material. The main physics behind NIM is that the refractive index taking for this material has to be negative square root of the permittivity and permeability [1].

Zeroth-order resonator (ZOR) is one of the novel applications of LHMs. Such LHM-based resonators can have arbitrary designed resonant frequencies that are independent on their physical lengths [5]. The concept of ZOR has been used for enhancing the bandwidth and beamwidth of microstrip patch antenna whose demand is increasing very rapidly in the field of wireless communication because of their low-profile structure [9]. Microstrip patch antenna based on ZOR has small in size, light in weight, low cost, ease of integration, and simplicity of manufacture.

In this paper, a novel ZOR-inspired patch antenna is proposed which consists of hexagonal resonator coupled with feed line. The simulation results exhibit that it has low return loss (high radiation), higher bandwidth of 450 MHz centered at 6.92 GHz and wide beam width which leads to find its various applications such as communication systems, especially for rooftop antenna of vehicle.

2 Design and Simulation

In this article, a novel design of hexagonal-shaped patch antenna having concentrate circular cuts on the ground is proposed. The dimension of the proposed antenna is $16 \times 20 \times 1.6 \text{ mm}^3$. The hexagonal-shaped proposed antenna is considered for C-band application which gives better performance compared to other patch antennas. Radiation pattern with wide beam width angle in C-band of frequency spectra and applicable for wireless communication is proposed. The structure consists of FR4 substrate which acts as a dielectric medium between the ground copper plane and the patch placed at the top of the substrate, FR4 has relative permittivity of 4.4 and height is taken as 1.6 mm.

Figure 1a shows the top view and bottom view of the proposed design. The proposed antenna consists of hexagonal-shaped resonator coupled with small gap of 0.1mm to the feed line. Bottom layer is shown in Fig. 1a. It has three concentric rings with a circular patch placed at the center. The design parameters of the antenna are given in Table 1.

3 Result and Discussion

Proposed antennas are designed and simulated using ANSYS-HFSS 19.0 software which is based on finite element method (FEM). The simulated return loss is shown in Fig. 2. It can be seen that it has single resonance with return loss value of -19.23 dB at 6.92 GHz frequency. The lower value of return loss shows perfect resonance. The -10 dB bandwidth is found to be 450 MHz from 6.69 GHz to 7.14 GHz. The VSWR in the proposed structure is 1.90 at 6.8 GHz as shown in Fig. 3. The value of VSWR is less than two which reflects the perfect matching with the input network to the proposed antenna.

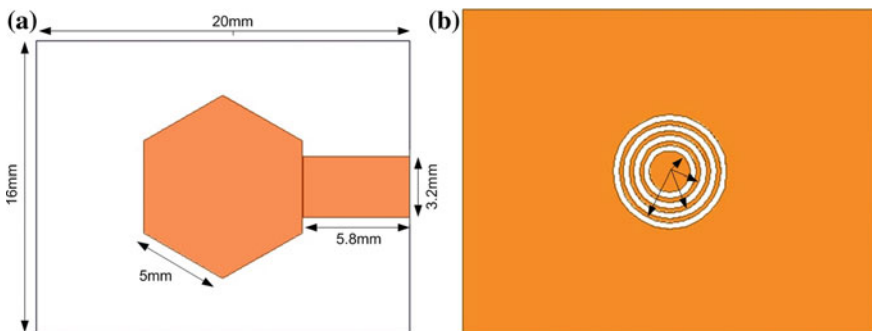


Fig. 1 Structure of proposed hexagonal-shaped patch antenna **a** top view and **b** bottom view of the proposed antenna

Table 1 Parameter of proposed structure

S. No	Parameter name	Value of dimension (mm)
2	Length of side of the patch	5
3	Length of the substrate	16
5	Thickness of the substrate	1.6
6	Thickness of the patch	0.035
7	Length of the feed line	5.8
8	Width of the feed line	3.32
9	Radius of 1st inner circle	1.3
10	Radius of 2nd inner circle	1.8
11	Radius of 3rd inner circle	2.3
12	Radius of outer circle	2.8

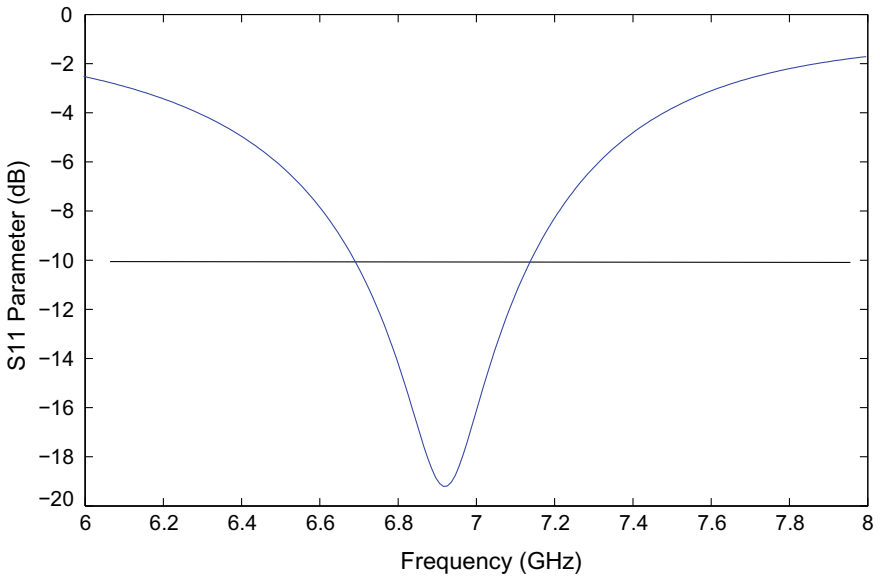


Fig. 2 S_{11} parameter at 6.92 GHz

The directivity of the proposed structure is shown in Fig. 4. The 3D radiation plot is shown in Fig. 5. The radiation plot of the proposed antenna has very less value of back lobe and nearly half spherical radiation pattern which is a type of ZOR antenna. This type of radiation pattern is suitable for various applications such as mount on top of the vehicle. The gain of the proposed antenna after simulation is found to be 4.374 dB at 6.92 GHz.

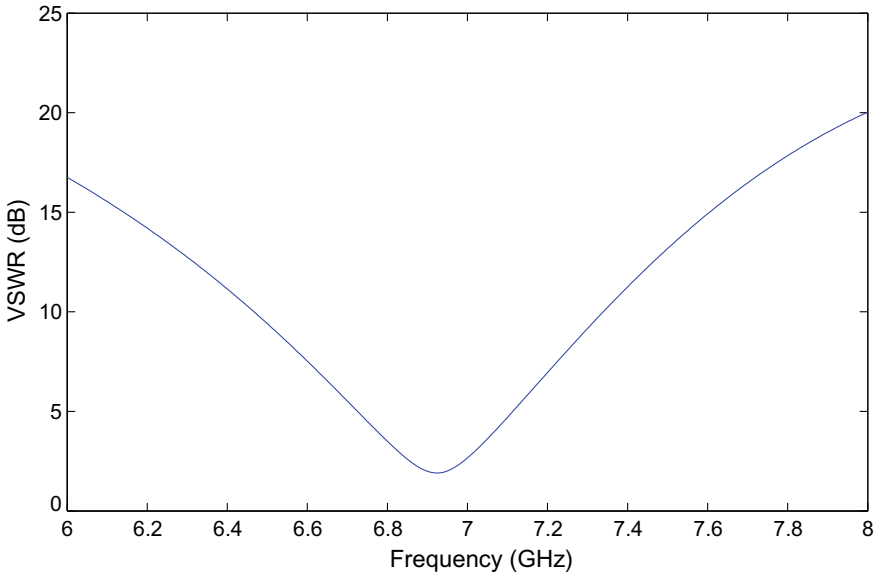


Fig. 3 VSWR at 6.92 GHz

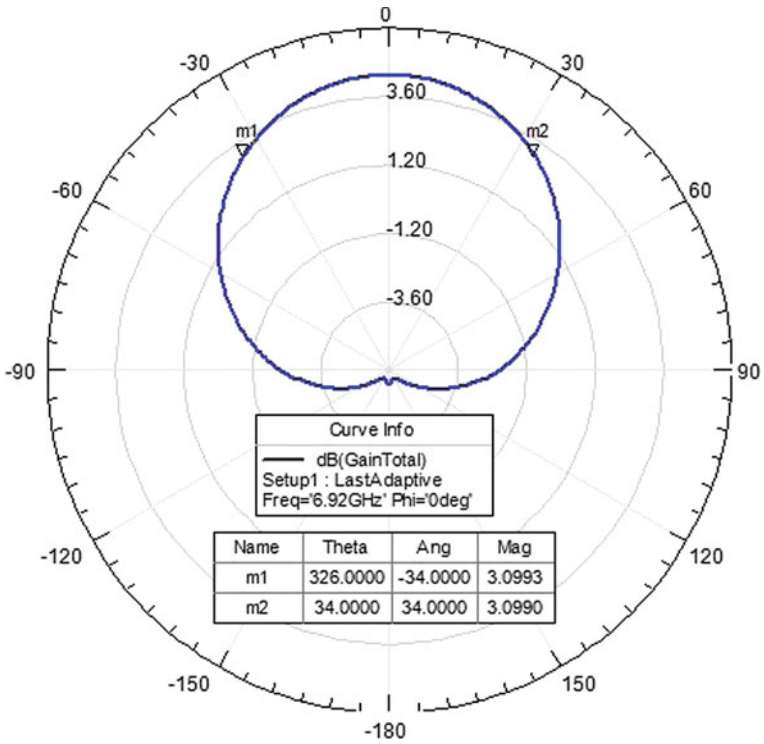


Fig. 4 Directivity at 6.92 GHz

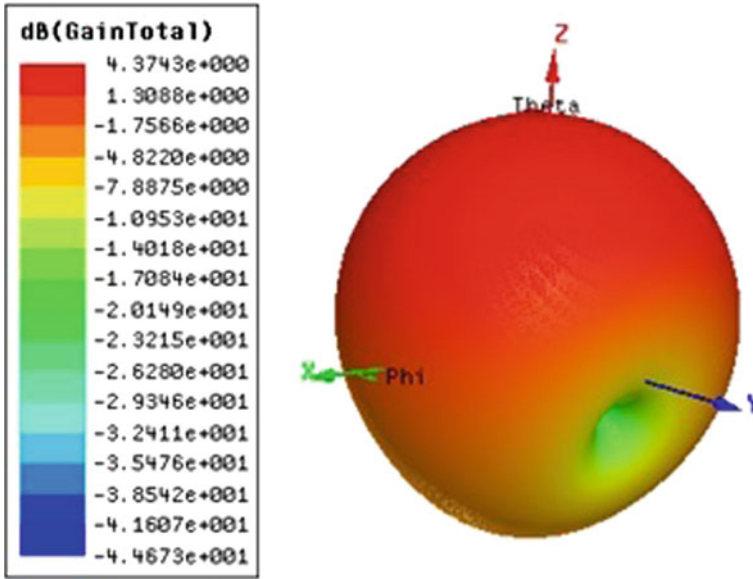


Fig. 5 Gain plot at 6.92 GHz

4 Conclusion

A novel hexagonal patch antenna based on the concept of ZOR has been presented. Simulation results of different parameters such as reflection coefficient, bandwidth, voltage standing wave ratio, directivity, and gain plots are shown and discussed. From the discussion, it is concluded that antenna gives satisfactory performance at 6.92 GHz frequency with gain of 4.374 dB and bandwidth equal to 450 MHz. This antenna finds its various applications in communication system, especially for mounting on rooftop of vehicle.

References

1. Ziolkowski, R.W., et al.: Composite medium with simultaneously negative permeability and permittivity. *IEEE Trans. Antennas Propag.* **51**(7) (2003)
2. Kafesaki, M., et al.: Design of a two-dimensional metamaterial cloak with minimum scattering using a quadratic transformation function. *J. Opt. A: Pure Appl. Opt.* **7**, 12–22 (2005)
3. Constantine, A.B., et al.: A Compact Tunable Metamaterial Filter Based on Split-Ring Resonators. *Antenna Theory and Design*, vol. 13, pp. 120–122. Wiley, New York (1997)
4. Landy, et al.: Perfect metamaterial absorber. *Phys. Rev. Lett.* **100** (2008)
5. Lin, G., Guangming, W., Chenxin, Z.: CRLH T-lines form small antenna. *Algorithms* (2018)
6. Deshmukh, A.A., et al.: Compact broadband slotted rectangular microstrip antenna. *IEEE Antennas Wireless Propag. Lett.* **8**, 410–413 (2009)
7. Waterhouse, R., et al.: Small microstrip patch antenna. *Electron. Lett.* **31**(8), 604–605 (1995)

8. Quevedo-Teruel, O., Pucci, E., Rajo-Iglesias, E.: Compact loaded PIFA for multifrequency applications. *IEEE Trans. Antennas Propag.* **58**(3), 656–664 (2010)
9. Yeh, S.H., et al.: Dual-band planar inverted F antenna for GSM/DCS mobile phones. *IEEE Trans. Antennas Propag.* **51**(5), 1124–1126 (2003)
10. Chang, F.S., et al.: Compact vertical patch antenna for dual-band WLAN operation. *Electron. Lett.* **44**(10), 612–613 (2008)
11. Peng, L., et al.: A novel compact broadband microstrip antenna. In: *Proceedings of APMC*, pp. 1–4 (2007)
12. Lau, K.L., et al.: Dual-band stacked folded shorted patch antenna. *Electron. Lett.* **43**(15), 789–790 (2007)
13. Chiu, C.Y., et al.: Small dual-band antenna with folded-patch technique. *IEEE Antennas Wireless Propag. Lett.* **3**, 108–110 (2004)

A Nascent Approach for Noise Reduction via EMD Thresholding



Rashi Kohli and Shubhi Gupta

Abstract This paper presents and highlights the analysis of data using a novel approach and method called empirical mode decomposition for noise reduction in nonstationary and nonlinear signals. Here, noise reduction is done via thresholding process using this fully data-driven technique. To begin with the process of EMD, the first step is to break down the incoming signal (generally consists of so many frequency components) into a number of monotone signals called intrinsic mode functions (IMFs) and then thresholding is used to reduce the noise from decomposed IMFs. The research paper objective is to suppress the noise signals using the appropriate threshold level in the process of creating and proposing the reduction approach in empirical mode decomposition. Summation of filtered IMFs gives the original signal. All the simulations are done by the MATLAB to verify the expected results.

Keywords EMD · IMFs · Soft thresholding · Hard thresholding · AWGN

1 Introduction

In real world, signals (in raw format) are represented as time domain signals. But, in many cases, only time domain description of signals is not sufficient. So, we have to analyze the signal in frequency domain also. Historically, there are various methods to do the frequency domain analysis; among them, Fourier and wavelet analyses are the most commonly used method. But these conventional methods have some restrictions, i.e., the signal must be strictly periodic, stationary, and the system must be linear. Although wavelet analysis has some advantages as compared to Fourier analysis (as it is not restricted to linear systems) it also has some restrictions. It uses

R. Kohli (✉) · S. Gupta
Amity University, Greater Noida, India
e-mail: rashikohli.amity@gmail.com

S. Gupta
e-mail: sgupta1@gn.amity.edu

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_6

fixed basis function and method is not adaptive. Therefore, to overcome the limitation of existing methods, this paper has proposed, implemented, and modified the data exploration technique chiefly called as empirical mode decomposition (EMD), which has been initially developed by Huang et al. in 1998, to perform comprehensive study and analysis of nonlinear and nonstationary data. This method is completely adaptive in nature because the basic functions, which are used for analysis are derived from the signals directly, so it is highly efficient method. The key part of this method is breaking the signal in the form of “intrinsic mode function” based on the confined attributes and the specific properties of the signal itself. Analyzing the signals and suppressing the noise to determine the nonstationary and stationary signal is relevant in order to improvise the security. In addition to this, the approach of EMD is new and is relevant to analyze the linear and nonstationary time series.

The approach proposed and used in the paper has significant advantages over the previous methods like effective multiresolution analysis capacity. Moreover, the original signal can be reconstructed with improvised security by decomposing the EMD signals and their corresponding IMF signals. There are two modes in which the IMF works. During the process of frequency domain, it performs the process of filtering and during the process of time domain, the IMF signals work in the intrinsic domain. Further, the other important parameter relevant in the process of creating the EMD is the threshold selection. The threshold selection is critical in the process of suppressing the noise signals, as if the threshold is too little, the processing of IMF will have more noise level. Also, if the threshold level is too much, then it will distort the signal accuracy to a greater extent. So, it is vital to appropriately select the threshold level during the IMF level [1–3].

2 Literature Survey

2.1 Empirical Mode Decomposition

EMD splits the given multitone signal into a mono-tone signal called intrinsic mode function (IMFs). The process of splitting the signals is done in order to decompose and separate the signals of high frequency with that of low frequency. This is done to ensure the completeness of the process of multitone signal. We can see from MATLAB simulation that by summing all the IMFs, the signal can be reconstructed. The process is used for filtering the signals using the complicated data set to generate decomposed and refined waveform using the method of intrinsic mode composition and functions (Fig. 1).



Fig. 1 Basic concept of EMD

2.2 EMD Algorithm Process

The following steps [1] ensure the principle of IMFs extraction. The process is called *Sifting* process. The sifting can be summarized as follows:

- (1) First step is to calculate a mean envelope for the signal denoted by $m_1(t)$ and the signal is represented as $x(t)$.
- (2) Next is to solve the residue in the process and set it according to $h_1(t) = x(t) - m_1(t)$.
- (3) Since we are decomposing the signals of high frequency and low frequency so, the important step is to validate whether $h_1(t)$ is an **IMF** or not, if the signal is IMF then STOP; else, repeat the process and treat $h_1(t)$ (with its extrema) as a new signal to obtain $h_{1, 1}(t)$.
- (4) Lastly, it is important to verify again the frequency of the signal and again if $h_{1, 1}(t)$ is an **IMF** if yes then STOP; else, continue the same process and all the above steps repeatedly to decompose the signals.

$$\begin{aligned}
 h_{1, 1}(t) &= h_1(t) - m_{1, 1}(t) \\
 &\quad \dots \\
 h_{1, k}(t) &= h_{1, k-1}(t) - m_{1, k}(t).
 \end{aligned}$$

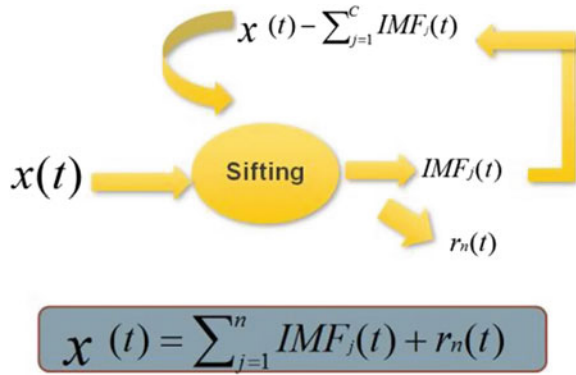
Assuming that after a finite number of iterations say I_1 times, $h_{1, K_1}(t)$ will be an IMF. Denote by $IMF_1(t)$, the first decomposed IMF.

$$= \text{Separate the values of signal and set } r_1(t) = x(t) - IMF_1(t).$$

This process is called as ‘‘Sifting process’’ and it is repeated until all signals are decomposed. The process of repetition includes

$$\begin{aligned}
 \text{STEP 1 : } & -r_2(t) = r_1(t) - IMF_2(t) \\
 & \dots \\
 \text{STEP (n) : } & -r_n(t) = r_{n-1}(t) - IMF_n(t).
 \end{aligned}$$

Fig. 2 Sifting process



The process is STOPPED when r_n (n -total number of decomposed IMF) has at most one extrema. As a result, signal $x(t)$ is decomposed into finite numbers of IMFs. This is described in Fig. 2

$$x(t) = \sum_{i=1}^C IMF_i(t) + r_c(t)$$

The sifting in algorithm is run for several times in order to guarantee that the computed IMF fulfills the required conditions as given below:

- (a) In a complete length of an IMF, the total quantity of zero crossings along with the extrema values must be equal or it can be fluctuated by at most one value.
- (b) Mean values of upper envelope and lower envelope which are designed by taking the extreme values of local maxima and local minima values, respectively, should be zero.

The benefits of sifting are that it can eradicate riding waves and thereby it can level the uneven amplitudes using the process of decomposition [4-6].

3 Proposed Method

3.1 Noise Reduction Using EMD

The purpose of any noise reduction method is mainly to minimize the level of the noise in the signals and its relevant components from the signal in a way such that it should not degrade the signal quality while ensures the minimum information loss. In signal processing, noise extraction from a signal is always a classical problem. Historically, we have so many noise reduction methods. Along with that, numerous amount of filtering methods has been developed previously for the very initial noise model, i.e., the AWGN “Additive white Gaussian noise.” Some of the other well-

known linear method is Wiener filtering, etc. All the methods above has some limitations and these can only be applied in scenarios where during the decomposition process signals contain sharp edges, spikes of small time period, and nonstationary decomposed signals. To overcome these problems, nonlinear methods such as wavelets thresholding have been proposed. The concept of thresholding depends on the assumption that the magnitude of the signal dominates the magnitude of the noise in a wavelets representation, so that coefficients of wavelets can be assigned to zero if their magnitudes are below from a predefined threshold level. But, the limitation of this method is that it has only a fixed number of basis functions, and thus is not applicable to all real signals.

Empirical mode decomposition, whose principle and method of decomposition which we have already seen, is a relatively new tool in signal processing. The result of empirical mode decomposition (the sifting procedure) is that $x(t)$ will be decomposed into $\text{IMF}_i(t)$, $i = 1, \dots, C$ and residual $r_c(t)$ $x(t) = \sum^C \text{IMF}_i(t) + r_c(t)$.

- Noise Reduction Method

Let $f_i(t)$ and IMF_i be a noiseless IMF and noisy IMF, respectively. Consider $y(t)$ be a deterministic signal distorted by AWGN, $n_i(t)$, with a noise level $\sigma_j(t)$ as follows:

$$\text{IMF}_i(t) = f_i(t) + n_i(t), \quad (1)$$

where $i = \{1, \dots, C\}$. An estimation ($\tilde{f}t$) of $f_i(t)$ based on the noisy observation $\text{IMF}_i(t)$ is given by

$$(\tilde{f}t) = \Gamma[\text{IMF}_i(t)], \quad (2)$$

where $\Gamma[h_i, \tau]$ is a preprocessing function, defined by a set of parameters τ_i , and is applied to signal h_i . Let $\tilde{x}(t)$ be a denoising signal and is given by

$$\tilde{x}(t) = \sum^c (\tilde{f}t) + r_c(t) \quad (3)$$

- EMD Thresholding

Denoising of the signal is done by first thresholding the noisy IMFs and then summing the clean IMF to get back the original signal [5, 7]. $\Gamma[., \tau_i]$ and τ_i are the thresholding function and the threshold parameter, respectively. Kim and Oh [4] have proposed a universal threshold for removing AWGN, τ_j given by

$$\tau_i = \tilde{\sigma}_i \sqrt{2 \cdot (N)}, \quad (4)$$

$$\sigma = \text{MAD}_i / .00125, \quad (5)$$

$$\text{MAD}_j = \text{Median}\{|\text{IMF}_i(t) - \text{Median}\{\text{IMF}_i(t)\}|\}, \quad (6)$$

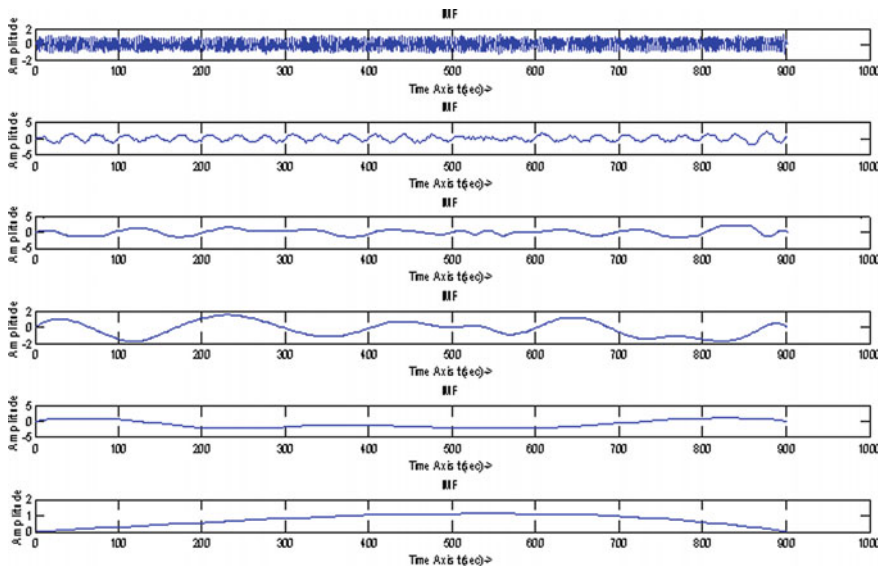


Fig. 3 Set of decomposed IMFs

where $\tilde{\sigma}_i$ —estimation of the noise level of the i th IMF(scale level), MAD_i —the absolute median deviation of the i th IMF.

The soft thresholding shrinks the IMF samples by τ_i toward zero as follows [2]:

$$\tilde{f}(t) = \begin{cases} \text{IMF}_i(t) - \tau_i & \text{if } \text{IMF}_i(t) \geq \tau_i \\ 0 & \text{if } |\text{IMF}_i(t)| < \tau_i \\ \text{IMF}_i(t) + \tau_i & \text{if } \text{IMF}_i(t) \leq -\tau_i \end{cases} \quad (7)$$

Hard thresholding is done in following manner [2]:

$$\tilde{f}(t) = \begin{cases} \text{IMF}_i(t) - \tau_i & \text{if } |\text{IMF}_i(t)| \geq \tau_i \\ 0 & \text{if } |\text{IMF}_i(t)| < \tau_i \end{cases} \quad (8)$$

- A. Soft thresholding result: The process of soft thresholding method relies on the sparse modeling technique which is typically applied to wavelet signals for suppressing the noise and denoising in statistical signal processing and analysis. This process is slightly different from the other hard thresholding technique in terms of wavelets and signals. The process has a single parameter which is used to control a threshold level on wavelet coefficients. The corrupted signal is shown in Fig. 3 and the other interfering signals are shown simultaneously.

Case(i): Signal is corrupted by other than AWGN

In Fig. 3, the noisy signals were detected and simulated using the MATLAB, where the approach of EMD process was applied. In Fig. 4, the original signal, noisy

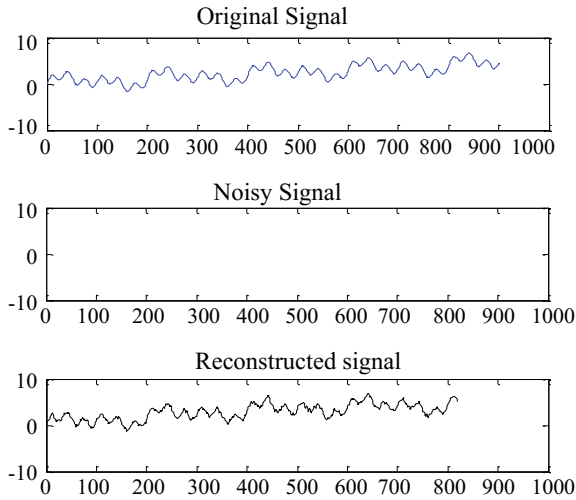


Fig. 4 Represent the reconstruction of signal in the presence of other than AWGN

signal, and reconstructed signal are shown in the presence of other AWGNs. The approach suppresses the noise using the denoising of the signal which is further done by first thresholding the noisy IMFs and then summing the clean IMFs to get back the original signal.

Case(ii): Signal is corrupted by AWGN

Figure 5 reflects the interference of original signals by the decomposed IMFs and in the presence of other AUWN signals. The results are simulated by the MATLAB to reflect the impact of original signal when there is presence of other signals at different time intervals and in both the time and frequency domains (Fig. 6).

- B. Hard thresholding result: The process of hard thresholding is shown in Fig. 7, where the signals and noise reduction process are simulated on the MATLAB and to suppress the noise we apply the following nonlinear transformation in the process of EMD transformations with the help of wavelet coefficients, where the threshold limit is applied. As discussed previously, it is very critical to select the threshold value as it directly impacts the real or original signals. The corresponding Fig. 8 reconstructs the same signal in the presence of other interfering signals and using the time domain approach each IMF level signal is reconstructed to suppress the noise level in the signal and to generate back the reconstructed signal (Figs. 9 and 10).

Case(i): Signal is corrupted by other than AWGN.

Case(ii): Signal is corrupted by AWGN.

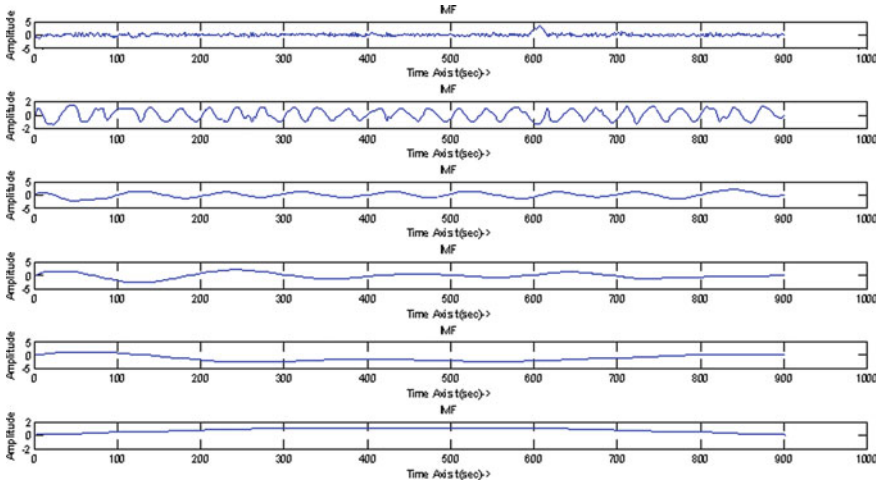


Fig. 5 Set of decomposed IMFs

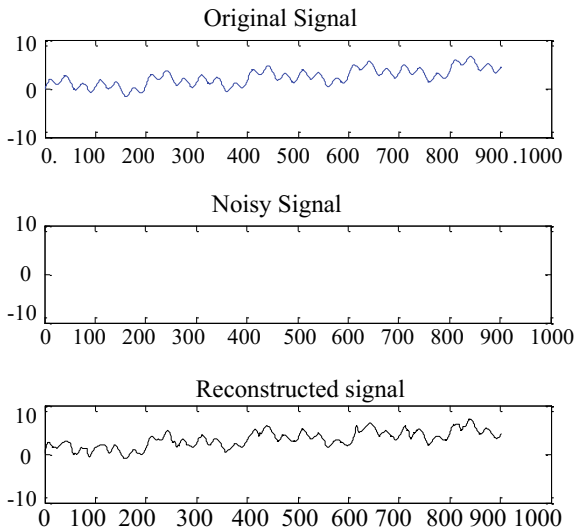


Fig. 6 Signal reconstruction in the presence of AWGN

4 Conclusion and Future Work

Thus, the proposed methods are very effective in noise removal. The MATLAB simulation can verify results. We showed that the new approach with relevant thresholding level is useful for removing noise and can improve the denoised results of soft and hard thresholding significantly. However, since thresholding is also applied to the

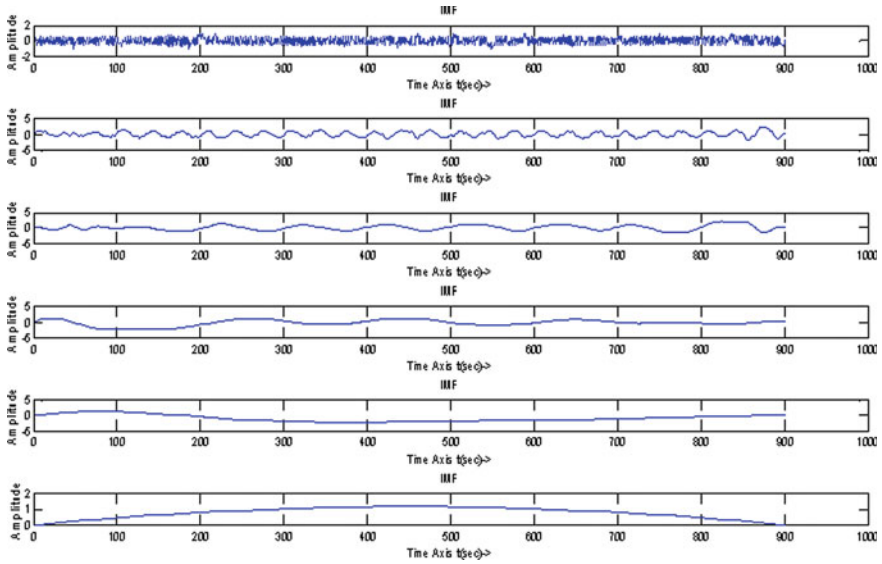


Fig. 7 Set of decomposed IMFs

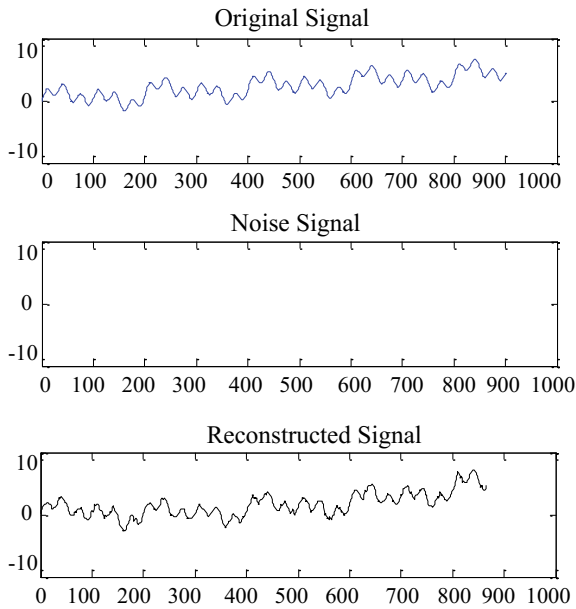


Fig. 8 Reconstruction of signal when it is corrupted by other than AWGN

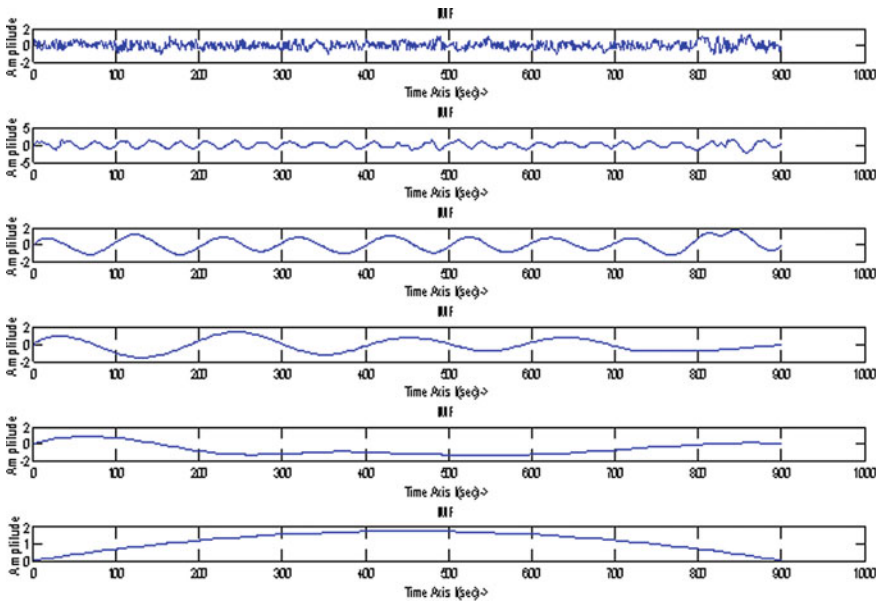


Fig. 9 Set of decomposed IMFs

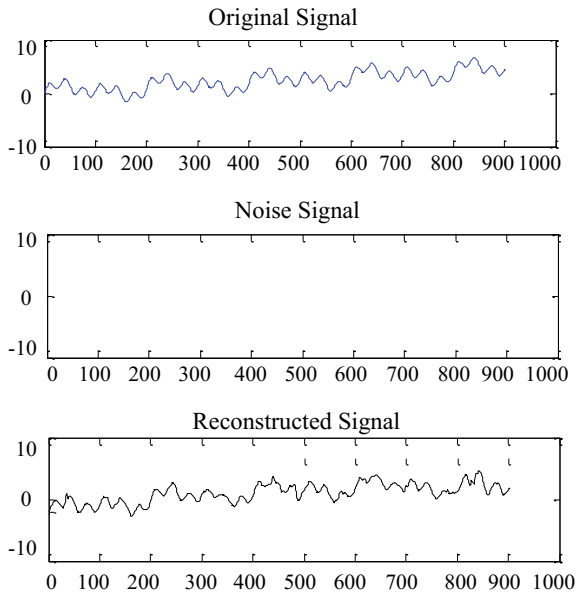


Fig. 10 Reconstruction of signal when it is corrupted by AWGN

signal dominant frames, there is a reasonable degradation in the signal component which affects the signal quality. In future, this process can be extended to apply noise reduction method to speech signals, edge detection techniques, and image processing and try to reduce the degradation level in signal components. The reconstruction of the signal and its degradation level can be improved and can be applied in various security and cryptographic algorithms in the case of biometrics to improve the security level aspects. The above method in the paper was tested on the MATLAB; however, this proposed method in future can be tested on real ECG signal and simulated signals (Doppler, blocks, bumps heavy sine, and piece regular) [8–10].

References

1. Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H.: The empirical mode decomposition and Hilbert spectrum for non-linear and non-stationary time series analysis. *Proc. R. Soc. London A* **454**, 903–995 (1998) <http://rspa.royalsocietypublishing.org/content/454/1971/903.full.pdf+html>
2. Flandrin, P., Goncalves, P., Rilling, G.: Detrending and denoising with empirical mode decomposition. In: *Proceedings of the 12th European Signal Processing Conference (EUSIPCO'04)*, pp. 1581–1584, Vienna, Austria (September 2004)
3. She, L., Xu, Z., Zhang, S., Song, Y.: De-noising of ECG based on EMD improved-thresholding and mathematical morphology operation, vol. 2, pp. 838–842 (2010) <https://doi.org/10.1109/bmei.2010.5639920>
4. Kim, D., Oh, H.-S.: EMD: a package for empirical mode decomposition and hilbert spectrum. *Contributed Res. Art. R. J.* **1**(1), (2009) ISSN2073-4859 [URL: http://journal.r-project.org/2009-1/RJournal_2009-1_Kim+Oh.pdf]
5. Donohue, D.L.: De-noising by soft-thresholding. *IEEE Trans. Inform. Theory*, **41**(3), 613–627 (1995)
6. Kopsinis, Y., McLaughlin, S.: Improved EMD using doubly-iterative sifting and high order spline interpolation. *J. Adv. Sig. Process. (JASP)*, vol. 2008, Article ID 128293, 8 pages, 2008. <https://doi.org/10.1155/2008/128293>
7. Boudraaand, A.O., Cexus, J.C.: Denoising via empirical mode decomposition. In: *Proceedings of the IEEE International Symposium on Control, Communications and Signal Processing (ISCCSP '06)*, p. 4, Marrakech, Morocco (March 2006)
8. Srivastava, D., Kohli, R., Gupta, S.: Implementation and statistical comparison of different edge detection techniques. In: Bhatia, S., Mishra, K., Tiwari, S., Singh, V. (eds.) *Advances in Computer and Computational Sciences. Advances in Intelligent Systems and Computing*, vol. 553. Springer, Singapore (2017)
9. Arora, G., Bibhu, V., Kohli, R., Pavani, P.: Multimodal biometrics for improvised security. In: *2016 International Conference on Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, pp. 1–5 (2016)
10. Mohguen, W., Bekka, R.E.: Empirical mode decomposition based denoising by customized thresholding. *World Acad. Sci. Eng. Technol. Int. Electron. Commun. Eng.* **11**(5) (2017)

Congestion Control Network Coding Scheme in Delay-Tolerant Network



Uroosa Zaidi, Pranee Saurabh, Ritu Prasad and Pradeep Mewada

Abstract Delay-tolerant network (DTN) allows nodes to follow, store, and forward packets through different delivery mechanisms. These data bundles are created by sender to send data to multiple destinations. DTN supports multicast approach for communication. DTN enables the capability of transmission that involves long interruption, wait, and interoperable communications among these networks. Congestion in wireless network remains a challenging issue as enhancing the available bandwidth of network is not possible. Multicast communication in DTN network might be an intricate subject because nodes generously move within the surroundings that often become insecure due to the absence of centralized architecture. This paper proposes a network-coding-based congestion control scheme for delay-tolerant network communication (NCCCS-DTN). NCCCS-DTN using network coding functions is better than its state of the art (CASPaR). The proposed NCCCS-DTN reduces the bandwidth consumption and also lowers load on the link. NCCCS-DTN approach provides best route and also utilizes bandwidth and offers efficient congestion-free communication.

Keywords Congestion · Bundles · Routing · Network coding · Delay-tolerant network

U. Zaidi (✉) · R. Prasad · P. Mewada
Technocrats Institute of Technology Advance, Bhopal 462021, Madhya Pradesh, India
e-mail: naqvi_danish@yahoo.com

R. Prasad
e-mail: rit7ndm@gmail.com

P. Mewada
e-mail: pradeepmewada07@gmail.com

P. Saurabh (✉)
Technocrats Institute of Technology, Bhopal 462021, Madhya Pradesh, India
e-mail: praneetsaurabh@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_7

1 Introduction

Delay-tolerant networks (DTNs) [1] are the networks in which the communication is controlled in dynamic environment and delay occurs due to first store and then forward data to multiple receivers approach. Different routing protocols of DTN follows the mechanism of first store then forward instead of forward data to facilitate greater communication control [2]. DTN is sometimes also called opportunity network as a result of the transmutation node invariably looking for opportunity to transmit bundle data from sender to receiver [3]. This scheme needs intermediate nodes to participate within the routing procedure to nearby nodes, e.g., accumulate data in local buffer space for others [4]. This mechanism of storing before forwarding results in consuming time and subsequently creates a delay and reduces network performance [5]. Once DTN node has to send a data packet then all the nearby nodes are the potential candidates to relay this information. Thus, relaying choice and assuming call have to be compelled to be created by the current node based on convinced routing strategy [6]. There are many routing approaches in DTN that work on different metrics. Successful delivery percentage and bandwidth consumption, and available bandwidth and storage capabilities are the metrics that determine the routing schemes. However, in shared communication environment, limited availability of resources extensively impedes and restricts planning of a perfect forwarding mechanism. Therefore, several routing schemes are designed to manage buffer management to outline the precedence for message carried [7]. Unavailability of link space further complicates and leads in refusal of the arriving data packets [8]. A network-coding-based congestion control scheme for delay-tolerant network (NCCCS-DTN) is developed using network coding that outperforms its state of the art (CASPaR). The proposed NCCCS-DTN reduces the bandwidth consumption and also lowers load on the link. NCCCS-DTN approach provides best route and also utilizes bandwidth and offers efficient congestion-free communication. This paper is organized in the following manner. Next section covers the theoretical aspects and the related work while Sect. 3 presents the proposed work. Section 4 puts forward the experimentation and result analysis followed by conclusion in Sect. 5.

2 Background and Related Work

2.1 Theoretical Aspects

This section presents the theoretical background of the various concepts related to this research. Delay-tolerant networks (DTN) works on the principle of first store and then forward data to multiple receiver approaches. DTN supports multicast communication that enables the nodes to send single message to multiple destinations thus optimally utilizes the bandwidth capacity [8]. The multicast communication is also very efficient in clustering technique to communicate the nodes having connected to

sender. The DTN in group communication improves the reliability of communication and also improves the energy utilization in network [9]. This approach leads to congestion in context of DTN. The various keywords involved in congestion in DTN are given as.

2.1.1 Free Buffer Size

The buffer space in network stores large amount of data, to mitigate the probability of congestion in the network. The storing and forwarding capability of node depends on the rate of incoming and outgoing data in the network.

2.1.2 Message Size

It specifies the message size that means less possibility hold on messages, and therefore the possibility of successful data communication and subsequent delivery are higher if the message size remains small.

2.1.3 Message TTL

Time to live defines the life of a packet in the network and higher limit increases the probability of a packet to survive in the network.

2.1.4 Message's Node Delay

This mechanism ensures that the message keeps time on every node. It determines efficient use of space for storing messages and monitors buffer space consumption.

2.2 Related Work

This section reviews the various explores in delay-tolerant networks. Wei et al. developed multiple attribute call making (MADM) based on the theme to manage congestion that establishes a gaggle of forward messages and its transmission order on each encounter event. MADM worked mostly on congestion management mechanism, routing commonplace, and congestion management module of form forwarding choices [8]. In another work, Hur and Kang [9] planned quota-based multicast approach and exclusively attained a better rate of delivery. This approach adapted to network conditions and their planned approach involved free members. Fall [10] highlighted the security analysis and proposed DTN RFCs for communication networks based on space. The main focus is on bundle security whereas cluster

communication is involved. Wan et al. [11] put forward and then compared different routing protocols for delay-tolerant networks in their work with performance evaluation. Burgess et al. [12] presented a relatively indispensable approach that aims to “restore” underlying link in DTN to remain relevant in existing TCP/IP protocols. Subsequently, it provided a precise variety of network service and it overcomes the congestion challenge. Stewart et al. [13] proposed a congestion shunning shortest path routing protocol (CASPaR). CASPaR is outlined by biological process that replicated and then it has done management of the overhead packets that were nearer to their destinations. Recently, bio-inspired advances [14, 15] also gained attention in realizing different goals on this domain [16, 17]. All these limitations actually point out the need for a mechanism that can overcome the problem of congestion in delay-tolerant network. Next section discusses the proposed network-coding-based congestion control scheme for delay-tolerant network.

3 Proposed Work

This section presents network-coding-based congestion control scheme for delay-tolerant network (NCCCS-DTN) using network coding functions to overcome problem of congestion. Congestion is not an easy problem for any communication due to limited bandwidth, channels, and frequencies. In the existing research work, numerous congestion control and avoidance method were defined and implemented where the congestion was detected through queue utilization and bandwidth utilization.

3.1 Algorithm

This subsection formally presents and subsequently describes the algorithm to overcome problem of congestion in DTN through shared link detection and data XOR-ing method and other group communication through DTN method. NCCCS-DTN approach gives the result in the form of average latency, overhead, hop count, and queue deviation.

Algorithm

Congestion Avoidance in DTN multicast routing.

Input: $G: \{g_i \text{ where } i \text{ 1 to } n\}$ groups

M : mobile nodes

S_n : sender nodes

R_n : receiver node

I_k : intermediate nodes

D_i : data packets

$C_{m,n}$: connection m to n node

DTN_B : dtn bundle

$Q_u(t)$: queue utilization at time t

Output: Average latency, overhead, average hop count, queue deviation

Function1: data send by shared connection

While $c_{i,k} \geq 1$ shared by $s_{k \geq 2}$ node

$C_{i,k}$ send ack to $\in s_k$ node for sharing the path and channel

I node perform ($d_k = d_i \text{ xor } d_j$)

$s_{k \geq 2} \in S_n$ send d_i data packet

k node perform $d_i = d_k \text{ xor } d_j$ and $d_j = d_k \text{ xor } d_i$ for data separation

If $Q_u(t)$ to i node $> 50\%$ than

i sends ack to $\in s_k$ node for data rate minimization

s_k minimize data rate & utilized shared connection

End if

Function2: DTN bundle for group communication

While $m > 2$ is under g_i

S_i sender & $R_i > 1 \{S_i, R_i \in g_i\}$

If R_i receives common message from S_i than

S_i create bundle of d_i

$DTN_B \leftarrow (d_i, R_i > 2 \text{ addresses})$

S_i sends (DTN_B) to $R_{i>1}$ nodes

Separate d_i with R_i addresses

Receives d_i message by R_i nodes

End if

3.2 Performance Parameters and Metrics

3.2.1 Performance Parameters

The performance of proposed network-coding-based congestion control scheme for delay-tolerant network (NCCCS-DTN) using network coding functions is evaluated under following performance parameters mentioned in Table 1. The simulation of both the proposed NCCCS-DTN and CASPaR is done in NS-2 simulator version 2.31. Number of mobile nodes mobility is considered random and the maximum range of nodes communication and sensing is about 550 m. The bundle of messages is delivered to multiple destinations using ODMRP protocol in DTN. The rest of the parameters are mentioned below.

Table 1 Performance parameters

Simulation parameters	Values
Area of communication	800 m × 800 m
Propagation mode	Two ray ground
Nodes consider for communication	100
Buffer size	10, 20, 30
Simulation time	100 (in seconds)
Application layer protocols	FTP, CBR
Radio range	550 m
MAC layer	802.11
Packet size	512, 1024 Bytes
Routing protocol	ODMRP
Nodes mobility	Random
Maximum node speed	25 m/s

3.2.2 Performance Metrics

Following are the performance metrics on which both the proposed NCCCS-DTN and CASPaR are evaluated to determine the impact of the new integrations in NCCS-DTN.

- (i) **Average hop count:** Average number of hops required for routing data between sources to destination through shortest path. If link is strong then average hop count is less otherwise more.
- (ii) **Delivery probability:** It is the possibility of data packets delivered with respect to data packets transmitted in dynamic network. It is percentage ratio of successful data delivery.
- (iii) **Routing overhead:** The number of routing packets sent in network to establish connection with receiver.
- (iv) **Throughput:** It is the ratio of data delivered versus data packets sent in DTN.
- (v) **Average latency:** It is the count in term of time consumed by receiver or sender or intermediate node during transmission.

4 Experimentation and Result Analysis

This section presents the various results obtained after different experiments carried to evaluate both the proposed network-coding-based congestion control scheme for delay-tolerant network (NCCCS-DTN) and congestion shunning shortest path rout-

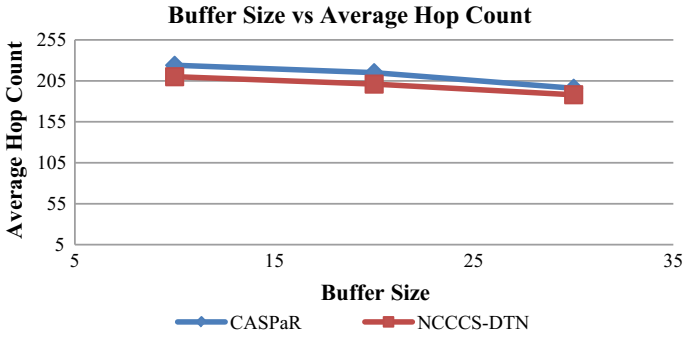


Fig. 1 Average hop count analysis

ing protocol (CASPaR). After experimentation, comparisons have been drawn to demonstrate the effect of new integrations in the proposed NCCCS-DTN.

4.1 Average Hop Count Performance Analysis

Mechanism of store and forwarding of data are time-consuming and adds to complexity of DTN. In this experiment, average hop count is measured between proposed NCCCS-DTN and CASPaR.

In Fig. 1, hop count of proposed NCCCS-DTN is lower than the existing CASPaR at different buffer sizes which is desirable. This reflects the positive impact of networking coding incorporated in the NCCCS-DTN. Lower hop count in NCCCS-DTN provides the strong connectivity and does not need to establish connection again and again in dynamic network.

4.2 Delivery Probability Performance Analysis

Congestion is persistent challenge in wireless network because of limited bandwidth availability that gets more complicated with multicast or unicast communication using the same available bandwidth. This experiment is carried to measure the delivery probability of packets in the proposed NCCCS-DTN and CASPaR. Results presented in Fig. 2 very clearly states that the proposed NCCCS-DTN reported higher packet delivery as compared to CASPaR under different buffer sizes. Results also indicate that NCCCS-DTN successfully handles load in network efficiently and reduces the multiple copies of data to nearby nodes.

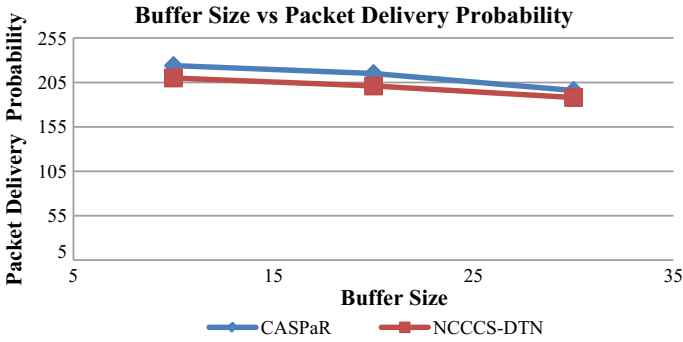


Fig. 2 Delivery probability analysis

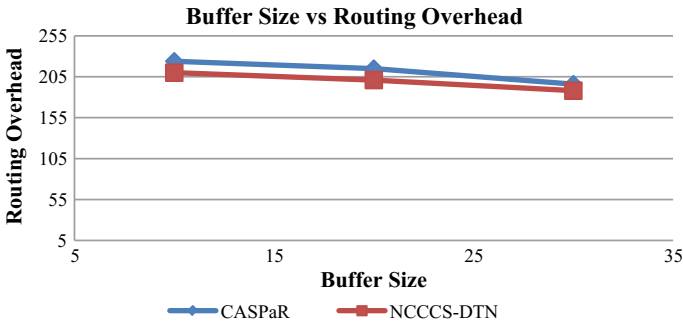


Fig. 3 Overhead analysis

4.3 Routing Overhead Performance Analysis

Routing overhead is a bitter truth that plays a part in lowering the performance of any routing mechanism. This experiment is performed to determine the router overhead and its subsequent impact in the proposed NCCCS-DTN and CASPaR.

Figure 3 puts forward the fact that the proposed NCCCS-DTN multicasting routing approach in DTN with coding scheme clocks improved routing the performance and reported lower routing overhead as compared to CASPaR under different buffer sizes.

4.4 Throughput Performance Analysis

Throughput is a key parameter in determining the performance of any routing algorithm. This experiment is performed to observe the performance of both NCCCS-DTN and CASPaR. Figure 4 reveals the importance of network coding in NCCCS-

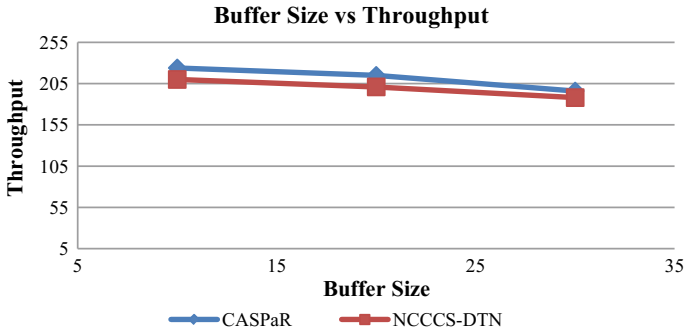


Fig. 4 Throughput analysis

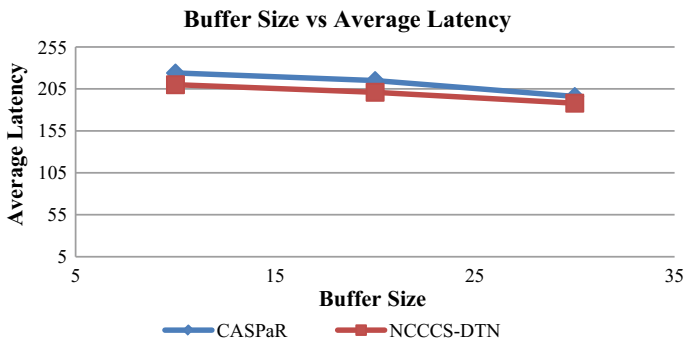


Fig. 5 Average latency analysis

DTN as it recorded higher throughput than CASPaR. This result also states more and proper bandwidth utilization that improved the capacity of multicasting communication and handled bundles in DTN proficiently.

4.5 Average Latency Performance Analysis

Latency is the measure of time consumed by receiver, sender, or intermediate node during transmission. Results of the experiments carried in Fig. 5 shows that NCCCS-DTN observed lower latency as compared to CASPaR. This illustrates that the proposed NCCCS-DTN is relatively more successful in congestion control.

5 Conclusion

The multicast routing in DTN sends bundle of messages to multiple destinations through different protocols often creating congestion in the network. This paper introduced NCCCS-DTN that incorporated network coding scheme to overcome congestion in DTN. NCCCS-DTN based on network coding efficiently used the link capacity and subsequently demonstrated efficient bandwidth consumption, high delivery ratio, and low latency. NCCCS-DTN due to new integrations reduced multiple message copies and consequently improved bandwidth utilization reported low packet drop and handled congestion more proficiently. Experimental results demonstrated that NCCCS-DTN outperformed CASPaR state-of-the-art method under different performance metrics like throughput, latency, and packet delivery ratio.

References

1. Zhu, H., Du, S., Gao, Z., Dong, M., Cao, Z.: A probabilistic misbehavior detection scheme toward efficient trust establishment in delay- tolerant networks. *IEEE Trans. Parallel Distrib. Syst.* **25**(1), 22–32 (2014)
2. Wei, K., Liang, X., Xu, K.: A survey of social-aware routing protocols in delay tolerant networks: applications, taxonomy and design-related issues. *IEEE Commun. Surv. Tutor.* **16**(1), First (2014) 556–578
3. Burgess, J., Gallagher, G., Jensen, D.: MaxProp: routing for vehicle- based disruption tolerant networking. In: *IEEE INFOCOM* (April, 2006)
4. Burleigh, S., Jennings, E., Schoolcraft, J.: Autonomous congestion control in delay-tolerant networks. In: *American Institute of Aeronautics and Astronautics* (2007)
5. Wei, K., Guo, S., Zeng, D., Xu, K.: A multi-attribute decision making approach to congestion control in delay tolerant networks. In: *IEEE Mobile and Wireless Networking Symposium*, pp. 2742–2747 (2014)
6. Santiago, J., Casaca, A., Pereira, R.P.: Multicast in delay tolerant networks using probabilities and mobility information. *Ad Hoc Sens. Wirel. Netw.* **7**(1–2), 51–68 (2009)
7. Zhao, W., Ammar, M., Zegura, E.: Multicasting in delay tolerant networks: semantic models and routing algorithms. In: *Proceedings of the 2005 ACM SIGCOMM workshop on Delay-tolerant networking*, pp. 268–275. ACM, New York (2005)
8. Wei, K., Guo, S., Zeng, D., Xu, K.: A multi attribute decision making approach to congestion control in delay tolerant networks. In: *IEEE ICC 2014—Mobile and Wireless Networking Symposium*, pp. 2742–2747 (2014)
9. Hur, J., Kang, K.: Secure data retrieval for decentralized disruption-tolerant military networks. *IEEE/ACM Trans. Netw.* **22**(1), 16–26 (2014)
10. Fall, K.: A delay tolerant network architecture for challenged internets. In: *Proceedings of SIGCOMM '03*, pp. 27–34 (2003)
11. Wan, L., Liu, F., Chen, Y., Zhang, H.: Routing protocols for delay tolerant networks: survey and performance evaluation. *Int. J. Wirel. Mob. Netw. (IJWMN)* **7**(3), 5–69 (2015)
12. Burgess, J., Bissias, G., Corner, M.D., Levine, B.: Surviving attacks on disruption-tolerant networks without authentication. In: *Proceedings of the 8th ACM international symposium on Mobile ad hoc networking and computing*, pp. 61–70 (2007)
13. Stewart, F., Kannan, R., Dvir, A., Krishnamachari, B.: CASPaR: congestion avoidance shortest path routing for delay tolerant networks. *IEEE Int. Conf. Comput., Netw. Commun., Wirel. Ad Hoc Sens. Netw.* pp. 1–5 (2016)

14. Saurabh, P., Verma, B.: An efficient proactive artificial immune system based anomaly detection and prevention system. *Expert. Syst. Appl.* **60**, 311–320 (2016)
15. Saurabh, P., Verma, B., Sharma, S.: An immunity inspired anomaly detection system: a general framework a general framework. In: *Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012)* vol 202 of the series *Advances in Intelligent Systems and Computing*, pp. 417–428. Springer, Berlin (2012)
16. Saurabh, P., Verma, B., Sharma, S.: Biologically inspired computer security system: the way ahead. *Recent. Trends Comput. Netw. Distrib. Syst., Secur. Commun. Comput. Inf. Sci.* **335**, 474–484 (2011)
17. Saurabh, P., Verma, B.: Immunity inspired cooperative agent based security system. *Int. Arab J. Inf. Technol.* **15**(2), 289–295 (2018)

IoT: Architecture, Technology, Applications, and Quality of Services



Vidhyotma and Jaiteg Singh

Abstract A decade back “Internet of Things” (IoT) has transmogrified the automation technologies which comprises the physical entity, sensors, actuators, and controllers with Internet connectivity. However, several technological applied issues and challenges are pertaining to the interoperability of humans, things, and machines. Ultimately diversified application areas of IoT such as health care, transport management, smart home, and smart cities have increased the scope of this technology in future. The paper highlights the layered architecture of IoT, communication protocols used by each layer as well as applications and its technologies.

Keywords IoT · RFID · Smart objects · Wireless sensors · WPAN · WSN

1 Introduction

The term Internet of Things (IoT) was coined by “Kevin Ashton” in 1999 at “Procter and Gamble” (P&G) while connecting the latest scheme of RFID in the supply chain of that company [1, 2]. A decade later, the IoT was launched as new technology [3]. Road map of IoT as per “SRI consulting business intelligence” is given in Fig. 1. It shows when the number of objects linked with the Internet would be exceeded the number of persons in the universe [4]. The road map also shows the increase in application areas of IoT with the duration of time. IoT had started from the retail supply chain and later on benefitted medical and health, transport, manufacturing, pharmaceuticals and safety and security application areas.

The IoT has made possible to interact between Things to Human (T2H), Things to Things (T2T), Human to Human (H2H), Human to Things (H2T), at a virtual level

Vidhyotma (✉) · J. Singh
Chitkara University Institute of Engineering and Technology, Chitkara University,
Chandigarh, India
e-mail: Vidhyotma.gandhi@chitkara.edu.in

J. Singh
e-mail: jaiteg.singh@chitkara.edu.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_8

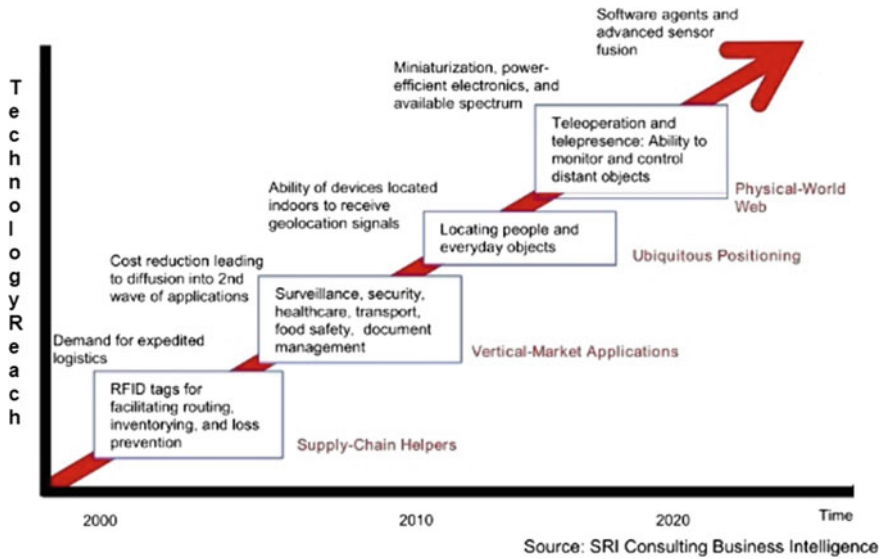


Fig. 1 Road map of the Internet of Things [35]

in daily routine life [5]. The IoT has added the latest dimension to this procedure by allowing interactions within smart objects. The omnipresence of communication network connects the objects “anytime, anyplace, any medium (wired or wireless) and anything or human” [6]. Figure 2 represents the diversified connectivity of IoT.

The IoT is a novel archetype which is hastily achieving height due to the recent advancement in wireless communications. The simple idea behind IoT is the prevalent existence of various objects or things with tags of Radio-Frequency Identification (RFID), different kinds of sensors, mobile phones, actuators, and so on as given in Fig. 3 [7].

It is implemented by assigning the address to the objects or human using any addressing scheme [5]. Different available technologies can be used to implement IoT. The basic structure of the IoT is based on smart devices, i.e., objects built with communication capabilities of Machine-to-Machine (M2M) [8]. The object can be of any type. It is necessary to assign an IP address to every device/object to provide the ability to transfer or receive data over the network [9]. Huge availability of IPV6s addresses is the major factor in the development of IoT technology. These smart things are able to make communication among them as well as to interact with the surrounding environment by interchanging, information, and data sensed about the environment [10]. During responding independently to the events of the real world and affecting them by the running processes which excite operation and allocate services along with or without the straight human intervention [11]. Actually, the IoT is a network of uniquely addressable interlinked objects, which is worldwide established on the low-power communication protocols [12]. IoT paradigm is represented in Fig. 4. The process consists of a large number of heterogeneous objects

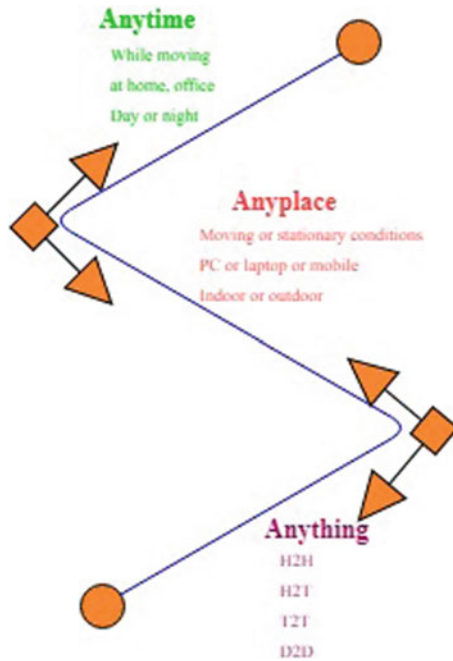


Fig. 2 Connectivity of IoT [36]

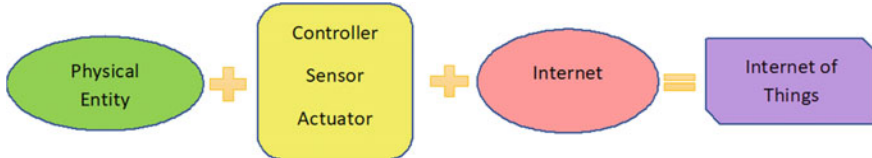
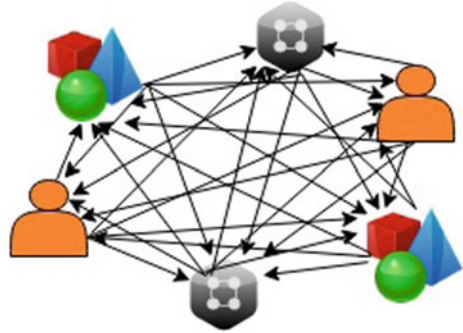


Fig. 3 Internet of Things

[13]. Any wireless communication technology can be used to communicate within objects but it is very costly to provide each object with its personal id on the web [3]. The general IoT deployment architecture consists of four layers, but the underneath and intermediate technologies of each layer vary according to the application. IoT can be implemented with different technologies like Wireless Private Area Network (WPAN), RFID, IPV6, etc. and has diversified application areas.

This paper is divided into seven sections. The introduction of IoT was given in Sect. 1. Section 2 describes the architecture of IoT and Sect. 3 explores the technologies used in IoT. Sections 4 and 5 discuss the applications of IoT and quality of services, respectively. Section 6 concludes the paper and Sect. 7 leads to the future scope in the field of IoT.

Fig. 4 Internet of Things paradigm [18]



2 Architecture of IoT

IoT works in the bottom to top approach. Sensors gather data using the smart devices, process this data using controllers to take decisions, and then communicate to the devices or persons for implementation [14]. User or applications are in the topmost layer and the technology, assigning addresses to smart objects, and communication media used are in the bottom layer. Smart objects, sensors, and software like micro-controllers used to process data are in the middle layer. The architecture of IoT consists of four layers as shown in Fig. 5. The functionalities of each layer have been briefly discussed below.

2.1 Application Layer

It is the user interface in the top of the architecture which is responsible for delivering several applications to the different users or industries in IoT [5]. Increasing usage of RFID technology in numerous applications increases the scope of IoT [7]. This application can belong to different industry verticals such as logistics, manufacturing, public safety, environment, food and drug, health care, retail marketing chain, and so on [5]. Session generation for every user and designing of the user interface for each application is also the part of this layer.

2.2 Edge/Transport Layer

The edge/transport layer utilizes the interface between bottom hardware surface and top application surface [12]. It is responsible for critical function that is device management as well as information management. It has responsibilities of all the issues like filtering and preprocessing of data, access control, analyzing of semantic, discovery of information, i.e., Electronic Product Code (EPC), information facility data analytics, and Object Naming Service (ONS). Middleware surface works in bidirectional mode [15].

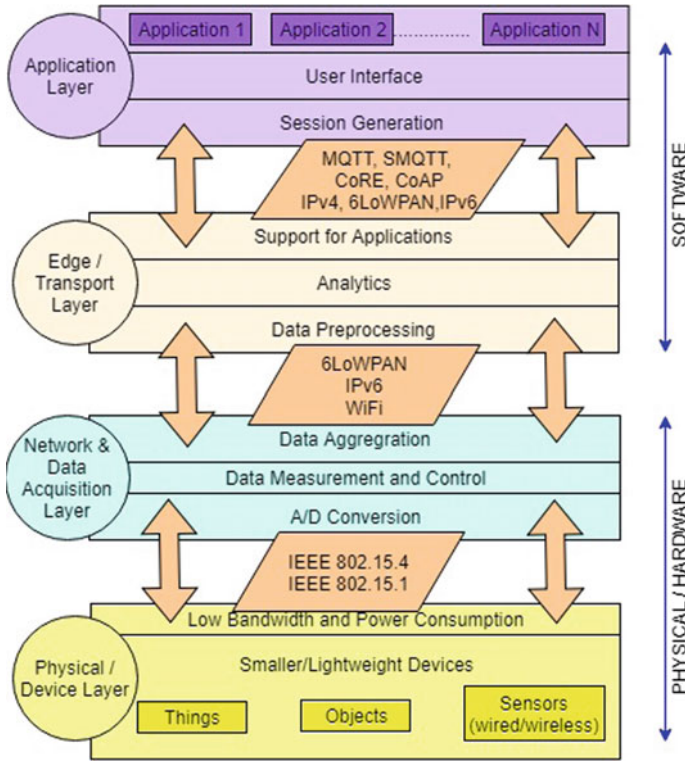


Fig. 5 Layered architecture of IoT

2.3 Network and Data Acquisition Layer

This layer is responsible for publishing, message routing, and subscribing, and based on the requirement it can perform the cross-platform communication. Data handling and conversion happen in the first stage of this layer [12]. Data measurement, aggregation, and control are the major functions of this layer.

2.4 Physical/Device Layer

Physical layer consists of sensors/constrained devices (sensors/actuators). According to Internet Engineering Task Force (IETF) publication RFC7228 constrained devices can be classified into three classes and are shown in Table 1 [16].

Sensors can be wearable or implantable and can record continuous or discrete signals. Figure 6 shows the categorization of different types of physiological sensors. RFID tags and readers, sensor networks, embedded systems, or other sensors form device layer [10]. The layer is responsible for primary data generated by these

Table 1 Constrained devices according to IETF (RFC7228)

	Class 0	Class 1	Class 2
RAM (KB)	10 (less than)	~10	~50
Flash (KB)	100	~100	~240
Protocols	MQTT, CoAP, EXI	UDP, CoAP, TLS, DTLS, HTTP	All protocol stacks
Cost	Low	Low	High

sensors, things, or objects. RFID tags and sensor networks provide identification. Embedded systems are used to collect the information, to store information, processing of information, actuation, control, and communication [17].



Fig. 6 Classification of sensors

3 Technologies Used in IoT

IoT can be implemented with the integration of several technologies. As the name reflects, IoT is the technology to make information or data available on the Internet to remotely control the real-time application [18]. The IoT syntactically consists of two terms; the first one depicts the vision of IoT, i.e., “network-oriented,” and the other focuses on the common “objects” to be merged into a regular structure [3]. The actual definition of IoT is obtained from a “Objects oriented” [8]. The underneath technologies used in IoT are given below.

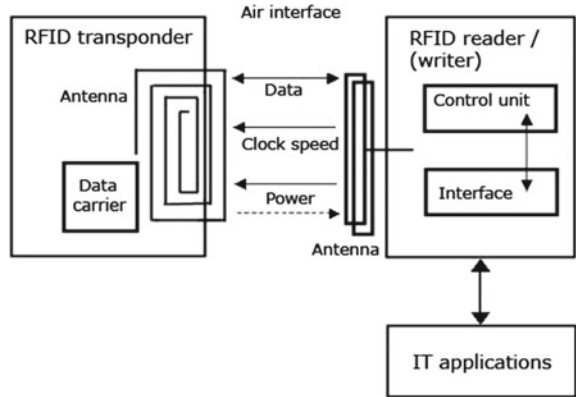
3.1 Wireless Private Area Network (WPAN)

The WSN can be used to create wireless private area network. These can be RF modules, Zigbee modules [9]. WPAN follows the IEEE 802.15.4 standards. It works in three frequency bands 868, 915, and 2.4 GHz. The parameter details of these frequency bands are given in Table 2 [19, 20]. In WPAN, devices are denoted as nodes. It can sense surroundings and communicate data using wireless connections to and from the monitored field [21]. The WPAN can control the environment and makes it able to interact with computers or human beings/different nodes and the surrounding environment [22]. Possibly, multiple hops forward data to the controller functioning as a sink node which can use the data locally or it can be linked to the other networks [22]. The nodes might be mobile or immobile and can be heterogeneous or homogeneous [23].

Table 2 WPAN frequency bands

Bands	Parameter							
	No. of channels	Frequency spectrum	Modulation scheme	Symbol rate/s	Symbol type	Chips/s	Unlicensed	FCC band
868 MHz	1	868 MHz	BPSK	20	Binary	0.3 Kb	Europe	–
915 MHz	10	902–928 MHz	BPSK	40	Binary	0.6 Kb	US	Industrial
2.4 GHz	16	2.40–2.48 GHz	O-QPSK	250	16-bit array	2 Kb	Worldwide	Scientific

Fig. 7 Working within “RFID” technology



3.2 Radio-Frequency Identification (RFID)

RFID is a wireless mechanics that use identification code to scan or recognize any person, animal, substance, product, car, or any other object that contains a tag [24]. This technology is the combination of three articulates: tag, reader, and antenna. A tag is known as a transponder, which has a printed circuit board and/or semiconductor. A reader is further known as a transceiver, which decodes existing information by the transponder or tag. Usually, an antenna is established on the tag, which receives or transmits the power in the form of the radio frequency. The data is transmitted through antenna and reader receives the data [6]. RFIDs are available in different frequency ranges and are chosen according to the applications as shown in Fig. 7 [7, 22].

- Low-density up to 30 MHz → up to 0.01 m for Close-joint
 - Lower than 135 KHz → up to 1 m 500 cm
 - 13.56 MHz(closeness) → 0.1 m
 - 13.56 MHz(environs) → 0.5–3 m
 - 433, 868 or 915 MHz → 0.5–50 m
 - 2.45 GHz → 10–100 m
 - 5.8 GHz → 10–1000 m
 - (Under expansion)
- } Remote Joint
- } Long Range

3.3 Internet Protocol Version 6 (IPv6)

IPv6s huge increase in address space played a major role in the development and growth of IoT. According to the research community, the address space expansion means that every atom in the universe can be assigned an IPV6 address and still

100+ piles of earth addresses are left to assign [25]. Samsung Electronics, Silicon Labs, Nest Labs, Yale Security, Free scale Semiconductor Big Ass Fans, and ARM have joined hands together to form the “Thread Group” to enhance the application capabilities of “Internet of Things” connectivity [26]. The thread will avail the maximum outcome of the number of existing standards including 6LoWPAN (Little-capability wireless personal area networks), IETF IPv6, and IEEE 802.15.4 [26]. It is already used inside the nest-based devices, i.e., Wi-Fi makeable smoke detectors and thermostats made by Google [24, 27]. US National Power NIC predicts “by 2025; Internet nodes may reside in everything in daily life like vehicles, paper documents, furniture, food packages, medicines, and more” [2].

4 Applications of IoT

Potentiality of applying the IoT in different domains like industry, domestic, retailing, defense, health, education, and other various fields is very high; rather in some domains it is already in use and research is going on. Few industrial and domestic applications that come under the umbrella of IoT are given below.

4.1 Retail Supply Chain

The retail supply chain is managed well with the help of RFID identification of goods and the smart-shelf concept is very useful in placing and picking up the things in retail stores [3]. The owner is always aware of stock status and sales and can easily maintain the purchase orders for goods.

4.2 Medical and Health

Wearable sensors can be used to locate, in the hospital, both doctors and nurses along with the patients at any point in time [26]. Some vital functions of the human body, i.e., cholesterol levels, sugar level, blood pressure, temperature, and the heartbeat rate can be monitored through the IoT system. In case of a heart attack, stimulate the heart muscle through the combination of the sensors implantation. Data can be recorded wirelessly using IoT technology [28].

4.3 Aid for Aged and Physically Disabled Persons

IoT is very beneficial for an aged and physically disabled person using wearable smart identity sensors. These wireless sensors transmit the health status of a person and can also alarm for help in case of an emergency condition [29].

4.4 Transport

RFID technology is already in use in vehicles production units and traffic control systems. Luxury vehicle, i.e., buses, cars, trains, and airplanes are equipped with actuators with increased processing powers and advanced sensors which help in reporting pressure in tires, the location of a vehicle. Deployment of IoT increases quality control, improves logistics, and improves customer services [15].

4.5 Manufacturing

In manufacturing units, IoT is used to remotely optimize production processes and keep the record of manufactured items on the daily, monthly, or yearly basis to calculate the turnover of a company using unique identity for each product. Ingredient ratio in a product can also be tracked through sensors [3].

4.6 Safety, Security

The home safety and security are must in today's life. There are lots of wireless sensors which are integrated with the IoT to provide safety by providing information on the Internet through a mesh network. One can remotely control the power system, locking system, and can also be aware of the water status of own home [30].

4.7 Automation

Internet of Things is playing a major role in the embedded industry. Smart homes, smart cities, and industry 4.0 are the main application areas of automation [17]. Low-power and cost-effective embedded devices are developed with the help of smart sensors which create the mesh network and transmit data on the Internet [31].

4.8 *Pharmaceutical*

In pharmaceutical companies, it is the time of demand to make customer's life risk free. It can be possible by IoT smart labeling technique. By providing smart labels to drugs it will be helpful to keep track of its quality, quantity, and also aware the patient about its storage and dosages details [32].

Few IoT applications along with their IoT properties are given in Table 3.

5 **Quality of Services (QoS)**

The IoT has an excess of applications, resources, and the network of components which inherently consist of a complex and shared system. Lots of devices in it are dynamic and heterogeneous in terms of energy, communication, and computation. Due to this, many IoT applications are critical in nature. This motivates to provide QoS across multiple dimensions [21]. Network and resource providers of IoT application connect them carefully so that multiple computing applications can coexist in it. The environmental attributes like temperature and location, the characteristics of the network like latency, bandwidth, power, and battery life play a vital role in application's QoS. IoT applications need responsiveness and precision smart objects are active members of IoT to collaborate and to communicate with each other by interchanging information of sensed data and act as a self-ruling to the physical world. It creates the services with or without interacting with any person [21]. Different QoS models are required for different IoT applications to limit obligatory factors to satisfy the needs of those services.

WSN is a major component of the IoT, QoS provided by optimizing the resource utilization by joining multiple sensor nodes with the global Internet. This protocol provides the best quality of control in the applications of IoT [33].

6 **Conclusion**

This paper concludes that "IoT" is an emerging technology. The integration of embedded systems and the Internet paved the researchers in new directions and the technology was developed. Use of wireless sensors to create the mesh of network and send data or information over the Internet enhances the utility of IoT in every field. The general architecture of IoT is given in this paper and can vary from application to application. To enhance the interoperability between human, thing, and machine still lots of work are needed to be done on its quality of services, communication protocols, and standardized architecture of "Internet of Things."

Table 3 IoT applications and their properties

	Retail supply chain	Medical and health	Aid for aged and physically disabled person	Transport	Manufacturing	Automation
Network size	Small	Small	Small	Big	Medium	Medium
Users	Region based	Region based	Region based	Region based	Within industry	Within premises
Energy	Rechargeable	Rechargeable	Rechargeable	Rechargeable	Rechargeable	Rechargeable
Internet	Wi-Fi, GSM	Wi-Fi, GSM	Wi-Fi, GSM	Wi-Fi, GSM	Wi-Fi, GSM	Wi-Fi, GSM
Topology	Bus	Star	Star		Mesh	Bus
Data management	Local server	Shared server	Shared server	Global	Local	Local
IoT devices	RFID, WSN	WBSN	WBSN	RFID, WSN	RFID, WSN	RFID, WSN, WBSN
Bandwidth	Small	Medium	Medium	Large	Medium	Small

7 Future Scope

The architecture of IoT is still application dependent and, similarly, protocols of communication are medium specific [5]. Both areas still have lots of research scope for development in future. Another area is to develop the technologies to revocable pseudonymity. Anonymity addresses the global identity schemes, encryption/encoding, and repository authority with the help of recognition, authentication of parties, authentication, and addressing plan as well as building the discovery services and global directory lookup services. In the field of robotics, IoT should be used to design collaborative robots [34].

References

1. Maney, K.: Meet Kevin Ashton, Father of the Internet of Things (2015) (Online). <http://www.newsweek.com/2015/03/06/meet-kevin-ashton-father-internet-things-308763.html>. Accessed 21 Dec 2017
2. Ashton, K.: That ‘Internet of Things’ Thing—2009-06-22—Page 1. RFID J. (2009) (Online). <http://www.rfidjournal.com/articles/view?4986>. Accessed 14 Dec 2017
3. Ibarra-Esquer, J.E., González-Navarro, F.F., Flores-Rios, B.L., Burtseva, L., Astorga-Vargas, M.A.: Tracking the evolution of the internet of things concept across different application domains. *Sensors* **17**(6), 1379 (2017)
4. Adibi, S.: *Mobile Health: A Technology road Map*. Springer, Cham (2015)
5. Said, O., Masud, M.: Towards Internet of Things: survey and future vision. *Omar Said Mehedi Masud Int. J. Comput. Netw.* **5**(1), 1 (2013)
6. Atzori, L., Iera, A., Morabito, G.: The Internet of Things: A survey. *Comput. Netw.* **54**(15), 2787–2805 (2010)
7. Welbourne, E., et al.: Building the Internet of Things using RFID. *IEEE Internet Comput.* **13**(3), 48–55 (2009)
8. Kortuem, G., Kawsar, F., Sundramoorthy, V., Fitton, D.: Smart objects as building blocks for the Internet of things. *IEEE Internet Comput.* **14**(1), 44–51 (2010)
9. Buratti, C., Conti, A., Dardari, D., Verdone, R.: An overview on wireless sensor networks technology and evolution. *Sensors* **9**(9), 6869–6896 (2009)
10. Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S.: Vision and challenges for realising the Internet of Things. The meaning of things lies not in the things themselves, but in our attitude towards them. Antoine de Saint-Exupéry (2010)
11. O’Neill, S.: *Generalized Hyperalgesia in Chronic Low-Back Pain*. River Publishers (2013)
12. Ray, P.P.: A survey on Internet of Things architectures. *J. King Saud Univ. Comput. Inf. Sci.* (2016)
13. Fortino, G., Russo, W., Savaglio, C., Viroli, M., Zhou, M.: Modeling opportunistic IoT services in open IoT ecosystems. *Comput. Sci. Eng.* **19**(5), 68–76 (2017)
14. Celandroni, N., et al.: A survey of architectures and scenarios in satellite-based wireless sensor networks: system design aspects. *Int. J. Satell. Commun. Netw.* **31**(1), 1–38 (2013)
15. Bandyopadhyay, D., Sen, J.: Internet of Things: applications and challenges in technology and standardization. *Wirel. Pers. Commun.* **58**(1), 49–69 (2011)
16. Zurich Tian, E.H.: Energy-efficient features of Internet of Things protocols (2018)
17. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of Things (IoT): a vision, architectural elements, and future directions. *Future Gener. Comput. Syst.* **29**(7), 1645–1660 (2013)
18. Chowdhury, M.N., Bhuiyan, M.M.H., Islam, S.: IOT: detection of keys, controlling machines and wireless sensing via mesh networking through internet. *Glob. J. Res. Eng.* (2013)

19. Karl, H., Willig, A.: *Protocols and Architectures for Wireless Sensor Networks*. Wiley (2005)
20. Hersent, O., Boswarthick, D., Elloumi, O.: *The Internet of Things: Key Applications and Protocols*. Wiley InterScience (Online service), Wiley (2012)
21. Nef, M., Perlepes, L., Stamoulis, G.I.: Enabling QoS in the Internet of Things. In: *CTRQ 2012 : The Fifth International Conference on Communication Theory, Reliability, and Quality of Service*, no. c, pp. 33–38 (2012)
22. Madakam, S., Ramaswamy, R., Tripathi, S.: Internet of Things (IoT): a literature review. *J. Comput. Commun.* **3**(3), 164–173 (2015)
23. Watt, J.H., Lynch, M.: Using the Internet for audience and customer research. In: *IPCC 99. Communication Jazz: Improvising the New International Communication Culture*. Proceedings 1999 IEEE International Professional Communication Conference (Cat. No. 99CH37023), pp. 121–130
24. Sole, M., Musu, C., Boi, F., Giusto, D., Popescu, V.: RFID sensor network for workplace safety management. In: *2013 IEEE 18th Conference on Emerging Technologies & Factory Automation (ETFA)*, pp. 1–4 (2013)
25. Miorandi, D., Sicari, S., De Pellegrini, F., Chlamtac, I.: Internet of Things: vision, applications and research challenges. *Ad Hoc Netw.* **10**(7), 1497–1516 (2012)
26. Darwish, A., Hassanien, A.E.: Wearable and implantable wireless sensor network solutions for healthcare monitoring. *Sensors* **11**(6), 5561–5595 (2011)
27. Kerner, S.M.: The Internet of Things thread network protocol debuts (2014) (Online). <http://www.enterprisenetworkingplanet.com/netsp/the-internet-of-things-thread-network-protocol-debuts.html>. Accessed 15 July 2014
28. Gravina, R., et al.: Enabling multiple BSN applications using the SPINE framework. In: *2010 International Conference on Body Sensor Networks (BSN)*, pp. 228–233. IEEE (2010)
29. Atallah, L., Lo, B., Yang, G.-Z., Siegemund, F.: Wirelessly accessible sensor populations (WASP) for elderly care monitoring. In: *2008 Second International Conference on Pervasive Computing Technologies for Healthcare*, pp. 2–7 (2008)
30. Jacobsson, A., Boldt, M., Carlsson, B.: A risk analysis of a smart home automation system. *Future Gener. Comput. Syst.* **56**, 719–733 (2016)
31. Ghasemzadeh, H., Fallahzadeh, R., Jafari, R.: A hardware-assisted energy-efficient processing model for activity recognition using wearables. *ACM Trans. Des. Autom. Electron. Syst.* **21**(4), 1–27 (2016)
32. Koop, C.E., et al.: Future delivery of health care: Cybercare. *IEEE Eng. Med. Biol. Mag.* **27**(6), 29–38 (2008)
33. Cao, H., Leung, V., Chow, C., Chan, H.: Enabling technologies for wireless body area networks: a survey and outlook. *IEEE Commun. Mag.* **47**(12), 84–93 (2009)
34. Industrial Internet of Things: Cobots and Connectivity (Online). <http://www.roboglobal.com/industrial-internet-of-things-cobots-and-connectivity>. Accessed 19 Dec 2017
35. A roadmap for the Internet of Things/Internet of objects (Online). <https://internetofobjects.wordpress.com/2011/03/31/a-roadmap-for-the-internet-of-things/>. Accessed 19 Dec 2017
36. Digital Lifestyle Malaysia—Internet of Things—overview (Online). <http://dlm.skmm.gov.my/Internet-of-Things/Overview.aspx>. Accessed 19 Dec 2017

High-Speed Optical Mode Division Multiplexing of Hermite–Gaussian Modes in Multimode Fiber



Saumya Srivastava, Kamal K. Upadhyay and Nar Singh

Abstract Mode division multiplexing (MDM) is an auspicious technology for reducing the congestion of network traffic and accomplish the future demands of the network infrastructure. Multimode fiber has the amazing data growth, multiplexing, and modulating of the data which have been distributed in terms of phase, intensity, wavelength, time domains, and phase. This paper designs the MDM of spiral phased Hermite–Gaussian modes in multimode fiber. The performance of the system is evaluated by total received power, Q-factor, and eye-diagram at different input power and the range of the fiber length 20 km distance is achieved at 10 Gbps data rate with acceptable bit error rate.

Keywords Mode division multiplexing (MDM) · Hermite–Gaussian modes (HG) · Multimode fiber (MMF) · Q-Factor

1 Introduction

Many industrial and academic members have developed a new technology for high data rate transmission and increase the capacity of optical fiber network. The solution for it is the use of mode division multiplexing (MDM) with the multimode fiber (MMF) which is a promising approach to increase the capacity of the system and fulfill the future demands. Many parallel data are transported at a time by the use of modes of MMF which is a new technology of MDM [1].

S. Srivastava (✉) · K. K. Upadhyay · N. Singh
Department of Electronics and Communication, University of Allahabad, Allahabad,
Uttar Pradesh, India
e-mail: srisaumya1088@gmail.com

K. K. Upadhyay
e-mail: kamal.kishoresiet@gmail.com

N. Singh
e-mail: nsjk53@rediffmail.com

In MDM, independent data signals are transmitted through single or group of modes in MMF and control the total amount of coupled power into each modes and delay of propagating modes [2]. Each channel of existing modes is multiplexed by the incident signals at MMF and at the receiver, and the signals are distributed by modal field matching so each channel has optimized impulse response. MDM has been exhibited in optical fiber communication such as spatial light modulators to generate mode profiles [3–6], photonic crystal fiber [7], few mode fiber [8, 9], optical signal processing [10–12], fiber Bragg grating [13, 14], and offset launch techniques [15–17]. MDM has a chance to outperform to the product of bandwidth-distance and barriers of spectral efficiency [1].

Various types of modes had been explored for MDM. Spatial light modulators and fiber Bragg grating generate the Laguerre–Gaussian modes [18]. Hermite–Gaussian modes are developed by silica of fused substrate and passive beam shaper is formed [19–21]. For donut beam generation, phase plates [22], etched fiber [23], and deflecting mirrors [24] had been accepted.

In this paper, we present on the MDM of a unique combination of a spiral phased Hermite–Gaussian modes with MMF and the MDM system performance is evaluated at different input power and fiber length. The rest of the paper is proceeds as in Sect. 2 is system design where MDM-MMF system for Hermite–Gaussian modes. Result and discussion are discussed in Sect. 3 and finally the paper is concluded in Sect. 4.

2 System Design

Figure 1 featured the MDM of Hermite–Gaussian mode in MMF that was designed in OptiSystem™ software. The proposed architecture comprises of three non-return to zero (NRZ) data are encoded and modulated with spatial laser at different input power and generate the three Hermite–Gaussian modes HG00, HG01, and HG10 at 10 Gbps data rate. Mathematical description of HG mode is given as [25]:

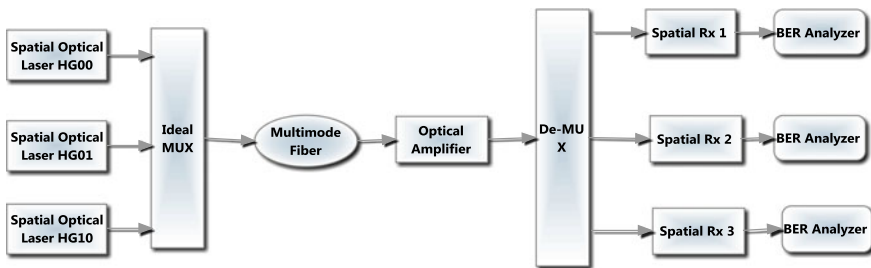


Fig. 1 The proposed HG-based MDM-MMF System

Table 1 Parameters used in this setup

Parameters	Values
Input power	0–40 dBm
Transmission wavelength	1550 nm
Multimode fiber length	5–20 km
EDFA length	8 m
Pump power 1 & 2	150 and 75 mW
Pump wavelength	980 nm
Responsivity of PIN photodetector	1 A/V
Dark current of photodiode	10 nA

$$\begin{aligned} \psi_{m,n}(r, \phi) = & H_m \left(\frac{\sqrt{2}x}{w_{0,x}} \right) \exp \left(-\frac{x^2}{w_{0,x}^2} \right) \exp \left(j \frac{\pi x^2}{\lambda R_{ox}} \right) \\ & \times H_n \left(\frac{\sqrt{2}y}{w_{0,y}} \right) \exp \left(-\frac{y^2}{w_{0,y}^2} \right) \left(j \frac{\pi y^2}{\lambda R_{oy}} \right) \end{aligned} \quad (1)$$

From the above equation, mode dependencies on x -axis and y -axis are m and n . Radius of curvature is R and size of spot is w_0 . H_m and H_n are the Hermite polynomials.

The output these channels is multiplexed and transmitted to the MMF. After the fiber, post amplifier is used to amplify the transmitted signal. At the receiver side, the transmitted modes are retrieved through non-interferometric modal decomposition [26]. The output mode is transmitted to a spatial PIN detector. To retrieve the original baseband signal, low-pass Gaussian filter is used. The parameters of the proposed design are shown in Table 1.

3 Results and Discussion

This section presents the proposed HG MDM based MMF system and discusses the findings from it. Performance of the system is evaluated from the total received power and quality-factor at the different input power and the length of fiber is varied.

From Fig. 2, total output power is calculated at the length of the fiber is 5, 10, 15, and 20 km for different channels, i.e., HG00, HG01, and HG10. For the channel 1, the received output power is 15.25 dBm for channel 2 13.24 dBm and 11.08 dBm for channel 3 at 5 km fiber length. If the length of fiber increases then the total received power decreases linearly for all the three channels. At 10 km fiber length, the output power is 14.35, 12.08, and 9.32 dBm for channel 1, 2, and 3, respectively. Increases the fiber length to 20 km, the total received power is 5.02 dBm for channel 1, 4.21 dBm for channel 2 and -2.226 dBm for channel 3.

Fig. 2 Output power versus fiber length

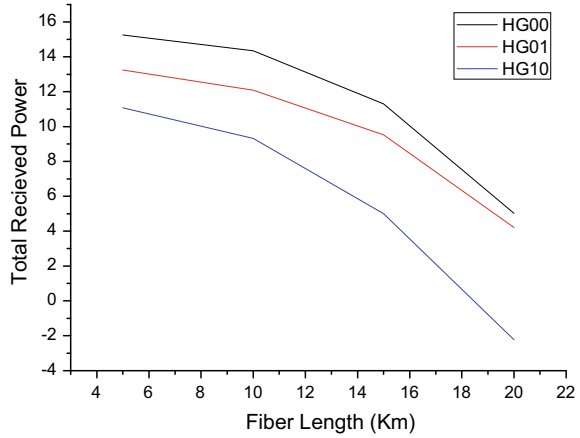
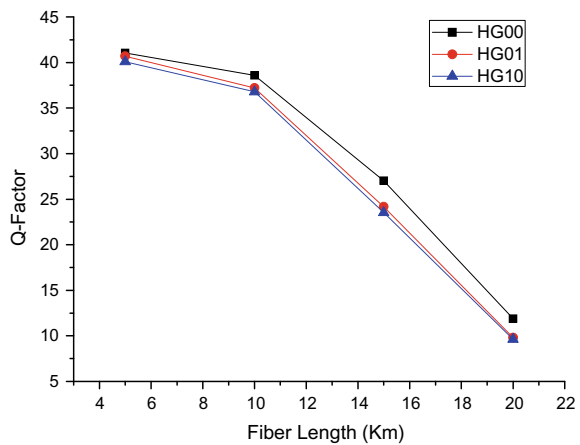


Fig. 3 Q-factor versus fiber length



As shown in Fig. 3, the value of quality-factor is analyzed with the length of fiber for different channels. For channel 1, Q-factor value is 41.05, for channel 2, it is 40.69 and for channel 3, the value of Q-factor is 40.09 at 5 km fiber length. Increases the fiber length, the Q-factor value is decreases linearly. So at 15 km fiber length, Q-factor is 27.02 for HG00, 24.18 for HG01, and 23.53 for HG10.

Figures 4 and 5 show the total received power and Q-factor at the different power. Increasing the input power, the value of output power and Q-factor are also increases for all three channels. At the 0 dBm input power, output power is 4.16 dBm received and Q-factor value is 10.43 for channel 1, 3.41 dBm output power is received and Q-factor is 8.68 for channel 2. At the 20 dBm input power, the total received power is 14.52 dBm and the value of Q-factor is 39.01 for channel 1. So again change the input power to 40 dBm, output power is goes to 16.09 dBm and Q-factor value is increases to 41.25.

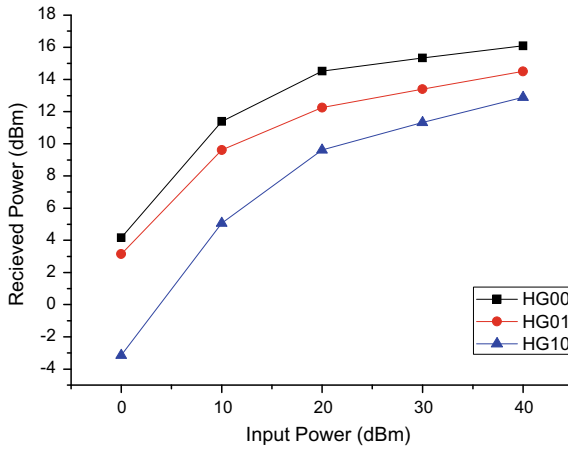


Fig. 4 Output power versus input power

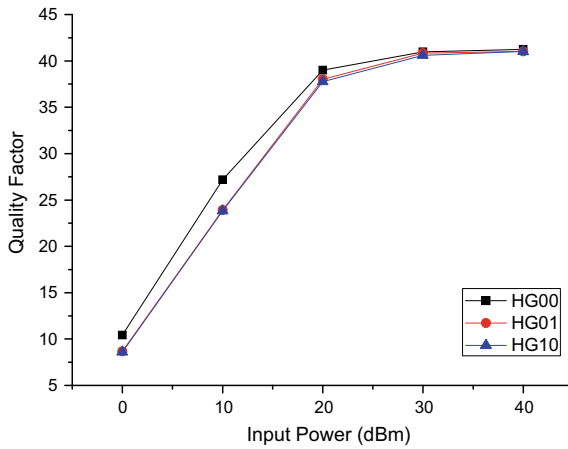


Fig. 5 Q-factor at different input power

Figures 6, 7, and 8 show the eye-diagram of channel 2 at different input power. The eye-diagram is not clear if the input power is 0 dBm is shown in Fig. 6. From Fig. 7, if the input power is changed to 20 dBm, the eye-diagram is much clear as compared to the previous input power. Again change the input power to 40 dBm the eye-diagram is more clear and wide in it among all the cases is shown in Fig. 8. So it shows the successful transmission of the data through MDM with HG modes.

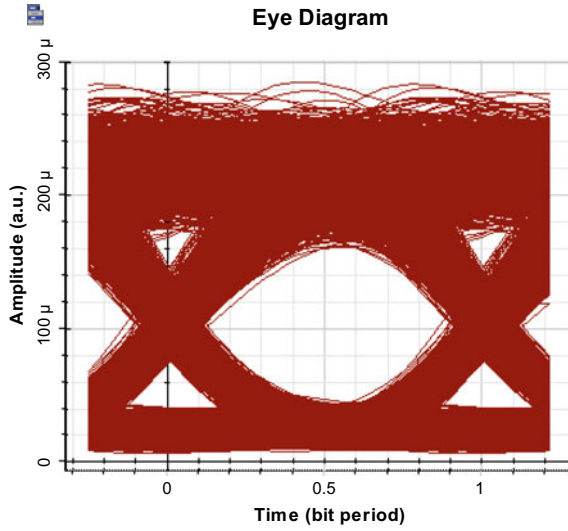


Fig. 6 Eye-diagram at 0 dBm input power

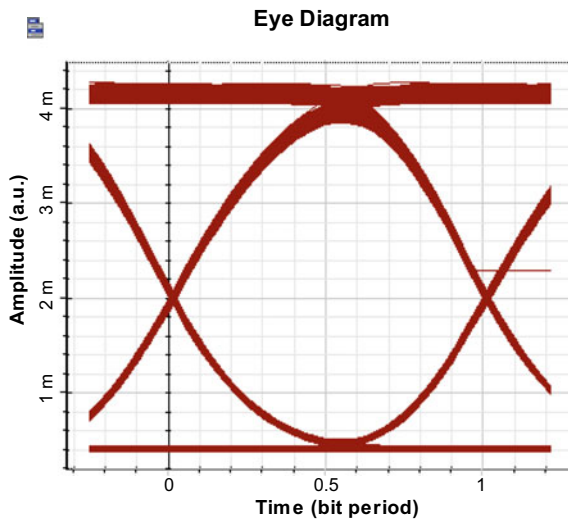
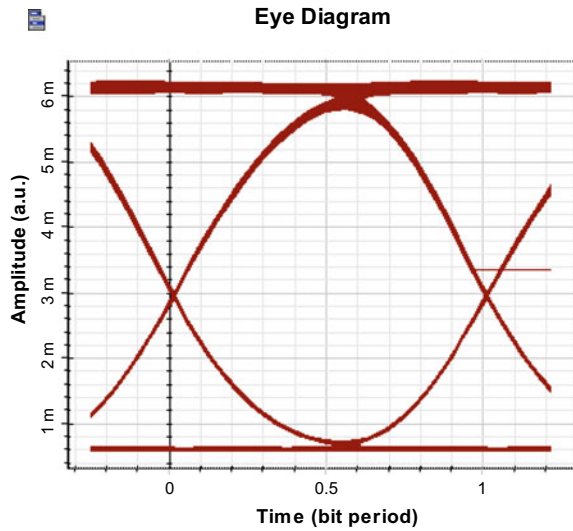


Fig. 7 Eye-diagram at 20 dBm input power of channel HG01

Fig. 8 Eye-diagram of Channel 2 (HG01) at 40 dBm input power



4 Conclusion

A new optical MDM with Hermite–Gaussian modes in MMF was designed. HG modes were used for MDM. HG00, HG01, and HG10 channels are used to analyze the system performance by the performance parameter which are quality-factor and output power at different input power and fiber length. The output power is measured 16.09 dBm at 40 dBm input power and 5.02 dBm output power is calculated at the length of fiber which is 20 km for channel HG00. The value of Q-factor is computed 41.03 at 40 dBm input power and the Q-factor is determined to 9.79 at 20 km fiber length for channel 2. Wide eye opening was achieved which shows the successful transmission of the data.

References

1. Bozinovic, N., Yue, Y., Ren, Y., et al.: Terabit-scale orbital angular momentum mode division multiplexing in fibers. *Science* **340**(6140), 1545–1548 (2013)
2. Amphawan, A.: Review of optical multiple-input–multiple-output techniques in multimode fiber. *Opt. Eng.* **50**(10), 102001-102001-6 (2011)
3. Amphawan, A.: Holographic mode-selective launch for bandwidth enhancement in multimode fiber. *Opt. Exp.* **19**(10), 9056–9065 (2011)
4. Amphawan, A.: Binary encoded computer generated holograms for temporal phase shifting. *Opt. Exp.* **19**(23), 23085–23096 (2011)
5. Amphawan, A.: Binary spatial amplitude modulation of continuous transverse modal electric field using a single lens for mode selectivity in multimode fiber. *J. Mod. Opt.* **59**(5), 460–469 (2012)

6. Amphawan, A., Brien, D.O.: Holographic mode field generation for a multimode fiber channel. In: IEEE International Conference on Photon 2010 (ICP2010), pp. 1–5 (2010)
7. Amphawan, A., Nedniyom, B., Samman, N.M.: AI: selective excitation of LP₀₁ mode in multimode fiber using solid-core photonic crystal fiber. *J. Mod. Opt.* **60**(20), 1675–1683 (2013)
8. Jung, Y., Chen, R., Ismael, R., Brambilla, G., Alam, S.-U., Giles, I., et al.: Dual mode fused optical fiber couplers suitable for mode division multiplexed transmission. *Opt. Express* **21**, 24326–24331 (2013)
9. Tsekrekos, C.P., Syvridis, D.: All-fiber broadband mode converter for future wavelength and mode division multiplexing systems. *IEEE Photon. Technol. Lett.* **24**, 1638–1641 (2012)
10. Arik, S.O., Askarov, D., Kahn, J.M.: Adaptive frequency-domain equalization in mode-division multiplexing systems. *J. Lightwave Technol.* **32**(10), 1841–1852 (2014)
11. Shemirani, M.B., Wilde, J.P., Kahn, J.M.: Adaptive compensation of multimode fiber dispersion by control of launched amplitude, phase, and polarization. *J. Lightwave Technol.* **28**(18), 2627–2639 (2010)
12. Amphawan, A., Mishra, V., Nedniyom, K.N.B.: Real-time holographic backlighting positioning sensor for enhanced power coupling efficiency into selective launches in multimode fiber. *J. Mod. Opt.* **50**(20), 1745–1752 (2012)
13. Jiangli, D., Kin Seng, C.: Mode-locked fiber laser with transverse-mode selection based on a two-mode FBG. *Photon. Technol. Lett. IEEE* **26**(17), 1766–1769 (2014)
14. Yam, S.S.-H., Gu, X., Mohammed, W., et al.: Multimode fiber Bragg grating wavelength filter in a 10-Gb/s system. *IEEE Photon. J.* **20**(8), 584–586 (2008)
15. Carpenter, J., Wilkinson, T.D.: Holographic offset launch for dynamic optimisation and characterisation of multimode fibre bandwidth. *J. Lightwave Technol.* **30**(10), 1437–1443 (2012)
16. Amphawan, A., Payne, F., O'Brien, D., et al.: Derivation of an analytical expression for the power coupling coefficient for offset launch into multimode fiber. *J. Lightwave Technol.* **28**(6), 861–869 (2010)
17. Carpenter, J., Wilkinson, T.D.: Adaptive enhancement of multimode fibre bandwidth by twin-spot offset launch. In: Conference on Lasers and Electro-Optics (CLEO), pp. 250–252 (2011)
18. Franz, B., Buelow, H.: Spatial multiplexers and demultiplexers for mode group division multiplex. In: 2013 15th International Conference on Transparent Optical Networks (ICTON), pp. 1–4 (2013)
19. Kwok, C., Penty, R.V., White, I.H., et al.: Novel passive launch scheme for ultimate bandwidth improvement of graded-index multimode fibers. In: Optical Fiber Communication Conference, pp. 1–3 (2010)
20. Geng, L., Kwok, C., Lee, S., et al.: Efficient line launch for bandwidth improvement of 10 Gbit/s multimode fibre links using elliptical Gaussian beam. In: 36th European Conference and Exhibition on Optical Communication (ECOC), pp. 1–3 (2010)
21. Li, Y., Ingham, J.D., Olle, V., et al.: 20 Gb/s mode-group-division multiplexing employing Hermite-Gaussian launches over worst-case multimode fiber links. In: Optical Fiber Communication Conference, W2A. 3 (2014)
22. Chen, J., Kuang, D.-F., Gui, M., et al.: Properties of Fraunhofer diffraction by an annular spiral phase plate for sidelobe suppression. *Chin. Phys. Lett.* **26**(9), 094210–094211 (2009)
23. Mayeh, M., Farahi, F.: Tailoring Gaussian Laser beam shape through controlled etching of single-mode and multimode fibers: simulation and experimental studies. *IEEE Sens. J.* **12**(1), 168–173 (2013)
24. Jones, P.H., Stride, E., Saffari, N.: Trapping and manipulation of microscopic bubbles with a scanning optical tweezer. *Appl. Phys. Lett.* **89**, 081113 (2006)
25. Ghatak, A., Thyagarajan, K.: An Introduction to Fiber Optics. Cambridge University Press, Cambridge (1998)
26. Mahmoud, S.W., Wiedenmann, D., Kicherer, M., Unold, H., Jager, R., Michalzik, R., et al.: Spatial investigation of transverse mode turn-on dynamics in VCSELs. *IEEE Photon. Technol. Lett.* **13**, 1152–1154 (2001)

Part II
Intelligent Computing Techniques

A Novel Algorithm for Automatic Text Summarization System Using Lexical Chain



Ashima Tiwari and Deepak Dembla

Abstract In the field of text classification and information retrieval, the process of text summarization has always been an important aspect. It decreases the size of text and preserves its information content by providing a shorter illustration. This illustration has the major portion of data, i.e., the most vital information and it is no longer than the half of source data. In addition, it is the best solution for information overloading problem as we do not have to scan through each line of long length documents and still receive the foremost important information. For the betterment of this process, we have proposed an algorithm utilizing lexical chain calculation and it is implemented using Eclipse Java Development Tool, enterprise edition for web developers. Along with WordNet API, this method also included the nouns and proper nouns in the computation of lexical chains. MAX tagger is used for part-of-speech tagging and statistical calculations. By the execution of proposed methodology, it is clear that it is a better approach which resulted in better output in terms of (i) Execution time, as compared to the existing algorithm; (ii) Improved matching of words between the human-generated summary and proposed algorithm-generated summary; (iii) Better recall, that are commonly used criteria for summary evaluation.

Keywords Lexical chains · Summary · Part-of-speech tags

A. Tiwari (✉)
JECRC University, C-198, Manu Marg, Tilak Nagar, Jaipur 302004, Rajasthan,
India
e-mail: ashima3018@gmail.com

D. Dembla
Department of IT and Computer Applications, JECRC University, Jaipur,
Rajasthan, India
e-mail: hod.it@jecrcu.edu.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_10

1 Introduction

Text summarization is the method of generating a shorter form of a given source file [1]. This shorter version occupies less space, has the most important information, and saves a lot of time in reading or surfing through any information. It can be helpful in various scientific fields and can facilitate quick indicative notes on any general topic.

Automatic text summarization offers an effective solution to the problem of information overloading, accelerating the surfing process by significantly compressing information content [2]. As a result, this improves productivity and enables the user to read less data and still receive the most relevant information to build solid conclusions. Through the technique of text summarization, the key facts of a text document are absorbed instantly that eases the selection of required information.

The information available on the Internet has grown up to be so huge in quantity that searching for any content leads to either a lot wastage of time by going through each and every document, or the chances are that we miss out important content. Automatic text summarization is an effective solution for filtering the foremost important data in time.

1.1 Text Summarization System

Text summarization systems are generally characterized into extractive and abstractive summarization [3]. Figure 1 shows the classification of text summarization technique.

- **Abstractive summarization:** Abstractive text summarization attempts to understand the basic concept of the documents information and after that, it specifies those

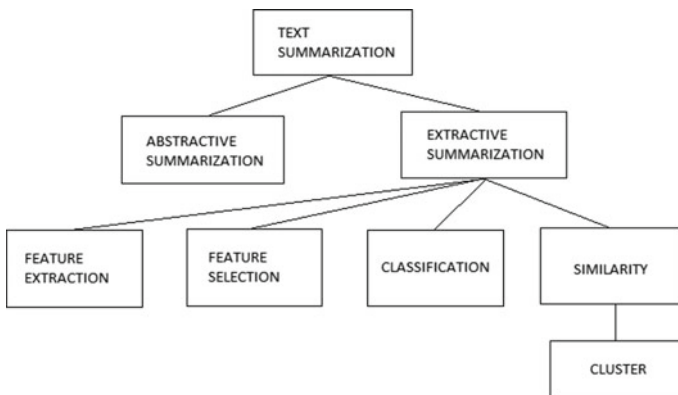


Fig. 1 Classification of text summarization

concepts or ideas in clear natural language as a resultant summarized document. In other words, it comprehends the key facts in each input document to provide the user with a shorter form of underlying information in natural language.

- **Extractive Summarization:** It depends on the analysis of data, done statistically, i.e., how frequently a word is used, its location, etc. This analysis helps in identifying the sentences to be extracted. These extracted sentences work as the key text segments. This type of summarization is conceptually easy and simple to implement.

2 Related Work

Kulkarni et al. [1] used the concept of lexical chains for summarization of text along with the use of correlation between sentences. The summary generated through this method was of better quality as analysis of document is done semantically with the help of lexical chains. Relation of a particular sentence with its preceding and succeeding sentence is considered in generation of summary using correlation of sentences. Keyword strength and other features like tf-idf are used for calculating the score of each chain.

Suman et al. [2] discussed the problem of information overloading and as a solution, and customized algorithms are explained. Instead of going through each sentence, it is better to generate a shorter version or summary. Here, the main objective is to reduce the size of a text and preserving its information content.

Jayarajan et al. [3] presented that better results in summarization process can be obtained if lexical chains are created instead of using bag of words. The comparison between new and old representation is done on a clustering task. An alternative representation is proposed for the document using lexical chains.

Sarkar et al. [4] presented a method for text summarization, which improves the performance of summarization in medical domain. It is shown by the experiments that the summarization performance can be improved by including the domain-specific features along with some other features like title and position, and term features in summarization process.

Patil et al. [5] presented extractive summarization technique for single document. The methods used for single document text summarization here calculate the significance of each sentence in the given document by using the graph based algorithm. Thus, the most significant sentences are used to create the summary.

Ercan et al. [6] presented the use of lexical chains for extraction of keywords to generate key text segments. The supervised learning task in this work is the issue of automatic keyword extraction. The most significant and semantically related words are combined to form the lexical chain or we can say that lexical chains contain the most relevant information of the particular document, which gives empirical results.

Cutting et al. [7] presented hidden Markov model based part-of-speech tagger along with its implementation. Unlabeled training text and a lexicon is used for

the implementation. These few resources are required for accurate part-of-speech tagging and robustness of process.

Barzilay et al. [8] created an algorithm, which uses the concept of combining various knowledge sources that are robust in nature. These sources include a POS tagger, WordNet thesaurus, and shallow parser to identify the nominal groups along with an algorithm for segmentation. It resulted in the betterment of strong chains identification process. In addition, the significance of the sentences chose was improved.

3 Problem Description

From all perspectives, text summarization has an important role in our working fields. Whenever we explore any information, we have to scan through a lot of information and it is tough to gain the significant information quickly [4]. Text Summarization process is the best solution for this information overloading problem and is very less time consuming.

The existing automatic text summarization systems use lexical chains and WordNet to extract the key text segments and generate summary but there are still some limitations that current solutions suffer from, like recall value is very low, that is the ratio of total words matched in the human-generated summary to the total words in given text file. In addition, execution takes too much time as Stanford tagger is used for part-of-speech tagging. In this research, better results are obtained by eliminating these issues. Review of literature is done for better understanding of lexical chains and how they are generated.

In the process of text summarization, the significance of using lexical chains is understood along with its scoring [6]. The use of WordNet is learnt. WordNet is a library of words for English language along with their senses. A comparative study of various existing strategies in the field of text summarization and its applications is done through literature review.

4 Proposed Solution

In the proposed concept, we have used the concept of text summarization by making use of the WordNet Library and the concept of the lexical chain analysis using the noun and proper noun. In our approach, we have taken the concept of the significance and utility calculation for each chain so that the chains related to the documents are selected and used in the summary generation process.

The following algorithm is adopted for better automatic text summarization:

- (1) Select the normal text-containing document to be summarized.
- (2) Process the text file by examining the line by line the text file.

- (3) On each line, the process of tokenization is performed to remove ‘comma’, ‘semicolon’, and ‘dot’ and each character is then substituted by space character.
- (4) Subdivide the line into array of word to process each word individually.
- (5) Now, we will remove the stop word from this array.
- (6) The next step is related to getting the base form of the word, in order to get the base form we will make the use of the WordNet.
- (7) After that, the classification of word, like noun, proper noun, verb, adjectives, etc. is done by making use of the tagger and we have used MAX Tagger for this purpose.
- (8) Form all the words in the array, we will select only those words which are noun and proper Noun.
- (9) Only the single or the unique occurrence of each word is collected into another array.
- (10) After that by making use of the WordNet library, we will make the lexical chains of the word similarity.
- (11) Then, we will compute the chain length of each chain to be applied in the formula for the computation of significance and utility of each chain.
- (12) Then on the basis of the computation of the threshold value, we classify the chain as the accepted chain.
- (13) Then all the words from the accepted chains are fetched.
- (14) In addition, as per the percentage of matching criteria entered by the user, summary will be generated and will be matched with the manual summary to calculate the recall value.

5 Implementation of Proposed Algorithm

Eclipse Java Enterprise Edition for Web Developers version: Kepler Service Release 1 is used for the execution of the proposed algorithm. It offers an integrated development environment with Java compiler that is provided as a built-in feature and java source files are available as full model. This eases the code analysis.

For summary evaluation, most common grounds are recall and matching-words between human and algorithm generated summary. These two are taken as parameters. Along with these, execution time is also considered. Significance and utility of chains are used in the algorithm to find the strongest chains.

5.1 Execution

Based on the algorithm proposed, the following steps are followed for its execution:

Step-1: Open the file “WordNetGui3.java” in eclipse.

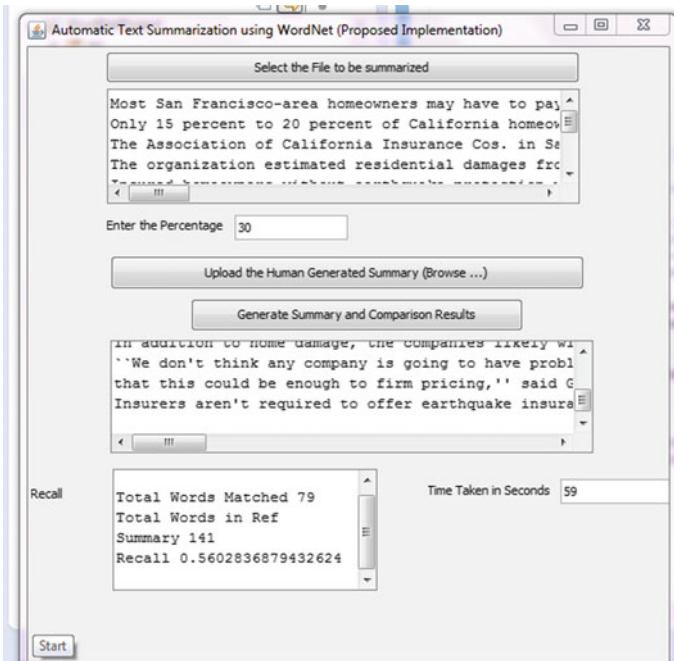


Fig. 2 Implemented GUI of the proposed system

Step-2: Run the Program: Go to the toolbar → Run option → Run as Option → Select java application.

Step-3: Select the file to be summarized.

Step-4: Open the sample Data.

Step-5: Enter the percentage of summary (Sampledata1.txt).

Step-6: Upload the human-generated summary.

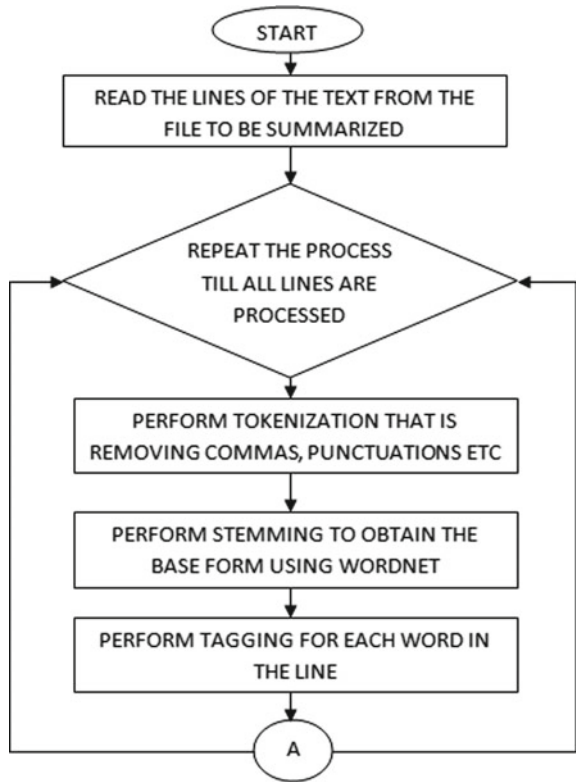
Step-7: Generate the summary and compare the results (Fig. 2).

5.2 Flow Chart for the Proposed Algorithm

This whole process of text summarization is divided into two parts. First is analysis of data that includes tokenization, stemming, and tagging while in the next part is the main processing part.

Thus, the above-explained algorithm is shown as a flow chart here in Figs. 3 and 4, respectively.

Fig. 3 Analysis of text summarization



6 Result Analysis

Along with WordNet API, this method also included the nouns and proper nouns in the computation of lexical chains, which improved recall.

This program is run on around 40 test outlines and from those, we have displayed around four tests and the aftereffect of the comparison is shown here, and records are brought with name SampleData.txt, SampleData2.txt, SampleData3.txt, and SampleData4.txt. Sample data are used to create the existing solution and the proposed algorithm base summary is contrasted with human-produced summary. The outcome is shown in Table 1 and Figs. 5, 6, and 7.

As discussed earlier, along with WordNet API, this method also included the nouns and proper nouns in the computation of lexical chains which resulted in improved matching of words between the human-generated summary and the summary generated by the use of proposed algorithm resultantly, improving the recall.

For part-of-speech tagging and statistical calculations, MAX tagger is used. This resulted in decreased execution time as compared to the existing algorithm.

Fig. 4 Process of text summarization

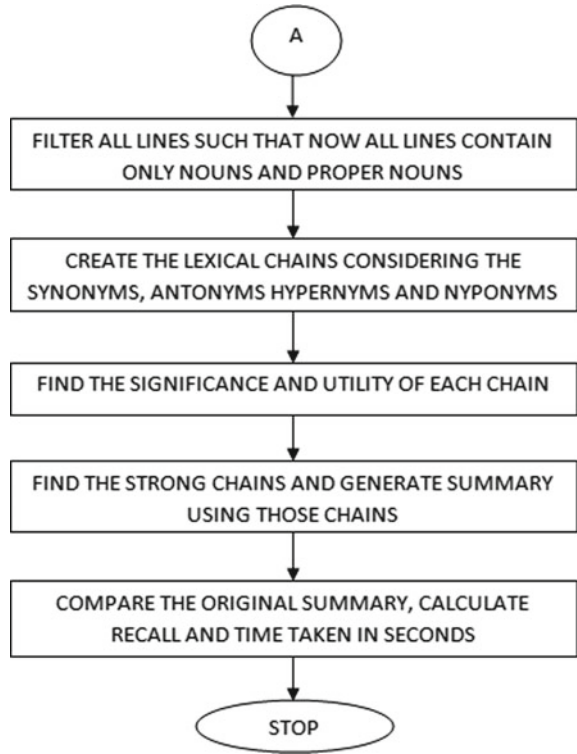


Table 1 Results of text summary generation

S. No.	Input file	Recall		Total words matched		Time taken to generate summary (in s)	
		Existing algorithm	Proposed algorithm	Existing algorithm	Proposed algorithm	Existing algorithm	Proposed algorithm
1	Sample data 1	0.45	0.48	65	69	103	47
2	Sampla data 2	0.52	0.57	42	56	132	39
3	Sampla data 3	0.43	0.64	61	67	119	35
4	Sampla data 4	0.41	0.45	108	119	210	163

Fig. 5 Recall value of summary generation

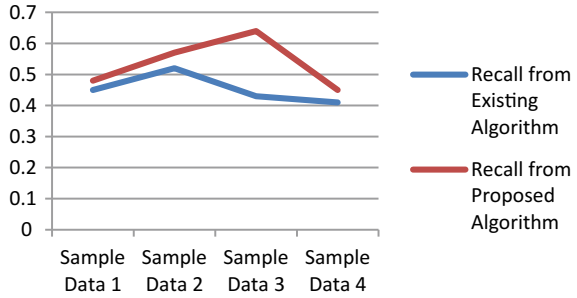


Fig. 6 Total words matched in summary generation

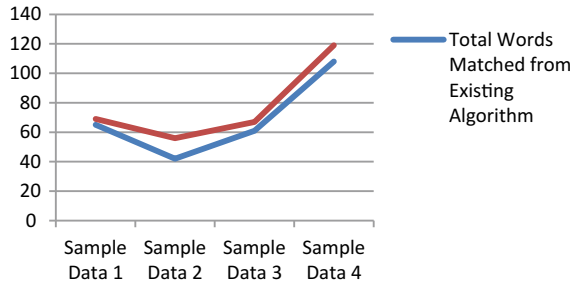
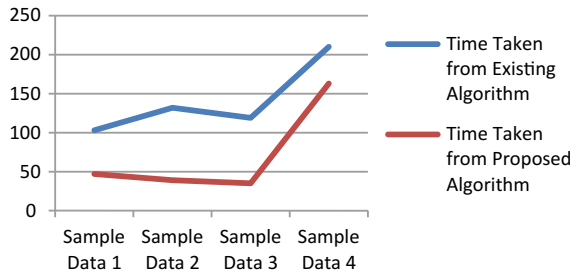


Fig. 7 Time taken in summary generation



7 Conclusion and Future Scope

Text summarization field is important due to its relevance with text classification and information retrieval processes and there are numerous algorithms and technologies available for summarizing the documents but there are still some improvements that are needed to be done for the sake of ease of process and growth as well. In this research work done, we have made the use of nouns and proper nouns for part-of-speech tagging purpose with WordNet API. In addition to this, statistical analysis is included in the proposed approach, which resulted in better outcomes in the form of execution time, recall value, and total words matched in the summary, which are common parameters of summary evaluation.

Further researches can be pursued in the area of multi-document summarization. This research can be extended to generate the summaries related to multiple documents at a time.

References

1. Kulkarni, A.R., Apte, S.S.: An automatic text summarization using lexical cohesion and correlation of sentences. *IJRET: Int. J. Res. Eng. Technol* (2014). eISSN: 2319-1163 | pISSN: 2321-7308
2. Suman, M., Maddu, T.: A new approach for text summarizer. *COMPUSOFT Int. J. Adv. Comput. Technol.* **4**(4) (2015). ISSN: 2320-0790
3. Jayarajan, D., Deodhare, D.: *Lexical Chains as Document Features* (2008)
4. Sarkar, K.: Using domain knowledge for text summarization in medical domain. *Int. J. Recent Trends Eng.* **1**(1) (2009)
5. Patil, A., Pharande, K., Nale, D.: Implementation of automatic text summarization. *Int. J. Res. Appl. Sci. Eng. Technol. (IJRASET)* **3**(5) (2015). ISSN: 2321-9653
6. Ercan, G., Cicekli, I.: Using lexical chains for keyword extraction. *Inf. Process. Manage.* **43**(2007), 1705–1714 (2007)
7. Cutting, D., Kupiec, J., Pedersen, J., Sibun, P.: *A Practical Part-of-Speech Tagger*. Xerox Palo Alto Research Center (1992)
8. Barzilay, R., Elhadad, M.: *Using Lexical Chains for Text Summarization* (1998)

Interactive User Interface to Advent HCI Artefact



Anil Kumar Dubey

Abstract HCI is the future for research. Much has been done in this area and a lot more needs to be explored. One of the prominent unexplored areas involves the previous HCI (human–computer interaction) design process and the two point the existing phases namely reliability of product and feedback and evaluation process are identified. Author proposes a novel approach for HCI design process, integration of previous design process that extends two more phases from previous as: testing for reliability of product and feedback and evaluation for user’s response and reality about them. The novelty in research is urged by the previously proposed modules in software engineering and streamlining the design process is them. Author contribute the new paradigm of design philosophy for human-centric design policies and introduce novel design models for exploitable design taking into account the concept of usability engineering. The new HCI waterfall model covers all stages of development for a human–computer interaction project. Each phase of the model has been well derived with graphical representation and shows their real existence. Similar parameters have been taken to design an HCI Interactive Incremental Model. The interactive incremental model of HCI involves the valued part for their verification of each phases, i.e. (what do you think to propose, Extended Version), known as evaluation. These two models open up the new era of HCI research to design few more models for future work and extend their importance in the development phase.

Keywords HCI (Human–computer interaction) · Design · Prototypes · Usability Engineering · Testing techniques

1 Introduction

ACM SIGCHI derived standard definition of HCI [1]. HCI has two basic intentions such as to facilitate better interactions between users and computers and to design computers in a user-friendly manner for maximum usability, also targeted to

A. K. Dubey (✉)

CSE, ABES Engineering College Ghaziabad, Ghaziabad 201009, Uttar Pradesh, India
e-mail: anildudenish@gmail.com

© Springer Nature Singapore Pte Ltd. 2019

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,

Advances in Intelligent Systems and Computing 904,

https://doi.org/10.1007/978-981-13-5934-7_11

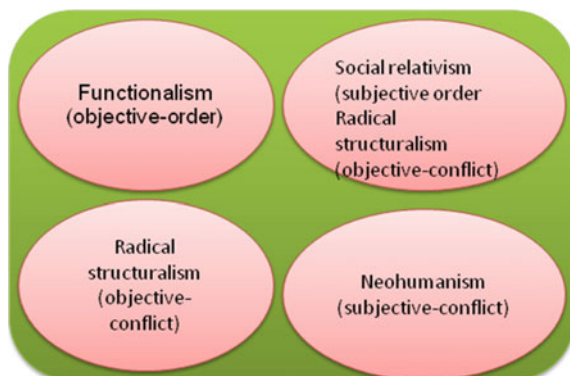
build system for achieving human’s cognitive model as per their requirements and to develop efficient systems for understanding the users’ task. In 1997, Steve Case remarked that “software engineering is the process incorporated with HCI phase of usability engineering.” For this remark, Charlie Kreitzberg expressed that software engineering should be a component of usability engineering. As individuals know that usability engineering is an important phase of HCI, the author can say that software engineering is indirectly a component of HCI [2]. Usable system product can be build up with the consideration of cost while could not be possible without the collaboration of two most technological era that is HCI and Software Engineering.

2 Historical Perspective and Reality of HCI in Systems Development

The historical perspective focused on object segments to integrate the design as well as social application. The object order describes the status quote of the functions that is depending on user’s choice and having the capability to integrate from one model to whole system. Another is the subjective order considering the social influence of user’s applications that may the user language selection or understanding the syntax of voice. After the Second segments of social relevant, the conversion and transformation of object data needed to the developer for structuring the layout of project. That involves radical Structure list, where status quote is transform or converted according to the design object. The fourth and last segment focused to analyse the variance into price and relationship to the buildup them (Fig. 1).

As many of the development life cycle of system changes or introduce by different researcher and nominating according to their changes. A lot of the SDLC model is accepted to design the structure of system development according to the required information. The same consideration of research work Hoffer et al. proposes the

Fig. 1 Four segment paradigms



development life cycle model for the usability of the system that is more applicable in the industries from last few decades [3–5] (Fig. 2).

Sutcliffe highlights the changes that influence the acceptance of human-centred development philosophy in comparison to escaped of traditional software development philosophy.

3 Problem Identification and Proposed Work

Different types of proposed processes are found to design more effective HCI [6, 7] products. As the society demands consider the user-friendly product in accordance with the usability engineering, the research organisation and industry are striving towards upgrading the design process for more interactive product. After examining the previous research, two basic problems have been identified into the previous proposed of design process: one of them is reliability problem and another is feedback and evaluation problem. Traditional design process simply uses the testing to test the product quality. As reliability is the main factor for current research computing era guaranteeing the life cycle duration of product application and quality, it is of prime importance for future researcher to proceed for HCI product design process considering reliability parameters. Another problem is that, in case of interactive design and prototyping, the feedback and evaluation process is missing from available design process and interface models. Feedback and evaluation is an essential point for

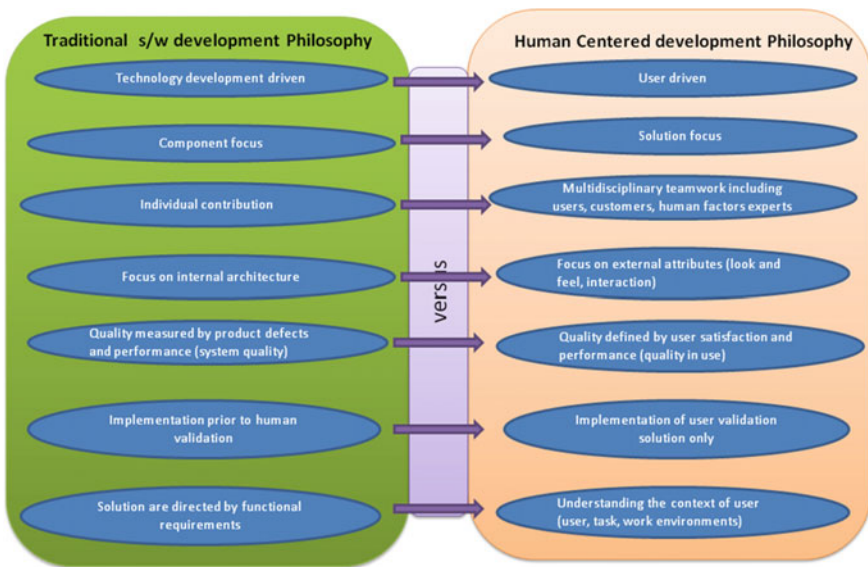


Fig. 2 Traditional software development versus human-centred development [11]

researcher to focus in HCI product design that is followed by usability engineering. Feedback and evaluation is an effective tool to serve diverse requirements of users and so it provides valuable input to evaluate the user’s response and evaluate the reality about them. To combat the above-stated problems, author proposes an approach to design more impressive HCI product. The proposed approach namely user interactive interface model is the integration of available previous models.

4 HCI User Interactive Interface Model

Author analysed the various aspects of software engineering and utilised them for the development of HCI phases. The initial development model of HCI follows the outlines of waterfall model of software engineering and hence the name HCI waterfall model. This model helps the developer to lay the fundamentals of designing an HCI product. It covers the users thought or preferences for HCI product building along with a regular updation process involved. There are eight phases for development life cycle of a product: (Fig. 3).

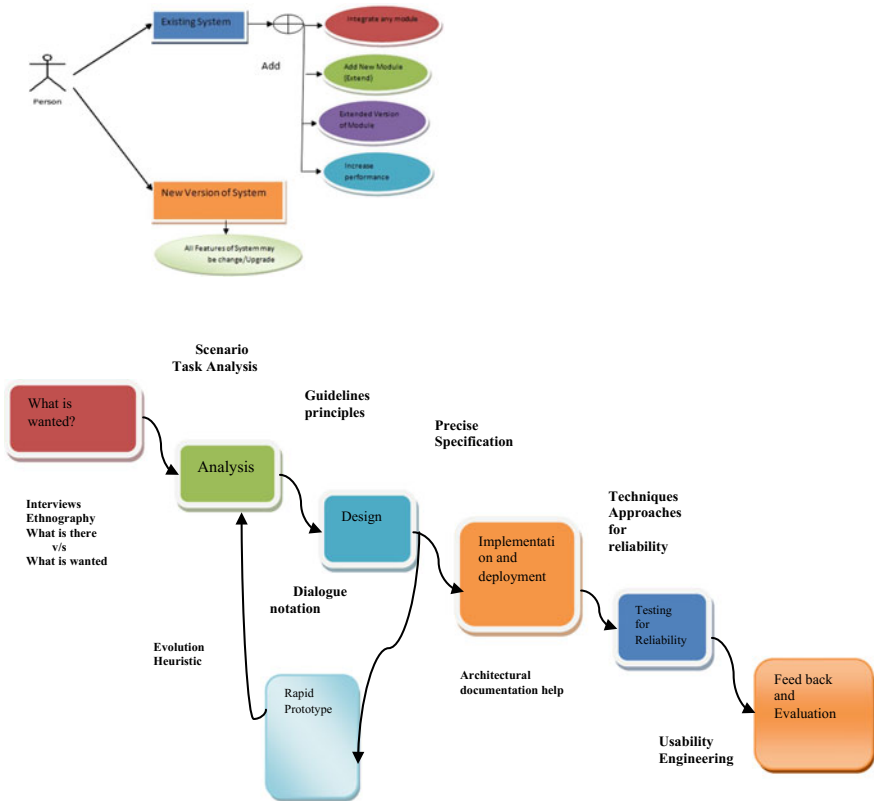
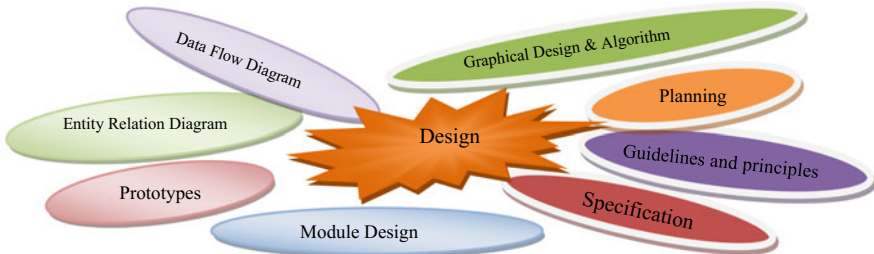


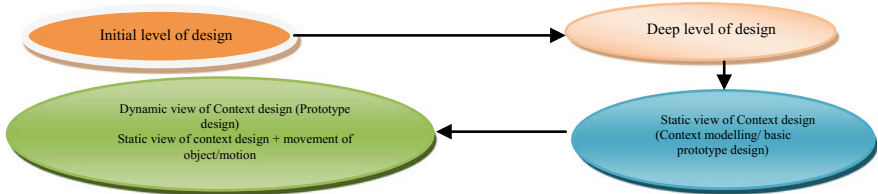
Fig. 3 HCI waterfall model

The initial phase of HCI waterfall model is described for user’s view as, what they think to plan or what they want. It is designed in UML model with an actor to set up their mind for ordering the product. Generally, they have two plans for ordering a product, i.e., existing system and new version of system. The human factors considered to analyse the task. These factors focus on the procedure and guidelines followed to gather initial data and how each model will be designed in future (Table 1).

Prototype Design and Dialogue Notation




Four types of design are used such as: Process-oriented design rationale, design space analysis, techniques for prototype, and interactive design.



- a. Initial level of design: It is the basic phase of design that covers the blueprint of realistic idea for acquired object, without highlighting the sub-attributes (artefacts) in the acquired domain.
- b. Deep level of design: in this second phase of designing complements the initial level of design fulfilled the required initial level. In this phase, designer focuses on sub-attributes (artefacts) of the acquired object model that consists all the possible ways of designing includes things and their appropriate relationships.
- c. Static view of Context design (Context Modelling): focuses the actual object, entities, artefacts, and things involved in the realistic projects the main parameters of this phase is to design the realistic of context for acquired object. Model rotated in the physical view and support to declare all the parameters of deep level of designing with verifying the reasons.
- d. Dynamic view of Context design (Prototype model): the designed model in static view of context designing is formulated with movement (action) to validate all the activates of designed prototype model that will easily unify from other’s proposal or similar prototype.

In realistic parameters, these phases validated all the artefacts, elements things attributes have bear taken to deploy such a prototype with valid reason.

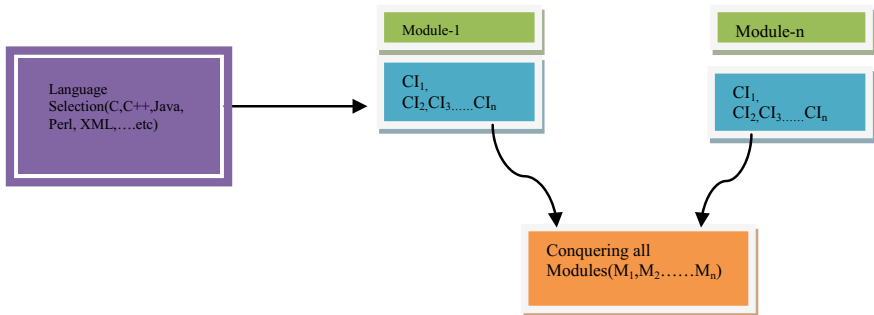
Table 1 Scenario task analysis of HCI waterfall model

 Person	Requirement gathering	Through primary collection	Interviews Viva Questionnaire
	Procedural follower	Through secondary collection	Previous development report Idea behind similar planning
		Structural follower	Step by step follower
		Sequential rules adopter	Minimise the procedure Cover theorem and lemma Basic principles with guidelines
	Task heuristic	One by one module analysis	Each gathered module is analysed through preference
		Expert committee	Select person having expertise in similar work
		Knowledgeable persons	Persons having well-suited knowledge of feasibility study

These phases continue till the acquired prototype has not been completely designed.

Implementation and deployment: Language Selection: developers choose an appropriate language from list of available language, i.e., C, C++, Java, Perl, CoBOL, HTML, XML, Pascal...etc.

- Implement configured item (CI): CI is the part of any module, which provides positive support to build overall system and is also known as a single module item.
- Conquering all Modules: After the implementation of all the configured item, it is essential to combine them and make a system. It is possible to help the conquering approach.



Algorithm for implementation and deployments:

- a. Structure view
- b. Behavioural view
- c. Architecture view (care both, structure view and behavioural view, then architect)
 - i. Initially, the design prototype has been taken from different views or phases
 - ii. As per the phases have been declared the deployment of individual model to collaborate to perform activities, sequences, etc.
 - iii. Management team must be discussed with the team leaders, implementers which type of models will be deployed by team member for, e.g., s/w modules, h/w modules, embedded modules, context modules, virtual modules, etc.
 - iv. Each individual part of single module has been discussed with implementation teams.
 - v. Each individual part of module prototype has been delivered to the implementation team process.
 - vi. Implementation team catches all the required items to deploy the module (environment, infrastructure, and team members).
 - vii. Team members deploy implemented module and review each and every step that progressing modules are following as the prototype structure(class, component, relations, objects, configured item, locations, etc.)
 - viii. Implementor also checks and verifies the physical view of progressed module in real and imaginations.

- ix. Implementer should follow the behavioural views in progressed module for their operations activities and bounded actions that were expected in the performance of progressed module.
- x. After both the process implementer focus to the integration of different modules implemented by different team members in a multiple and single architecture.
- xi. If the acquired project is implemented and fulfil all the conditions as environmental factor as views and performance according to the standard then GOTO next step called reliability factor else Step 6.

Techniques method and approaches for Reliability: Traditional design process simply uses the testing to test the product quality. As reliability [8] is the main factor for current research, computing era guarantees the life cycle duration of product application and quality. Therefore, it is important for future researcher to incorporate reliability components into HCI product design process.

All the documents and authenticated data has been tested and verified before to start the bonding of agreement. The design rules, aesthetics, and relevant visual context are tested for ensuring the availability of interface features regards to user application. To ensuring the compatibility of interface, tester must be followed up according to human centred of usability [9].

Testing for reliability: Initially, follow SDLC testing approach and then it will follow the proposed HCI testing rules:

- a. Cases based on: fixed environments and varying temperature
- b. Cases based on: fixed temperature and varying environments
- c. Cases based on: varying temperature and varying environments

Reliability refers to the mathematical computation of ...

- (1) Temp constant and variation in environmental factors such as remove signal devices from system or insert extra devices/artefacts in system
- (2) Environment constant with variance in temperature (lowest, medium, highest).

Measure output performance of acquired object at minimum temperature and highest temperature. After this mention, the noted point with acquired object prototype model that the deployed model will be more effective with mentioned/pointed facts/parameters.

The feedback and reviewed comments are considered from experts, knowledgeable persons, and actual users regarding the mentioned factors. All the positive and negative comments will be mentioned in matrix form with number of rows N(R) and number of columns N(C). Usability factor of product utilisation is followed according to HCI Paradigms. If % number of positive comments is higher than number of negative feedback comments, it means the user has accepted the acquired object model designed with proposed 6 phases of HCI lifecycle model and finally author conclude that the deployment of 6 phase model is much more acceptable and robust as compared to existing model designed using four phases. Feedback and evaluation be able to serve diverse principle and so, it is valuable to imagine about what exactly the users response and evaluate the reality about them. Usability also concerned for improvement purpose of ease of use method design phase [10].

After these process five quality components, [7] described usability is applying: (Fig. 4).

5 Comparative Study from Existence

As the software engineering, the development life cycle of software project has been derived for the better design and implementation of software project and admires to use the application of these by developer for any software project. In case of the interaction with the human, its essential to follow up the inaction design model for better way of understanding the each individual phase of development in a proper way if the HCI developer followed the exiting software development process model than there may be the complexity to identify and recognise the perspective way (to design the embedded hardware) will be arises. It means the development process of interactive tools is required a new model that has been designed or proposed according

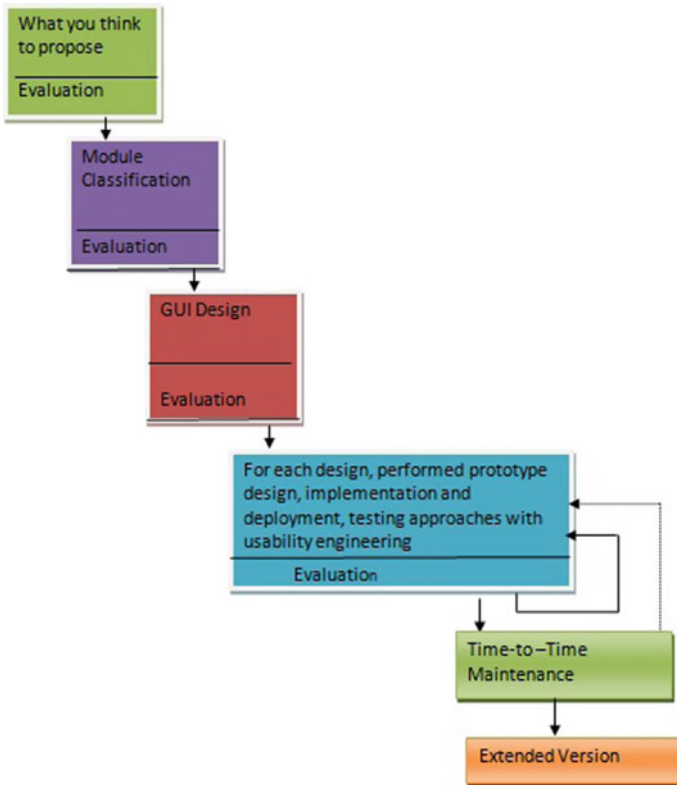


Fig. 4 HCI interactive incremental model

Table 2 Comparative traditional and human-centric design method

S. No.	Model	Tradition development life cycle	Proposed exploitable design	Effort
1.	Waterfall	It does not cover the usability engineering	Build the product according to usability engineering	Consider the precise specification for scientific term
		Traditional requirements analysis (documentary file), formal	Actual analysis of what is the need and what the user wants (face 2 face communication and ontology)	Accuracy is higher
		Well defined prototype is defined	Rapid prototype is designed	Negative response which may be improved upon
2.	Incremental model	It's does not support interactive design parameter	It's support interactive design parameter	Versioning has been admire in HCI interactive parameter
		Single phase of design	Collaborate with UML for design and modelling	Initial, deep level designed with modelling is performed

to the consideration of interactive parameters. To resolve such a problem, the author proposed the interaction development model (i.e., human–computer technological model). In case of interaction development process, the proposed HCI framework for exploitable design is better in comparison to the software development model (Table 2).

- A. Previous proposed design process simply uses the testing to test the product quality. As reliability is the main factor for current research computing era guarantees the life cycle duration of product application and quality. Therefore, it is important for future researcher to use into HCI product design process. The proposal considering these parameters via testing approach that is tested the deployed system.
- B. As the previous proposal of HCI process of design, the researcher missed the feedback and evaluation process, which are more effective parameters of HCI product design. The proposal fulfilled the important criteria of feedback and evaluation, which is best and more essential point for researcher to focus the HCI product design that is followed the usability engineering. Feedback and evaluation are able to serve diverse principle and so it is valuable to imagine

Table 3 Traditional interaction design process versus proposed approach

S. No.	Key points	Tradition interaction design process	Proposed approach
1.	Phases of design	Four main phases plus an interaction loop	Six main phases plus an interaction loop
2.	Testing approaches	Simple testing performed, missed the reliability status	Testing approach is used for system level testing of deployed system with reliability
3.	Feedback and evaluation	Missed and partial attraction	Strongly consider as a individual phase

about what exactly the users response and evaluate the reality about them as well as design process (Table 3).

6 Conclusion

From last three decades, the works have done on HCI is located their fundamental concept, graphical representation, and few of the measuring parameters. Minor work also has been done on their development process, but they are not sufficient to prove their existence according any reference model. Here, the author examined the previous HCI design process related work and pointed two problems into existing phases such as reliability of product and feedback & evaluation process. Author proposes a novel approach for HCI design process, integration of previous design process that have extends two phases more from previous as: testing for reliability of product and feedback and evaluation for user's response and reality about them. Author presents the novelty of research work into HCI and interface-based technology. Author's contributed the new paradigm of design philosophy for human-centric design policies and introduced novel design models for exploitable design considering the reliability, feedback, and evaluation phases. Author has developed HCI waterfall model and HCI interactive incremental model to cover the evaluation process. In future, author can also design few of other HCI development process model in similarity to software engineering and theoretically can prove them.

References

1. Hewett, S., et al.: ACM SIGCHI Curricula for Human-Computer Interaction. Chapter 2: Human—Computer Interaction. Last updated in 2009
2. Roger, S.: Software Engineering: A Practitioner's Approach, 6th edn. Pressman
3. Baguma, R.: Accessible Web Design through Web Accessibility Guidelines. 13 April 2010 at 15.30 hours. ISBN: 978-90-9025099-1

4. Marsden, G., Cook, D.: Human-Computer Interaction—3 Life cycles and Models. University of Cape Town (2006)
5. Zhang, P., et al.: Integrating human-computer interaction development into the systems development life cycle: a methodology. *Commun. Assoc. Inform. Syst.* **15**, 512–543 (2005)
6. Dubey, A.K. et al.: Empirical study to appraise consciousness of HCI technologies. In: ICSPCT, 12–13 July 2014 Ajmer. IEEE Xplorer. ISBN: 978-1-4799-3140-8/14
7. Dubey, A.K., Gulabani, K., Rathi, R.: HCI evaluation through scientific methods of computational perception. In: Presented at 2nd IEEE International Conference on Computer Communication and Systems (ICCCS 2014), 20th–21st February 2014, Savitha Engineering College, Thandalam, Chennai. 978-1-4799-3671-7/29, selected to published in IEEE Digital Explore, <https://doi.org/10.1109/icccs.2014.7068185>
8. Trivedi, P., Dubey, A.K., Pachori, S.: Reliability Tactics, 3rd ICECT, 8–10 Apr 2011. ISBN: 978-1-4244-8679-3, <https://doi.org/10.1109/icectech.2011.5941583>
9. Galitz, W.O.: The Essential Guide to User Interface Design, 2nd edn. Wiley Computer Publishing. ISBN: 0-471-084646
10. Shneidermann, B.: Text book on “Designing the User Interface”, 3rd edn. Published by Pearson Education Asia
11. Gulliksen, J., Boivie, I.: Usability Throughout the Entire Software Development Lifecycle A Summary of the INTERACT 2001 Workshop. Department of I.T. Technology, Uppsala University, Sweden, Technical report 2001-026, Nov 2001. ISSN 1404-3203

A Comprehensive Survey on Artificial Bee Colony Algorithm as a Frontier in Swarm Intelligence



Shiv Kumar Agarwal and Surendra Yadav

Abstract The nature is an intrinsic basis of idea for researchers continuously working in the area of optimization. The Artificial Bee Colony (ABC) algorithm imitates the foraging behavior of real honeybees and it is effectively used to solve multi-model and complex problems. Various strategies is developed on the behavior of honeybees but ABC is the most popular among all. The ABC algorithm is used to get rid of difficult real-world optimization problems that are not solvable by conventional methods. This paper presents a state-of-the-art study of ABC and its latest modifications with in-depth evaluation and analysis of recent popular variants of ABC.

Keywords Artificial bee colony algorithm · Engineering optimization · Swarm intelligence · Foraging behavior

1 Introduction

The ABC simulates foraging demeanor of real honeybees for solving multi-model and multi-dimensional problems. The intelligent behavior of honeybees attracted various researchers to develop some new strategies. There are number of algorithms based on their behavior while searching for food, partner formatting, social organization, etc. It is first identified by D. Karaboga in the year 2005 [15] that they follow a systematic process while collecting nectar in order to produce honey. It is a stochastic in nature as some arbitrary element play a crucial role in process of optimization and also depicts swarming behavior. The honeybee shows extraordinary behavior while probing for proper food sources and then exploiting nectar from best food sources. The ABC algorithm is very efficient in comparison to other competitive algorithms but it has few drawbacks also similar to other algorithms, like premature convergence and stagnation. That is the reason that since its inception, it has been modified by a number of researchers. The ABC algorithm is very simple in implementation as it

S. K. Agarwal (✉) · S. Yadav

Department of Computer Science and Engineering, Career Point University, Kota, India
e-mail: agarwalshiv83@yahoo.co.in

© Springer Nature Singapore Pte Ltd. 2019

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_12

125

has only three control parameters and follows three simple steps while moving toward the optimum. This paper considered recent modification in ABC that improved position update process while other survey and study conducted on ABC are general and not focused on the position update process.

2 Artificial Bee Colony Algorithm

The ABC algorithm [15] is inspired by amazing behavior of honeybees while searching for better food sources. Analogous to natural honeybees, the ABC algorithm divided all the bees into three different groups according to their behavior and nature of task performed. The whole population is composed of three types of bees: Employed Bee (EB), Onlooker Bee (OB), and Scout Bee (SB). The EBs are accountable for probing for new food sources and providing information on the subject of food sources to the bees that are residing in the hive (onlooker bees). Based on the information received from the employed bee, the onlooker bees start exploiting these food sources. If a food source is exhausted due to exploitation, it is considered as abandoned and is replaced by scout bee. The bees are continuously trying to improve solutions using greedy search strategy, until the termination criteria meet and memories the best solution established thus far. The success of ABC algorithm depends on the balance between these two processes. Initialization of swarm also plays an important role in deciding direction of solution search process.

The behavior of natural honeybees can be described by three parameters: Food Source, Bees, and Dance. The bee identifies a specific food source (flower) while searching for food. Then, it starts collecting information about that particular food source (i.e., quantity of nectar that may be exploitable from it), its direction, and distance from the hive. A food source is analogous to the solution in search space. In order to get a better outcome, it is essentially required to use proper initialization strategy and an exploitation mechanism that can accelerate good solutions in the next iteration. The group of bees that voluntarily starts searching for food sources is known as employed bee. These bees collect information with reference to food sources and share with OBs that are waiting in the beehive. This information exchange takes place using special kind of dance, like round, waggle, and tremble dance. Bees that observe the dance of EBs are known as OBs and identify the best food sources rich in nectar. Now, the onlooker bees start exploiting these food sources.

2.1 Phases of ABC Algorithm

The ABC algorithm has three phases, namely EB phase, OB phase, and SB phase. These three steps are iterated to find optimal solutions after initialization. The key steps of the ABC algorithm are explained in the subsequent sections.

Initialization: The first step in ABC is initialization of parameters and set up an initial population randomly using the following equation.

$$X_{ij} = LB_j + rand(0, 1) \times (UB_j - LB_j) \quad (2.1)$$

where $i = 1, 2, \dots, (ColonySize/2)$ and $j = 1, 2, \dots, D$. Here, D represents dimension of problem. X_{ij} denotes location of i th solution in j th dimension. LB_j and UB_j denotes lower and upper boundary values of search region correspondingly. $rand$ is a randomly selected value in the range $(0, 1)$.

Employed Bee Phase: This phase try to detect superior quality solutions in proximity of current solutions. If quality of fresh solution is enhanced than present solution, the position is updated. The position of employed bee is updated using Eq. 2.2.

$$v_{ij} = x_{ij} + \phi \times (x_{ij} - x_{kj}) \quad (2.2)$$

where $\phi \in [-1, 1]$ is an arbitrary number, $k \in 1, 2, (ColonySize/2)$ is a haphazardly identified index such that $k \neq i$.

Onlooker Bee Phase: The OB choose a food source using probability function and become an EB. The new EB starts searching for some innovative solution in proximity of present solution. The fitness and probability is computed using fitness of solution with the help of following equations.

$$Prob_i = \frac{fitness_i}{\sum_{i=1}^{colonySize/2} Fitness_i} \quad (2.3)$$

Scout Bee Phase: An EB become a scout when the solution value not updated till the predefined threshold limit. This scout bee engenders new solution instead of rejected solution using Eq. 2.1.

The performance of ABC algorithm measure is compared by Karaboga et al. [16, 19] with other competitive algorithms for critical analysis. Some recent hybrids of ABC are compared and analyzed by Kumar et al. [25]. A detailed survey on ABC algorithm is carried out in [7, 22, 36] and list of recent modifications and applications of ABC. Bansal et al. [3] performed a stability analysis of ABC using von Neumann stability criterion. New stability criteria suggested by Bansal et al. [3] for parameters to bind the inaccuracy in succeeding iteration and supported by experimental results. The computational complexity of ABC depends on population size and dimension of the problem. If N is the size of the considered population, then each loop (in ABC it have two loop one for each onlooker and employed bee phase) iterates for $N/2$ times (as number of EB and OB are $N/2$) and every phase having loop over D times (dimension of the problem). The algorithm stops when it meet the termination criteria (maximum number of iteration).

2.2 Modifications in ABC Algorithm

Since its development, it has undergone numerous changes as it has some drawbacks, like stagnation and early convergence. To overcome these drawbacks and to improve the balancing between exploration and exploitation, variants of ABC are continuously popping up over time. Researchers are trying to improve its performance in order to make it one of the best choices for complex optimization. Karaboga et al. [18] compared the performance of ABC with three other competitive algorithms and concluded that ABC has the capability to break out local optima and can be used to solve diverse engineering optimization problems but still has low convergence rate. Karaboga et al. [17] used ABC to train feedforward Neural Network (NN) and make use of three different benchmark problems to train the NN. Zhu and Kwong [40] developed a new version of ABC with motivation from particle swarm optimization named as Gbest-guided ABC (GABC) algorithm. It is experienced that the performance of ABC depends on two contradictory processes namely exploration and exploitation. Proper balancing between these two processes improves the rate of convergence. The position update equation in basic ABC [15] driven by current position and arbitrarily generated step size. In order to improve this process, a new position update equation suggested in [40] as shown in Eq. 2.4.

$$x'_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \psi_{ij}(y_j - x_{ij}) \quad (2.4)$$

where $\psi_{ij}(y_j - x_{ij})$ is a newly added term and named as gbest. ψ_{ij} is an unvaryingly dispersed arbitrary number in the range $[0, C]$ for some positive constant C and y_j stands for global best solution in present swarm. This equation improves the exploitation of best feasible solution as it tries to improve y_i like Particle Swarm Optimization (PSO) [21]. In GABC [40] position update equation (Eq. 2.4) is similar to Differential Evolution (DE) [31]. The GABC is very efficient variant of ABC and further modified number of times to solve complex optimization problems. Like Jadhav et al. [13] implemented GABC for economic load dispatch problem with inclusion of wind energy. The probability calculation process improved for better exploitation capability as depicted in Eq. 2.5.

$$P_i = \frac{0.9 \times fit_i}{Fit_{best}} + 0.1 \quad (2.5)$$

where fit_i and fit_{best} denotes the fitness of i th solution and highest fitness. A discrete gbest ABC anticipated in [11] for optimal cloud service composition with time attenuation function. A new version of gbest ABC proposed in [8] with a couple of alteration in original GABC. The EB and OB phases are modified and recommended a new position update process as shown in Eqs. 2.6 and 2.7 correspondingly.

$$x_{new_{ij}} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \frac{(1-t)}{T} \times \psi_{ij}(y_j - x_{ij}) \quad (2.6)$$

$$x_{new_{ij}} = x_{ij} + \frac{(1-t)}{T} \times \phi_{ij}(x_{ij} - x_{kj}) + \psi_{ij}(y_j - x_{ij}) \quad (2.7)$$

where t and T denotes current and total count of iterations respectively. Rest all symbols have their usual meaning. These equations has two major components, first component $\phi_{ij}(x_{ij} - x_{kj})$ maintains the stochastic nature in proposed algorithm and second component, i.e., $\psi_{ij}(y_j - x_{ij})$ drives toward global optimum and speed up the convergence. The weightage of these components iteratively changes that are responsible to maintain proper balance between intensification and diversification. Sharma et al. [33] recently developed a variant of GABC by considering local best and global best solutions. The EB and OB phases are modified as shown in Eqs. 2.8 and 2.9, respectively.

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \psi_{ij}(x_{Lbestj} - x_{ij}) \quad (2.8)$$

$$v_{ij} = x_{ij} + \phi_{ij}(x_{ij} - x_{kj}) + \psi_{ij}(x_{Gbestj} - x_{ij}) \quad (2.9)$$

where x_{Lbestj} denotes local best value in present population and x_{Gbestj} denotes global best value in current population, rest all symbols have their usual meaning.

Bansal et al. [6] incorporated a memetic process in ABC to improve the performance of ABC. This approach also makes use of GSS strategy and employed it as a local search in addition to existing steps and named as memetic search. This strategy updates only best solutions in current population as it assumes that there are more chances to get optimum results in proximity of the fittest solution. The step size in position update equation (Eq. 2.2) decided by two components: first is a random number F and second is the difference of current and a randomly selected entity. In [6], the random element F is decided by GSS.

$$x_{newj} = x_{bestj} + F(x_{bestj} - x_{kj}) \quad (2.10)$$

The MeABC [6] algorithm modified by various researchers as it is very efficient. Kumar et al. [24, 26–28, 30] proposed modifications in MeABC to improve balancing between exploration and exploitation. These variant modifies the onlooker bee and employed bee phase [26], incorporated fitness-based position update with memetic search [28], levy flight search in memetic search strategy [24], improvement in onlooker bee phase in MeABC [27], and randomized MeABC [30]. The Opposition-based Learning (OBL) introduced in ABC algorithm by El-Abd [10] to get better performance of ABC. El-Abd [10] applied the concept of OBL during initialization and position update phase. The concept of OBL based on assumption that counterpart solution of an individual is also equally important for the purpose of optimization and there are 50% chances that opposite solutions are better fitted then original solutions. The opposite solution of x_{ij} is denoted by Eq. 2.11 for initialization phase by El-Abd [10].

$$ox_{ij} = UB_j + LB_j - x_{ij} \quad (2.11)$$

where x_{ij} denotes i th solution in j th direction and lower bound and upper bounds are represented by LB and UB correspondingly. Recently, Sharma et al. [32] incorporated OBL with a local search strategy in ABC. Hussain et al. [12] used shrinking hypersphere local search strategy in ABC and position update strategy modified as follows

$$x_{newj} = x_{bestj} + \phi \times (x_{bestj} - r_j) \quad (2.12)$$

where x_{bestj} is best solution in current swarm and r denotes radius of hyperspace with x_{best} as center. It is initialized with Euclidian distance and then it is updated using Eq. 2.13.

$$r_j = r_j - \left(\frac{x_{bestj} - x_{worstj}}{LSI} \right) \quad (2.13)$$

Sharma et al. [34] incorporated Fully Informed Learning (FIL) at onlooker bee stage in ABC. Each individual collects information about best solutions in the current swarm and in its proximity. It assumes that best solutions update their position according to information collected in their neighborhood with the purpose of enhancement in exploitation capability. Bansal et al. [4] used global and local search in ABC to solve the optimal power flow problem.

After analyzing the performance of considered variant, it may be concluded that Gbest ABC [40] and MeABC [6] are most popular variants among researchers and provide better performance.

3 Applications of ABC Algorithm

The ABC algorithm is the most efficient stochastic algorithms in the field of optimization. Since its inception, the popularity of ABC is increasing day by day. At first, it was developed for unconstrained optimization problems and then applied for constrained optimization problems. The ABC is very popular due to less number of parameters and one can easily implement it. To begin with, it was concerned to numerical optimization problems [15] and then, it was explored for constrained optimization problems. The applications of ABC algorithms is listed in [7, 19] with subject area of use. The ABC algorithm has application in the field of electrical, computer science, mechanical and electronics engineering. The ABC applied for feature selection, multi-cast routing, single machine scheduling, balancing in assembly line, designing of controller and to find solution of traveling salesman problem. The ABC algorithm has endless list of applications, there are numerous field including science, engineering and management where ABC algorithm is in use with priority over other competitive optimization algorithm.

Table 1 List of recent modifications in ABC algorithm

Year	Modifications in ABC algorithm	Remark
2000	Gbest-guided ABC updated solution search equation using global best solution [40]	New position update strategy directed by current best solution introduced
2011	BSF-ABC modified OB by comparing current and best solutions found so far [2]	New position update strategy directed by best solution found so far introduced
2011	Opposition-Based learning incorporated in ABC algorithm [32]	Solution initialization and updating process used opposition-based learning
2013	A memetic search phase added in ABC as a local search process [6]	A new local search phase introduced stimulated by golden section search
2014	A hybrid ABC by using one operator of GA [29]	A new operator introduced in ABC motivated by crossover operator in GA
2014	A modified golden section search mechanism in ABC [23]	Onlooker bee phase modified using golden section search mechanism
2014	A self-adaptive ABC that uses best individuals for exploitation and worst fit solutions for exploration [5]	The step sizes and "limit" parameter both are adaptively modified based on fitness of individual
2015	Discrete gbest-guided ABC algorithm for composition of cloud services [11]	Discrete version of Gbest- guided ABC introduced
2015	Accelerating ABC to enhance convergence rate [14]	An additional local search phase added using greedy logarithmic decreasing step size
2016	The EB and OB phase are modified using local and global best solution [33]	Local and global best solutions provide direction to search strategy
2016	Disruption operator initiated in ABC [35]	A new phase based on natural phenomenon of disruption added in ABC
2016	A Self-Perturbing ABC proposed to find global optimum solution [39]	Natural trend of ailing fish being eaten by strong one was simulated
2016	A hybrid ABC proposed that is adaptive in nature for ANFIS [20]	Two new parameters introduced: Arithmetic crossover rate and adaptivity coefficient
2016	The SB phase modified using rate of change in place of limit [1]	A new parameter rate of change introduced in ABC
2016	Shuffled ABC algorithm [37]	Hybrid of ABC and Shuffled Frog Leaping algorithm
2016	Position of individuals updated by weight driven approach [38]	Modification proposed in EB phase
2017	A rank based ABC that changes their ranking adaptively [9]	A ranking-based search strategy added in ABC
2018	A modified Gbest-Guided ABC [8]	Certain modifications suggested in [40]

4 Conclusion

The ABC Algorithm is most appreciable nature-inspired algorithm for optimization. This paper provides a detailed introduction to ABC algorithms and selected modifications discussed in detail. A summary of modifications in ABC is tabled in Table 1. ABC algorithm is preferred over other NIAs and gives better performance for complex optimization problems. The best thing of ABC is that it is not affected by the initial value of parameters. It has some drawbacks like lack of balancing between exploration and exploitation process due to premature convergence.

References

1. Anuar, S., Selamat, A., Sallehuddin, R.: A modified scout bee for artificial bee colony algorithm and its performance on optimization problems. *J. King Saud Univ.-Comput. Inf. Sci.* **28**(4), 395–406 (2016)
2. Banharsakun, A., Achalakul, T., Sirinaovakul, B.: The best-so-far selection in artificial bee colony algorithm. *Appl. Soft Comput.* **11**(2), 2888–2901 (2011)
3. Bansal, J.C., Gopal, A., Nagar, A.K.: Stability analysis of artificial bee colony optimization algorithm. *Swarm Evol. Comput.* (2018)
4. Bansal, J.C., Jadon, S.S., Tiwari, R., Kiran, D., Panigrahi, B.K.: Optimal power flow using artificial bee colony algorithm with global and local neighborhoods. *Int. J. Syst. Assur. Eng. Manag.* **8**(4), 2158–2169 (2017)
5. Bansal, J.C., Sharma, H., Arya, K.V., Deep, K., Pant, M.: Self-adaptive artificial bee colony. *Optimization* **63**(10), 1513–1532 (2014)
6. Bansal, J.C., Sharma, H., Arya, K.V., Nagar, A.: Memetic search in artificial bee colony algorithm. *Soft Comput.* **17**(10), 1911–1928 (2013)
7. Bansal, J.C., Sharma, H., Jadon, S.S.: Artificial bee colony algorithm: a survey. *Int. J. Adv. Intell. Parad.* **5**(1–2), 123–159 (2013)
8. Bhambu, P., Sharma, S., Kumar, S.: Modified gbest artificial bee colony algorithm. In: *Soft Computing: Theories and Applications*, pp. 665–677. Springer, Berlin (2018)
9. Cui, L., Li, G., Wang, X., Lin, Q., Chen, J., Lu, N., Lu, J.: A ranking-based adaptive artificial bee colony algorithm for global numerical optimization. *Inf. Sci.* **417**, 169–185 (2017)
10. El-Abd, M.: Opposition-based artificial bee colony algorithm. In: *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation*, pp. 109–116. ACM, New York (2011)
11. Huo, Y., Zhuang, Y., Gu, J., Ni, S., Xue, Y.: Discrete gbest-guided artificial bee colony algorithm for cloud service composition. *Appl. Intell.* **42**(4), 661–678 (2015)
12. Hussain, A., Gupta, S., Singh, R., Trivedi, P., Sharma, H.: Shrinking hyper-sphere based artificial bee colony algorithm. In: *2015 International Conference on Computer, Communication and Control (IC4)*, pp. 1–6. IEEE, New York (2015)
13. Jadhav, H.T., Roy, R.: Gbest guided artificial bee colony algorithm for environmental/economic dispatch considering wind power. *Expert. Syst. Appl.* **40**(16), 6385–6399 (2013)
14. Jadon, S.S., Chand Bansal, J., Tiwari, R., Sharma, H.: Accelerating artificial bee colony algorithm with adaptive local search. *Memetic Comput.* **7**(3), 215–230 (2015)
15. Karaboga, D.: An idea based on honey bee swarm for numerical optimization. Technical Report TR06, Erciyes Univ. Press, Erciyes (2005)

16. Karaboga, D., Akay, B.: A comparative study of artificial bee colony algorithm. *Appl. Math. Comput.* **214**(1), 108–132 (2009)
17. Karaboga, D., Akay, B., Ozturk, C.: Artificial bee colony (ABC) optimization algorithm for training feed-forward neural networks. In: *International Conference on Modeling Decisions for Artificial Intelligence*, pp. 318–329. Springer, Berlin (2007)
18. Karaboga, D., Basturk, B.: A powerful and efficient algorithm for numerical function optimization: artificial bee colony (ABC) algorithm. *J. Glob. Optim.* **39**(3), 459–471 (2007)
19. Karaboga, D., Basturk, B.: On the performance of artificial bee colony (ABC) algorithm. *Appl. Soft Comput.* **8**(1), 687–697 (2008)
20. Karaboga, D., Kaya, E.: An adaptive and hybrid artificial bee colony algorithm (ABC) for anfis training. *Appl. Soft Comput.* **49**, 423–436 (2016)
21. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: *Proceedings of IEEE International Conference on Neural Networks*, Proceedings, vol. 4, pp. 1942–1948. IEEE, New York (1995)
22. Kumar, D., Mishra, K.K.: Artificial bee colony as a frontier in evolutionary optimization: a survey. In: *Advances in Computer and Computational Sciences*, pp. 541–548. Springer, Berlin (2017)
23. Kumar, S., Bhambu, P., Sharma, V.K.: New local search strategy in artificial bee colony algorithm. *Int. J. Comput. Sci. Inf. Technol.* **5**(2), 2559–2565 (2014)
24. Kumar, S., Kumar, A., Sharma, V.K., Sharma, H.: A novel hybrid memetic search in artificial bee colony algorithm. In: *2014 Seventh International Conference on Contemporary Computing (IC3)*, pp. 68–73. IEEE, New York (2014)
25. Kumar, S., Sharma, V.K., Kumari, R.: Comparative study of hybrids of artificial bee colony algorithm. *Int. J. Inf. Commun. Comput. Technol.* **1**(2), 20–28 (2014)
26. Kumar, S., Sharma, V.K., Kumari, R.: An improved memetic search in artificial bee colony algorithm. *Int. J. Comput. Sci. Inform. Technol.* (0975–9646) **5**(2), 1237–47 (2014)
27. Kumar, S., Sharma, V.K., Kumari, R.: Improved onlooker bee phase in artificial bee colony algorithm. *Int. J. Comput. Appl.* **90**(6), 20–25 (2014)
28. Kumar, S., Sharma, V.K., Kumari, R.: Memetic search in artificial bee colony algorithm with fitness based position update. In: *Recent Advances and Innovations in Engineering (ICRAIE)*, 2014, pp. 1–6. IEEE, New York (2014)
29. Kumar, S., Sharma, V.K., Kumari, R.: A novel hybrid crossover based artificial bee colony algorithm for optimization problem. arXiv preprint [arXiv:1407.5574](https://arxiv.org/abs/1407.5574) (2014)
30. Kumar, S., Sharma, V.K., Kumari, R.: Randomized memetic artificial bee colony algorithm. arXiv preprint [arXiv:1408.0102](https://arxiv.org/abs/1408.0102) (2014)
31. Price, K.V.: Differential evolution: a fast and simple numerical optimizer. In: *Proceedings of Biennial Conference of the North American on Fuzzy Information Processing Society (NAFIPS)*, pp. 524–527. IEEE, New York (1996)
32. Sharma, H., Bansal, J.C., Arya, K.V.: Opposition based Lévy flight artificial bee colony. *Memetic Comput.* **5**(3), 213–227 (2013)
33. Sharma, H., Sharma, S., Kumar, S.: Lbest gbest artificial bee colony algorithm. In: *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pp. 893–898. IEEE, New York (2016)
34. Sharma, K., Gupta, P.C., Sharma, H.: Fully informed artificial bee colony algorithm. *J. Exp. Theor. Artif. Intell.* **28**(1–2), 403–416 (2016)
35. Sharma, N., Sharma, H., Sharma, A., Bansal, J.C.: Modified artificial bee colony algorithm based on disruption operator. In: *Proceedings of Fifth International Conference on Soft Computing for Problem Solving*, pp. 889–900. Springer, Berlin (2016)
36. Sharma, S., Bhambu, P.: Artificial bee colony algorithm: a survey. *Int. J. Comput. Appl.* **149**(4) (2016)
37. Sharma, T.K., Pant, M.: Shuffled artificial bee colony algorithm. *Soft Comput.* **21**(20), 6085–6104 (2017)
38. Tiwari, P., Kumar, S.: Weight driven position update artificial bee colony algorithm. In: *International Conference on Advances in Computing, Communication, & Automation (ICACCA)* (Fall), pp. 1–6. IEEE, New York (2016)

39. Zhou, S., Feng, D., Ding, P.: A novel global ABC algorithm with self-perturbing. *J. Intell. Syst.* **26**(4), 729–740 (2017)
40. Zhu, G., Kwong, S.: Gbest-guided artificial bee colony algorithm for numerical function optimization. *Appl. Math. Comput.* **217**(7), 3166–3173 (2010)

Modeling and Simulation of Al6082 MMC of Gravity Die Casting for Solidification Time



Harendra Pal, Dinesh Kumar Kasdekar and Sharad Agrawal

Abstract The solidification of cast part remains a process of excellent interest. It immediately impacts the production rate, casting defects, and mechanical property of casting. The phenomenon of solidification of cast part is complicated in foundry industry as well as the modeling and simulation needed in industry in advance than it is far in reality undertaken. This research specializes in the impact of casting method parameters of Al6082 MMC. Modeling and simulation examine the casting solidification time in the foundry in gravity die casting technique. The design of experiment is done with the help of full factorial design (FFD). The casting technique parameters are pouring temperature, pouring rate, and die preheat temperature (DHT) on solidification time that has been studied. This paper explains the optimization of casting method parameters using genetic algorithms. It also gives the information about the generation of optimization models, simulation, and methodology used and obtains the optimum process parameters. The predicted trials have been used to comparatively compare with simulation and experimental results, and the simulated comparison results are observed in a proper way.

Keywords Gravity die casting · Solidification time · Click2CAST · ANOVA · Genetic algorithms

H. Pal (✉) · D. K. Kasdekar · S. Agrawal
Department of Mechanical Engineering, Madhav Institute of Technology and Science,
Gwalior 474005, Madhya Pradesh, India
e-mail: harendrapalji@gmail.com

D. K. Kasdekar
e-mail: Dkasdekar82@gmail.com

S. Agrawal
e-mail: Sharad.mits03@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_13

1 Introduction

Casting is the process that is used to produce metallic components. This process produces metallic items by heating of metal at certain temperature and then liquid is injected into a prepared die cavity, permitting it to solidify [1, 2]. Different types of engineering materials are present for the production of metal object. Simulation is a very essential tool for any production company mainly in aircraft spare parts, automobile industry and mould making manufacturing company. Maximum costs are related with the manufacturing, and such a casting component further undergoes to the machining process that reduced the life of the component so that its application is also limited. Therefore, casting method gives various difficulties in any foundry industry so that defects can be decreased using modeling and simulation. The simulation program selection helps for foundryman for better selection of different parts of casting process such as gating system, runner and riser, core, and their particular location. It also provides the best suggestion to change the predefined design of casting to enhance in addition to decrease production expenses. Therefore, simulation program permits to carry out better result earlier than the making of any product by casting. It determines mould filling and solidification solutions to gain excessive yield at the popular high-quality stage and involving most efficient method plans, which is well appropriate with each product's necessities and foundry functionality [3]. The aim of this research is to carry out simulation for solidification time with experimental data and predicted trials to decrease above defects. Simulation for solidification time and evaluation had been examined by different simulation software. And boundaries have been taken under consideration while solving the case examined. An available literature gives the following information about the casting simulation program. In Oza et al. [4], an attempt has been made with the comparison of FE evaluation with real experimental facts of shell core drum provided for the validation. In Reddy et al. [5], an attempt has been made in the gravity die casting technique with the pouring of liquid metal into a non-expendable metal cavity under the force of gravity. The solidification time of Al–Si–Mg alloys for die casting will increase with higher pouring temperature of liquid metal and decrease with lower inside the preheat temperature of metal cavity. In Tiwari et al. [6], an attempt has been made to provide that the material composition, degasification, and molten metal filing temperature can have an effect on the fluidity property of Al–Si–Mg alloys. In Babington et al. [7], an attempt has observed the effect of die casting method variables on the casting defects and their impact on the mechanical properties. In Choudhari et al. [8], an attempt has observed the redesign and make a good quality casting free from defects, especially shrinkage cavity that occurs in casting part. The input parameters used for modeling and simulation examined give higher shrinkage defects in produced cast part which becomes the predominant reason for the rejection of casting components in the foundry. The software outcome and experimental data had been compared and difference between them is found to be in good agreement. In Choudhari et al. [9], an attempt has been made to perform the complete simulation and modeling with auto-cast software for optimization that purely depends on Vector Gradient Method (VGM). In Dabade

et al. [10], an attempt has been made to design the experiments with simulation program that are merged to find out the method-associated and sand-associated defects that generate in green sand casting. In Hussainy et al. [11], an attempt has been made to research the motives for casting defects taking place in cast parts and reduce these defects by simulation approach along with experimental result. In Anglada et al. [12], an attempt has been made for casting simulation consisting of the mould, and the heat is rejected by the molten metal that is dissipated through the mould wall along with the mould filling and solidification of the alloys. In Venik et al. [13], an attempt has been made to the metallic liquid filling behavior in mould cavity that is affected by the liquid pouring temperature, mould preheat temperature, and the viscosity of alloy to assess the impact of gravity die casting process. A proper selection of casting parameters for the gravity die casting process is heavily on the foundry technologies and experienced because of their numerous and diverse range. Casting parameters provided by the casting builder cannot meet by the foundries requirements. To solve this task, software prediction (Click2CAST Trial version 4.0), regression model, and genetic algorithm model are developed as an efficient approach to determine the optimum casting process parameters in gravity die casting process for Al6082 metal matrix composite. In this paper, pouring temperature (PT), pouring rate (PR), and die preheat temperature (DHT) on solidification time have been studied. For verification of the empirical models, experiments have been carried out in the foundry shop.

2 Experimentation and Simulation for Gravity Die Casting

The simulation program in foundry industry proved to be an important technique for minimizing casting defect and produced good quality casting and optimization of casting process. It provides better solution and gives information about the casting part without physical trial [14]. Manufacturing rate is increased by producing higher amount of defect-free casting, and information about good casting has improved production rate and it is also trained for new foundry engineers [15]. Therefore, the Click2CAST (trial version) casting simulation software is employed for predicting solidification time of Al6082 MMC for gravity die casting. In gravity die casting, three input parameters are taken for simulation study. Proper analysis has been done for input parameter to find out the solidification analysis in the gravity die casting. Five steps of simulation analysis have been carried out using simulation software. In this study, the solidification time evaluation is carried out through the usage of simulation and experimentation. Simulation studies are on solidification time of aluminum composite, which have been conducted using the Click2CAST casting commercial software trial version. The solid model of cast part and C.I die is created using the CATIA software and imported casting part into the casting simulation software [14–16]. Recommended casting simulation steps are shown in Fig. 1.

Experimental setup is designed and experiment is conducted in a small-scale foundry shop producing ductile Al6082 MMC cast components. During casting, the rate of solidification of casted piece mostly influences the strength, hardness, and

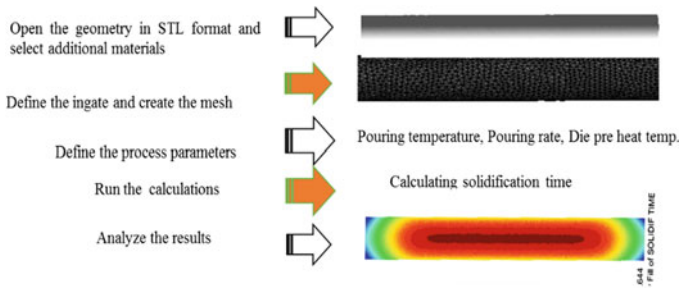


Fig. 1 Click2CAST (trial version) simulation steps

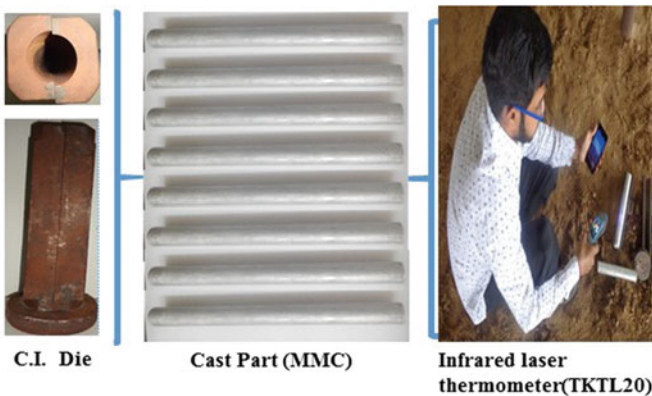


Fig. 2 Die and produced cast part MMC

machinability of the casting part [17]. For the cast iron die, the time needed for solidification is about 26.06 s for the casting part. Casting does not have any hotspot that shows casting process is better. The die open time is dependent on solidification time of casting part. After solidification, casting part is ejected from the die. The solidification time is calculated through stopwatch at solidus temperature 542 °C. Infrared laser thermometer (TKTL20) is used to measure temperature of solid casted part. Its temperature range using probe is from -64 to 1400 °C. The simulation program provides information about the solidification time in the form of layer by layer and also gives the information regarding the particular layer as shown in the last stage of Fig. 1. It can be observed that the solidification time obtained by simulation is correct and does not affect casting quality greatly. The die and cast part for circular shape casting is shown in Fig. 2.

In this study, Al6082 composite is taken for modeling and simulation. Table 1 gives the information about the chemical composition of this material. MSME, Agra helps us to identify the composition of material which uses as mentioned alloy in the casting. So it is possible for us to make identical alloy for simulation purposes. In this examine, to determine the approach to obtain this trade-off, the optimum sequence of

Table 1 Chemical composition of Al6082 MMC

S. No.	Composite designation	Hybrid composite material
1	Composite (1)	98%Al6082 + (1%SiC + 1%Grp.)

input casting variables has been obtained by GA. The high aspect ratio of casting is demonstrated with excellent performance of the gravity die casting process to yield minimum solidification time. The optimization results will be best for direct use for foundry shop and casting technology to utilize in research center and many casting industries. Also results can be enhanced using higher order regression equation for solidification time in gravity die casting process.

2.1 Mathematical Formulation

For most suitable casting method in the product range, the simulation of casting parameters is important. Mathematical modeling tool of regression is observed comparatively cheap to better discover the useful nonlinearity and interaction functions concerned within the practical casting outcomes. The different techniques for obtaining such mathematical models are available. In this paper, practical outcomes have been utilized by full factorial design (FFD) for modeling purposes. Three input parameters such as pouring temperature (PT), pouring rate (PR), and DHT are chosen in this paper. These are the essential casting parameters.

2.2 Full Factorial Design (FFD)

An FFD can reduce the given problems using the method of optimization of all input process parameters together at the same time with an aim to obtain the suitable comprehensive optimization of the casting process parameters. FFD is used to minimize the total number of experiments. The FFD modifies the improvement of each factor on response parameter along with the effect of each factor change with respect to the value of another parameter [18]. FFD includes the more accurate result in comparing the complete fundamental component only with effects and combination of various factors could be accomplished the usage of DOE only. In FFD, every combination of each element emerges with all framework of each alternative element, rather than organizing a sequence of freelance exercise we are able to integrate this analysis into one. A simple experimental style is one with all input parameters set at two levels each. These level measures are known as “high” and “low” or “+1” and “-1” severally. If there are k parameters, everyone at a pair of levels, an FFD has 2k runs. In this analysis, three decision variables of 2-level FFD (8 runs) are taken for the

Table 2 Process parameters for the experiment

S. No.	Parameter	Unit	Level	
1	Pouring temperature (PT)	°C	670	690
2	Pouring rate (PR)	cm ³ /s	3.21	3.53
3	Die preheat temperature (DHT)	°C	150	200

simulation of solidification time (ST) within the acceptable range of decision factors and actual values as shown in Table 2.

2.3 Analysis of Variance (ANOVA)

After making the trial, the most significant parameter affecting the solidification time (ST) has been determined by analysis of variance (ANOVA). Sum of squares (SS) of every part shows a control within the method because the worth of the R2 will increase. The importance of the equal component in gravity die casting method also increases. The model factors and interaction effects of each process parameter having F values <0.05 are observed as potentially significant.

3 Methodology

3.1 Evolutionary Algorithm Using GA

GA is a soft computing technique totally based on natural selections and natural genetics of hypothesis process as well as it is a parallel and worldwide search approach used for emulating natural genetic operations [19]. It can provide a worldwide solution after enough iteration; however, it has an excessive computational burden. GA-primarily based techniques for method optimization of casting process parameters have numerous benefits. These are not simplest treat to discrete factor, however, additionally overcome the spatial property drawback. Similarly, they have the functionality to look for the worldwide top-quality or quasi-optimums inside an inexpensive computation time. On the contrary, research on evolutionary algorithms has shown that these techniques are often with success wont to take away most of the above-named difficulties of classical techniques [20]. The choice of most appropriate gravity die casting parameters like pouring temperature (PT), pouring rate (PR), and die preheat temperature (DHT) is totally an important issue for each casting manner. In foundry industry, input factor is decided based on previous casting data or specialized casting books but the input parameter values provided by these sources are approximate value that does not have the appropriate values [21]. In

any optimization technique, it becomes aware of the output of chief importance so, known as optimization goal or optimization criterion. Gravity die casting method performance is especially passionate about correct alternative of the input method parameters. Moreover, in high aspect ratio set solidification time with casting process, the standard of created casting is critically laid low with the method parameters. The selection of variable is for (i) minimum solidification time results in good quality of invented casting and (ii) minimum casting defects. Therefore, during this work, to attain an answer to the present trade-off, the optimum combination of input method parameters has noticed victimization GA-based mostly optimisation. The optimum sequence of method parameters has been obtained exploitation genetic algorithmic rule optimisation and verified through an experiment.

4 Results and Discussions

4.1 Analysis of Variance

ANOVA method is applied to determine the relative significance of each input parameter on solidification time, and the ANOVA result for solidification time is presented in Table 3. The ANOVA is accomplished at 95% confidence stage, i.e., at 5% level of importance. It is found that all the casting factors are significant factor that affects the solidification time. *F*-value for the maximum significant factor is examined to be much less than 0.0001. ANOVA predicts the model error that is usually and independently allotted with the identical variance in each component level. These assumptions may be obtained by calculating the residuals value. Residual value is obtained by comparing the particular observation data and data acquired from the ANOVA model with respect to the input parameters. By means of constructing an ordinary probability plot of the residuals, the normality assumptions can be checked. The analysis changed by using the popular software that is particularly used for design of experiment programs, called design expert 6.0.8. It is also important that residuals can be usually allotted so the regression evaluation to be applicable [22].

Residuals are excellent for calculating of error. The individual deviations of the examination Y_i from their equipped values are called residuals. Residual plots are used to find out the assumptions for the making regression model. The model errors are generally predicted by ANOVA, and model errors are freelance distributed. These assumptions may be obtained by means of residuals. Solidification time is represented by the first- and second-order response surface model and can be expressed as a characteristic casting process parameters. The relationship among the solidification time and casting process input parameters has been represented as follows [23]:

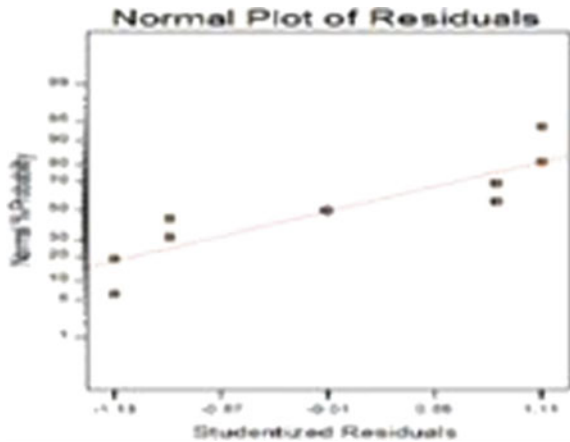
$$\begin{aligned}
 ST = & +23.85 + 0.26 * PT + 0 * PR + 1.98 * DHT + 0 * PT * PR \\
 & - 0.028 * PT * DHT + 0 * PR * DHT + 0 * PT * PR * DHT \quad (1)
 \end{aligned}$$

Table 3 ANOVA results of solidification time

Source	DF	Sum of square	Mean square	Prob > F
Model	7	31.82	4.55	<0.0001
A:PT	1	0.53	0.53	0.0182
B:PR	1	0.00	0.00	<0.0002
C:DHT	1	31.28	31.28	0.0293
Residual	0	0.00	0.00	0.00
Total	7	31.82		

S—0.03478; R-Squared—0.9998; R-Squared (adj.)—0.9997; R-Squared (pred.)—0.9995

Fig. 3 Normal probability plot of residuals



The normal probability plots of the outcome result and solidification time (ST) are shown in Figs. 3. The graph gives the information about normal distribution indicated by straight lines.

Figures 4 plot the residual versus predicted and residual versus run the levels of PT, PR, and DHT, respectively. These are some indications for process parameters that levels of 1.50, -1.50, and 3.00, -3.00 of pouring temperature, pouring rate, and DHT have slightly lower variability in response.

4.2 Response Surface Analysis for ST

The fitness model is determined by founding comparison of linear and nonlinear outcomes equations. Figure 5 shows the model having three-dimensional response surfaces (Eq. 1). The curvature along with response surface is because of the interaction impact; in impact, the plane is twisted. The graphs are generated by taking the

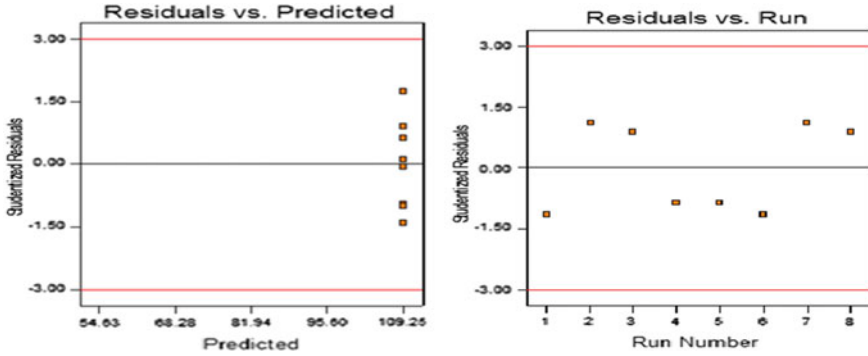


Fig. 4 Plot the residual versus predicted and residual versus run

DESIGN-EXPERT Plot

ST

X = A: pouring temp

Y = B: pouring rate

Actual Factor

C: Die pre heat temp = 175.00

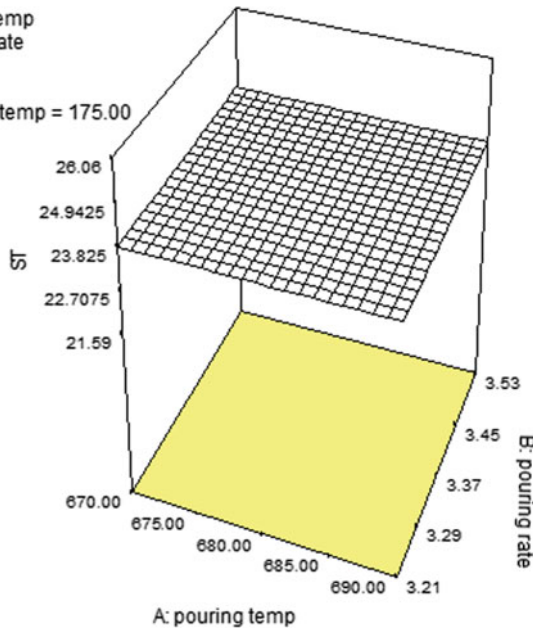


Fig. 5 3-D response surface interaction on solidification time

center level values as the hold values of the freelance parameters which include PT, PR, and DHT. The response surface plot reveals the interface. Each contour corresponds to specific altitude of the response surface. The levels of independent factors are analyzed by the graph that shows that result changes in response surface height or shape. Additionally, these graphs are useful in ANOVA analysis.

Table 4 Comparison between measured value and predicted value

Trial No.	PT	PR	DHT	ST (s) experiment	ST (s) software	Predicted value	Prediction accuracy %	Prediction accuracy %
1	690	3.53	150	25.5	24.26	25.44	95.3	99.4
2	670	3.21	200	24.5	23.60	24.52	96.2	99.7
3	670	3.21	150	23.4	22.06	23.39	94.3	99.8
4	690	3.21	200	26.3	25.12	26.33	95.4	99.7
5	670	3.53	200	24.5	23.15	24.57	94.2	99.9
6	690	3.21	150	25.5	24.56	25.50	96.3	99.9
7	690	3.53	200	26.0	25.01	26.03	96.0	99.8
8	670	3.53	150	23.5	22.02	23.51	93.6	99.9

4.3 Model Accuracy

The result on solidification time is obtained for eight experimental trials of gravity die casting method, and the values given by the FFD model (Eq. 1) are presented in Table 4.

Gravity die casting of Al6082 MMC using models is accomplished. FFD of experiment is followed to study the effect of various casting variables, viz., PT, PR, and DHT that are input parameters keeping with carried out experimental design. The FFD modeling is found accurate in predicting solidification time for the duration of gravity die casting method of Al6082 MMC with mean error and forecast accuracy is recorded as in Table 4. The experimental values of solidification time obtained are in the range of 23.42–26.06 s. The model exhibited great goodness of fit within the present examined as evident from statistical hypothesis exams carried out at 5% significance level. It is found that the expected values by simulation are very near the experimental values of FFD.

4.4 Genetic Algorithm-Based Optimization

The primary aim of this examination is to determine the optimum process factor that reduces the solidification time. For this examine, MATLAB genetic algorithm (GA) tool cabinet is employed to work out the optimum casting method issue. A lot of potentialities are obtained at numerous combos of input parameters [24]. It is found that the optimized settings of genetic tool are as follows: selection–roulette, reproduction crossover fraction –0.8, mutation crossover uniform and ratio 0.01, and migration forward. At this optimized setting, the best fitness is shown in Fig. 6. Function tolerance is achieved after 52 iteration number. The optimum value obtained

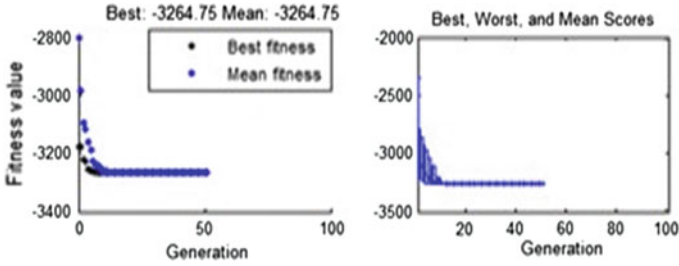


Fig. 6 Fitness curve in solidification time

using the genetic algorithm technique is pouring temperature at 690 °C, pouring rate at 3.25 cm³/s, and DHT at 200 °C. Hence, the optimal solidification time obtained from GA is 26.06 s.

5 Conclusions

This examination has confirmed the solidification time minimized by casting simulation techniques. The simulation-based technique allows the feeder place by means of brief solidification evaluation. GA technique offers a scientific and efficient tool in modeling and the optimization. The following conclusions are acquired from these studies:

- (i) Costs of foundry trials are much higher than the simulation trial while presenting defect-free casting with optimization. Therefore, it is concluded that experimental outcomes had been compared with simulation outcomes.
- (ii) The regression model of solidification time advanced through FFD provides tons of knowledge with a little amount of experimentations. The model advanced by means of FFD turned into rationally suitable and can be applied for the prediction of responses with restriction of factors.
- (iii) It is observed that the model obtained by ANOVA is considerable.
- (iv) AI method, i.e., genetic algorithm, increases the variability inside the iteration of experiments, i.e., mutation, crossover, crossover fraction, population, etc. It is used for locating of highest quality mixture of independent factor for a lower value of solidification time.
- (v) The mean fitness and vary decreases with growing range of iterations throughout the genetic process.

References

1. Choudhari, C.M., Narkhede, B.E., Mahajan, S.K.: Finite element simulation of temperature distribution during solidification in cylindrical sand casting with experimental validation. In: 4th International and 25th All India Machine Tool Design and Research (AIMTDR), Jadavpur University, Kolkata, India, vol. 1, pp. 3–8 (2012)
2. Choudhari, C.M., Padalkar, K.J., Dhumal, K.K., Narkhede, B.E., Mahajan, S.K.: Defect free casting by using simulation software. *Appl. Mech. Mater.* **313**, 1130–1134 (2013)
3. DeGarmo, E.P., Black, J.T., Kohser, R.A.: *Materials and Processes in Manufacturing*. Prentice-Hall Inc., New Jersey (1997)
4. Oza, A.D., Patel, T.M.: Analysis and validation of gravity die casting process using pro-cast. *Int. J. Appl. Innov. Eng. Manag.* **2**(4), 46–52 (2013)
5. Reddy, A.C., Rajanna, C.: Design of gravity die casting process parameters of Al–Si–Mg alloys. *J. Mach. Form. Technol.* **1**(½), 1–25 (2009)
6. Tiwari, S.N.: On fluidity characteristics of metals and alloys. *Indian Foundry J.* **44**(5), 77–82 (1998)
7. Babington, W., Kleppinger, D.H.: Aluminum die castings—the effect of process variables on their properties. In: *Proceeding of ASTM*, pp. 169–174 (1951)
8. Choudhari, C.M., Narkhede, B.E., Mahajan, S.K.: Methoding and simulation of LM 6 sand casting for defect minimization with its experimental validation. *Procedia Eng.* **97**, 1145–1154 (2014)
9. Choudhari, C.M., Narkhede, B.E., Mahajan, S.K.: Casting design and simulation of cover plate using auto cast-X software for defect minimization with experimental validation. *Procedia Mater. Sci.* **4**, 786–797 (2014)
10. Dabade, U.A., Bhedasgaonkar, R.C.: Casting defect analysis using design of experiments (DOE) and computer aided casting simulation technique. *Procedia CIRP.* **7**, 616–621 (2013)
11. Hussainy, S.F., Mohiuddin, M.V., Laxminarayana, P., Krishnaiah, A., Sundararajan, S.: A practical approach to eliminate defects in gravity die cast Al-Alloy casting using simulation software. *Int. J. Res. Eng. Technol.* **04**(1), 114–124 (2015)
12. Anglada, E., Meléndez, A., Vicario, I., Arratibel, E., Aguillo, I.: Adjustment of a high pressure die casting simulation model against experimental data. *Procedia Eng.* **13**, 966–973 (2015)
13. Venik, A.I.: Analysis of metal flow in die castings. *Machinery* **99**, 1501–1505 (1961)
14. Ravi, B.: Casting simulation and optimization: benefits, bottlenecks, and best practices. *Tech. Paper Indian Foundry J.* **54**, 47 (2008)
15. Hussainy, S.F., Mohiuddin, M.V., Laxminarayan, P., Krishnaiah, A.: A practical approach to eliminate defects in gravity die cast al-alloy casting using simulation software. *Int. J. Res. Eng. Technol.* **4**, 114–124 (2015)
16. Ravi, B.: Casting simulation—best practices. *Indian Foundry Congr.* **58**, 19–29 (2010)
17. Manjunath Swamy, H.M., Nataraj, J.R.: Design optimization of gating system by fluid flow and solidification simulation for front axle housing. *Int. J. Eng. Res. Dev.* **4**(6) (2012)
18. Kushwah, S.S., Kasdekar, D.K., Agrawal, S.: Mathematical and prediction modeling of material removal rate for evaluating the effects of process parameters. In: *Ambient Communications and Computer Systems*, Springer Nature Singapore, Chapter No. 44 (2018). Book ISBN: 978-981-10-7385-4
19. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Publishing Company Inc., MA (1989)
20. Soodamani, R., Liu, Z.Q.: GA based learning for a model—based object recognition system. *Int. J. Approximate Reasoning* **23**, 85–109 (2000)
21. Chen, P.H., Chang, H.C.: Large-scale economic dispatch by genetic algorithm. *IEEE Trans. Power Syst.* **10**(4), 1919–1926 (1995)
22. *Design and Analysis of Experiments*. Montgomery John Wiley and Sons, vol. 18, pp. 163 (2001)

23. Palanikumar, K.: Application of Taguchi and response surface methodologies for surface roughness in machining glass fiber reinforced plastics by PCD tooling. *Int. J. Adv. Manuf. Technol.* **36**, 19–27 (2008)
24. Kasdekar, D.K., Parashar, V., Soni, P.: Optimization of machining parameters in electro discharge machining of Al6061–Cu–SiC–graphite metal matrix composite. *Mater. Sci. Forum* **860**, 61–64 (2016)

A Survey on the Detection of Windows Desktops Malware



Sanjay K. Sahay and Ashu Sharma

Abstract An important feature of malware is that it can self-replicate. It is not known who created the first self-replicating program in the world, but it is clear that the first malware/virus (Creeper) was created by the Bible Broadcasting Network engineer Robert (Bob) H. Thomas, probably around 1970, and since then malware are evolving continuously to evade the known detection techniques from early-day signature-based to the date machine learning methods. The complexity of the malware is continuously growing using sophisticated obfuscation techniques not only to attack individual computational devices but also for the military espionage, to disrupt industries, ransomware, etc. Thus researchers are motivated to find an effective anti-malware to detect the known as well as new or previously unseen malware. Hence, time-to-time to defend the attacks/threats from the advanced malware, a number of static and dynamic methods are proposed by the researchers. Therefore, in order to understand various techniques proposed/used for the detection of new or previously unseen Windows Desktops malware in this paper, we present the survey conducted by us on the work done by the researchers in this field.

Keywords Security · Windows OS · Malware · Metamorphic · Signature matching · Machine learning

1 Introduction

The development of computer security has a military origin, and since 1950 it is a major concern. Initially, because of national defense and intelligence, the US government was “a major force behind security research and technology” [1]. Till 70s,

S. K. Sahay (✉)

Department of CS & IS, BITS, Pilani, Sancoale, Goa Campus, India
e-mail: ssahay@goa.bits-pilani.ac.in

A. Sharma

C3i, CSE, IIT Kanpur, Kanpur, India
e-mail: ashush@cse.iitk.ac.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_14

computers users were small; hence, protection of data was easier. But, in 80s, Personal Computers (PC) came into being which was small enough to fit on a desk, and in 90s the Internet has made a revolutionary impact on the PC user, basically due to its near-instant communication. Today, Internet is ubiquitous, through which most of the computational devices share the information among the users in the different networks and organizations. However, these computational devices which are connected to the Internet have rendered countless invasions to the users/organizations from the initial static malware (**malicious software**) to second-generation highly sophisticated customize dynamic malware [2]. An important property of this malware is that it can self-replicate. It is not known who created the first self-replicating program in the world, but it is clear that the first malware/virus (Creeper) was created by the Bible Broadcasting Network engineer Robert (Bob) H. Thomas, probably around 1970, and since then malware are evolving continuously with high complexity to evade the known detection techniques from early-day signature-based to the date machine learning techniques. The complexity of the malware is continuously growing using sophisticated obfuscation techniques not only to attack individual computational devices but also for the military espionage, to disrupt industries, ransomware, etc. [3]. In this, the two highest data breaches in 2013 were \$5.4 million and \$4.8 million in the United States and Germany, respectively [4]. Also, intrusion in the Google systems shows that how organized attacks can be designed for long-term access to an organization's networks [5]. Thus, researchers are motivated to find an effective anti-malware to detect the known as well as new or previously unseen malware.

Despite the advancement in the anti-malware, cyberattacks are on an uptrend. An estimate by Symantec shows that the rate of creation of new instances of malware was 41%, with a total of over 400 million existing new malware instances [6]. According to FireEye survey [7], there were 47% security incident/network breaches from the malware in the organization. In the last quarter of 2015, a 26% increase in new ransomware samples has been reported [8], and in the first quarter of 2016, the Quick Heal Threat Research Lab received more than 340 million malware samples running into hundreds of thousands of devices [8]. Symantec reported 54 zero-day vulnerabilities, and it is doubling each year [9]. McAfee reported a new malware which can infect the firmware of the solid-state storage and hard drive devices and cannot be removed by either by reinstalling the OS or formatting the storage devices [10]. In this on 2014, 11 zero-day vulnerabilities were reported, in which six were from the industrial control systems [11]. McAfee report says that in 2014 around two hundred million malware samples are known [12]. The Symantec Internet Security Threat report of 2014 says that there were 62% more breaches than the previous year [4]. In the first quarter of 2017, the highest number (295 million) of malware samples was detected in the systems of Quick Heal users. However, compared with the first quarter of 2016, there was a drop of 13.61% in the detection count.

In the last couple of years, cybercriminals had focused mainly on the bank customers that in a single attack, how to steal millions of dollars. In 2016, Symantec uncovered the most effective bank robbers (Banswift group) who stolen the US \$81 millions from the Central bank of Bangladesh [13]. Later on, some more cases were reported in South Asia banks, which were attacked with the same malware code that

was used in Bangladesh. In this, ransomware is continuously creating a problem for both businesses and consumers, with random campaigns pushing out mass. In 2016, the average ransom demand rises to \$1077, i.e., a 366% increased from the previous year. A growing reliance on cloud services also creates vulnerabilities for organizations which cause the ability to hackers to hijack databases of MongoDB for ransom [13]. The new ransomware families known during 2016 were 101, i.e., more than three times compared to the year 2015, and in 2017 Symantec reported a 36% increase in ransomware infections [13], while Quick Heal Labs detected ten new ransomware families in the same quarter [14]. The uptrend shows that attackers are jumping on the popular ransomware to modifying the existing ones or create a new ransomware family.

To counter/defend the Windows Desktops (~90% market share, an attractive target for computer hackers and cybercriminals) malware, there are many anti-malware defense systems from early-day signature matching to the date machine learning methods [2]. Some malware can be detected by the traditional signature-based anti-malware. But it cannot detect advanced/second-generation unknown malware as it uses advanced obfuscation techniques and can generate millions of malware variant to bypass the security protections and eventually evade detection. Also, due to the availability of intelligent software to create variants of malware [2], it appears that malware developer is always ahead of the anti-malware group. In this, throughout the globe, Microsoft analyzes tens of millions of data files daily by its real-time anti-malware which presents ~160 Million computing devices [4]. According to the Computer Security Institute survey, on an average, there is \$345,000 loss by the security attack per incident [15]. As daily new malware variants are evolving, the defense system to counter malware is becoming a difficult task to protect the computational devices from them [10]. Therefore, if adequate measures have not been taken to counter the malware, then the consequence will be devastating, in particular from the advanced metamorphic malware, which poses a big threat to the endpoint security, because it uses advanced obfuscation techniques to evade detection by the available detection methods. Hence, to identify effectively the advanced/metamorphic malware, it is a need of the time to design new methods to analyze the behavior of malware for the effective detection of malware with minimum false positives. In this time-to-time, a number of static and dynamic methods have been proposed by the researchers. Therefore, in order to understand various techniques proposed/used for the detection of new or previously unseen Windows Desktops malware in this paper in Sect. 2, we briefly discuss the types of malware, and in Sect. 3, we present the survey conducted by us on the work done by the researchers in this field. Finally, in Sect. 4, we summarize our conclusion.

2 Types of Malware

Malware can be categorized as first- and second-generation malware. The first-generation or static malware are basically classified on how it infects the devices as Viruses, Worms, and Trojans. Few other notable first-generation malware are

rootkits, spyware, crime-ware, adware, etc. This type of malware exhibits different sorts of malicious behavior on the target systems; however, its structure does not change. But in the second-generation or dynamic malware after each infection and while keeping the action same, it changes its structure to create new variant [2]. On the basis of the strategy by which either the code or the structure is obfuscated to conceal the signature of the malware, the second-generation malware are further classified as encrypted, oligomorphic, polymorphic, and metamorphic malware.

Encryption is the first concealment technique in which malware body consists of an encrypted malicious code, key, and an algorithm/decryptor. In this, body of the malware is XORed with the generated key to make it difficult to detect. The main objective to create encrypted malware was to evade the static code analysis and traditional signature-based detection technique. However, encrypted malware can be identified by analyzing the used invariant algorithm/decryptor, and this limitation of the encrypted malware makes the anti-malware group find the various concealment techniques to evade the detection mechanism, in which *Oligomorphic* malware, the structure of the decryptors gets changed from one variant to other, i.e., it provides a set of obfuscated decryptors. Oligomorphic malware at most can create few hundreds of decryptors, e.g., Win95/Memorial [16].

Polymorphic malware is similar to oligomorphic malware, but it can generate millions of decryptors by mutating the instructions using different obfuscation techniques in the variant of malware. Mark Washburn in 1990 created the first polymorphic malware known as 1260 [17]. While *Metamorphic* malware instead of mutating the decryptors the malware body is mutated itself (i.e., body-polymorphic) to create a new variant using various sophisticated obfuscation techniques without changing its actions to evade the detection [17]. The first metamorphic malware Win95/Regswap was created in 1998, and after that 3,628,800 different variants of Win32/Ghost were created [18], and W32/NGVCK was designed in 2001 using Next Generation Virus Creation Kit which was one of the strongest metamorphic malware.

3 Detection of Windows Desktops Second-Generation Malware

In 1970, the first virus came into existence [19], and since then there is a strong fight between the malware and anti-malware developer, which led to the continuous evolution of the techniques to detect the malware from early-day signature-based to the date machine learning methods. Consequently, the complexity of the malware is continuously growing, basically using the sophisticated obfuscation techniques not only to attack individual computational devices but also for the military espionage, to disrupt industries, ransomware, etc. Hence, to defend the threats/attacks from the advanced malware, time-to-time, a number of static and dynamic methods have been proposed by the researchers [20–24]. In the static analysis, the executables are analyzed to find the unique pattern, viz., *n*-grams, Application Program Interface

(API) sequences, opcodes, data flow, resources, Dynamic Link Library (DLL) usage, etc. without executing the codes, whereas, in the dynamic analysis, the programs are investigated by monitoring its dynamic behaviors, viz., system calls, network connections, usage of the resources, etc. However, in both the static and dynamic methods, generally, first classifiers are trained with a dataset to distinguish benign from the malware programs. In this Schultz et al. in 2001 [25], using data mining methods claimed that their framework could detect the new unknown malware twice than the detection rate of traditional signature-based techniques. For the experiment analysis, they used 3265 malicious and 1001 benign programs. The classification has been done by extracting the features from DLL residing in the PE files, and strings and bytes sequences are extracted from the executables. They applied repeated Incremental Pruning to Reduce Error Reduction (RIPPER) [26] to find patterns in the DLL and strings. Finally, multinomial naive Bayes algorithm has been applied with n -grams byte sequences input data to obtain 97.76% detection rate.

In 2004, Kolter and Maloof [20] applied the Information Retrieval techniques to identify the unknown malicious programs. They experimented with 1971 benign and 1651 malware programs, selected 255 million distinct n -grams with different inductive methods, including decision trees (DT), naive Bayes (NB), and SVM. They achieved an area under the receiver operating characteristic curve of 0.996 from the boosted decision tree. Karim et al. in 2005 designed a malware phylogeny model using n -perms as a feature to match the possible permuted code [27]. The experiment has been performed using n -perms and n -grams to compare the relative effectiveness of the similarity measures and the permuted variants programs, respectively. They observed that n -perms provide comparable phylogeny models and high similarity scores for the permuted programs. Also, it can differentiate the related/unrelated in a set of permuted variants. Hence, a good choice for making phylogeny models for the malware is evolved through permutations. Their analysis suggests that the phylogeny models obtained by this method may be able to reconcile the name inconsistencies and can help in the investigation of unknown malware.

In 2006, Henchiri et al. [21] proposed a search method to select generic features for the detection of the computer virus of different families. They used 1512 labeled viruses, 1488 benign program, and Iterative Dichotomiser-3, J48, Naive Bayes, Sequential minimal optimization classifier for the evaluation. Their result outperforms the traditional search methods, and the best detection accuracy obtained was 92.56% by the J48 classifier. They claimed that using a family of non-specific features, their approach performs very well compared to the existing method to detect the unknown malware. Blair in 2007 [28] discusses the opcodes as a predictor of malware, he disassembled 67 malware, and 20 benign executables to study the distributions of malware opcode occurrence. He observed that the distribution of the opcodes between benign and malware programs statistically differ significantly, and the rare opcodes in them appear to be a good predictor of the executables.

In 2008, Moskovitch et al. [29] based on the text categorization concepts proposed a methodology for the detection of unknown malware. They performed an extensive evaluation with $\sim 30,000$ malware and benign programs to investigate the imbalance problem. The accuracy achieved by them was up to 95% by training the model

with a dataset which has below 20% malicious file content. Later on, his group [30] proposed a method to detect the unknown malware using n -grams (3, 4, 5, or 6) of the opcodes as features, generated by disassembling the executables. They took the top 1000 features after applying the document frequency measure from the features which range from thousands to millions. They found that boosted DT provides the best accuracy (94.43%).

To detect the polymorphic/metamorphic malware, Ye et al. [31] in 2008 analyzed the Windows API called by program executables to design an intelligent malware detection system using Objective-Oriented Association (OOA). Their system consists of three major modules, viz., OOA rule generator, PE parser, and rule-based classifier. They conducted a comprehensive experiment on a collection of programs (636 malicious and 1207 benign executables) taken from the Kingsoft Corporation for the comparison of different malware detection techniques. According to them, the efficiency and accuracy of the IMDS system outperform the popular anti-malware, viz., McAfee VirusScan, Norton Antivirus, and other data mining methods, such as SVM, Naive Bayes, and DT. Tian et al. [32] discussed *function length as a tool for malware classification* and claimed that classifying the Trojans on the lengths of their functions will be fast, simple, and scalable. Their result indicates that for the classification of malware, function length plays an important role, and if combined with other features, may lead to an inexpensive, efficient, and scalable technique for the classification. But they also showed that it would be “unrealistic to expect function length information on its own to produce perfect accuracy to distinguish the families.” Siddiqui et al. [33] applied data mining techniques for the extraction of variable length instruction sequence to identify the worms from the benign programs. Their analysis was based on program control flow information contained in the instruction sequences. From these instruction sequences, they formulated a binary classification problem and built tree-based classifiers. For experimental analysis, a dataset of 2774 (1330 benign programs and 1444 worms) is used and detected 95.6% malware in the dataset (not used in building the model).

In 2009, Tabish et al. [34] proposed a non-signature method to analyze the byte-level file content. They claimed that their approach provides implicit robustness against common obfuscation techniques, and also the framework can classify the given malware family. Their approach uses 13 different statistical and information-theoretic features, computed on 1–4g of each file. They have tested their approach with 37,420 malware dataset from VX heaven, and 1800 benign files collected from the different desktops and achieved 90% detection accuracy. Mehdi et al. [35] advocated the need of an accurate, efficient, and sophisticated method for the detection of malware on the very first day of its appearance, called In-Execution Malware Analysis and Detection (IMAD). They claimed that in addition to detect the malware with 90% accuracy while in execution, it can also detect the zero-day malware without a priori knowledge. Their results indicate that IMAD has low false alarm rate and can achieve better accuracy. Later on, they [36] proposed hyper-grams, a variable-length system call method, which uses IMAD for zero-day malware detection. Their experimental analysis with 72 benign and malware files shows that malware detection

with hyper-gram processing overheads is low, and improves the detection accuracy compared to the conventional n -grams.

In 2011, Santos et al. pointed out that supervised machine learning is an effective method to detect the unknown malware, but are not efficient because it requires a significant number of labeled benign and malware files. Therefore, they proposed a single-class learning approach to detect the unknown malware based on the occurrence of opcodes, which does not need a large number of labeled data [37]. They did an empirical study with the dataset comprising 1000 benign and malware executables each [38]. Their analysis shows that the single-class learning reduces the detection cost of the unknown malware. Additionally, they observed that it is more important to obtain labeled benign and malware samples, and shown that by labeling 60% of the legitimate programs, one can achieve $\sim 85\%$ accuracy. In 2012, Ravi et al. [24] proposed an association mining based classification which yields higher detection accuracy than previous data mining methods, by employing NB, SVM, and DT techniques. Their detection system uses the API call sequence modeled by the third-order Markov chain. They compared the accuracy of the proposed system with the existing data mining methods and claimed that their proposed system outperforms (90% of accuracy) the existing malware detection systems.

In 2013, Liangboonprakong et al. [22] proposed a technique to classify the malware family based on n -grams sequential pattern features. They conducted the experiment with four different sizes of n -grams and three classification models (SVM, C4.5 DT, and ANN). From the analysis, they concluded that the larger n -gram size gives higher accuracy, and the proposed feature extraction methods achieved up to 96.64% accuracy with SVM and 4-g. Santos et al. based on the opcode sequence occurrence proposed a technique to identify the unknown malware, and also described a method to find the importance of each opcode. They experimented with a malware dataset of 13,189 [38], and 13,000 benign executables collected from different systems and applications and claimed that their approach gives a good detection ratio with a low false positive ratio. In addition, they provided an empirical validation of the method and found that with one opcode and two opcode sequence length SVM can give an accuracy up to 92.92 and 95.90%, respectively [39].

In 2014, to identify the malicious file Salehi et al. [23] monitored the log files including the names of API, input arguments and combination of this files to generate three feature sets based on runtime behavior of benign and malware executables. They reduced the features by removing the redundant and unnecessary attributes which do not contribute significantly to model the binaries. The experiment was conducted with 385 benign and 826 malware file and used RF, J48, Rotation RF, FT, and NB classifiers. They obtained highest true positive rate (94.6%) from the Random forest by taking only API calls as features, whereas when only argument was taken as features, they found 98.1% true positive rate [23]. In 2015, Jikku Kuriakose et al. presented a non-signature-based detection system to detect metamorphic malware. Their applied feature ranking methods (TF-IDF, TF-IDF-CF, GSS, OR, CMFS, and MOR) was highly successful because MWORM and NGVCK viruses were able to detect with 100% accuracy by using top ten discriminant malware bi-grams. However, experimentally it can be shown that metamorphism exhibited by MWORM and

NGVCK is weak. But the obfuscation techniques such as dead code in the malware can evade the sequence alignment-based technique. Thus, approaches similar to the proposed can find such dead code because it can be effectively synthesized during the feature selection phase. Hence, the developed statistical methods can be used to detect complex obfuscated metamorphic malware [40].

In 2015, Mansour Ahmadi et al. analyzed almost a half terabytes of data which contains more than 20,000 malware samples (Microsoft malware classification challenge, 2015). They proposed an effective learning-based system to categorize the variants of malware into their real family by emphasizing the phases related to the extraction, and selection of prominent features for the malware representation. They extracted the features from the information of the structural characteristics of portable executables and were grouped on the basis of characteristics of malware behavior. With Malware Challenge dataset, the proposed approach detected the malware up to 99.8% accuracy [41].

In 2016, Ashu et al. observed that the size of malware generated from the malware generator kits is within 5 KB range. Therefore, initially using Malicia dataset [42] and NB classifier they investigated the group-wise classification to detect the unknown malware in the 5 KB range and obtained 8.7% more accuracy than the classification done without grouping the data [43]. Later on, classifying the data by random forest, they found that the unknown malware can be detected up to 97.5% accuracy [44]. Further, they grouped the data by optimal k-means clustering algorithm and found that with NBT classifier the detection accuracy can be up to 99.11% [45].

In 2017, Zhixing Xu et al. proposes an online framework to detect the malware using machine learning technique and based on virtual memory access patterns. The technique collects and summarizes the system-call/per-function memory access patterns and a two-level classification architecture. The experimental analysis was focused on the kernel rootkits and memory corruption attacks on user programs by the malware. The framework used logistic regression and random forest classifier and performed on the RIPE benchmark suite for the experiments. The random forest outperformed with a true positive rate of 99% with less than 5% false positive rate [46]. Wang et al. [47] proposed a novel adversary-resistant method that can obstruct attackers from creating adversarial samples by nullifying features randomly within the samples. They evaluated the technique using 17,399 benign and 14,679 malware variants and theoretically validated the robustness of the technique. Their empirical result shows that the technique while maintaining high classification accuracy significantly boosts deep neural networks robustness to adversarial samples. Also, they conducted experiments using the CIFAR-10 & MNIST datasets to show the applicability of their method and obtain maximum 98.61% accuracy on MNIST and 81.07% accuracy on CIFAR-10 datasets.

In 2018, Kotov et al. introduced a static analysis technique which allows generic de-obfuscation of Windows API calls. They used API calls and their arguments used in a binary to model Hidden Markov model and detected the malware with 87.6% accuracy [48]. Burnap et al. proposed a Malware Operational Plot Review (MOPR) model using metrics of the machine activity to automatically differentiate the benign executables and malicious samples. For classification, rather using the raw

data for the feature selection, they have used similar behavior unsupervised cluster using self-organizing feature maps (reduces the overfitting). Their MOPR model obtained 93.76% of malware detection accuracy which they claim an increase of 7.24–25.68% compared to a set of machine learning methods (RF, BayesNet, MLP, and SVM) [49]. A virtual time control mechanics-based method has been proposed by Li et al. [50] utilizing a modified Xen hypervisor, which efficiently identifies the nontrivial anomalous codes which may be passed by the conventional sandboxing techniques. In their approach, a defined speed ratio virtual clock source is generated to accelerate the modified hypervisor in which sandbox systems are running, i.e., it does not modify the kernels of the OS nor intercept the function calls, and therefore can be used in the different OS.

4 Summary

Time-to-time, a number of static and dynamic methods are proposed by the researchers for the detection of second-generation Windows Desktops malware has been discussed in this paper. However, to evade the available detection methods, the malware writers adopted new tactics and techniques to change the behavior of the malware. From the literature survey, we understand that all the malware are not built with the same functionality. Therefore, if one can separate/group the malware family with its functionality, then the malware detection accuracy can be improved, and we hope that in future researcher will investigate more and more on group-wise detection of malware, not only to increase the accuracy but also for its efficient detection.

References

1. Lehtinen, R., Gangemi, G.T.: Computer Security Basics: Computer Security. O'Reilly Media (2006)
2. Sharma, A., Sahay, S.K.: Evolution and detection of polymorphic and metamorphic malwares: a survey. *Int. J. Comput. Appl.* **90**(2), 7–11 (2014)
3. Stone, R.: A call to cyber arms. *Science* **339**(6123), 1026–1027 (2013)
4. Aimoto, S., AlKhatib, T., Coogan, P., Corpin, M., DiMaggio, J.: Internet security threat report. Technical report, Symantec (2014)
5. Daly, M.K.: Advanced persistent threat. *Usenix*, Nov, 4, 2009
6. Aimoto, S., AlKhatib, T., Coogan, P., Corpin, M., DiMaggio, J.: Internet security threat report. Technical report, Symantec Corporation, April 2012
7. The need for speed: 2013 incident response survey. Technical report, FireEye (2013)
8. Quick heal quarterly threat report q2: Technical report, p. 2015. Quick Heal, Feb. 2015
9. Aimoto, S., AlKhatib, T., Coogan, P., Corpin, M., DiMaggio, J.: Internet security threat report 2016. Technical report, Symantec (2016)
10. Beek, C., Frosst, D., Greve, P., Gund, Y., Moreno, F.: McAfee labs threats report. Technical report, McAfee (2015)
11. Aimoto, S., AlKhatib, T., Coogan, P., Corpin, M., DiMaggio, J.: Internet security threat report. Technical report, Symantec Corporation, 2015

12. Beek, C., Frosst, D., Greve, P., Gund, Y., Moreno, F.: McAfee labs threats report. Technical report, McAfee (2014)
13. Aimoto, S., AlKhatib, T., Coogan, P., Corpin, M., DiMaggio, J.: Internet security threat report. Technical report, Symantec Corporation (2017)
14. Ladkat, A., Zure, D., Mathew, L., More, P., Moon, P., Dhasade, P., Kadam, S., Khedkar, S., Girme, T., Chaudhari, L., Sudame, P., Temgire, S., Borse, S., Pharate, S.: Quick heal quarterly threat report | q1 2017. Technical report, Quick Heal (2017)
15. Richardson, R.: 14th annual csi/fbi computer crime and security survey-2009. Technical report (2019)
16. Shah, A.: Approximate Disassembly using Dynamic Programming. Ph.D. thesis, Citeseer (2010)
17. Rad, B.B., Masrom, M., Ibrahim, S.: Camouflage in malware: from encryption to metamorphism. *Int. J. Comput. Sci. Netw. Secur.* **12**(8), 74–83 (2012)
18. Szor, P., Ferrie, P.: Hunting for metamorphic. In: *Virus Bulletin Conference*, pp. 123–144 (2001)
19. Szor, P.: *The Art of Computer Virus Research and Defense*. Pearson Education (2005)
20. Kolter, J.Z., Maloof, M.A.: Learning to detect malicious executables in the wild. In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 470–478. ACM (2004)
21. Henchiri, O., Japkowicz, N.: A feature selection and evaluation scheme for computer virus detection. In: *Sixth International Conference on Data Mining, 2006. ICDM'06*, pp. 891–895. IEEE (2006)
22. Liangboonprakong, C., Sornil, O.: Classification of malware families based on n-grams sequential pattern features. In: *2013 8th IEEE Conference on Industrial Electronics and Applications (ICIEA)*, pp. 777–782. IEEE (2013)
23. Salehi, Z., Sami, A., Ghiyasi, M.: Using feature generation from api calls for malware detection. *Comput. Fraud Secur.* **9**, 9–18 (2014)
24. Ravi, C., Manoharan, R.: Malware detection using windows api sequence and machine learning. *Int. J. Comput. Appl.* **43**(17), 12–16 (2012)
25. Schultz, M.G., Eskin, E., Zadok, E., Stolfo, S.J.: Data mining methods for detection of new malicious executables. In: *2001 IEEE Symposium on Security and Privacy, 2001. S&P 2001. Proceedings*, pp. 38–49. IEEE (2001)
26. Fürnkranz, J., Gamberger, D., Lavrač, N.: Rule learning in a nutshell. In: *Foundations of Rule Learning*, pp. 19–55. Springer (2012)
27. Karim, MdE, Walenstein, A., Lakhota, A., Parida, L.: Malware phylogeny generation using permutations of code. *J. Comput. Virol.* **1**(1–2), 13–23 (2005)
28. Bilar, D.: Opcodes as predictor for malware. *Int. J. Electron. Secur. Digit. Forensics* **1**(2), 156–168 (2007)
29. Moskovitch, R., Elovici, Y., Rokach, L.: Detection of unknown computer worms based on behavioral classification of the host. *Comput. Stat. Data Anal.* **52**(9), 4544–4566 (2008)
30. Moskovitch, R., Feher, C., Tzachar, N., Berger, E., Gitelman, M., Dolev, S., Elovici, Y.: Unknown malcode detection using opcode representation. In: *Intelligence and Security Informatics*, pp. 204–215. Springer (2008)
31. Ye, Y., Wang, D., Li, T., Ye, D., Jiang, Q.: An intelligent pe-malware detection system based on association mining. *J. Comput. Virol.* **4**(4), 323–334 (2008)
32. Tian, R., Batten, L.M., Versteeg, S.C.: Function length as a tool for malware classification. In: *3rd International Conference on Malicious and Unwanted Software, 2008. MALWARE 2008*, pp. 69–76. IEEE (2008)
33. Siddiqui, M., Wang, M.C., Lee, J.: Detecting internet worms using data mining techniques. *J. System. Cybern. Inform.* **6**(6), 48–53 (2008)
34. Tabish, S.M., Shafiq, M.Z., Farooq, M.: Malware detection using statistical analysis of byte-level file content. In: *Proceedings of the ACM SIGKDD Workshop on CyberSecurity and Intelligence Informatics*, pp. 23–31. ACM (2009)

35. Mehdi, S.B., Tanwani, A.K., Farooq, M.: Imad: in-execution malware analysis and detection. In: Proceedings of the 11th Annual Conference on Genetic and Evolutionary Computation, pp. 1553–1560. ACM (2009)
36. Mehdi, B., Ahmed, F., Khayyam, S.A., Farooq, M.: Towards a theory of generalizing system call representation for in-execution malware detection. In: 2010 IEEE International Conference on Communications (ICC), pp. 1–5. IEEE (2010)
37. Santos, I., Nieves, J., Bringas, P.G.: Semi-supervised learning for unknown malware detection. In: International Symposium on Distributed Computing and Artificial Intelligence, pp. 415–422. Springer (2011)
38. webmaster@vxheaven.org. Viruses don't harm, ignorance does. <http://vx.netlux.org> (2017)
39. Santos, I., Brezo, F., Ugarte-Pedrero, X., Bringas, P.G.: Opcode sequences as representation of executables for data-mining-based unknown malware detection. *Inform. Sci.* **231**, 64–82 (2013)
40. Kuriakose, J., Vinod, P.: Unknown metamorphic malware detection: modelling with fewer relevant features and robust feature selection techniques. *IAENG Int. J. Comput. Sci.* **42**(2), 139–151 (2015)
41. Ahmadi, M., Ulyanov, D., Semenov, S., Trofimov, M., Giacinto, G.: Novel feature extraction, selection and fusion for effective malware family classification. In: Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy, pp. 183–194. ACM (2016)
42. Nappa, A., Rafique, M.Z., Caballero, J.: Driving in the cloud: an analysis of drive-by download operations and abuse reporting. In: Detection of Intrusions and Malware, and Vulnerability Assessment, pp. 1–20. Springer (2013)
43. Sharma, A., Sahay, S.K.: An investigation of the classifiers to detect android malicious apps. In: Proceedings of ICICT 2016 Information and Communication Technology, vol. 625, pp. 207–217. Springer (2017)
44. Sharma, A., Sahay, S.K., Kumar, A.: Improving the detection accuracy of unknown malware by partitioning the executables in groups. In: Advanced Computing and Communication Technologies, pp. 421–431. Springer (2016)
45. Sahay, S.K., Sharma, A.: Grouping the executables to detect malwares with high accuracy. *Procedia Comput. Sci.* **78**, 667–674 (2016)
46. Xu, Z., Ray, S., Subramanyan, P., Malik, S.: Malware detection using machine learning based analysis of virtual memory access patterns. In: 2017 Design, Automation & Test in Europe Conference & Exhibition (DATE), pp. 169–174. IEEE (2017)
47. Xing, X., Giles, C.L., Zhang, K., Ororbica, A.G., Liu, X.: Adversary resistant deep neural networks with an application to malware detection. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining INeural Networks with an Application to Malware Detection, pp. 1145–1153. ACM (2018)
48. Kotov, V., Wojnowicz, M.: Towards generic deobfuscation of windows api calls. arXiv preprint [arXiv:1802.04466](https://arxiv.org/abs/1802.04466) (2018)
49. Burnap, P., French, R., Turner, F., Jones, K.: Malware classification using self organising feature maps and machine activity data. *Comput. Secur.* **73**, 399–410 (2018)
50. Li, Z.-q., Qiao, Y.-c., Hasan, T., Jiang, Q.-s.: A similar module extraction approach for android malware. *DEStech Trans. Comput. Sci. Eng. (mso)* (2018)

Machine Learning Based Approach for Detection of Lung Cancer in DICOM CT Image



Chethan Dev, Kripa Kumar, Arjun Palathil, T. Anjali and Vinitha Panicker

Abstract Lung cancer is one of the leading causes of cancer among all other types of cancer. Thus, an early and effective identification of lung cancer can increase the survival rate among patients. This method presents a computer-aided classification method in computerized tomography images of lungs. In the proposed system, MATLAB has been used for implementing all the procedures. The various stages involved include image acquisition, image preprocessing, segmentation, feature extraction and support vector machine (SVM) classification. First, the DICOM format lung CT image is passed as input which undergoes preprocessing. Then, a threshold value is calculated and image is segmented into left lung and right lung. After that 33 features of each segmented lung are taken and passed as input to the SVM. Finally, the image is classified as cancerous or non-cancerous based on the training data. This method aims to give more satisfactory results when compared to other existing systems.

Keywords Preprocessing of image · Segmentation · Extraction of features · Support vector machine

C. Dev (✉) · K. Kumar · A. Palathil · T. Anjali · V. Panicker
Department of Computer Science, Amrita Vishwa Vidyapeetham, Amrita School of Engineering,
Amritapuri, Kollam, Kerala, India
e-mail: devchethan@gmail.com

K. Kumar
e-mail: kripakumar24@gmail.com

A. Palathil
e-mail: arjunpalathil2014@gmail.com

T. Anjali
e-mail: anjalit@am.amrita.edu

V. Panicker
e-mail: vinithapanicker@am.amrita.edu

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_15

1 Introduction

Lung cancer is caused due to uncontrolled multiplication of abnormal cells, finally causing tumour. The malignant tumour grows locally and can also spread into the bloodstream or lymphatic system and spread to the other areas of the body. People of ages above 40 are the ones who are most affected. About 7.6 million people died because of lung cancer every year according to world health organization. The number of deaths caused due to lung cancer is more when compared to any other type of cancer. More than 25% of entire cancer-associated deaths are due to lung cancer alone. Lung cancer is divided into two types, i.e. small cell lung cancer and non-small-cell lung cancer (most common lung cancer). It consists of four stages. In the first stage, cancer is caused in the lung region and in stages 2 and 3, it is spread onto the chest. At stage 4, it spreads to all parts of the body. This disease mostly occurs in the people who smoke frequently. The mortality rate of cancer may become around 17% around the entire world. The survival rate can definitely increase by 47% if we successfully detect cancer in its premature stage itself. In most cases, doctors tend to detect cancer based on the symptoms of a patient which starts to arise only when it reaches its advanced stage. The symptoms are paining of bone or chest, weight loss, coughing blood and getting no proper breath. It is highly unpredictable when it comes to detection of the lung nodules in its premature stage. There are various methods present in hospitals that detect cancer when it reaches its advanced stage like X-ray, magnetic resonance imaging scan, computed tomography scan, etc. The main drawback of these methods is that they are expensive and take more time for detection. And doctors mostly tend to detect cancer based on the symptoms of the patient which arises only its advanced stages. The survival rate of lung cancer is about 15%. Therefore, to improve the cure rates, there is an urgent need to auto-detect lung cancer in its premature stage itself.

2 Related Work

Lung cancer detection is one of the major issues the world is facing today. To improve the detection rate of cancerous nodule, many works are going in and around the world. Estimating the cancerous nodule in its premature stage is an important factor. Currently, there exist various methods for cancer detection. But the drawback in all these methods is that they have certain amount of false positive rate associated with it. In lung cancer detection method [1], the authors have suggested a method that integrates a strong feature extraction method in which all the features are passed through an artificial neural network (ANN) followed by training the system for classification purpose. In an enhanced k -nearest neighbour method for detecting and classifying MRI lung cancer images for large amount data [2], they have developed a method to improve the classification performance using enhanced K-NN where the images are first trained. While testing an image, the Euclidian distance is calculated

between the two images and the minimum distance k is labelled as one class. In [3] after the preprocessing, the segmentation is done using ROI region extraction method. The features are extracted using grey-level co-occurrence matrix (GLCM). The features are area, convex area, equivalent diameter, solidity, energy, contrast, homogeneity, correlation and eccentricity. The values obtained are passed as inputs to support vector machine (SVM) with maximum margin hyperplane with linear classifier and classify as cancerous or not. An evaluation of features extraction from lung CT images for the classification stage of malignancy [4] has proposed a system in which the lung image undergoes preprocessing which includes grayscale conversion, image enhancement, histogram equalization to stretch the low contrast image and noise removal. Next segmentation is done by thresholding method and finally features such as area, perimeter, shape complexity, mean, standard deviation and circularity are obtained. In the paper Lung Cancer Detection and Classification [5], the authors have suggested various preprocessing stages to get more accurate results based on various enhancement and segmentation techniques and the area, perimeter, average intensity, etc. are the features extracted and the lung nodules are classified based on the size of the lung nodule [6]. Bayesian classifier is based on Bayes theorem. Bayesian classifier makes use of the class membership probabilities, such as the probability that a given sample belongs to a particular class. In [7], the authors have proposed a system for depression detection based on facial expressions. They have used Viola–Jones face detection algorithm to detect faces. A Gabor filter bank of 40 filters is used to find the facial features of different faces. A set of feature vectors are obtained and tested using SVM classifier. The feature set used in [8] is based on texture, oriented gradient's histogram and colour moments, and is classified to find out which plant the leaf belongs to and thereby knowing its medicinal value [9]. It presents a survey on deep convolution neural network in image processing where a neural network performs well for classification of large data and how it is possible to concatenate supervised learning technique with deep architecture for enhancement of the learning process. In [10], author has proposed an improved SVM classifier to detect leukaemia cancer using fast-correlation-based filter to select genes which are most prominent and not correlated. Here, the image is normalized to reduce the computation time and leads to more accurate results. Detection of lung cancer stages on CT scan images using various image processing techniques [11] has used Gabor filter and watershed segmentation to give satisfying results in the preprocessing stage, and they have measured the dimensions of the tumour to detect the lung cancer stage accurately [12]. Image processing-based detection of lung cancer on CT scan images has used binarization technique to detect if the lung is normal or abnormal. Lung cancer detection using SVM algorithm and optimization techniques [13] uses a particle swarm optimization method, i.e. a multilevel threshold method is used for segmenting the images. Lung cancer detection and classification using machine learning and multinomial Bayesian lung cancer detection and classification by using machine learning and multinomial bayesian [14] has used kernalised Bayesian technique for classification. In lung cancer detection using image processing techniques [15], pixel percentage and mask labelling with high accuracy and robust operation are used as main detected features for image comparison. Lung cancer detection using

neural networks [16] comprises image enhancement techniques such as fast Fourier transform and log-Gabor filter. Lung cancer detection by using ANN and fuzzy clustering methods [17] presents two segmentation methods, i.e. Hopfield neural network and fuzzy C-mean clustering algorithm. Prashant Naresh Int. Journal of Engineering Research and Applications [18] has used structural and textural features and also SVM with RBF kernel to obtain high accuracy.

The proposed system has been successful in reducing the false positive rate and can also detect cancer nodules present at the edges of the lung. Compared to the existing systems, our system consists of fewer and quality steps in preprocessing that helped us save the time required for training and testing phase and we have been able to acquire high accuracy for the given dataset. The main contribution of our proposed system is that we have combined the strong feature selection method for the DICOM format images and passed it to the SVM which has not been experimented in any of the existing work.

3 Proposed System

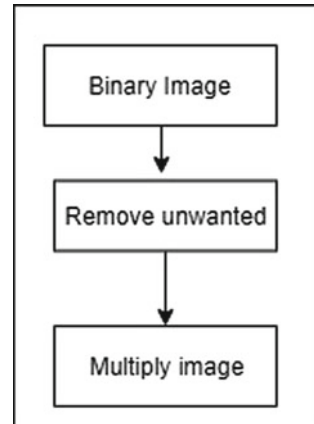
This process of detection of lung cancer is divided into various stages—Image acquisition, image preprocessing, segmentation, feature extraction and SVM classification.

3.1 Image Acquisition

The input to our system is lung computerized tomography image. The reason to choose CT image is that it has less noise or distortion and high clarity. The file format used for this system is Digital Imaging and Communications in Medicine (DICOM) format. The software called RadiAnt has been used to view these DICOM images.

3.2 Preprocessing of Image

This step is done to obtain a clear image and also removes unnecessary information present in the background of the image (Fig. 1).

Fig. 1 Preprocessing steps

3.3 Convert to Binary Image

To convert the image into 0(black) and 1(white) form, we use the Matlab function `imbinarize`. The advantage of doing this step is that it promotes fast processing and needs less memory space.

3.4 Eliminate Unwanted Region

To remove the unwanted portion of the image, we use the function `bwareafilt`. It extracts all connected component of white pixels and assigns a label to each. It has n largest objects and only the largest connected component which is assigned the first label is displayed and the rest is removed. Hence, the two white lines appearing at the bottom of the binarized image are removed, and further processing is made easier.

3.5 Multiply Image

Now the noise reduced grayscale image is multiplied with the binarised image in which all the unwanted regions are removed. This step is done using the Matlab function `immultiply`. The resultant image is a grayscale image which is used for further processing of the image (Figs. 2 and 3).

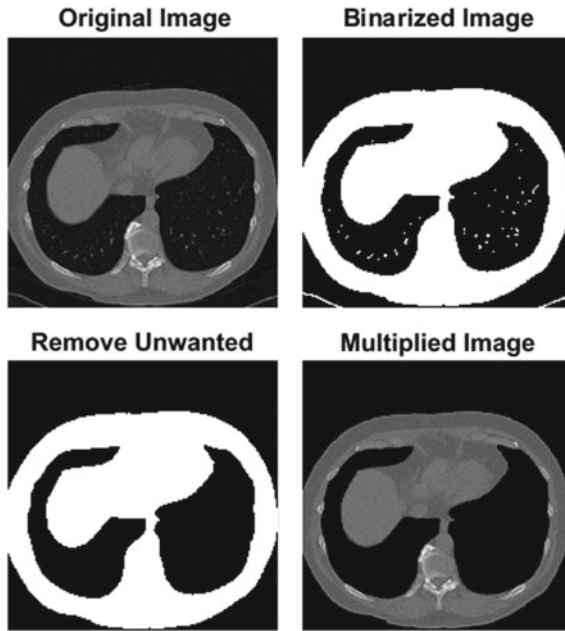


Fig. 2 Preprocessing steps

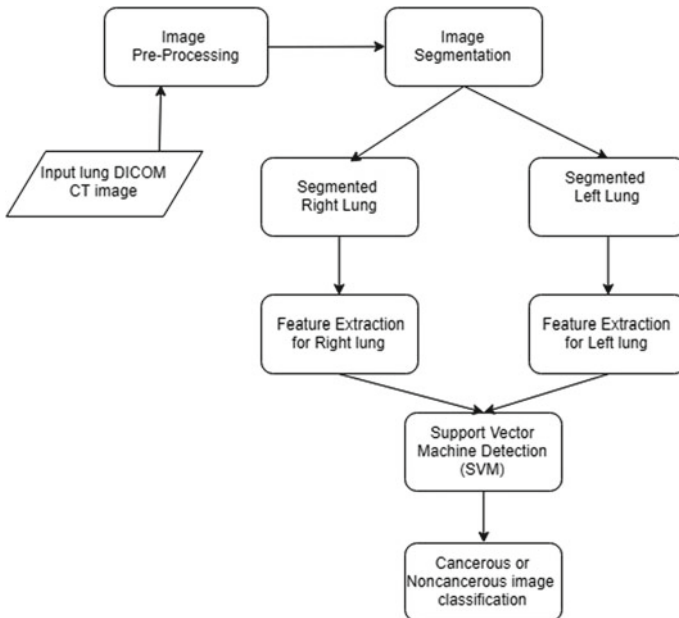


Fig. 3 Stages of lung cancer detection

3.6 Segmentation

Image segmentation is the technique of dividing the DICOM image into various parts. The main aim of segmentation is to identify objects or other relevant information from the digital image so that we can examine the image more precisely. It is usually used to locate objects and borders in images. The process of segmentation is composed of different steps. It starts by clearing the borders of the original image followed by filling up the holes present in the image. Then, a process called dilation is performed on the image. Each pixel is assigned a label such that the pixel with same label shares specific attributes. Finally, left and right lungs are segmented from the original image (Fig. 4).

3.7 Extraction of Features

The system has been introduced with a strong feature extraction method to extract 33 features for each type of the segmented lung CT image.

Feature 1

The proportion of height to width of segmented lung is taken.

$$\text{The 1st feature is : feature1} = \text{height/width} \quad (1)$$

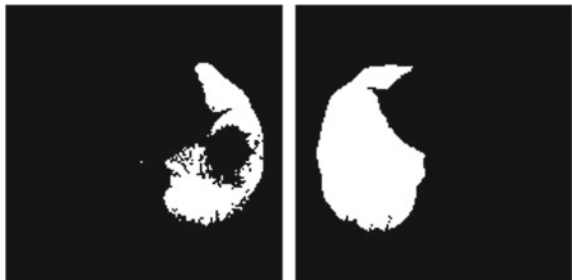
Feature 2–25

These features are used to find the distribution of the white pixels in the image, and for this we have to find the total pixels of the image.

$$\text{totalpixels} = \text{height} * \text{width} \quad (2)$$

The features 2–5 are taken to find the distribution of white pixels located in the upper, down, left and right, respectively.

Fig. 4 Segmented left lung and segmented right lung



$$\text{feature2} = \text{uppixels}/\text{totalpixels} \quad (3)$$

$$\text{feature3} = \text{downpixels}/\text{totalpixels} \quad (4)$$

$$\text{feature4} = \text{leftpixels}/\text{totalpixels} \quad (5)$$

$$\text{feature4} = \text{rightpixels}/\text{totalpixels} \quad (6)$$

The rest of the features from 6 to 25 are to find the distribution of white pixels in every subregion of the image (Fig. 5).

$$\text{feature} = \text{subareapixel}/\text{totalpixels} \quad (7)$$

Feature 26

To determine the mean distance from the entire white pixels to the central point, where (x, y) -coordinates of central.

$$\text{feature26} = 1/\text{totalpixels} * \sum_y \sum_x \sqrt{(x - i)^2 * (y - j)^2} \quad (8)$$

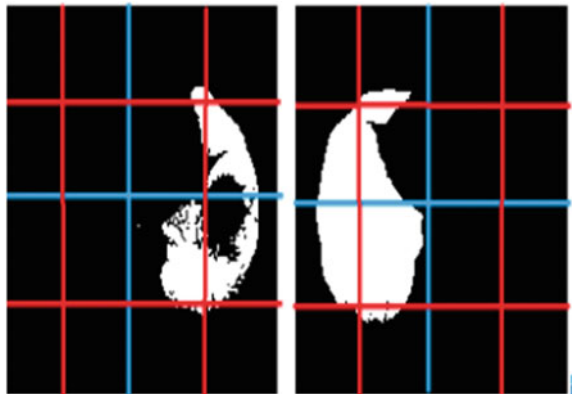
point (i, j) -each pixels in the image

Feature 27–33

Determine the image moment of the segmented lung. Image moments give description of an object that exceptionally represents its shape.

$$m_{pq} = \sum_{x=1}^M \sum_{y=1}^N (x)^p (y)^q f(x, y) \quad (9)$$

Fig. 5 Feature extraction of subarea pixel



As the image moment changes according to the position of the image, there is a need for normalization and it is done by calculating the central moment which is done by subtracting the x and y with their respective mean values.

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q f(x, y) \quad (10)$$

For scaling normalization central moment changes as

$$\eta_{pq} = \mu_{pq} / \mu_{00}^r \quad (11)$$

$$r = [(p + q)/2] + 1 \quad (12)$$

seven of the shape descriptor values that are computed from central moments that are independent to object translation, scale and orientation are

$$C_1 = \eta_{20} + \eta_{02} \quad (13)$$

$$C_2 = (\eta_{20} + \eta_{02})^2 + 4\eta_{11}^2 \quad (14)$$

$$C_3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \quad (15)$$

$$C_4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \quad (16)$$

$$\begin{aligned} C_5 = & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03}) \\ & + [3(\eta_{20} - \eta_{12}^2 - (\eta_{21} + \eta_{03})^2) \end{aligned} \quad (17)$$

$$\begin{aligned} C_6 = & (\eta_{20} - \eta_{20})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\ & + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \end{aligned} \quad (18)$$

$$\begin{aligned} C_7 = & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & - (\eta_{30} + 3\eta_{30})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \quad (19)$$

4 Support Vector Machine Classification

A SVM is used for classification purpose of cancerous and non-cancerous images. It takes a set of inputs and during testing phase it decides which of the two inputs it belongs to as it is a binary linear classifier. Here, we use the maximum margin hyperplane. The hyperplane learns through the training phase and to maximize the margin

it has an optimization procedure. All the above-mentioned modules are implemented on the training dataset images and their features are obtained. Thirty-three features of right and left lung each are stored in an array and are passed as inputs through the SVM to prepare the system for classification purpose. After the training of SVM is completed, the features of the test image are passed on to the machine and the SVM classifies the image based on its training. A label is assigned to both cancerous and non-cancerous images as 0 and 1, respectively, during the training phase. The values of the test data are compared with the values of the train data and as per the predicted output if the answer is 0 it is cancerous and if it is 1 then non-cancerous.

5 Testing and Training

The features extracted are then used to train the system for classification. The proposed system detects whether the right or left lung is affected. The dataset used in the proposed solution consists of total 200 DICOM lung CT images. First, the system is trained with 120 images including 60 cancerous and 60 non-cancerous. After the training stage, we test the images and specify if the lung is affected or not. To detect the lung cancer accurately, the system tests about 80 lung CT images which include 40 cancerous and 40 non-cancerous images. Out of the 40 cancerous images, there are 24 images whose right lung is affected and 16 images whose left lung is affected. As after segmentation, the total lung image gets divided into two separate images as shown in Fig. 4. Features of right lung and left lung are passed separately such that SVM classifies which out of the two is affected specifically.

6 Experimental Results

The dataset used in the proposed solution having DICOM lung CT images is of size 512×512 . Tables 1 and 2 are the result for confusion matrix which includes true positive, true negative, false negative and false positive rate. Therefore, we obtain an accuracy of 86.25% for the given set of images. If we give more training samples, the accuracy will increase to an even higher extent.

Table 1 Confusion matrix

	Predicted No	Predicted Yes
Actual No	35	5
Actual Yes	6	34

Table 2 Experimental results

Accuracy	86.25%
True positive rate	0.85
False positive rate	0.125
Specificity	0.875
Precision	0.871

7 Comparison with the Existing System

The proposed system comprises a strong feature selection method when compared to classification of lung tumour using SVM that uses GLCM features. The proposed system consists of a process of conversion to binary image which needs less memory and promotes fast processing of training and testing the images. The existing method provides only with 25 diseased lung CT JPEG images. This system is trained using 60 diseased images and tested using 40 diseased images of DICOM CT image which provides more information compared to a CT image.

8 Implementation

This computer-aided classification method is developed using the software MATLAB. It comprises of a language that is highly efficient and used for technical computation. It puts together computations, visuals and programming where problems and solutions are given in well-known mathematical expressions. Toolbox is the main feature of MATLAB which has a huge collection of functions. It helps to learn and apply specialized technology to provide solutions to problems of a specific class. It includes SVM and image processing toolbox.

9 Conclusion

The proposed method helps physicians to discover cancer in its early stage to increase the survival rate among patients.

Compared to JPG format used in most of the existing methods, DICOM format images used in the proposed system prove to provide more efficient results. We had also implemented the same method using probabilistic neural network where it provided low accuracy compared to SVM. In the first phase, the input DICOM lung CT image undergoes preprocessing which includes noise reduction and binarization of the image. In the second stage, image is segmented into right and left lung. The third stage comprises a strong feature selection method. The extracted features are passed to the SVM for classification of cancerous or non-cancerous images. The overall

accuracy of the system is 86.25%. For further future work, a dynamic thresholding method can be used and a segmentation method that segments only the tumour region can be used to improve the accuracy of the system.

10 Source of CT Images Used

The dataset of DICOM images used in the proposed system has been downloaded from the free dataset in Internet and also obtained from Amrita Institute of Medical Sciences hospital, AIMS Kochi. Figures 2 and 4 displayed in the paper are the screenshots of the obtained output of preprocessing and segmentation, respectively, from MATLAB.

References

1. Miah, Md. B.A., Yousuf, M.A.: Detection of lung cancer from CT image using image processing and neural network. In: 2nd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT) (2015)
2. Thamilselvan, P., Sathiaseelan, J.G.R.: An enhanced k nearest neighbor method to detecting and classifying MRI lung cancer images for large amount data. *Int. J. Appl. Eng. Res.* **11**, 4223–4229 (2016)
3. Tidke, S.P., Chakkarwar, V.A.: Classification of lung tumor using SVM. *Int. J. Comput. Eng. Res. (IJCER)* **2**(5)
4. Singh, S., Singh, Y., Vijay, R.: An evaluation of features extraction from lung CT images for the classification stage of malignancy. *IOSR J. Comput. Eng. (IOSR-JCE)*, 2278–8727
5. Kanitkar, S., Thombare, N.D., Lokhande, S.S.: Lung cancer detection and classification: a review. *Int. J. Eng. Res. Technol. (IJERT)* **2**(12) (2013)
6. Rocky, H., Jereesh, A.S.: Lung cancer detection using Bayesian classifier. *Int. J. Adv. Inf. Sci. Technol. (IIAIST)*, 2319, 2682
7. Venkataraman, D., Parameswaran, N.S.: Extraction of facial features for depression detection among students. *Int. J. Pure Appl. Math.*
8. Venkataraman, D., Mangayarkarasi, N.: Support vector machine based classification of medicinal plants using leaf features. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2017)
9. Aarthi, R., Harini, S.: A survey of deep convolutional neural network applications in image processing. *Int. J. Pure Appl. Math.*
10. Kavitha, K.R., Gopinath, A., Gopi, M.: Applying improved SVM classifier for leukemia cancer classification using FCBF. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI) (2017)
11. Gajdhane, A.V., Deshpande, L.M.: Detection of lung cancer stages on CT scan images by using various image processing techniques. *IOSR J. Comput. Eng. (IOSR-JCE)*
12. Abdillah, B., Bustamam, A., Sarwinda, D.: Image processing based detection of lung cancer on CT scan images. In: Asian Mathematical Conference 2016 (AMC 2016)
13. Asuntha, A., Brindha, A., Indirani, S., Srinivasan, A.: Lung cancer detection using SVM algorithm and optimization techniques. *J. Chem. Pharm. Sci.*
14. Dwivedi, S.A., Borse, R.P., Yametkar, A.M.: Lung cancer detection and classification by using machine learning and multinomial Bayesian. *IOSR J. Electron. Commun. Eng. (IOSR-JECE)*

15. AL-Tarawneh, M.S.: Lung cancer detection using image processing techniques. *Leonardo Electron. J. Pract. Technol.* ISSN 1583-1078
16. Sehgal, R., Gupta, S.: Lung cancer detection using neural networks. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* ISSN: 2277 128X
17. Taher, F., Werghi, N., Al-Ahmad, H., Sammouda, R.: Lung cancer detection by using artificial neural network and fuzzy clustering methods. *Am. J. Biomed. Eng.* **2**(3), 136–142 (2012)
18. Naresh, P., Shettar, R.: Early detection of lung cancer using neural network techniques. *Int. J. Eng. Res. Appl.*

Extractive Summary: An Optimization Approach Using Bat Algorithm



Anshuman Pattanaik, Santwana Sagnika, Madhabananda Das
and Bhabani Sankar Prasad Mishra

Abstract A summary is the shorter version of the existing long text which elaborates the whole idea of the document. The most conventional and easy version of summarization is extractive text summarization. Extractive approach selects sentences from the original text based on some features. In automatic text summarization (ATS), it is a difficult task to select such sentences with high accuracy to meet the optimum meaning of the full text as compared to manual approaches. In this paper, author tries to represent extractive summary as an optimization problem where the objective is to cover maximum topics of the document and simultaneously minimize the redundancies between the sentences of the summary. Bat algorithm (BA) is used as an optimization technique which provides efficient result in creating an extractive summary.

Keywords Text summarization · Extractive · Sentence scoring · Multi-objective optimization · Bat algorithm · TF-ISF

1 Introduction

Automatic text summarization [1] (ATS) is one of the oldest fields of text processing which started in 1958 [2]. The author elaborated generation of abstracts out of long literature. Due to excessive use of Internet and World Wide Web, availability of text

A. Pattanaik (✉) · S. Sagnika · M. Das · B. S. P. Mishra
School of Computer Engineering, Kalinga Institute of Industrial Technology, (Deemed-to-be University), Bhubaneswar, India
e-mail: anshumanpattanaik21@gmail.com

S. Sagnika
e-mail: santwana.sagnika@gmail.com

M. Das
e-mail: mnd_prof@kiit.ac.in

B. S. P. Mishra
e-mail: mishra.bsp@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_16

data grows exponentially, which boosts the interest of researchers in this area. In the last two decades, various ATS techniques have been evolved. Before the age of the digital world, précis writing is a general practice between the philologists. Précis is the summarized version of the actual text or audio. Text processing summarization plays a vital role while it comes to document understanding and information retrieval. In general, summarization is classified into two major categories, such as extractive and abstractive.

An extractive [1] summary is the most primitive way of summarization. In this approach, simple random sentences from the original document are selected in order to explain the whole idea of the document. It seems to be a simple yet complicated process. In order to choose the sentences which will give the effective result for the summary, many concepts and parameters are taken into account by various researchers in due times.

An abstractive [1] summary can be explained as a paraphrased compressed data of the existing document. What it really means a shorter paragraph or group of sentences which elaborate the total meaning of the original document. Abstractive summarization is way different from the extractive approach; there are no sentence similarities between original document and summary.

In this paper, the author focuses on extractive summary. In this method, various sentence scoring and selection mechanisms are taken into consideration for summary generation. Basically, it extracts words, phrases or sentences from the original text to represent the whole idea of the article. The author tries to represent this as an optimization problem, while maximizing sentence coverage and minimizing redundancies between them. Term frequency–inverse sentences frequency [3] is used for sentences scoring. Bat algorithm (BA) is implemented for optimization purpose, which is an echolocation-based algorithm. This algorithm mimics the behaviours of bats and how they avoid obstacles and find their food particle through their voice frequencies.

This paper covers the related works, followed by basic summarization concepts, BA, then proposed method, implementation criteria, result analysis and conclusion.

2 Related Work

ATS is a vital field to cover due to its vast use and numerous data available in the digital world. ATS is classified into different formats due to styles, approaches, input type, output style and many more, but the basic approach is dependent on output approach. Abstractive summarization is a bit more complicated than the extractive type to generate. The extractive approach is the basic approach where the sentences/phrases are selected which elaborate the whole article. Various researchers implemented so many techniques over the years to create an effective extractive summary.

In creating extractive summaries, various sentence selection techniques have been implemented over the time period, such as term frequency, tf-idf, cue words, noun, proper noun, title word, etc.

Alguliyev et al. [4] in 2016 focused on sentences scoring and sentences selection for creating extractive summary. Sentences scoring can be done in various ways. In this paper, author took term frequency–inverse sentences frequency as scoring technique. After generating score for each sentence, the second phase is sentence selection; here, one optimization algorithm taken into consideration is named as human learning optimization algorithm. Objective of the algorithm to maximize the coverage and minimize the redundancies of the summary generated.

García-Hernández et al. [5] in 2013 applied the genetic algorithm in extractive text summarization. In this paper, author considers sentences selection as an evolutionary problem. Let the summary have $k\%$ of compression rate. Then, the chromosome is selected in such a way that it must satisfy the fitness function. Here, author considers highly frequent words and sentences position as the key to the generating summary.

Binwahlan et al. [6] in 2009 approached for extractive summary generation using swarm intelligence as solving domain. In this paper, the author considered five different sentence scoring techniques such as sentence centrality, title feature, word sentence score, key words and sentence similarity. The total algorithm is classified into three major sections such as sentence feature generation, identification process using PSO and swarm-based approach for summary generation.

We can consider extractive text summarization as an optimization problem. This problem can be explained by maximization of coverage of topics in the document. Many optimization techniques are available for various problems. Many heuristics and evolutionary bio-inspired optimization algorithms are also there such as GA, PSO, CSO, ABC, ant-colony optimization, etc.

BA [7] is one of the many bio-inspired meta-heuristic approaches which mimics the food-finding technique of bat which is echolocation based.

Sagnika et al. [8] in 2018 introduce BA for workflow scheduling in cloud computing. In the paper, the author focuses on processing cost along with coverage and load distribution. Their experimental analysis clearly states that BA provides better result as compared to PSO and CSO.

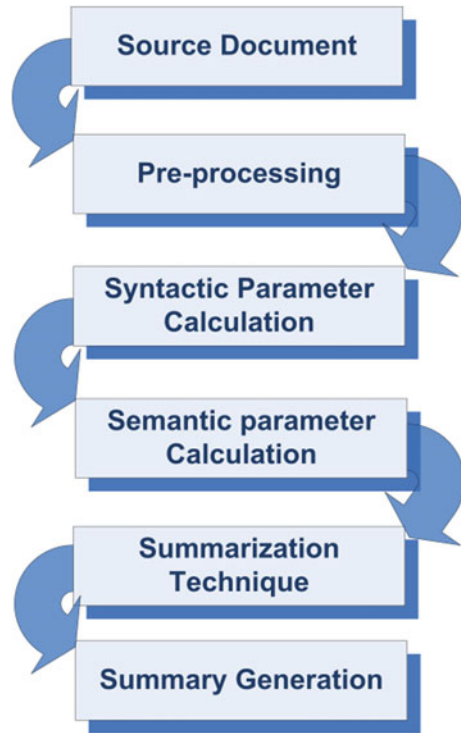
In this paper, the author focuses on utilizing the BA for optimizing the extractive summary by maximizing the coverage and minimizing the overlapping between the sentences semantics. In upcoming sections, all the concepts, as well as experiments, are well explained.

3 Summarization Concepts

Summarization is the method of representing original article in a compact form which reflects the original idea of the document. Generally, ATS is classified into two categories such as extractive and abstractive. In this paper, we will focus on extractive summarization.

Extractive summarization is the common approach to text summarization. In this approach, sentences are first assigned to some scores and ranked accordingly. Then best-fit sentences are picked from the list to form the summary. Sentences are picked

Fig. 1 Basic summarization mechanism



in such a way that it covers most of the topics of the whole document. In this one, more factor comes into account, i.e. ‘Compression rate’. The compression rate is the ratio between the number of words in the summary to the number of words in the original text. Generally, it varies from 20 to 40% of the original text. Figure 1 gives the brief idea how summarizer works.

Before generating the summary, source data undergoes various phases such as pre-processing, syntactic parameter extraction, semantic parameter extraction, applying summarization mechanism and generation of the summary. All these phases are default phases for both extractive and abstractive type of summarization.

Pre-processing phase is the initial phase where the raw data undergoes cleaning, which includes omitting stop words, tokenization, case folding, lemmatization, name entity reorganization, POS tagging, etc.

Semantic parameter calculation is one of the vital phases in ATS. In this phase sentences, scoring process is taken into consideration. There are several sentences scoring techniques available such as term frequency [2], sentences location [9–11], cue words [10], title similarity [10, 11], proper noun [11], word co-occurrence [12], term frequency–inverse document frequency [11], positive keyword, negative keyword, name entity, etc.

Semantic parameters are the parameters which focus on the linguistic relationship between sentences as well as words and phrases. Malik et al. [13] explained the sentences similarity calculation using cosine similarity parameters. Word ordering [14] is one of the features in semantic parameter calculation as two sentences having same words but a different order can change the linguistic meaning of it.

After generation of syntactic and semantic parameters, various techniques are implemented over and over to find the best-fit sentences according to score and rank of sentences. As an output to this phase, a summary is generated as a final outcome of the source document.

4 Bat Prey Hunting Behaviour

Bats are nocturnal birds. In dark, they move and hunt for their prey. In low light, it is very difficult to find food but they manage to achieve their food. They use some sound waves for that. They emit specific wavelength sound waves and frequencies, and then wait for the echo. They analyse the variation in loudness and frequencies to find the distance of the prey. This behaviour is known as echolocation-based behaviour [7, 15, 16]. Bat can emit very short-lived sound pulses whose frequency can vary from 25 to 100 kHz. They emit more sound pulses; while nearing the prey, it ranges from 10 to 200 per seconds. These are ultrasonic sounds with very high dB (nearly 110 dB). By the use of echolocation mechanism, they determine size, speed, position and location of the prey ahead. By mimicking this behaviour of bats, a search technique was designed by Yang [7] in 2010. This search technique is mainly used in optimization problems to find the optimal solution for the problem domain.

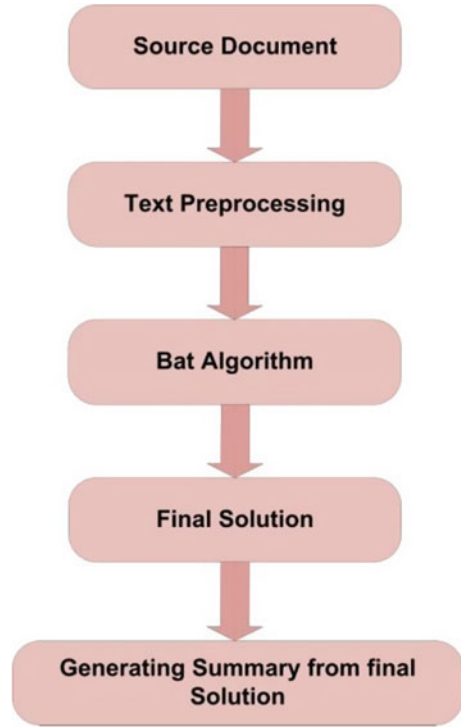
5 Proposed Method

In this paper, extractive summarization is considered as an optimization problem which is addressed by BA to understand how the complete process works we have to understand the process phase by phase. Figure 2 explains how the proposed algorithm works.

In basic summarization techniques, pre-processing indicates the removal of stop words, case folding, lemmatization and many more. In our approach for summarization, we first split the total data into sentences and then into words. Then, all the stop words are removed from the bundles of the words. All the words are converted to lower case. Then, we designed our objective function in such a way that we addressed two major issues in summaries. First, it focuses on coverage of all the topics and second, it tries to minimize the redundancies between sentences of the summary. This objective function can be represented as stated below.

Objective function

Fig. 2 Schematic diagram of the proposed algorithm



$$f(y) = \alpha \cdot f_{\text{coverage}}(y) + (1 - \alpha) \cdot f_{\text{redundancy}}(y), \quad (1)$$

Subject to

$$\sum_{i=1}^n \text{len}(S_i) y_i \leq L_{\max}, \quad (2)$$

$$y_i \in \{0, 1\}, \quad i = 1, \dots, n \quad (3)$$

where y_i defined as follows:

$$y_i = \begin{cases} 1 & \text{if } S_i \text{ is selected} \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

L_{\max} is the predefined length of the summary and $\text{len}(S_i)$ is the length of the sentence S_i

$f_{\text{coverage}}(y)$ and $f_{\text{redundancy}}(y)$ refer to the function which covers all the topics and minimize sentence to sentences redundancies, respectively. Parameter α works as a binding factor between the two functions in Eq. 1. $f_{\text{coverage}}(y)$ and $f_{\text{redundancy}}(y)$ is explained as follows:

$$f_{\text{coverage}}(y) = \sum_{i=1}^n \text{Score}(S_i) y_i \quad (5)$$

$$f_{\text{redundancy}}(y) = \sum_{i=1}^{n-1} \sum_{j=i+1}^n (1 - \text{Similarity}(S_i, S_j)) y_i y_j \quad (6)$$

Equations 1–3 work in such a way that they will produce a summary of a length at $\max L_{\max}$, which satisfies both the cases of Eqs. 5 and 6 by providing high coverage and less redundant. In Eq. 5 to solve coverage, we need to find out how to find score of a sentence. Score of a sentence refers to its frequency value. Score of a sentence can be calculated by the following equation:

$$\text{Score}(S_i) = \sum_{j=1}^{I_i} \text{Mass}(S_j) \times \exp\left(-\left(\frac{1 - \text{Similarity}(S_i, S_j)}{\sigma}\right)^2\right) \quad (7)$$

where $i = 1, \dots, n$; S_j is a sentence having similarity with the sentence S_i , i.e. $\text{Similarity}(S_i, S_j) \neq 0$ and I_i is the number of sentences having similarity with S_j . $\text{Mass}(S_j)$ is the ‘mass’ of sentence S_j . σ controls the influence scope of the sentence. The optimal value can be found from the method proposed by Gan et al. [17] in 2009.

Mass of the sentence is one of the important factors which has vital impact on score of the sentence. Mass can be calculated as follows:

$$\text{Mass}(S_i) = \frac{\text{Similarity}(S_i, P)}{\sum_{j=1}^n \text{Similarity}(S_j, P)} \quad (8)$$

where S_i is the sentence. P is the weight of the sentence compared to the whole data (sentences) available in the document and can be calculated as follows:

$$P = \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^m W_{il} \quad (9)$$

where

$$i = 1, \dots, n$$

$$l = 1, \dots, m$$

n is the number of sentences in the document and m is the number of words in the sentence.

Similarity is the function to find similarities between two sentences in natural language processing. There are several methods available to find similarities such as Jaccard [18] similarity and cosine [19] similarity. In this work, authors have utilized cosine similarity. The equation can be devised as follows:

$$\text{Similarity}(S_i, S_j) = \frac{\sum_{l=1}^m W_{il} W_{jl}}{\sqrt{(\sum_{l=1}^m W_{il}^2 \cdot \sum_{l=1}^m W_{jl}^2)}}, \quad i, j = 1, \dots, n \quad (10)$$

where S_i and S_j are the sentence vector of weights of the words.

Weight of a word in a sentence is nothing but the term frequency–inverse sentence frequency (TF-ISF)

$$\text{TF} - \text{ISF} = \text{tf}_{ij} \times \text{isf}_{ij} \quad (11)$$

In Eq. (11)

$$\text{tf}_{ij} = M_{ij}/M \quad (12)$$

where $i = 1, \dots, n$; $j = 1, \dots, m$

n is the number of sentences in document and m is number of words in a sentence.

$$\text{isf}_{ij} = \log(n/n_j), \quad j = 1, \dots, m \quad (13)$$

where M_{ij} is the number of occurrence of t_j in sentence S_i and n_j is the number of sentences containing term t_j .

BA is implemented in this to find the optimal solution for Eq. (1). In various works, author has found out that BA provides better result in terms of time complexity as well as optimal solution as compared to other existing meta-heuristic techniques such as GA, PSO, HLA, etc. As it is a meta-heuristic echolocation-based algorithm, it helps in finding the optimum solution in the range. N bats are employed to find the best food in the given region. That means N solutions are initially generated to find the optimal solution for the summary of a given document. First fitness of each individual bat is checked. Bat takes random walk to generate fittest bat among them which represent the optimal solution that accordingly provides the best summary. The operative steps of the BA are as stated below.

Algorithm : Steps for Bat Algorithm

```

begin
  BAT Algorithm;
  Data: Randomly Initialized Population
  Result: Maximaizing the Solution
  Set population size=N;
  initialization of stopping criteria;
  Initialize the velocities  $v_i$  for each bat with random values;
  Initialize the frequencies of pulse  $f_i$  for each bat;
  Loudness A and pulse rate r are given values;
  while Termination criteria is not fulfilled do
    Perform calculation of fitness function for each bat;
    Using given equation update the valocity and position of each bat;
    Select a random solution with some probability;
    Random local walk is used to find a new local solution near to this solution;
    if The frequency and amplitude of a new solution is greater then
      | new solution is accepted and A is reduced and r is increased
    end
    if The new solutions are compared with the existing solutions and if better then
      | old solutions are replaced with the new solutions
    end
  end
end

```

Solving this problem, in reality, using BA Eq. (1) is alone not responsible, to satisfy the fitness function, both Eqs. (1) and (2) are taken into consideration. The penalty method [20] is applied to unconstrained problem for and converting it to constrained-based optimization problem [21].

$$F(y) = f(y) + \gamma \cdot \ln \left(L_{\max} - \sum_{i=1}^n \text{len}(S_i) \cdot y_i \right) \tag{14}$$

where $\gamma > 0$ is the penalty parameter.

$F(y)$ is valid only if $L_{\max} - \sum_{i=1}^n \text{len}(S_i) \cdot y_i > 0$

However, for a feasible solution, the penalty term is non-zero, but it becomes an ‘anti-penalty’ if $L_{\max} - \sum_{i=1}^n \text{len}(S_i) \cdot y_i \geq 1$. Penalty parameter is dependent on current iteration.

So,

$$\gamma = \gamma^- + (\gamma^+ - \gamma^-) \frac{t}{t_{\max}} \tag{15}$$

In Eq. (15), t is the current iteration and t_{\max} is the maximum iteration value. γ^- and γ^+ are start and end points of penalty parameter and $0 < \gamma^- < \gamma^+ < 0.5$.

6 Implementation

Natural language processing can be implemented easily in python, as python is open source and loaded with various libraries. In this paper, author works on ATS which can be easily implemented over python by the help of various modules such as numpy, pandas, sklearn and many more. Numpy handles mathematical issues, pandas focuses on dataset and sklearn helps in statistical as well as reporting issues. An Indian news dataset [22] is being used for the summarization. The news dataset consists of 4516 news data and corresponding gold summaries (human-written). The author implemented the proposed method on the dataset and compared it with standard gold summary and summary generated from MS word office package. Then, the efficiency of the algorithm is calculated using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [23] package. Rouge-1 and rouge-2 are calculated. The average result of F-measure, precision and recall for all these is given in the table below.

Tables 1 and 2 provide the efficiency major analysis for *proposed method* and *MS word Summariser*. Efficiency is calculated through recall and precision in ROUGE package.

Recall in ROUGE package indicates the overlapping of system generated words as compared with reference summary.

$$\text{Recall} = \frac{\text{Number of overlapped words}}{\text{Number of words in Reference Summary}} \quad (16)$$

The higher the recall value, the higher the similarity between system generated summary and the gold summary. In above tables, one can easily distinguish the values of recall are higher in case of the proposed method as compared to Ms word summarizer. In general, extractive summaries can be able to achieve more than 40–45%

Table 1 ROUGE values for the proposed method

Dataset	Rouge-1			Rouge-2		
	<i>f</i>	<i>p</i>	<i>r</i>	<i>f</i>	<i>p</i>	<i>r</i>
1	0.432836	0.568627	0.583333	0.219512	0.321429	0.259259
2	0.419753	0.54167	0.515152	0.285714	0.22807	0.382353
3	0.442836	0.578627	0.593333	0.249512	0.351429	0.292595

Table 2 ROUGE values for Ms word summarizer

Dataset	Rouge-1			Rouge-2		
	<i>f</i>	<i>p</i>	<i>r</i>	<i>f</i>	<i>p</i>	<i>r</i>
1	0.36619	0.25490	0.56	0.1315	0.0892	0.24
2	0.4660	0.5	0.4363	0.285714	0.3157	0.2608
3	0.37619	0.26491	0.53	0.1323	0.0802	0.22

of gold summary in ROUGE-1. Therefore, it is clearly a good result after all. In ROUGE-1, recall indicates the availability of individual words, but not their occurrence and position. In case of ROUGE-2, it focuses on each pair of words instead of individuals. Relevance of a summary can be judged by the precision.

$$\text{Precision} = \frac{\text{Number of overlapped words}}{\text{Total number of words in System Summary}} \quad (17)$$

Precision provides the idea of how much the system summary is worth. The more the precision value it shows, the more the relevance of it. As per the above tables, it can clearly state that this method provides a better result.

7 Conclusion and Future Work

In this paper, the authors applied a new optimization technique—BA, to solve the extractive summarization. Basic objective was to generate an extractive summary which has high coverage of topics and less redundant in nature. The applied method provides the desired output in minimal amount of time. Many compression rates have been tested for the applied algorithm, and it provides better result for 35% compression rate. Hence, it can be said that BA provides promising results for extractive text summarization. In future work, we can consider other feature extraction mechanism along with TF-ISF to gain better results and also can consider other standard datasets for testing purpose.

References

1. Gambhir, M., Gupta, V.: Recent automatic text summarization techniques: a survey. *Artif. Intell. Rev.* **47**(1), 1–66 (2017)
2. Luhn, H.P.: The automatic creation of literature abstracts. *IBM J. Res. Dev.* **2**(2), 159–165 (1958)
3. Jafari, M., et al.: Automatic text summarization using fuzzy inference. In: 2016 22nd International Conference on Automation and Computing (ICAC). IEEE (2016)
4. Alguliyev, R., Aliguliyev, R., Isazade, N.: A sentence selection model and HLO algorithm for extractive text summarization. In: 2016 IEEE 10th International Conference on Application of Information and Communication Technologies (AICT). IEEE (2016)
5. García-Hernández, R.A., Ledeneva, Y.: Single extractive text summarization based on a genetic algorithm. In: Mexican Conference on Pattern Recognition. Springer, Berlin, Heidelberg (2013)
6. Binwahlan, M.S., Salim, N., Suanmali, L.: Swarm based text summarization. In: International Association of Computer Science and Information Technology-Spring Conference, 2009. IAC-SITSC'09. IEEE (2009)
7. Yang, X.-S.: A new metaheuristic bat-inspired algorithm. In: Nature Inspired Cooperative Strategies for Optimization (NICSO 2010), pp. 65–74. Springer, Berlin, Heidelberg (2010)
8. Sagnika, S., Bilgaiyan, S., Mishra, B.S.P.: Workflow scheduling in cloud computing environment using Bat Algorithm. In: Proceedings of First International Conference on Smart System, Innovations and Computing. Springer, Singapore (2018)

9. Baxendale, P.B.: Machine-made index for technical literature—an experiment. *IBM J. Res. Dev.* **2**(4), 354–361 (1958)
10. Edmundson, H.P.: New methods in automatic extracting. *J. ACM (JACM)* **16**(2), 264–285 (1969)
11. Nobata, C., et al.: Sentence Extraction System Assembling Multiple Evidence. *NTCIR* (2001)
12. Matsuo, Y., Ishizuka, M.: Keyword extraction from a single document using word co-occurrence statistical information. *Int. J. Artif. Intell. Tools* **13**(01), 157–169 (2004)
13. Malik, R., L. V. Subramaniam, S. Kaushik. Automatically selecting answer templates to respond to customer emails. *IJCAI* **7**(1659) (2007)
14. Sharifzadeh, M., Bashash, K., Bashokian, S.: A Comparison with Two Semantic Sensor Data Storages in Total Data Transmission. *arXiv preprint arXiv:1401.7499* (2014)
15. Yang, X.-S.: Bat Algorithm: literature review and applications. *Int. J. Bio-Inspired Comput.* **5**(3), 141–149 (2013)
16. Yılmaz, S., Küçükşille, E.U.: A new modification approach on bat algorithm for solving optimization problems. *Appl. Soft Comput.* **28**, 259–275 (2015)
17. Gan, W.-Y., et al. Community discovery method in networks based on topological potential. *J. Softw.* **20**(8), 2241–2254 (2009)
18. Niwattanakul, S., et al.: Using of Jaccard coefficient for keywords similarity. In: *Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Vol. 1. No. 6 (2013)
19. Nguyen, H.V., Bai, L.: *Cosine Similarity Metric Learning for Face Verification*. Asian Conference on Computer Vision. Springer, Berlin (2010)
20. Alguliyev, R.M., Aliguliyev, R.M., Isazade, N.R.: An unsupervised approach to generating generic summaries of documents. *Appl. Soft Comput.* **34**, 236–250 (2015)
21. Yeniay, Ö.: Penalty function methods for constrained optimization with genetic algorithms. *Math. Comput. Appl.* **10**(1), 45–56 (2005)
22. News Dataset Available.: https://github.com/sunnysai12345/News_Summary/blob/master/news_summary.csv
23. Lin, C.-Y.: Rouge: a package for automatic evaluation of summaries. *Text summarization branches out* (2004)

Phylogenetics Algorithms and Applications



Geetika Munjal, Madasu Hanmandlu and Sangeet Srivastava

Abstract Phylogenetics is a powerful approach in finding evolution of current day species. By studying phylogenetic trees, scientists gain a better understanding of how species have evolved while explaining the similarities and differences among species. The phylogenetic study can help in analysing the evolution and the similarities among diseases and viruses, and further help in prescribing their vaccines against them. This paper explores computational solutions for building phylogeny of species along with highlighting benefits of alignment-free methods of phylogenetics. The paper has also discussed the application of phylogenetic study in disease diagnosis and evolution.

Keywords Phylogenetics · Cancer evolution · Sequence analysis

1 Introduction

Phylogenetics can be considered as one of the best tools for understanding the spread of contagious disease, for example, transmission of the human immunodeficiency virus (HIV) and the origin and subsequent evolution of the severe acute respiratory syndrome (SARS) associated coronavirus (SCoV) [1]. Earlier, morphological traits were used for assessing similarities between species and building phylogenetic trees. Presently, phylogenetics relies on information extracted from genetic material such as deoxyribonucleic acid (DNA), ribonucleic acid (RNA) or protein sequences [2]. Methods used for phylogenetic inference have changed drastically during the past two decades: from alignment-based to alignment-free methods [3]. This paper has

G. Munjal (✉) · S. Srivastava
The NorthCap University, Gurugram, India
e-mail: geetika@ncuindia.edu; munjalgheetika@gmail.com

S. Srivastava
e-mail: sangeetsrivastava@ncuindia.edu

M. Hanmandlu
IIT Delhi, New Delhi, India
e-mail: mhmandlu@gmail.com

reviewed various methods under phylogenetic tree construction from character to distance methods and alignment-based to alignment-free methods. A brief review of phylogenetic tree applications is also given in cancer studies.

2 Literature Review

A phylogenetic tree can be unrooted or rooted, implying directions corresponding to evolutionary time, i.e. the species at the leaves of a tree relate to the current day species. The species can be expressed as DNA strings which are formed by combining four nucleotides A, T, C and G (A—adenine, T—thymine, C—cytosine and G—guanine). In literature, various string processing algorithms are reported which can quickly analyse these DNA and RNA sequences and build a phylogeny of sequences or species based on their similarity and dissimilarity. A high similarity among two sequences usually implies significant functional or structural likeliness, and these sequences are closely related in the phylogenetic tree. To get more precise information about the extent of similarity to some other sequence stored in a database, we must be able to compare sequences quickly with a set of sequences. For this, we need to perform the multiple sequence comparison. Dynamic programming concepts facilitate this comparison using alignment methods, but it involves more computation. Moreover, the iterative computational steps limit its utility for long length sequences [3]. Alignment-free methods overcome this limitation as they follow alternative metrics like word frequency or sequence entropy for finding similarity between sequences.

3 Methods of Phylogenetic Tree Construction

Phylogenetic tree generation consists of sequence alignment where the resulting tree reveals how alignment can influence the tree formation. Alignment-based methodologies are probably the most widely used tools in sequence analysis problems [4]. They consist of arranging two sequences: one on the top of another to highlight their common symbols and substrings. An alignment method is based on alignment parameters including insertion, deletions and gaps which play a pivotal role in the construction of the phylogenetic tree. A phylogenetic tree is formed as an outcome of sequence analysis performed on the DNA or RNA strings [5]. Sequence comparison reveals the patterns of shared history between species, helping in the prediction of ancestral states. The comparison of sequences also helps in understanding the biology of living organisms which is required to find similarity and relationship among species. For sequence comparison, we can follow alignment-based or alignment-free methods [3, 6, 7].

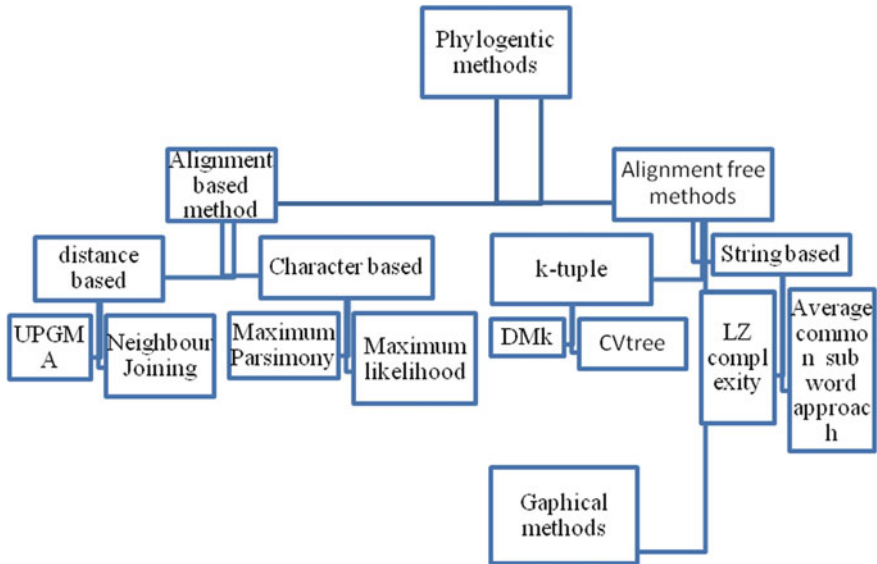


Fig. 1 Hierarchical view of phylogenetic methods²

3.1 Sequence Alignment

Sequence alignment is a method to identify homologous sequences. It is categorized as pairwise alignment in which only two sequences are compared at a time whereas in multiple sequence alignment more than two sequences are compared. Alignment-based can be global or local [8, 9]. These alignment-based algorithms can also be used with distance methods to express the similarity between two sequences, reflecting the number of changes in each sequence. Figure 1 gives a hierarchical view of various methods for phylogenetic tree building.

3.2 Character-Based Methods

The character-based methods compare all sequences simultaneously considering one character/site at a time. These are maximum parsimony and maximum likelihood. These methods use probability and consider variation in a set of sequences [10]. Both approaches consider the tree with the best score tree, which requires the smallest number of changes to perform alignment. Maximum parsimony method suffers badly from the long-branch attraction and gives the least information about the branch lengths [10]. In such cases, if two external branches are separated by short internal branches, it leads to the incorrect tree. Some of the salient features of character-based methods are mentioned in Table 1.

Table 1 Comparison of different phylogenetic tree construction methods

Method	Advantage	Disadvantage	Other information
Maximum parsimony	Appropriate for very similar sequences and a small number of sequences	Very time-consuming as it tests all possible trees Parsimony may fail for diverged sequences Suffers from the long-branch attraction	Predict the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences It is built with the fewest changes required to explain (tree) the differences observed in the data
Maximum likelihood	Suitable for very dissimilar sequences We can formulate hypothesis about evolutionary relationships More accurate phylogenetic trees can be constructed for a small number of taxa in a reasonable time frame	A slow search algorithm will lead to slow response Takes a long time for large datasets	It tries to find a model that has the highest probability to generate the input sequence under a given evolutionary model
Neighbour joining	Faster than the character-based method They are fast and can be used with a variety of models	Conversion from sequence data to distance data leads to loss of information	Provides an unrooted tree and a single resultant tree
UPGMA	Reliable for related sequences	Evolution rate is constant in all branches	UPGMA provides rooted tree
Fitch Mangrolish	Less sensitive to variations in evolutionary rate	Dependent on the model used to obtain the distance matrix	

3.3 *Distance-Based Methods*

Distance-based methods use the dissimilarity (the distance) between the two sequences to construct trees. They are much less computationally intensive than the character based methods are mostly accurate as they take mutations into count. For tree generation, generally, hierarchical clustering is used in which dendrograms (clusters) are created. Table 1 briefly compares various phylogenetic tree construction methods.

4 Alignment-Based Versus Alignment-Free Sequence Comparison

Multiple alignments of related sequences may often yield the most helpful information on its phylogeny. However, it can produce incorrect results when applied to more divergent sequence rearrangements [3]. Some computationally intensive multiple alignment methods align sequences strictly based on the order in which they receive them. Multiple sequence alignment methods emphasize that more closely related sequences should be aligned first. In cases of sequences being less related to one another, however, sharing a common ancestor may be clustered separately [11]. This implies that they can be more accurately aligned, but may result in incorrect phylogeny. Alignment can provide an optimized tree if a recursive approach is followed; however, this will increase the complexity of the problem. If the differences among the lengths of sequences are very high, the alignment performance significantly impacts tree generation.

The use of dynamic programming in alignment makes computation more complicated, and iterative steps limit their utility for large datasets. Therefore, consistent efforts have been made in developing and improving multiple sequence alignment methods for supporting variable length sequences with high accuracy and also for aligning a larger number of sequences simultaneously. Because of the problems associated with alignment-based phylogeny the importance of alignment-free methods is apparent [3]. Hence, the alignment quality affects the relationship created in a phylogenetic tree based on the consideration discussed above.

4.1 Alignment-Free Methods for Sequence Analysis

Alignment-free methods proposed in recent years can be classified into various categories as shown in Fig. 1. These include k-tuple based on the word frequencies, methods that represent the sequence without using the word frequencies, i.e. compression algorithms probabilistic methods and information theory-based method. In the k-tuple method, a genetic sequence is represented by a frequency vector of fixed length subsequence and the similarity or dissimilarity measures are found based on the frequency vector of subsequence. The probabilistic methods represent the sequences using the transition matrix of a Markov chain [12] of a pre-specified order, and comparison of two sequences is done by finding the distance between two transition matrices. Graphical representation comprising 2D or 3D or even 20D methods provides an easy way to view, sort and compare various sequences. Graphical representation further helps in recognizing major characteristics among similar biological sequences.

As discussed k-tuple method uses k-words to characterize the compositional features of a sequence numerically. A biological sequence is numerically converted into a vector or a matrix composed of the word frequency. The k-word frequency pro-

vides a fast arithmetic speed and can be applied to full sequences. The problem with k -tuple is a big value of k that poses a challenge in the computing time and space, and k -word methods underestimate or even ignore the importance of its location. The string-based distance measure uses substring matches with k mismatches.

5 Application of Phylogenetics in Cancer Studies

Cancer research is considered one of the most significant areas in the medical community. Mutations in genomic sequences are responsible for cancer development and increased aggressiveness in patients [13, 14]. The combination of all such genes mutations, or progression pathways, across a population can be summarized in a phylogeny describing the different evolutionary pathways [9]. Application of the phylogenetic tree can be explored for finding similarities among breast cancer subtypes based on gene data [14, 15]. Discovery of genes associated in cancer subtype help researchers to map different pathways to classify cancer subtypes according to their mutations. Methods of phylogenetic tree inference have proliferated in cancer genome studies such as breast cancer [13]. Phylogenetic can capture important mutational events among different cancer types; a network approach can also capture tumour similarities.

It has been observed from the literature that in cancer disease, the driver genes change the cancer progression, and it even affects the participation of other genes thus generating gene interaction network. Phylogenetic methods can solve the problem of class prediction by using a classification tree. Phylogenetic methods give us a deeper understanding of biological heterogeneity among cancer subtype.

6 Conclusion

The research focuses on the various methods of sequence analysis to generate phylogenetic trees. The limitations associated with sequence alignment methods lead to the development of alignment-free sequence analysis. However, most of the existing alignment-free methods are unable to build an accurate tree so more refinement is required in alignment-free methods. The phylogenetic study is not limited to species evolution, but disease evolution as well. Extending phylogenetic to disease diagnosis can give birth to new treatment options and understanding its progression.

Acknowledgements The research is funded by Department of Science and Technology, Delhi, under the sanction number SR/WOS-A/ET-1015/2015.

References

1. Lam, T.-Y., Hon, C.-C., Tang, J.W.: Use of phylogenetics in the molecular epidemiology and evolutionary studies of viral infections. *Crit. Rev. Clin. Lab. Sci.* **47**(1), 5–49 (2010)
2. Moret, B.M.E., Warnow, T.: Reconstructing optimal phylogenetic trees : a challenge in experimental algorithmics. In: *Experimental Algorithmics, LNCS*, pp. 163–180 (2002)
3. Vinga, S.: Editorial: alignment-free methods in computational biology. *Brief. Bioinform.* **15**(3), 341–342 (2014)
4. Geetika, Hanmandlu, M., Gaur, D.: Analyzing DNA strings using information theory concepts. In: *ICTCS-16, ACM Conference, Udaipur*, no. 9 (2016)
5. Munjal, G., Hanmandlu, M., Saini, A., Gaur, D.: Modified k-Tuple method for the construction of phylogenetic trees. *Trends Bioinform.* **8**(3), 75–85 (2015)
6. Schwartz, R., Schäffer, A.A.: The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* **18**(4), 213–229 (2017)
7. Chan, R.H., Chan, T.H., Yeung, H.M., Wang, R.W.: On maximum entropy principle for sequence comparison. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **9**(1), 79–87 (2012)
8. Needleman, W.: A general method applicable to the search for similarity in the amino acid sequence of two proteins. *J. Mol. Biol.* **1970**(48), 443–453 (1969)
9. Smith, T.F., Waterman, M.S.: Comparison of biosequences. *Adv. Appl. Math.* **2**(4), 482–489 (1981)
10. Alon, N., Chor, B., Pardi, F., Rapoport, A.: Approximate maximum parsimony and ancestral maximum likelihood. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **7**(1), 1–7 (2010)
11. Burr, T.: Phylogenetic trees in bioinformatics. *Curr. Bioinform.* **5**(1), 40–52 (2010)
12. Bai, F., Xu, J., Liu, L.: Weighted relative entropy for phylogenetic tree based on 2-step Markov Model. *Math. Biosci.* **246**(1), 8–13 (2013)
13. Somarelli, J., Ware, K., Kostadinov, R., Robinson, J., Amri, H., Abu-Asab, M., Fourie, N., Diogo, R., Swofford, D., Townsend, J.: PhyloOncology: understanding cancer through phylogenetic analysis. *Biochimica et Biophysica Acta (BBA)—Rev. Cancer* **1867**(2), 101–108 (2017)
14. Desper, R., Khan, J., Schäffer, A.: Tumor classification using phylogenetic methods on expression data. *J. Theor. Biol.* **228**(4), 477–496 (2004)
15. Munjal, G., Hanmandlu, M., Srivastava, S.: Novel gene selection method for breast cancer classification. *J. Biochem. Technol.* **8**(4), 1116–1120
16. Borozan, I., Watt, S., Ferretti, V.: Sequence analysis Integrating alignment-based and alignment-free sequence similarity measures for biological sequence classification. *Bioinformatics* **31**(January), 1396–1404 (2015)
17. Nemeth, C.: Hidden Markov models with applications to DNA sequence analysis. *STOR-i*, Lancaster University
18. Hohl, M., Ragan, M.A.: Is multiple-sequence alignment required for accurate inference of phylogeny? *Syst. Biol.* **56**(2), 206–221 (2007)
19. Potter, R.M.: Constructing phylogenetic trees using multiple sequence alignment. University of Washington (2008)
20. Burr, T.: Phylogenetic trees in bioinformatics. *Curr. Bioinform.* **5**(1), 40–52 (2010)
21. Cho, A.: Constructing phylogenetic trees using maximum likelihood. Ph.D. Thesis, Scripps women’s college Claremont (2012)
22. Felsenstein, J.: *PHYLIP*. University of Washington Seattle, WA (1993)
23. Sardaraz, M., Tahir, M., Aziz Ikram, T., Bajwa, H.: Applications and algorithms for inference of huge phylogenetic trees: a review. *Am. J. Bioinform. Res.* **2**(1), 21–26 (2012)
24. Dawyndt, P., De Meyer, H., De Baets, B.: UPGMA clustering revisited: a weight-driven approach to transitive approximation. *Int. J. Approx. Reason.* **42**(3), 174–191 (2006)
25. Potiny, S.: An improved phylogenetic tree comparison method. Thesis University of North Carolina (2010)
26. Bryant, D., Moulton, V.: Neighbor-net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.* **21**(2), 255–265 (2004)

27. Brinkman, F.S.L.: *Bioinformatics: a practical guide to the analysis of genes and proteins*. Publisher John Wiley and Sons (2001)
28. Munjal, G., Sharma, P., Gaur, D.: Sequence similarity using composition method. *Int. J. Data Sci.* **3**(1), 19–28. <https://doi.org/10.1504/IJDS.2018.090626>
29. Leimeister, C., Morgenstern, B.: Sequence analysis kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics* **30**(14), 2000–2008 (2014)

Comparative Study of Forecasting Model for Price Prediction of Rice



Kesari Verma, N. K. Nagwani and Shrish Verma

Abstract In this paper, we reviewed different forecasting models and proposed a neural network-based regression model for forecasting model price of rice. This study aims to prepare a neural network model that uses some of the existing data as feature and treat some data for defining target value. We also performed experimental evaluation of various forecasting models in real data of rice using time series data (price) of rice of India from 2001 to 2012. The algorithms are evaluated based on evaluation measures such as mean error (ME), root mean square error (MSE), mean absolute error (MAE), mean percentage error (MPE), mean absolute percentage error (MAPE) and mean absolute square error (MASE).

Keywords Data mining · Regression model · Neural network model · Forecasting price of rice

1 Introduction

Prediction in time series data is a challenging task for machine learning community. The price of rice can be represented as one-dimensional series of data called time series [1]. A time series is a sequence of observations $t_i \in R$, usually ordered in time.

K. Verma (✉)

Department of Computer Applications, National Institute of Technology, Raipur,
Chhattisgarh, India

e-mail: kverma.mca@nitrr.ac.in

N. K. Nagwani

Department of Computer Science & Engineering, National Institute of Technology, Raipur,
Chhattisgarh, India

e-mail: nknagwani.cs@nitrr.ac.in

S. Verma

Department of Electronics, National Institute of Technology, Raipur, Chhattisgarh, India
e-mail: shrishverma@nitrr.ac.in

© Springer Nature Singapore Pte Ltd. 2019

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,

Advances in Intelligent Systems and Computing 904,

https://doi.org/10.1007/978-981-13-5934-7_18

The task is to predict future values of a time series $y(t)$ from past historical values of a time series $x(t)$ [1].

$$y(t) = f(y(t-1), \dots, y(t-d), x(t-1), \dots, (t-d))$$

An observed discrete univariate time series be $t_1, t_2, t_3, t_4, \dots, t_n$ is distribution of price in different years. This represents the event that has occurred in the sequence ordered by year t . A general model for the time series can be defined by using equation as follows:

$$F(t, \phi) = g(t) + \phi_t \quad t = 1, \dots, n$$

The series has two parts:

- a. First part is $g(t)$, called a trend, which is a deterministic function of time.
- b. Second part is stochastic sequence: a residual term ϕ_t , which also called noise, which follows a probability law.

The learning procedure aims at main goals to choose a parametric family of hypothesis $g(t)$ which contains or gives a good approximation of the unknown function f . The best approximation is based on good parametric identification.

Definition 1 [Prediction] A supervised classification problem is defined in the form: $D = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$. Here, \mathbf{x} values are vector in the form: $\mathbf{x} = \{x_1, \dots, x_n\}$, x can be discrete or real valued. These components are called attributes of the database that helps to predict the class label y for the dataset. Prediction is based on interdependence of the variable in the past, and it is possible if it preserved for future also.

This study uses supervised learning that aims to devise a method to construct a model for assigning instances to one of a finite set of real numbers on the basis of a vector of variables measured on the instances. The accuracy of the model is dependent on appropriate extraction of instance from time series. In time series prediction to predict $F(t)$, the feature vector can be modelled $1 \dots t - 1$. In order to reduce the computational time, the proposed algorithm selected optimal set of data point as instance in time series input neurons.

The proposed approach has the following advantages, which we will be empirically demonstrated with some experiments:

- After learning the data presented to them (a sample), ANNs can often correctly infer the unseen part of a population and have the capability of prediction for non-linear data also.
- ANN works on the data-driven approach, it automatically creates an approximate model based on data, and it does not require a prior theory and information for prediction.

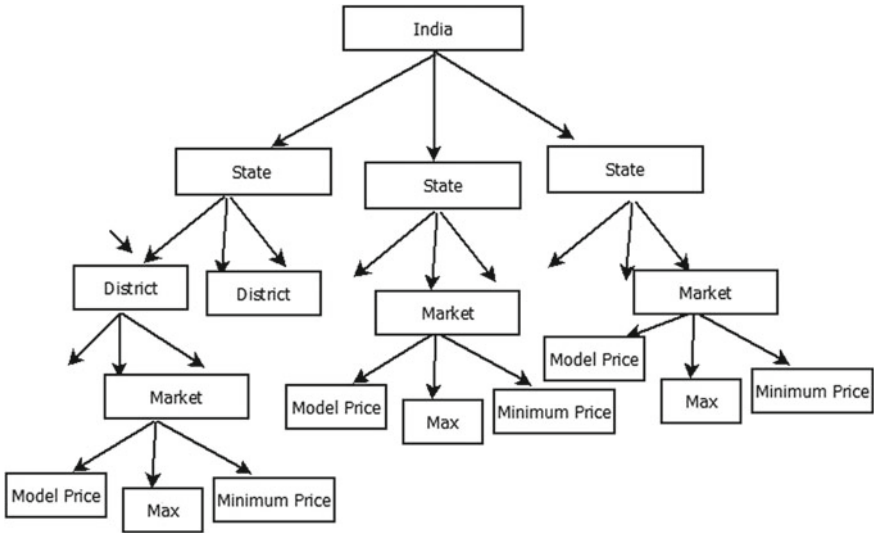


Fig. 1 Concept hierarchy of distribution of data

1.1 Dataset

In this study, dataset has been taken from Open Government Data (OGD) Platform India for rice and wheat price distribution information. The data is authentic data of the last 12 years for rice and wheat prices of different markets at district and state of India [2]. The data provided from OGD are multidimensional data which contain date, state (28), district and market in each district. The data for 1 year is more than 90,000 approximate after the year 2003. The data is difficult to represent with single time series. The concept hierarchy for data is shown in Fig. 1. The plot of model price of rice, maximum price and minimum price in different years is shown in Fig. 2. There are more than 100 varieties of rice are available in India, but in this study, we have taken the price of fine rice [3].

The rest of the paper is organized as follows. In Sect. 2, we discussed the related work in this area. In Sect. 3, the formulation of the proposed model is discussed. Section 4 gives the experimental analysis of existing and proposed model. Section 5 contains the concluding remarks.

2 Related Work

In this section, we discuss various forecasting model for time series. Initially, the statistical model has been used for forecasting of time series. Averaging is one of the

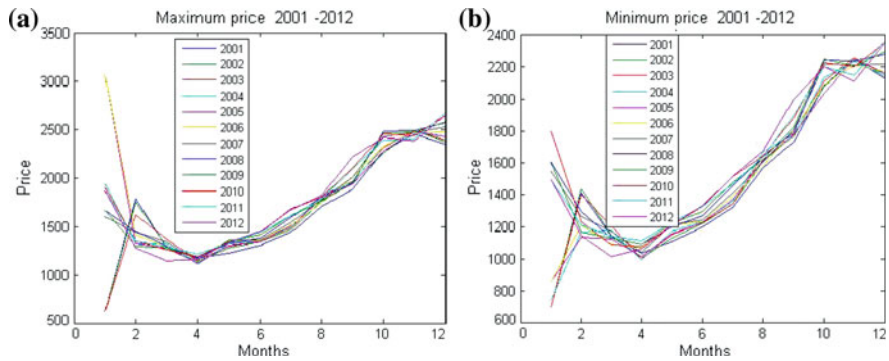


Fig. 2 **a** Maximum price of rice 2001–2012. **b** Minimum price of rice 2001–2012 price of rice in year 2004

oldest methods of forecasting. If historical data be denoted by t_1, \dots, t_n , then forecasts value can be obtained using Eq. 2 [4].

$$\hat{y} = \frac{y_1 + y_2 + \dots + y_n}{n}$$

Naïve-based method—This approach is another method of forecasting which successfully work for financial and economic time series. All forecasts are obtained from historical data value. It is based on probabilistic data from historical observations.

Sessional naïve approach—Sessional naïve approach is an enhancement over naïve approach that s useful for highly seasonal data like vegetation index in specific months of the year. In this case, each forecast is set to equal of the last observed value on same seasons. Seasonality can be obtained by additive and multiplicative decomposition of the series. The traditional approaches to time series prediction, such as the Box–Jenkins or ARIMA, assume time series are generated using linear processes [5].

ARIMA is one of the oldest techniques used for time series prediction. Yule [6] first formulated the concept of autoregressive (AR) and moving average (MA) models. Box and Jenkins [7] discussed the enormous impact on the theory and practice of modern time series analysis and forecasting, but not extended for non-linear exponential smoothing methods. Exponential smoothing model [8, 9] which work for linear data only further, Winters [10] introduced a new model that works on single, double and triple exponential concepts. The equation for single exponential smoothing is defined as follows. Single smoothing is useful for short-range forecasting:

$$F_{t+1} = F_t + \alpha(Y_t - F_t)$$

where Y_t is observed value, F_t denotes the forecast, the forecast error is computed by $(Y_t - F_t)$ and α is a constant between 0 and 1. It is extended as an adaptive approach

which is defined in Eq. (2). The extension in adaptive approach it uses αt which is varying whereas in above equation α is constant throughout the equation:

$$F_{t+1} = \alpha_t Y_t + (1 - \alpha_t) F_t$$

2.1 Double Smoothing

Exponential smoothing with a trend works similarly like simple smoothing except that two components are updated each period—level and trend. The Holts model [11] is derived using two smoothing constants, α and β (with value between 0 and 1) L_t denotes estimation of the level of the series at time t and b_t denotes estimation of slope at time t . The following equation adjusts L_t the trend for previous period b_{t-1} by adding it to the last smoothed value L_{t-1} :

$$\begin{aligned} L_t &= \alpha Y_t + (1 - \alpha)(L_{t-1} + b_{t-1}) \\ b_t &= \beta(L_t - L_{t-1}) + (1 - \beta)b_{t-1} \\ F_{t+m} &= L_t + b_t m \end{aligned}$$

Pegels [12] provided a useful classification of the trend additive (linear) or multiplicative (non-linear) which is effectively used for information extraction. Gardner [13] provided an exhaustive review and synthesis of work in exponential smoothing and discussed the standard smoothing methods. Kalman [14] proposed a recursive algorithm named as Kalman filter for computing forecasts. Shumway and Stoffer [15] ensemble the expected maximization algorithm with the Kalman filter to forecasting time series using state-space models.

Adya and Collopy [1] evaluated the ANN-based forecasting in various studies. De Gooijer et al. [16] proposed an extensive survey of different time series forecasting model from the period 1982–2005 and summarizing 940 papers, including about 340 papers published under the International Institute of Forecasters (IIF) journal.

An artificial neural network (ANN) is used as one of the effective techniques for non-linear processes that have an unknown functional relationship [17]. ANN model is able to fit the data with high degree polynomial in order to generate output using hidden layer and neurons. Qi [18] discussed that ANNs are better than other methods for regression problem. De Gooijer et al. [16] introduced the model for instrument investment analysis using the neural network and genetic algorithm.

Support vector machine is one of the most outstanding models, which is based on the statistical learning theory and applies the structural risk minimization theory, and has a global optimal solution. The classification of various time series forecasting model is shown in Fig. 3. Apart from this averaging method, naïve method, RWF method, sessional naïve, drift model, ARIMA model, Croston model, state-space smooth model, Holt winters model, linear regression model, STL model, structural time, neural network and Tthat models.

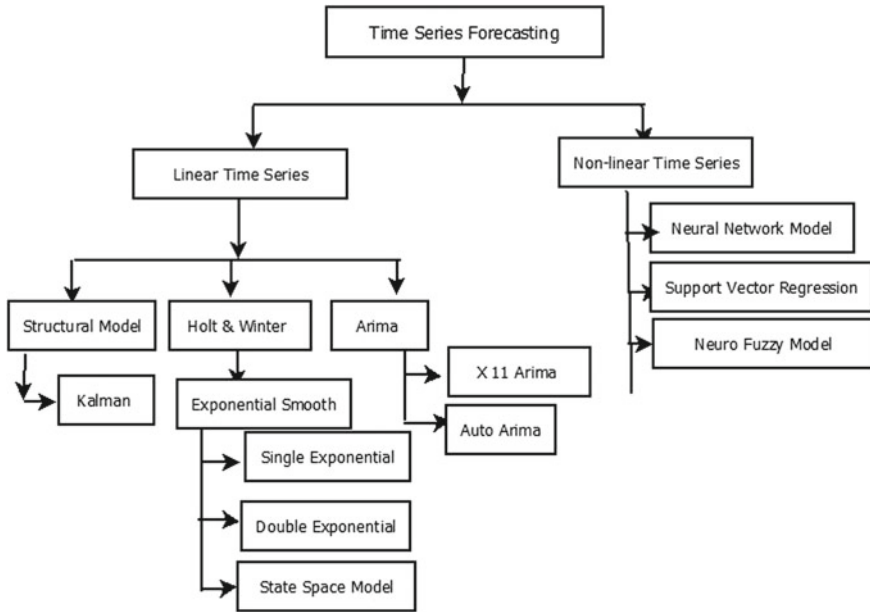


Fig. 3 Categorization of time forecasting model

3 Proposed Neural Network Regression Model

In this section, we discuss how artificial neural multilayer perceptron model is used for time series prediction. The idea is to convert the time series prediction task into a function approximation task. The task is to predict the value at $t(n)$ at time t , based on previous data. The function approximation model is shown in the figure. It is shown in Fig. 1 that time series $t(5)$ is predicted-based $t(1)$, $t(2)$, $t(3)$ and $t(4)$. The value for $t(n)$ is predicted based on $t_n \leftarrow f(t_{n-t-1}, \dots, t_{n-4}, t_{n-3}, t_{n-2}, t_{n-1})$. The task is to predict the price $t(s)$ at year t , given as much as 12 previous numbers. Suppose we used 4 time lags $t(s - 1)$; $t(s - 2)$; $t(s - 4)$; $t(s - 12)$ in order to predict(s). The proposed regression model for prediction is shown in Fig. 4. The neural network model based on this time lag is shown in Fig. 5.

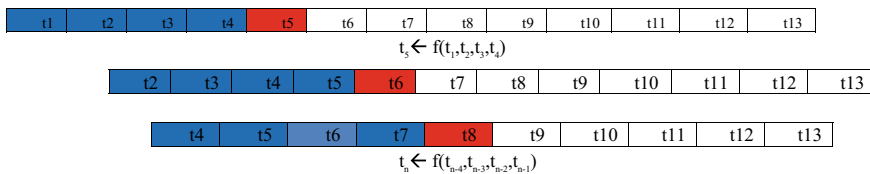


Fig. 4 Regression model for prediction

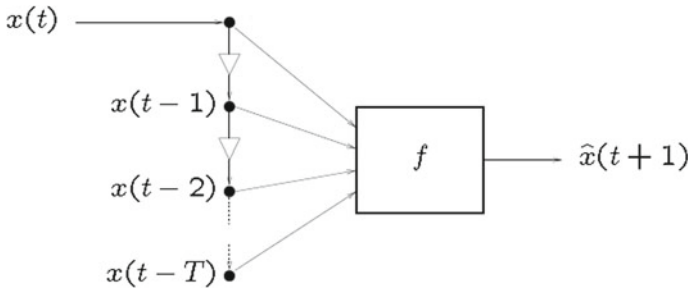


Fig. 5 Neural network model for prediction

Proposed Algorithm

The neural network-based forecasting method is shown in Algorithm 1.

Algorithm 1 Neural Network based Forecasting

- Input : Time Series Dataset D , $\text{Min} \leftarrow$ Minimum value for target, $\text{Max} \leftarrow$ Max value for target
1. $j=1$;
 2. **for** $i = \text{Min} : \text{Max}$
 3. $Dp(j,1) = D(i-12)$;
 4. $Dp(j,2) = D(i-6)$;
 5. $Dp(j,3) = D(i-2)$;
 6. $Dp(j,4) = D(i-1)$;
 7. $Dp(j,5) = D(i)$;
 8. $j = j+1$;
 9. **End**
 10. Divided the data for Training and Testing set $Dp_{\text{tr}} \leftarrow$ Training set and $Dp_{\text{ts}} \leftarrow$ Test Set
 11. initialize network weights $w \leftarrow$ random number
 12. **do**
 13. **forEach** training example D_{tr}
 14. prediction = neural-net-output(network, Dp_{tr}) // forward pass
 15. actual = target-output(D_{tr})
 16. compute error (prediction - actual) at the output units
 17. compute for all weights from hidden layer to output layer // backward pass
 18. compute for all weights from input layer to hidden layer // backward pass continued
 19. update network weights // input layer not modified by error estimate
 20. **until** all examples classified correctly or another stopping criterion satisfied or minimum error found
 21. **return** the network
 22. Testprediction = neural-net-output(network, Dp_{ts})
 23. actual = target-output(Dp_{ts}) // Compute the output for test data
 24. compute error (prediction - actual) at the output units Compute the error for test data $e_i = Y_i - F_i$
 25. Compute ME,MAE, MAPE, MSE,MASE
-

In the above algorithm, the line numbers 1–9 are used to divide the time series as input vector and target vector. In this algorithm, some of the input vectors are treated as a target value. It is also shown in Table 2.

4 Experimental Evaluation

The experiments were performed in Windows operating system with Pentium IV machine with 1 GB RAM. No other application was running while performance computation. Data from the year 2004–2012 of model price of rice, the price of rice and minimum price were used. Data from 2004 to 2011 were used for training purpose and 2012 was used for prediction purpose. The data of variety rice were aggregated for each month of the year.

Evaluation Parameter

All evaluation parameters such as mean error, mean absolute error, mean absolute percentage errors, mean square error and percentage errors are defined in the following equation. Here, Y_t is the actual value, and F_t is the measured value.

Error	$e_t = Y_t - F_t$
Mean error (ME)	$ME = \frac{1}{n} \sum_{t=1}^n e_t$
Mean absolute error	$MAE = \frac{1}{n} \sum_{t=1}^n e_t $
Mean absolute percentage error (MAPE)	$MAPE = \frac{1}{n} \sum_{t=1}^n ME $
Mean square error (MSE)	$MSE = \frac{1}{n} \sum_{t=1}^n e_t^2$
Root mean square error (RMSE)	$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n e_t^2}$
Percentage error	$p_t = 100 * e_t/Y_t$

In the proposed neural network regression model, the predicted result is shown in Fig. 6. The mean square error is 0.0313 for training data, but for testing data using single step prediction is 0.8984 and using iterative prediction method 0.9788.

For forecasting purpose, we have taken Jan–Dec, 2012, and performed using averaging method, naïve method, RWF method, sessional naïve, drift model, ARIMA model, Croston model, state space smooth model, Holt winters model, linear regression model, STL model, structural time, neural network and Tbat models. The results are shown in Fig. 7 and there estimated parameter values are shown in Tables 1 and 2. The black line shows the real data value and a blue line indicates the forecasted value. Most of the forecasting methods are applied from forecasting package in R [5] and Matlab 2012.

The error in the training data and test data are evaluated based on ME, RMSE, MAE, MPE, MAPE and MASE. The plot for error in model price prediction, maximum price prediction and minimum price is shown in Fig. 8. In experimentation, it is observed that averaging method performed potentially well with training error ME

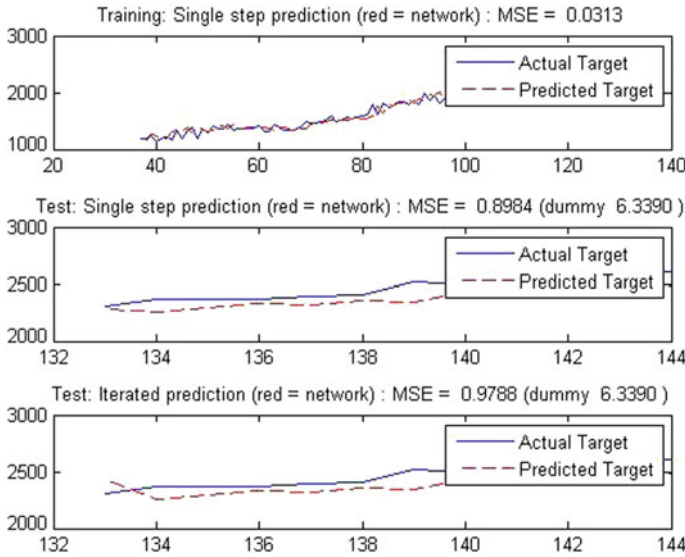


Fig. 6 Forecasting using ANN model

Table 1 Conversion of time series into matrix of input neurons and target

Input vector				Target
D(1)	D(7)	D(11)	D(12)	D(13)
D(2)	D(8)	D(12)	D(13)	D(14)
D(3)	D(9)	D(13)	D(14)	D(15)
D(4)	D(10)	D(14)	D(15)	D(16)

= 0, but it is maximum with testing data, thus averaging method is not appropriate for non-linear dataset.

The performance of different forecasting models for model, price, maximum price and minimum price is shown in Fig. 6. The error is minimal for structural time series model for training data of model price and maximum price. While for the testing data the performance of linear regression model, croston model and ARIMA model are quite interesting. The performance of the proposed model is not as good as structural time series model, but it is better than other models like naïve, seasonal naïve and Croston model. In experimentation, many times the value for ME, MPE is negative, which may give minimum value. These negative values were not considered due to this indicate variance from 0; thus, only minimum of positive value for ME and MPE was considered in error. The ME, RMSE, MAE, MPE, MAPE and MASE are shown in Tables 1, 2 and 3. The proposed time series model also performed well and shown good results. In the table, the value for MPE is negative also, but we considered only the positive value for analysis purpose.

Table 2 Error using different models for prediction of model price in training data

	ME	RMSE	MAE	MPE	MAPE	MASE
Averaging method	0.009	444.61	397.42	-6.354	23.41	6.006
Naïve method	11.814	86.468	66.173	0.556	3.901	1.000
RWF	11.814	86.468	66.173	0.556	3.901	1.000
Sessional naïve	169.47	216.78	180.26	8.946	9.550	2.725
Drift model	-0.007	85.657	65.305	-0.141	3.862	0.987
ARIMA model	-0.011	70.948	55.388	-0.136	3.179	0.307
Croston	123.82	158.51	130.60	6.493	6.957	1.974
State-space smooth	-0.019	72.083	56.249	-0.162	3.252	0.850
Holt winters	-1.031	73.675	57.005	0.127	3.349	0.854
Linear regression	0.006	103.46	86.371	-0.224	5.153	1.305
STL	13.957	66.783	52.742	0.679	3.157	0.790
Structural time	0.275	0.995	0.762	0.014	0.044	0.012
Neural network	-0.050	79.102	66.064	-0.646	3.660	0.998
Tbat	-0.267	74.194	57.194	0.002	3.325	0.864
Proposed method	7.499	0.017	68.33	-0.004	0.005	0.023

Table 3 Error in testing data for model price prediction

	ME	RMSE	MAE	MPE	MAPE	MASE
Averaging method	679.29	690.32	679.29	27.280	27.28	10.265
Naïve method	1.781	195.33	154.76	5.907	6.037	2.339
RWF	151.78	195.33	154.76	5.907	6.037	2.339
Sessional naïve	53.49	164.55	124.41	1.875	4.891	1.880
Drift model	74.988	114.78	80.602	2.871	3.114	1.218
ARIMA model	57.762	104.56	73.789	2.172	2.859	1.115
Croston	68.985	133.45	90.072	2.618	3.526	0.000
State-space smooth	23.675	88.499	70.934	0.795	2.800	1.072
Holt winters	136.55	192.46	145.70	5.265	5.662	2.202
Linear regression	-166.39	189.71	166.39	-6.917	6.917	2.514
STL	204.18	269.77	213.82	7.932	8.351	3.231
Structural time	133.16	181.25	139.25	5.153	5.417	2.104
Neural network	107.47	183.06	130.10	4.066	5.037	1.966
Tbat	168.80	220.16	177.66	6.561	6.946	2.685
Proposed	29.4	105.2	85.8	0.103	0.10	0.123

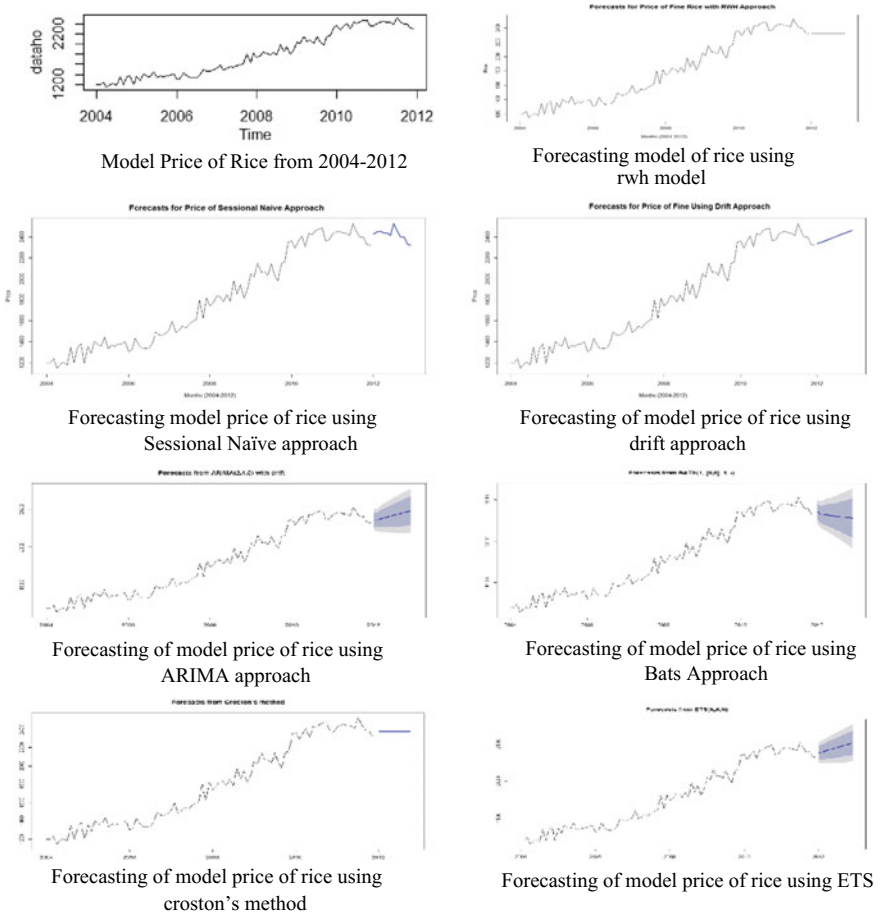


Fig. 7 Forecasting using different methods of forecasting package in R and proposed ANN method

5 Conclusion

In this paper, we discussed and experimented with various forecasting models for prediction of market price of rice. We also presented a new regression-based neural network model for time series prediction. NN is one of the popular models used for classification purpose; we used some data as feature and some data to decide the target class; in this way, it is used as regression model. All experiments were performed in fine quality rice data of India. In this paper, due to large size of data, we obtained the monthly aggregation of price of rice. The proposed neural network model provides a remarkable result with MPE, MAPE evaluation parameter. In the future, we will explore these prediction models for other variety of rice as well as weekly and on daily data.

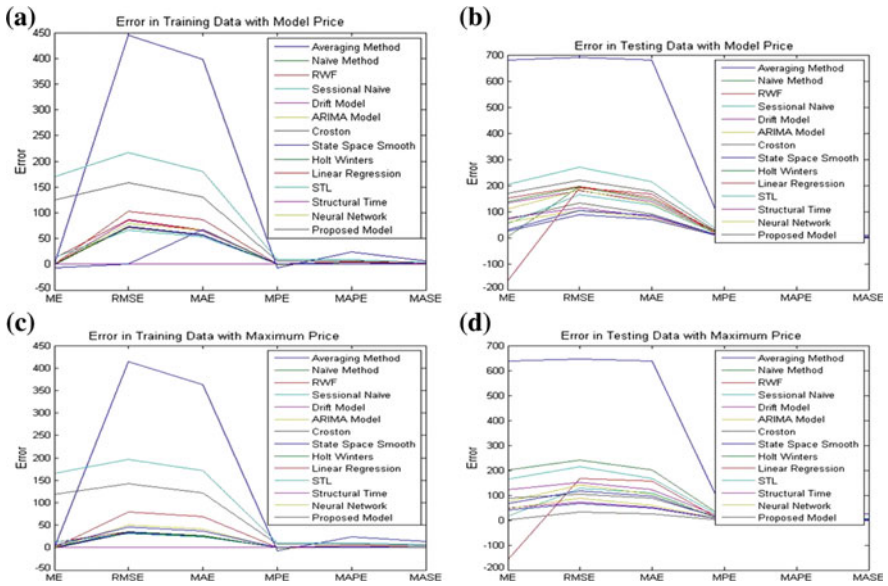


Fig. 8 Performance of different forecasting models. **a** Forecasting on training data of model price. **b** Forecasting on testing data of model price. **c** Forecasting on training data of maximum prices. **d** Forecasting on testing data of maximum prices

Acknowledgements This work is supported by Chhattisgarh Council of Science & Technology (CCOST), Raipur, Chhattisgarh under the grant No. 8057/CCOST/MRP/13, Raipur, dated 27.12.2013 and project titled, “Predictive and Visual Analysis of Price Distribution Information of Rice and Wheat across India”. We thank National Institute of Technology Raipur for providing necessary infrastructure and time to carry this research work.

References

1. Adya, M., Collopy, F.: How effective are neural networks at forecasting and prediction? a review and evaluation. *J. Forecast.* **17**, 481–495 (1998)
2. https://data.gov.in/catalog/variety-wise-daily-market-prices-data-rice#web_catalog_tabs_block_10. Accessed on Apr 2014
3. Nagwani, N.K., Verma, K., Verma, S.: On visualizing the price distribution information of rice. And wheat across India. *Int. J. Eng. Res. Sci. Technol.* **4**(1), 145–152 (2015)
4. Jamil, W., Kalnishkan, Y., Bouchachia, A.: Aggregation algorithm vs. average for time series prediction. In: Proceedings of the ECMLPKDD 2016 Workshop on Large-Scale Learning from Data Streams in Evolving Environments STREAMEVOLV-2016, September (2016)
5. <https://cran.r-project.org/web/packages/forecast/index.html>. Accessed on Aug 2014
6. Yule, G.U.: On the method of investigating periodicities in disturbed series, with special reference to Wölfer’s sunspotnumbers. *Philos. Trans. R. Soc. Lond. Ser. A* **226**, 267–298 (1927)
7. Box, G.E.P., Jenkins, G.M.: Time series analysis: forecasting and control. San Francisco, Holden Day (revised 1976) (1970)
8. Brown, R.G.: Statistical forecasting for inventory control. McGraw-Hill, New York (1959)

9. Brown, R.G.: Smoothing, forecasting and prediction of discrete time series. Englewood Cliffs, NJ, Prentice-Hall (1963)
10. Winters, P.R.: Forecasting sales by exponentially weighted moving averages. *Manage. Sci.* **6**, 324–342 (1960)
11. Holt, C.C.: Forecasting seasonals and trends by exponentially weighted moving averages. *Int. J. Forecast.* **20**(1), 5–10 (2004). <https://doi.org/10.1016/j.ijforecast.2003.09.015>
12. Pegels, C.C.: Exponential smoothing: some new variations. *Manage. Sci.* **12**, 311–315 (1969)
13. Gardner, E.S.: Exponential smoothing—the State of the Art. *J. Forecast.* **4**(1), 1–28 (1985)
14. Kalman, R.E.: A new approach to linear filtering and prediction problems. *Trans. ASME—J. Basic Eng.* **82D**(1), 35–45 (1960)
15. Shumway, R.H., Stoffer, D.S.: An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.* **3**(4), 253–264 (1982)
16. De Gooijer, Jan G., Hyndman, Rob J.: 25 years of time series forecasting. *Int. J. Forecast.* **22**, 443–473 (2006)
17. Darbellay, G.A., Slama, M.: Forecasting the short-term demand for electricity: do neural networks stand a better chance? *Int. J. Forecast.* **16**, 71–83 (2000)
18. Qi, M.: Predicting US recessions with leading indicators via neural network models. *Int. J. Forecast.* **17**, 383–401 (2001)

Wind Power Forecasting Using Hybrid ARIMA-ANN Technique



Pavan Kumar Singh, Nitin Singh and Richa Negi

Abstract The wind power forecasting along with the prior knowledge of wind speed has become very important for the efficient functioning of wind power generation and effective management of risk and revenue. Several single approach models are there for forecasting of wind power, i.e., ARIMA, support vector machine (SVM), artificial neural networks (ANN), extreme learning machine (ELM), etc., but hybridization of these models is considered as an effective alternative for forecasting. In the proposed work, the hybridized model combining ARIMA and artificial neural network (ANN) is presented in order to provide a better prediction of wind power. The wind speed data of Denmark is used for evaluation of the proposed model. From the result obtained, it becomes evident that the hybridization of ARIMA and ANN is better in forecasting the wind power as compared to the two models working separately for wind power forecasting.

Keywords Wind energy forecasting · Wind speed forecasting · Smart grid · ARIMA · ANN · Hybrid model · Renewable energy

1 Introduction

Sustainable and clean energy is becoming an interesting area in the research and industry sector. The most cited sustainable source of energy is wind energy. As wind energy does not release any atmospheric emission, so it is the most efficient and clean source of energy. The probabilistic nature of wind poses a challenge in the grid integration of wind energy sources to the existing grid, which leads to a large number of problems like unit commitment, power quality, and frequency deviation. The system operators in order to increase the reliability of the grid require accurate wind prediction methods to calculate the approximate amount of wind energy that can be obtained at a specific time. Precise methods for predicting wind power can

P. K. Singh · N. Singh (✉) · R. Negi
Department of Electrical Engineering, MNNIT Allahabad, Allahabad, India
e-mail: nitins@mnnit.ac.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_19

Table 1 Wind power forecasting existing approaches [1]

Approaches	Input	Example
Method based on artificial intelligence	Previous year data	Artificial neural network, support vector machine, etc.
Methods based on statistical techniques	Previous year data	ARIMA, GARCH, Kalman filter, etc.
Physical methods	Meteorological data	Numerical weather prediction (NWP), etc.

help mitigate the undesirable effects of wind energy, i.e., provide stability and a rate in the electricity market. Power from wind turbine relies upon the wind speed, which is variable in nature and also it depends upon the regional climatic conditions and the terrain type. For any wind turbine, the wind power and speed are related; the connection among the wind power $P(W)$ generated at any particular wind speed can be given as (1) [1]

$$P = 1/2(\rho A v^3) \quad (1)$$

where the air density in kg/m^3 is given as ρ ; it is based on atmospheric temperature and pressure. From Eq. (1), it is found that the relation between the wind power and speed is nonlinear in nature. As the wind power and speed correlate positively, hence the prediction of wind power and the speed is equivalent. Wind power forecasting can be done using individual models and hybrid models. There are different approaches in case of single approach forecast. These are computational intelligence, statistical, and physical approach [2]. A summary of the individual models for forecasting is given in Table 1.

Physical models use information like air pressure, air temperature, and the climatic information for the forecast of wind power. Meteorologists develop these models for large-scale climatic predictions but for short-term forecast the results obtained from these models is erroneous. This paper proposes an ARIMA-ANN hybrid model, which combines ARIMA and artificial neural network (ANN) to predict the wind power. These models are used with the wind speed data in the short and medium term forecast shelves (1, 6, and 12 h of forward forecast).

The rest of the document is organized into various sections, i.e., Sects. 2 and 3 give the detail description of the autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) models. Section 4 describes the feasibility of selecting the hybrid model and the methodology used, i.e., how these models were hybridized in detail. Finally, the results are discussed in Sect. 5, and the conclusion of the manuscript is given thereafter.

2 ARIMA

ARIMA is a famous tool for predicting and modeling time series [1–4], which can be used to predict wind power. ARIMA comprises three different operations, where AR stands for Autoregressive, I stands for Integrated, and MA stands for moving average.

The formulation of the ARIMA model from the time series is done by using the maximum probability estimate. The application of ARIMA model for forecasting involves the following steps: (i) determination of the appropriate model from the existing class of models; (ii) evaluation of the model parameters; (iii) validation of the obtained model; and (iv) obtaining the forecast. The mathematical expression for ARIMA (p, d, q) can be written as (2)

$$\nabla^d x_t = \underbrace{\sum_{i=1}^p \vartheta_i \nabla^d x_{t-i}}_{\text{AR Term}} + w_t + \underbrace{\sum_{j=1}^q \theta_j w_{t-j}}_{\text{MA Term}} \quad (2)$$

where the wind power time series is given as x , Gaussian white noise with zero mean and constant variance is shown as w , the AR coefficients are shown as ϑ_i , and θ_j is the MA coefficient. The nonstationary time series is made stationary by differencing the series which is shown by the operator ∇^d ; the operator ∇^d is given by (3)

$$\nabla^d x = (x_t - x_{t-1}) - (x_{t-1} - x_{t-2}) \dots (x_{t-d+1} - x_{t-d}) \quad (3)$$

2.1 Model Identification

The order of the ARIMA model is decided by the orders of p , d , and q terms, respectively. It is necessary to study statistics of wind power time series; it is done by studying the plots of autocorrelation function (ACF) and partial ACF (PACF) [2, 5]. The autocorrelation plots show that any two data points in the series, i.e., x_s and x_t , do have a positive correlation with each other. The equations for calculating ACF [$\rho(s, t)$] between x_s and x_t are given as (4) and (5), respectively:

$$\rho(s, t) = \frac{\gamma(s, t)}{\sqrt{\gamma(s, s)\gamma(t, t)}} = \frac{\gamma(s, t)}{|\sigma_s \sigma_t|} \quad (4)$$

$$\gamma(s, t) = E[(x_s - \mu_s)(x_t - \mu_t)] \quad (5)$$

PACF differs from ACF, because all the data points in case of ACF are considered between x_s and x_t , $h = |s - t|$, known as the distribution effect [6]. But PACF ignores intermediate points and shows the connection between x_s and x_t .

The order of the ARIMA model is found out by evaluating the histogram, ACF, and PACF plots. First, the order to differencing term, i.e., d is calculated by the fact that the histogram is distributed normally. Second, the order of p is determined by inspecting the last significant point of PACF; third, the order of the q or MA terms is decided by evaluating the break point ACF plot. This break point is generally taken as the order of q or MA terms. Then, the basic ARIMA models are chosen by taking the considering the p and q terms as zero, i.e., ARIMA ($p, d, 0$) which is AR model and ARIMA ($0, d, q$) which is MA model with stationary series.

2.2 Model Estimation

After the range of order p and q are obtained, the numerical values of $\varphi_1, \dots, \varphi_p$ is found by Yule–Walker estimation [5] method. Durbin estimates are used to find numerical values $\theta_1, \dots, \theta_q$.

However, there may exist several possible ARIMA models of candidates, as discussed in the last subsection. After calculating the parameter value, the Bayesian Information Criteria (BIC) or the Akaike information criteria (AIC) are used to select the best model for each candidate model [7, 8]. The equations for calculating AIC and BIC are given as (6) and (7)

$$\text{AIC} = \log \frac{\sum_{t=k}^n (x_t - \bar{x})^2}{n} + \frac{n + 2k}{n} \quad (6)$$

$$\text{BIC} = \log \frac{\sum_{t=k}^n (x_t - \bar{x})^2}{n} + \frac{k \log n}{n} \quad (7)$$

where the number of parameters is given by k and the sample size in the model is given by n . The ARIMA model with the estimated parameters can be used for forecasting the day ahead values of x_{t+m} based on the historical values of the data collected by the current $x_t, x_{t-1}, \dots, x_{t-n}$.

3 Artificial Neural Network Model

ANN is data processing model which is inspired by the biological nervous system, such as the process of processing of the human brain. It is made up of a large number of highly interconnected process elements (neurons) that work in conjunction with each other to solve specific problems [9–12]. NNs learn by examples like human being. Artificial neuron has many inputs and only one output as shown in Fig. 1.

To build an ANN model, the information of input vectors, neurons, layers, and output vector is essential. However, there is no general rule, it is important to start

Fig. 1 A simple neuron

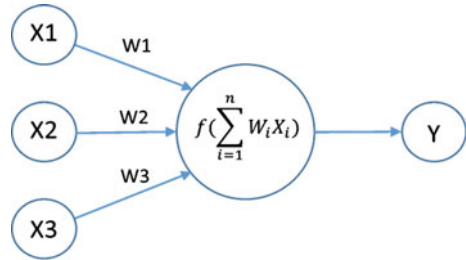
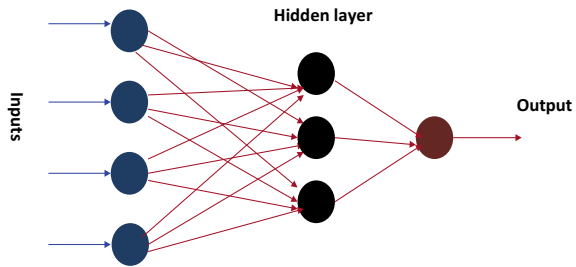


Fig. 2 Artificial neural network



with simple architecture, which is acceptable, and if not, then you must offer more complex configurations.

An artificial neurological network consists of various artificial neurons as shown in Fig. 2 that connect to each other according to the architecture of a particular network. The purpose of the neural network is to interpret meaningful information in the input data.

In order to predict the future value, an inherent strategy called the repetition method is used. For example, for a two delay forecast models that present the following statement given by (8). For such example in order to see two periods in advance, the model will be given as (9)

$$y_{t+1} = f(x_t, x_{t-1}) \tag{8}$$

$$y_{t+2} = f(y_{t+1}, x_t) \tag{9}$$

Similarly, for three period’s models the model will be as shown in (10)

$$y_{t+3} = f(y_{t+2}, y_{t+1}) \tag{10}$$

It is commonly used for short-term prediction that affects the delay ratio with respect to delay. Figure 3 shows the block diagram of ANN which provides the overall view of the working of ANN which adjusts the weights between neurons to give the desired output. Previous day same hour wind speed data, previous week same hour wind speed data and previous day 24 h average are used as input.

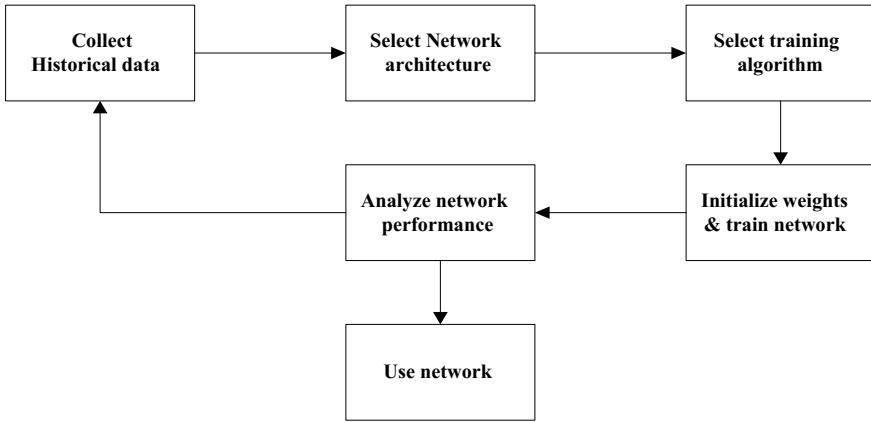


Fig. 3 Block diagram of ANN

4 Proposed Hybrid Model

Dynamic series of wind energy show linear and nonlinear functions. Modeling and forecasting of linear and nonlinear figures are considered as a separate hybrid model [13–15]. The general hybrid model is given in the following text.

First, the linear methodology such as ARIMA is used to model the linear part $\widehat{L}_t = f_l(x_{t-1}, \dots, x_{t-p})$ and then the residuals are calculated $\vec{E}_t = \vec{x}_t - \widehat{L}_t$. Second, a nonlinear method, similar to ANN, is applied to the remainders for modeling a nonlinear part $\widehat{N}_t = f_n(E_{t-1}, \dots, E_{t-n})$. Then, the two simulated parts are summed up in order to get the final forecast, i.e., $\widehat{x}_t = \widehat{L}_t + \widehat{N}_t$ (Fig. 4).

The equations of hybrid ARIMA-ANN model are given below as (11), (12), (13) and (14). All the equations are progressive in nature.

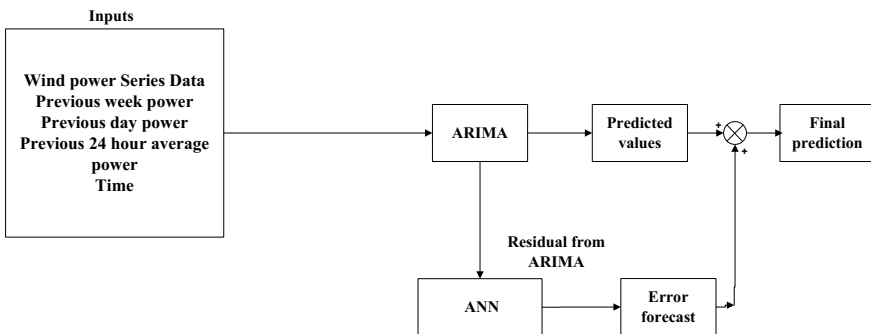


Fig. 4 Block diagram of hybrid model

$$\widehat{L}_t = ARIMA(x_{t-1}, \dots, x_{t-p}) \tag{11}$$

$$\overrightarrow{E_{t-i}} = \overrightarrow{x_{t-i}} - \widehat{L}_{t-1} \tag{12}$$

$$\widehat{N}_{t+m} = ANN(E_{t-1}, \dots, E_{t-n}) \tag{13}$$

$$\widehat{x}_{t+m} = \widehat{L}_{t+m} + \widehat{N}_{t+m} \tag{14}$$

5 Results and Discussion

From January 2016 to December 2016, the Danish wind power data was collected from the data center in Denmark having a resolution of 1 h. Monthly analysis of the wind power series is carried out, the first 70% of the monthly data training and with the rest 30% of the test data is performed (Figs. 5 and 6).

The ARIMA model is simulated using the Matlab, and the ANN model is simulated in Matlab using the toolkit. The order and parameters of the ARIMA model are found from the ACF and PACF diagrams. For the proposed hybrid ANN-ARIMA model, the residuals are found by calculating the variation between the actual wind energy and the results obtained from ARIMA. These residuals were given as input to the ANN. The remnants were taught using ANN, and output from ANN was added to the ARIMA output to form the expected final result (Figs. 7, 8 and 9).

To compare the models, some error estimation methods like mean absolute error (MAE), mean error (ME), mean square error (MSE), and mean absolute percentage error (MAPE) were used. If the actual data is y_t at given time t and the forecasted data

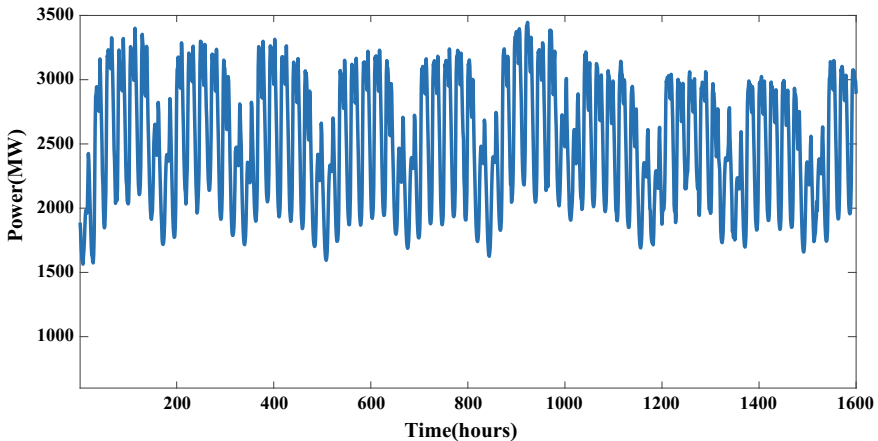


Fig. 5 Historical data

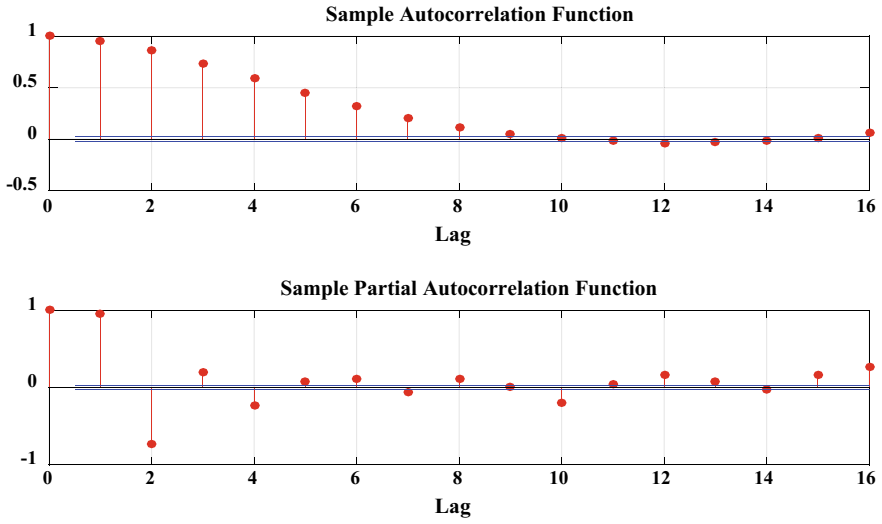


Fig. 6 ACF and PACF plot for p and q determination

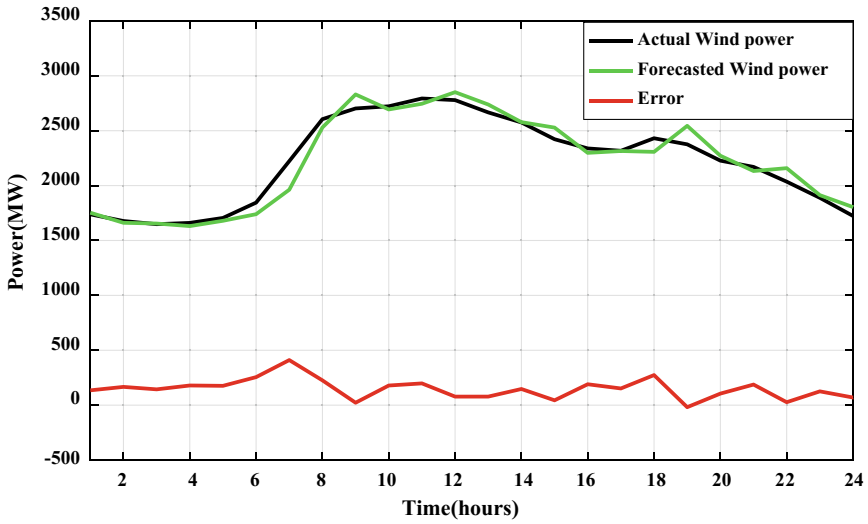


Fig. 7 Wind power forecast using ARIMA

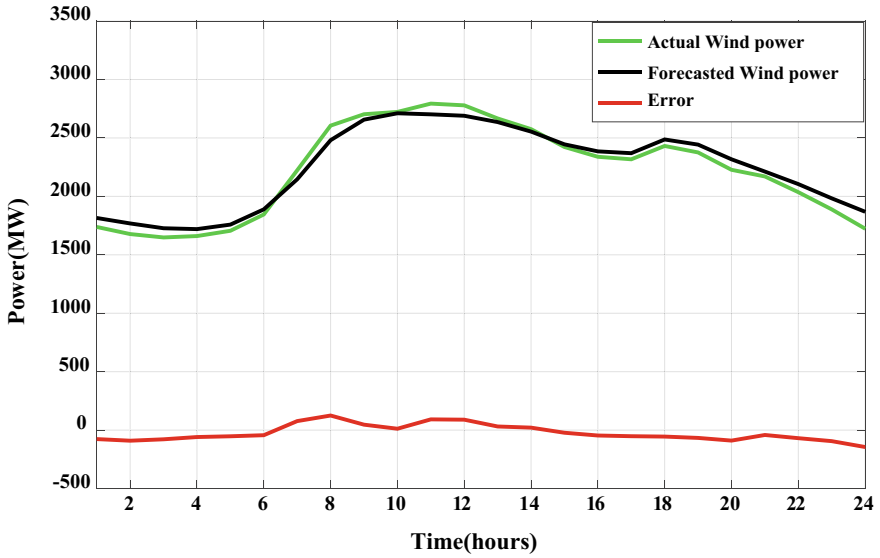


Fig. 8 Wind power forecast using ANN

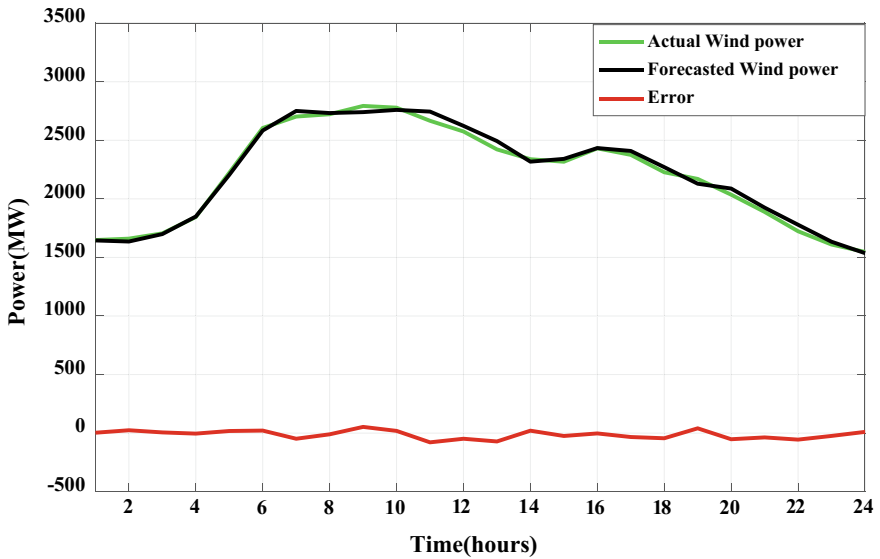


Fig. 9 Wind power forecast using hybrid model

is F_t for the same time, then the error is calculated by using the expression shown using (15)

$$e_t = y_t - F_t \quad (15)$$

Then the mean average error (MAE) is given as (16)

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n e_t \quad (16)$$

The mean square error (MSE) is given as (17)

$$\text{MSE} = \frac{1}{n} \sum_{t=1}^n e_t^2 \quad (17)$$

The mean absolute error (MAE) can be calculated as (18)

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |e_t| \quad (18)$$

The mean absolute percent error (MAPE) is given as (19)

$$\text{MAPE} = \frac{1}{n * y_t} \sum_{t=1}^n |e_t| * 100 \quad (19)$$

See Tables 2 and 3.

6 Conclusion

This paper proposes a hybrid ARIMA-ANN model for predicting the wind power for smart grids. The concept of ANN, ARIMA, and the feasibility of hybridizing the ARIMA-ANN model are discussed. The wind power data of Denmark is used for the evaluation of suggested hybrid model, and the result shows that all three models forecast the behavior of time series rationally. But when estimating the statistical errors, it is found that the hybrid model has a higher forecasting accuracy of 3.0662% than the ARIMA and ANN models, which have an accuracy of 7.8131 and 3.9103%, respectively. From this, it is found that the hybrid model can be a good option for predicting wind power. Moreover, the hybrid model improves the overall forecasting performance by taking care of the shortcoming posed by the individual model.

Table 2 Comparison of actual and forecasted data for 24 h

Actual values (MW)	Forecasted values (MW)		
	ARIMA	ANN	Hybrid ARIMA-ANN
1648.4	1654.454	1701.514	1672.499
1660.3	1631.136	1692.226	1648.4
1705.4	1679.386	1741.255	1693.624
1845	1739.64	1887.839	1829.907
2222.9	1962.31	2177.503	2166.511
2605.2	2529.403	2469.413	2589.78
2702.7	2830.33	2656.685	2749.369
2722.5	2693.621	2733.284	2713.8
2793.6	2745.402	2715.492	2744.706
2778.7	2850.855	2687.745	2779.618
2666.6	2738.584	2631.384	2752.198
2575.2	2578.025	2568.822	2614.927
2422.4	2528.238	2475.665	2484.491
2338.6	2298.192	2412.573	2323.719
2317.1	2314.984	2349.709	2312.776
2431.2	2307.276	2479.876	2454.329
2375.8	2544.172	2471.121	2412.82
2227.2	2271.947	2339.904	2260.834
2170.3	2132.717	2229.072	2110.782
2035.8	2158.839	2137.077	2076.868
1888.6	1912.951	2002.726	1923.212
1723.6	1804.04	1860.741	1772.721
1610.1	1604.786	1724.04	1620.162
1547.2	1555.783	1642.394	1540.463

Table 3 Comparison of statistical error measures

Model	MAE	MSE	RMSE	MAPE
ARIMA	150.6967	2.7452e+04	165.6864	7.8131
ANN	84.8139	1.4861e+04	121.9055	3.9103
HYBRID (ARIMA + ANN)	58.6423	5.1975e+03	72.0939	3.0662

References

1. Ye, R., Suganthan, P.N., Srikanth, N., Sarkar, S.: A hybrid ARIMA-DENFIS method for wind speed forecasting. *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6 (2013). <https://doi.org/10.1109/FUZZ-IEEE.2013.6622503>
2. Soman, S.S., Zareipour, H., Malik, O., Mandal, P.: A review of wind power and wind speed forecasting methods with different time horizons, *North American Power Symposium*. pp. 1–8 (2010). <https://doi.org/10.1109/NAPS.2010.5619586>
3. Radziukynas, V., Klementavičius, A.: Short-term wind speed forecasting with ARIMA model. *IEEE. 55th International Scientific Conference on Power and Electrical Engineering of Riga Technical University (RTUCON)*. pp. 145–149 (2014). <https://doi.org/10.1109/RTUCON.2014.6998223>
4. Hill, D.C., McMillan, D., Bell, K.R.W., Infield, D.: Application of auto-regressive models to U.K. wind speed data for power system impact studies. *IEEE Trans. Sustain. Energy* **3**(1), 134–141 (2012). <https://doi.org/10.1109/TSTE.2011.2163324>
5. Eshel, G.: The Yule Walker equations for the AR coefficients. University of South Carolina, 215. *Tech. Rep.* 1–8 (2010). [Online]. Available: http://www.stat.sc.edu/~vesselin/STAT520_YW.pdf
6. Eldali, F.A., Hansen, T.M., Suryanarayanan, S., Chong, E.K.: Employing ARIMA models to improve wind power forecasts: a case study in ERCOT, *IEEE. North American Power Symposium (NAPS)*, pp. 1–6 (2016). <https://doi.org/10.1109/NAPS.2016.7747861>
7. Pai, P.-F., Lin, C.-S.: “A hybrid arima and support vector machines model in stock price forecasting”. *Omega* **33**(6), 497–505 (2005)
8. Min, Y., Bin, W., Liang-Li, Z., Xi, C.: “Wind speed forecasting based on EEMD and ARIMA”. *IEEE*, pp. 1299–1302 (2015). <https://doi.org/10.1016/j.omega.2004.07.024>
9. Carolin, M., Fernandez, M.E.: “Analysis of wind power generation and prediction using ANN: a case study”. *Renew. Energy* **33**, 986–992 (2008) <https://doi.org/10.1016/j.renene.2007.06.013>
10. Catalão, J.P.S., Pousinho, H.M.I., Mendes, V.M.F.: “Short-term wind power forecasting in Portugal by neural networks and wavelet transform”. *Renewable Energy* **36**, 1245–1251 (2011) <https://doi.org/10.1016/j.renene.2010.09.016>
11. Catalao, J.P.S., Pousinho, H.M.I., Mendes, V.M.F.: “An artificial neural network approach for short-term wind power forecasting in Portugal”. *15th International Conference on Intelligent System Applications to Power Systems*, pp. 1–5 (2009) <https://doi.org/10.1109/isap.2009.5352853>
12. Kasabov, N.K., Song, Q.: “DENFIS: dynamic evolving neural-fuzzy inference system and its application for time-series prediction”. *IEEE Trans. Fuzzy Syst.* **10** (2), 144–154 (2002) <https://doi.org/10.1109/91.995117>
13. Kariniotakis, G.N., Stavrakakis, G.S., Nogaret, E.F.: “Wind power forecasting using advanced neural networks models”. *IEEE Trans. Energy. Conver.* **11**(4), 762–767 (1996) <https://doi.org/10.1109/60.556376>
14. Cadenas, E., Rivera, W.: “Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA-ANN model”. *Renew. Energy* **35**(12), 2732–2738 (2010) <https://doi.org/10.1016/j.renene.2010.04.022>
15. Chang, G.W., Lu, H.J., Hsu, L.Y., Chen, Y.Y.: “A hybrid model for forecasting wind speed and wind power generation”. *IEEE Power and Energy Society General Meeting (PESGM)*, pp. 1–5 (2016) <https://doi.org/10.1109/pesgm.2016.7742039>

Part III
Hardware and Software Design

On the Development of Feature-Based Sprint in AGILE



Sarika Sharma and Deepak Kumar

Abstract AGILE methodology is widely used throughout the information technology industry for software development. In AGILE methodology, the delivery cycle is broken down into sprints or iterations. Many iteration-based AGILE teams use a time-boxed 1-hour discussion midway through a 2-week iteration (the team selects an iteration duration that provides them frequent enough feedback) (Cohn in AGILE Estimating and Planning, 2005 [1]). A sprint or iteration is expected to deliver a piece of functionality within a fixed period of time. The length of AGILE sprint is measured in number of days. Hence, it can be concluded that the unit of sprint is time, measured in numbers of days. But since AGILE is feature-driven, the unit of sprint should not be limited to time. There is a need to explore other measures that can be used to define or determine the length of the AGILE sprint. The purpose of this paper is to find out any other unit to define or derive the length of sprint. The paper also explores the various ways in which the AGILE developer(s), AGILE tester(s), and AGILE end user(s) can be benefited by changing the unit of sprint length from time to something else. This paper particularly talks about the possibility of using the “feature” as unit of sprint and lists down the advantages and disadvantages of using “feature” over “time” as unit of sprint. The term “sprint” or “iteration” is used as synonyms throughout the discussion.

Keywords Software engineering · Software development · AGILE · Sprint · Iteration · Length of sprint · Sprint planning · Iteration · Planning · Fixed length sprint

S. Sharma (✉) · D. Kumar
Amity University, Noida, Uttar Pradesh, India
e-mail: sarika.s17@gmail.com

D. Kumar
e-mail: deepakgupta_du@rediffmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_20

1 Introduction

1.1 Definition—Length of Sprint

A sprint (or iteration) is the basic unit of development in AGILE. The sprint is a time-boxed effort; that is, it is restricted to a specific duration [2]. The duration is fixed in advance for each sprint and is normally between 1 week and 1 month, with 2 weeks being the most common [3]. Most AGILE teams work in iterations of 2–4 weeks. It is possible to go slightly longer, and some teams have experimented with even shorter iteration [4]. Hence, there is no standard way to determine the length of sprint or iteration under AGILE methodology. Based upon the current practices available in the market, the length of sprint can be defined as the time frame usually measured in number of days to deliver some piece of functionality to end user for providing required feedback to the AGILE team involved in the development of the software using AGILE methodology. Usually, the time frame chosen by AGILE teams as sprint length is selected keeping in mind that frequent customer or end user feedbacks can be obtained. That said, it is not necessary that at the end of a sprint or iteration the customer will have a view of a complete feature for providing review comments. Here, feature refers to a functionality that can be examined by the end user as a complete unit independently. This is context a feature can be represented by a set of user stories. Hence, even when the customer feedback is obtained on the output deliverable of a sprint, it might be of little use since at this point of time the customer has not analyzed a complete feature and therefore does not have a fair understanding of what is being developed and what will be offered to him at the end of the subsequent sprints or iterations. It has been seen that AGILE teams have come up with a common practice to choose different time-boxes for different sprints. Sometimes, as per the AGILE teams, every sprint is allocated the same time frame across the project and in some cases, the chosen time-boxes are different for different sprints or iterations running under a given project.

Where the chosen time-boxes are consistent across all the sprints or iterations under a given project, it can be termed as “fixed length sprints” and where the chosen time-boxes are different for at least one sprint under a given project; then it can be termed as “variable length sprints”. However, changing the time-boxes does not mean that at the end of a sprint or iteration the output of a given sprint will allow an end user to get a feature-based view of the product being developed. Hence, keeping in mind the purpose of AGILE, i.e., to ensure end user engagement in the most effective way so that AGILE team can ensure the product being developed is in line with end user(s) requirements. It has become important to ensure that the output of the sprint or iteration should provide a glimpse of the software product being developed to meet end user(s) requirement. Thus, there is a need to come up with a trustworthy approach or mechanism to ensure the output of the sprint is useful.

The proposed alternative that can be used to define the length of sprint is with the help of “feature”, since a software product is made up of several features, and features are a component that can be analyzed by end user to provide review comments about

Table 1 Tabular representation of fixed sprint length

Assigned sprint number	Sprint length (in days)
Sprint 1	SL_Fix
Sprint 2	SL_Fix
Sprint 3	SL_Fix
Sprint 4	SL_Fix
Sprint 5	SL_Fix

the software product being developed. By using “feature” for defining the length of sprint or iteration; one can ensure the output of the sprint can be evaluated by end user and other stakeholders in a much efficient manner. Hence, the proposed approach suggests the length of the sprint should be equal to the number of days taken to deliver at least one feature. The AGILE developer will benefit by this approach since the developer has to think around the complete delivery of a feature rather than completing a piece of work within some given time frame. Also, the AGILE testers will be able to test the complete feature as an output of a sprint and will know what has been tested and would benefit in planning the testing related to different features.

2 Industry Practice

2.1 Time as Length of Sprint

2.1.1 Fixed Length Sprint

The sprint is simply a time-box of between 1 and 4 weeks that provides a space for work to be done [5]. A fixed length sprint can be defined as a set of sprints for which the length of all the sprints is equal under a given project.

As you can see in Fig. 1, all the sprints are of equal size say “*t*” is the number of days. The technique of fixed length sprint can ensure the customer to have a glimpse of the product developed after every “*t*” number of days but unfortunately, it does not promise that the output of every sprint is a feature or a complete functionality that can be evaluated by the end user, since every feature to be developed for a product not necessarily takes equal number of days. The risk is if the length of the sprint is too short, then the delivery of one feature will spill over two or more sprints which will make the evaluation of the feature at customer end quite difficult. Also, if the length of the sprint is too long then two or more features will be delivered simultaneously which will add an unnecessary burden on the end user(s) since evaluating and providing feedback on so many features in one go will be quite tedious and may impact the quality of feedback provided.

Table 1 highlights the length of the sprint is sprints under a given project using fixed length sprint approach which is denoted by string “SL_Fix”.

Fig. 1 Graphical representation of a set of fixed length sprints where all the sprints are of equal length

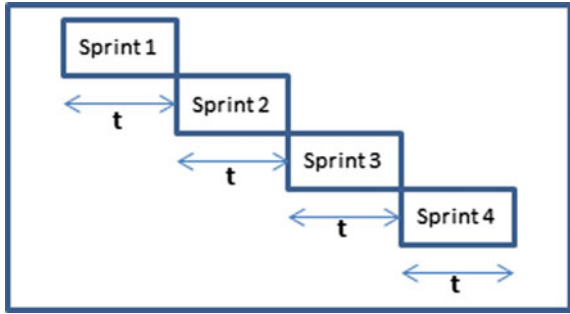
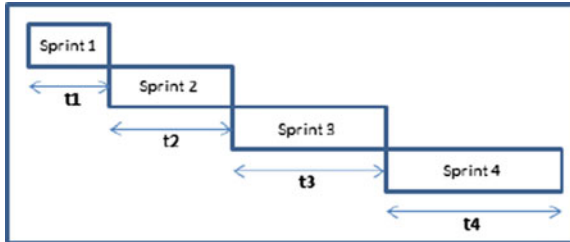


Fig. 2 Graphical representation of a variable length sprint where length of at least one sprint is different from other sprints



Now, by adding the length of all the sprints the total time estimated to deliver the project (say T_{EstFix}) can be calculated as

$$T_{EstFix} = SL_Fix + SL_Fix + SL_Fix + SL_Fix + SL_Fix \tag{1a}$$

Or it can be written as

$$T_{EstFix} = SL_Fix * \text{Number of Sprints} \tag{1b}$$

The above equation will be used to compare the various approaches defining the length of sprint in Sect. 4.

2.1.2 Variable Length Sprint

A variable length sprint can be defined as a set of sprints for which length of at least one sprint is different from other sprints or iterations for a given project.

As you can see in Fig. 2, the length of sprint is different for different iterations under a given project. The reason behind keeping the iterations of different lengths is neither defined nor standard; it may vary from one AGILE team to another for the same project under consideration. In this scenario, also it cannot ensure the part of the required software to be developed at the end of every sprint can be evaluated by the end user as a complete unit or feature.

Table 2 Tabular representation of variable sprint length

Assigned sprint number	Variable sprint length (in days)
Sprint 1	SL_VAR1
Sprint 2	SL_VAR2
Sprint 3	SL_VAR3
Sprint 4	SL_VAR4
Sprint 5	SL_VAR5

Table 2 highlights that the length of sprint, denoted by string “SL_Varn”, is different for various sprints defined under a project using variable length sprint approach.

Now, by adding the length of all the sprints, the total time estimated to deliver the project (say T_{EstVar}) can be calculated as

$$T_{EstVar} = SLVAR1 + SLVAR2 + SLVAR3 + SLVAR4 + SLVAR5 \quad (2)$$

The above equation will be used to compare the various approaches to define the length of the sprint in Sect. 4.

3 Proposed Approach

3.1 Feature as Length of Sprint

Let us assume a feature can be used to define unit of sprint or iteration measurement when each sprint or iteration output is at least a feature that can be evaluated by the end user in order to provide the required feedback to the AGILE team for a product being developed.

H_0 : Feature can be used as a unit of sprint measurement.

H_1 : Feature cannot be used as a unit of sprint measurement.

Here, H_0 is null hypothesis, and H_1 is alternate hypothesis.

As you can see in Fig. 3, all the sprints are defined in a way that the outcome of a sprint is always a feature. Hence, it ensures the output of every sprint is at least a feature that can be evaluated by the end user to provide relevant feedback to the AGILE team for the product being developed. Also, it tells us in advance how many defined iterations will be required to deliver a software product. In this case, on the completion of the fourth sprint, the software development should come to an end if no error or bugs found.

Using “feature” as unit of sprint will indirectly impact the length of a sprint when measured in number of days. This is because every feature may not take equal number of days for development. Hence, if one want to compare the “time” as length of sprint to “feature” as length of sprint then s/he will notice in “feature” based length of sprint,

Fig. 3 Graphical representation of a length of sprint where “feature” is the unit of sprint measurement

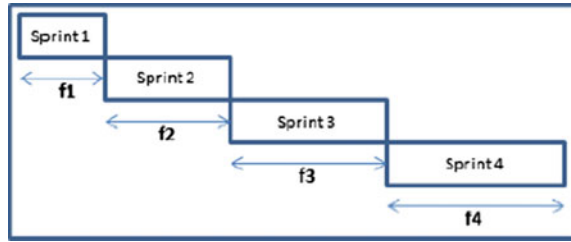


Table 3 Tabular representation of length of sprint as feature

Assigned sprint number	Feature to be delivered	Time taken to deliver a feature (in range)
Sprint 1	Feature 1	$R1_{min} - R1_{max}$
Sprint 2	Feature 2	$R2_{min} - R2_{max}$
Sprint 3	Feature 3	$R3_{min} - R3_{max}$
Sprint 4	Feature 4	$R4_{min} - R4_{max}$
Sprint 5	Feature 5	$R5_{min} - R5_{max}$

the length of sprint is expressed in time interval say range (in number of days) as shown in Table 1. When doing sprint planning for similar projects, one can refer to the history database to determine what should be the expected time interval to deliver a similar feature.

Table 3 highlights the length of sprint in range (denoted by string “ Rn_{min} ” and “ Rn_{max} ”) for various sprints defined under a project using feature as length of sprint approach, where Rn_{min} denotes minimum estimated time to deliver a feature, and Rn_{max} denotes maximum estimated time to deliver a feature.

Now, by adding the length of all the sprints, i.e., all lower limits and all upper limits separately, the total time estimated to deliver the project in minimum (say T_{min}) and maximum (say T_{max}) number of days can be calculated as

$$T_{min} = R1_{min} + R2_{min} + R3_{min} + R4_{min} + R5_{min} \tag{3}$$

$$T_{max} = R1_{max} + R2_{max} + R3_{max} + R4_{max} + R5_{max} \tag{4}$$

Also,

$$\text{Estimated Project Delivery Time Interval} = T_{min} \text{ to } T_{max} \tag{5}$$

From Eq. (5), it can be said that the estimated time interval (in days) for delivering the project will be between the minimum time estimated to deliver the project (T_{min}) and the maximum time estimated to deliver the project (T_{max}).

The above equation(s) will be used to compare the various approaches to define length of the sprint in Sect. 4.

4 Implementation Details

Let us compare all three approaches, i.e., fixed length sprint, variable length sprint, and feature as unit of sprint while developing the fund transfer utility for a bank (see Table 4).

Before executing the case study for examining all the three approaches, the below prerequisites were agreed to ensure the comparisons made are fair and acceptable:

- The given project is developed by the same AGILE team using a given sprint length approach at a time.
- The said project is analyzed by the same end user(s).
- All the provided conditions remain constant for all the three runs of software development; except the technique to define the sprint of length.
- For fixed length sprint and variable length sprint, the sprint length is a random number chosen by the AGILE team while for feature as length of sprint, the similar historic projects are referenced to find out time interval required to deliver a feature.
- After every sprint, the customer feedback was recorded.

The data used for analysis while writing this paper is sourced from Waves Zones Technologies Pvt. Ltd., Ghaziabad (U.P.), India.

4.1 Using Fixed Length Sprint

The above requirements were delivered using a fixed time-box sprint approach. From Table 1, it can be seen that for every sprint AGILE team have randomly allocated

Table 4 The features to be developed for the fund transfer utility for a bank

Assigned feature number	Feature to be developed
<i>F1</i>	Fund transfer between various accounts of same account holder
<i>F2</i>	Intra-bank fund transfer within a state
<i>F3</i>	Interstate bank fund transfer
<i>F4</i>	International fund transfer in same currency
<i>F5</i>	International fund transfer in different currencies

Table 5 The output of a sprint is a part of the software to be developed say D_n under fixed sprint length in number of days

Assigned delivery number	Part of software to be developed	Sprint number	Number of days
$D1$	Delivery 1	$S1$	15
$D2$	Delivery 2	$S2$	15
$D3$	Delivery 3	$S3$	15
$D4$	Delivery 4	$S4$	15
$D5$	Delivery 5	$S5$	15

15 days as length of sprint which is consistent across all the sprints under the project (Table 5).

Using Eq. (1a),

$$T_{EstFix} = SL_{Fix} + SL_{Fix} + SL_{Fix} + SL_{Fix} + SL_{Fix} \tag{1a}$$

$$T_{EstFix} = 15 + 15 + 15 + 15 + 15 = 75 \text{ days} \tag{1a}$$

Using Eq. (1b),

$$T_{EstFix} = SL_{Fix} * \text{Number of Sprints} \tag{1b}$$

$$T_{EstFix} = 15 * 5 = 75 \text{ days} \tag{1b}$$

Now, let us analyze the delivery at the end of every sprint:

- **At the end of sprint 1:** Delivery $D1$ consists of part of $F1$.
- **At the end of sprint 2:** Delivery $D2$ consists of remaining part $F1$ + part of $F2$.
- **At the end of sprint 3:** Delivery $D3$ consisting of remaining part of $F2$.
- **At the end of sprint 4:** Delivery $D4$ consists of some part of $F4$ and $F5$.
- **At the end of sprint 5:** Delivery $D4$ consists of remaining part of $F4$ and $F5$.
- **Additional sprint 6:** Here, all the required features were not delivered after sprint 5; therefore, the AGILE team added one more sprint of 15 days to deliver the functionality of $F3$.

In this case, the output of a sprint is a piece of software functionality which is not necessarily a feature. Thus, adding difficulty for AGILE developer, AGILE tester and the end user to track how much required functionality has been delivered and what part of the functionality is yet to be delivered. This had an impact on the ability of end user to analyze the output of sprint resulting in poor quality of feedback from end user. Also, the project was not delivered within expected time lines.

Table 6 Depicting sprint length in number of days

Assigned delivery number	Part of software to be developed	Sprint number	Number of days
<i>D1</i>	Delivery 1	<i>S1</i>	20
<i>D2</i>	Delivery 2	<i>S2</i>	15
<i>D3</i>	Delivery 3	<i>S3</i>	10
<i>D4</i>	Delivery 4	<i>S4</i>	20
<i>D5</i>	Delivery 5	<i>S5</i>	10

4.2 Using Variable Length Sprint

The given requirements were now delivered using the variable length sprint. Here, the AGILE team has chosen to use different time periods as lengths of various sprints for the given project (Table 6).

Using Eq. (2),

$$T_{EstVar} = SLVAR1 + SLVAR2 + SLVAR3 + SLVAR4 + SLVAR5 \quad (2a)$$

$$T_{EstVar} = 20 + 15 + 10 + 20 + 10 = 75 \text{ days}$$

Now, let us analyze the delivery at the end of every sprint:

- **At the end of sprint 1:** Delivery *D1* consists of part of *F1* and some part of *F2*.
- **At the end of sprint 2:** Delivery *D2* consists of remaining part *F1* + some other part of *F2*.
- **At the end of sprint 3:** Delivery *D3* consisting of remaining part of *F2* and some part of *F3*.
- **At the end of sprint 4:** Delivery *D4* consists of some part of *F4* and *F5*.
- **At the end of sprint 5:** Delivery *D4* consists of remaining part of *F4* & some part of *F5*.
- **Additional sprint 6 and 7:** Here, all the required features were not delivered after sprint 5 therefore; the AGILE team added two more sprints of 10 and 15 days, respectively, to deliver remaining part of *F3* and *F5*.

Again, the output of a sprint is a piece of software functionality which is not necessarily a feature. Thus, adding difficulty for AGILE developer, AGILE tester and the end user to track how much required functionality has been delivered and what part of functionality is yet to be delivered. This may impact the ability of end user to analyze the output of sprint resulting in poor quality of feedback from end user. Also, the delivery was not completed within expected time lines.

Table 7 Depicting sprint length in number of days

Assigned feature number	Feature to be developed	Sprint number	Number of days (in range)
F1	Fund transfer between various accounts of same account holder	S1	10–12
F2	Intra-bank fund transfer within a state	S2	15–17
F3	Interstate bank fund transfer	S3	10–12
F4	International fund transfer in same currency	S4	15–17
F5	International fund transfer in different currencies	S5	20–22

4.3 Using Feature as Length of Sprint

Now, applying the “feature” as unit of sprint to the same requirements discussed above for “fixed length sprint” and “variable length sprint” using Table 3 (Table 7).

Using Eq. (3),

$$T_{\min} = R1_{\min} + R2_{\min} + R3_{\min} + R4_{\min} + R5_{\min} \tag{3}$$

$$T_{\min} = 10 + 15 + 10 + 15 + 20 = 70 \text{ days}$$

Using Eq. (4),

$$T_{\max} = R1_{\max} + R2_{\max} + R3_{\max} + R4_{\max} + R5_{\max} \tag{4}$$

$$T_{\max} = 12 + 17 + 12 + 17 + 22 = 80 \text{ days}$$

Using Eq. (5),

$$\text{Estimated Project Delivery Time Interval} = T_{\min} - T_{\max} \tag{5}$$

$$\text{Estimated Project Delivery Time Interval} = 70 - 80 \text{ days}$$

Using Eq. (6),

$$\text{Estimated Variation in Project Delivery} = T_{\max} - T_{\min} \tag{6}$$

Table 8 List of observations made during the project execution

Property	Fixed length sprint	Variable length sprint	Feature as length of sprint
Sprint length	Equal for all sprints	Different for at least one sprint	Is dependent on feature
Sprint length expressed in	Exact number of days	Exact number of days	In range
Number of sprints	Not known at the beginning of the project	Not known at the beginning of the project	Known in advance at the beginning of the project
Sprint output	Some piece of functionality	Some piece of functionality	At least one feature
Delivery tracking	Complex	Complex	Simplified
On time delivery	No	No	Yes
Cost estimation of feature	Complex	Complex	Simplified
Ability of end user to understand sprint output	Low	Low	High
Quality of end user feedback after every sprint	Low	Low	High
Customer satisfaction	Low	Low	High

$$\text{Estimated Variation in Project Delivery} = 80 - 70 = 10 \text{ days}$$

Now, let us analyze the delivery at the end of every sprint:

- **At the end of sprint 1:** Delivery *D1* consists feature *F1*.
- **At the end of sprint 2:** Delivery *D2* consists feature *F2*.
- **At the end of sprint 3:** Delivery *D3* consists feature *F3*.
- **At the end of sprint 4:** Delivery *D4* consists feature *F4*.
- **At the end of sprint 5:** Delivery *D5* consists feature *F5*.

Here, the output of a sprint is clearly a feature which can be easily developed, tested, and analyzed by the end user. Hence, giving an opportunity to the end user to evaluate a feature as a unit and provide quality feedback to the AGILE team. Also, the project completed on time without any delivery tracking issues.

5 Observations

AGILE product and process metrics were employed during research which included burn down chart(s) and defect trend chart(s) for every sprint. Apart from this, the minutes of meeting from sprint review and sprint retrospective were carefully studied to arrive at below observations (Table 8).

6 Comparison of “Time” Versus “Feature” as Length of Sprint

This section lists the advantages of the proposed approach over the current approach and vice versa.

6.1 *Advantages of Feature Over Time as Unit of Sprint*

- Customer can evaluate the output of every sprint and can easily track the progress made on the product delivery.
- Feature specific customer feedback adds value to the quality of software product being developed.
- Any change in customer requirement can be adopted without much rework since one complete feature is being developed at appoint of time.
- It makes easy for the customer to estimate the development cost of every feature.
- Customer may choose the order of feature to be delivered by AGILE project team.
- The approach is customer friendly.
- Length of the sprint can be estimated in terms of range taking reference from similar historic projects, and brings this closer to real-world practices.

6.2 *Advantages of Time Over Feature as Unit of Sprint*

- Using time as unit of sprint, one can ensure that the sprint will come to an end after a fixed number of days.
- Length of the sprint is known in advance.
- Length of sprint is uniform across the project life cycle; therefore, managing a sprint and its associated ceremonies is quite easier.
- Cost is uniformly distributed across every sprint.

In spite of the above advantages, the key disadvantage of time over feature is even when the sprint has a definite end time it is not necessarily the output of the sprint is usable for the end user.

7 Conclusion

Throughout the research paper, the usage of “time-box” as length of sprint was examined by comparing it with newly proposed approach of using feature as length of sprint. After analyzing all the results it was concluded that “feature” is more efficient

when used to define the length of sprint as compared to “time-box” approach. The proposed approach will benefit the industry scenarios where the project requirement can be broken down into several independent features for delivery purpose. The industry scenarios where the features are very closely coupled the proposed approach may be difficult to employ. The proposed approach will help the end user to understand how much cost has been invested on every feature of the project and will make the tracking of project quite simplified for scrum master. That said, alternate hypothesis H_1 can be rejected in favor of H_0 .

8 Way Forward

Taking this concept further, a feature can be sized as small, medium, and large. It is worth exploring how using a feature can simplify the costing model for a sprint, since every feature can be priced differently depending upon the efforts required. Estimating the project cost will be much simplified since the end user will now know how much project budget is spent on every feature. The pricing model-based upon the feature will be explored in my future research work.

References

1. Cohn, M.: AGILE Estimating And Planning Is The Definitive, Practical Guide To Estimating And Planning AGILE Projects, AGILE Alliance Cofounder Mike Cohn Discusses The Philosophy Of AGILE Estimating [2] (2005)
2. AGILE Alliance “AGILE Practice Guide” [1] (2017)
3. Gangji, A., Hartman, B.: Agile SCRUM for Denver Web Development. Neon Rain Interactive. Retrieved September 25 (2015)
4. Dhir, S., Deepak, K., Singh, V.B.: An estimation technique in agile archetype using story points and function point analysis. *Int. J. Process. Manag. Benchmarking* 7(4), 518–539 (2017). *Inder-science*. ISSN online: 1741-816X, ISSN print: 1460-6739
5. Cole, R., Scotcher, E.: Brilliant AGILE Project Management [3] (2016)

A Reliable Novel Framework of User-Oriented Software Engineering



Gurpreet Singh Saini, Sanjay Kumar Dubey and Sunil Kumar Bharti

Abstract The reliability is one of the prominent factors for quality determination in any of the projects conducted. It has become a prime focus for the software developers and the respective teams to keep reliability as a goal of each completed module. There are some appropriate models that successfully produce intended results as well as maintain reliability in terms of functional and non-functional features. However, while achieving such results, they have surpassed the project cost estimations and even the development time. Still, almost 80% of software projects fail to complete or either miss their deadlines. Reliability is one great need of present and in this paper, we present a reliable framework for user-oriented software engineering which is based upon a hybrid approach developed with the help of fuzzy logic and the dynamic networks. This will result in reliable, fast and an early product out of the software development life cycle.

Keywords Resource allocation · Fuzzy logic · Soft computing · Reliability · Software quality · Software engineering

1 Introduction

The products developed today through engineering methodologies are curated with high cost and longer development durations; even then the maintenance and implementation costs are going higher day by day as evident from the reports by SIRRUSH Corporation [1]. These extended deadlines not only incur higher costs but also end

G. S. Saini (✉) · S. K. Dubey
Amity University, Noida, Uttar Pradesh, India
e-mail: g.saini4888@live.com

S. K. Dubey
e-mail: skdubey1@amity.edu

S. K. Bharti
Central University of Haryana, Mahendergarh, India
e-mail: sunilbharti@cuh.ac.in

up producing an unreliable product out of the development stable which needs constant monitoring. Reliability has always been kept at the top in quality checking process as evident by the famous quality models by McCall [2], Boehm [3], ISO [4] and CMMI [5]. Thus, reliability has one prime role in establishing the fact whether the development process followed is sufficient enough to produce a quality product which is defect free and will incur minimum maintenance cost. Though reliability can only be judged once the product is developed fully, however with right practices at the beginning, we can ensure both time and cost are saved, and a reliable product is delivered to client side. To answer these issues, multiple models were suggested and developed. However, every model failed in answering the exact timing for the fault detection and removal. The Musa model [6] has some assumptions, which lays down the scope of improving the reliability in the models to be developed in future, and the same is presented in this research paper. The paper provides an insight into the reliability feature of the model proposed by Saini et al. [7].

1.1 Literature Background

In accordance with the literature review [6–9], there are certain factors that affect the software reliability such as lack of resources, insufficient executive support, less involvement of user in the project specification, changing and unrealistic user expectations from the product, and it can be easily understood from Fig. 1. Thus, reliability measurement has become one major metric in understanding the quality of product delivered to the end user.

Reliability has been answered by as many as 40 models [10] till now under software reliability growth models and many more are in development. Reliability is one feature which lays down the foundation for quality model to be developed upon as it measures the functionality layer of the software development life cycle as evident in Fig. 2.

The basic formula for calculating the software reliability is stated as

$$\text{MTBF} = \text{MTTF} + \text{MTTR} \quad (1)$$

MTBF is mean time between failures

MTTF is mean time to fail

MTTR is mean time to repair

2 Development

While developing the novel framework for user-oriented software development model, reliability was kept as one key criterion for development of the algorithm

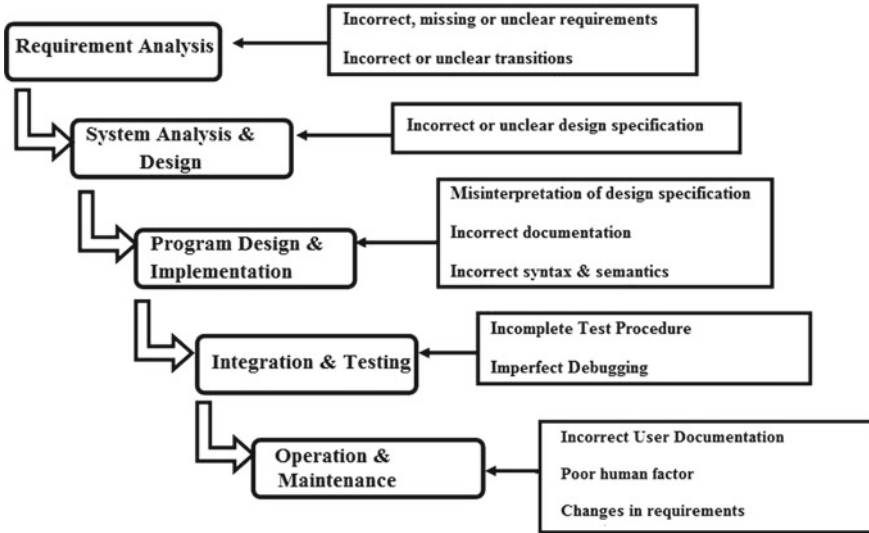


Fig. 1 Reasons for fault occurrence at different levels of SDLC

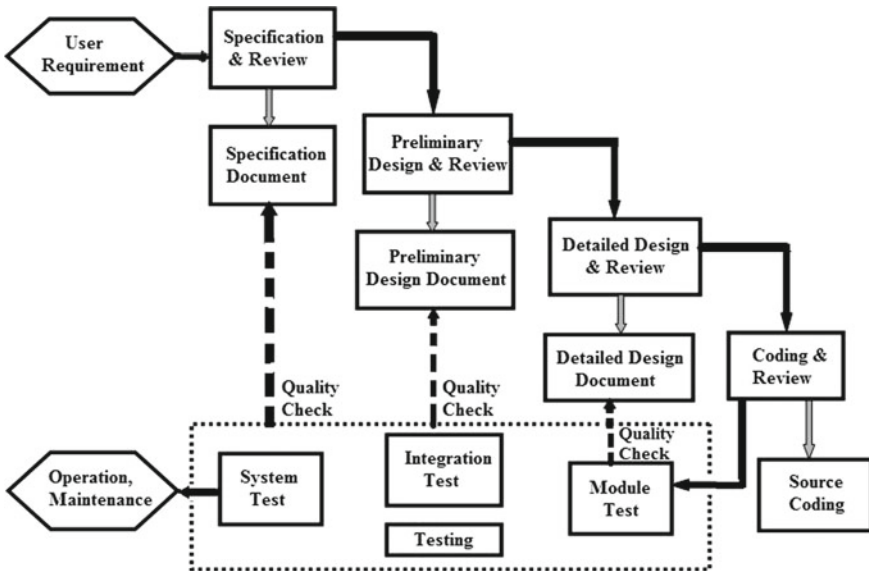


Fig. 2 Reliability verification at different levels of SDLC

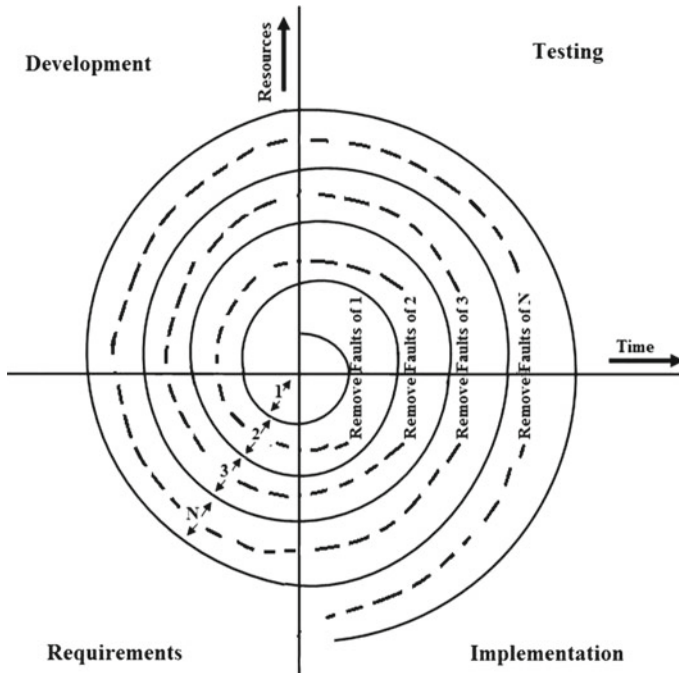


Fig. 3 Strategy of novel framework in improving the reliability

in Saini et al. [7]. The algorithm directly focused upon making the process efficient enough in answering the issues as stated in Fig. 1 with the implementation of the fuzzy logics and dynamic graphs. The hybrid approach of fuzzy logic together with dynamic graph theory for the division of requirements, their prioritization and forming dynamic graphs makes development process efficient steering towards project success. The generated dynamic graphs help in providing a direction to the development cycles and give early and reliable deliverables. Efficient planning was implemented in the algorithm to develop a procedure which could yield an early product with limited functionality and later on giving iterations to it for further functionalities as per prioritizations generated for the user stated requirements. These early developed components in smaller time span contribute towards the success of a project as it results in evolutionary design and development of software components. It involves users at early stage with the prototypes developed so that their expectations can be set realistically.

The novel framework could reduce the chances of failure and generate a product which will have less faults and a specific time-bound development cycle. This statement could be understood using Fig. 3 wherein the fault specific to one iteration is just answered before moving to the very next iteration or parallel to the development of the second iteration (Fig. 4).

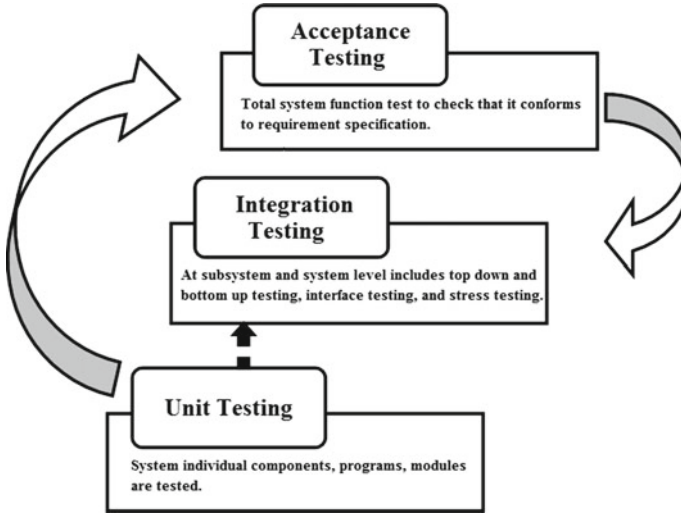


Fig. 4 Testing methodology used in novel framework

Also, it must be noted that the HARD and MUST requirements as marked in the novel framework algorithm must be tested at each iteration thoroughly as these requirements will hold or delay the future work and incur cost if not tested properly. In test case development, the focus must be clear to test all the units first followed by the module and finally for the first iteration product to be delivered to the end user. These iteration cycles must be followed until the final product is delivered to the client hence reducing the failure rate to be minimum as per the no. of iterations. Also, the reliability increases up to nth rate as the failures are detected early and hence reducing the MTBF up to 1/n factor.

Hence, Eq. 1 could be rewritten in the following manner:

$$MTBF_{new} = 1/n(MTTF + MTTR) \tag{2}$$

It could be noted that Eq. 2 holds true for the faults detected at the end of the software development cycle. However, with this user-oriented software development framework into action the faults occur at the following rate:

First cycle faults + second cycle faults + third cycle faults + ..., Nth cycle faults = Total no. of faults/No. of iterations

$$A/n + B/n + C/n + D/n \dots, +N/n = N/n \tag{3}$$

where A, B, C, D, etc., are no. of faults every iteration and n is the iteration no.

Hence, practically the no. of faults left to be answered at the END of the development cycle remains to be N which is only result of the requirements for the Module of nth iteration.

3 Result and Analysis

The results withdrawn using the above model show less no. of faults to be answered after the complete product is delivered to the user end. However, while computing the complete result few assumptions are made as follows:

- a. Once iteration is over and product is delivered to the user, its relative faults are detected and solved before the next iteration is delivered.
- b. All the iterations are independent modules with less coupling and more cohesion.
- c. The faults are of generic coding nature. The model assumes that the requisites are scheduled properly using the novel algorithm of resource allocation and prioritization.
- d. Also, the modules being worked upon are being developed parallel with different timelines. The one with less time period is delivered as first iteration followed by the second iteration timeline and so on.
- e. The data used for calculation is based on the above assumptions.

The algorithm focuses on reliability as one of the key points in improving the quality of the previous models [7, 11–13]. Table 1 shows the comparison between the faults generated in the novel framework and also the standard models of development. It clearly shows that the faults are discovered and answered at an early phase in the novel framework.

As the faults are discovered early in the development cycle, so is the case with their solving strategy. The rectification of the faults detected in the first iteration is solved while the second iteration development period and same is with the faults occurring in the following iterations. If compared, with the std. model, the no. of faults to be answered is very less as shown in Table 2.

Let us assume the time taken to answer (to resolve) one fault is \times . Hence while solving the complete faults, $23\times$ time is consumed. However, if the timeline of fault occurrence and solving is taken into consideration, the novel framework produces an early model with less faults at the end to be taken care of as visible in Table 2. Finally, the novel framework will take $2\times$ time to produce a quality product with

Table 1 Time of faults generated: novel versus std.

Novel framework		Std. framework	
Iteration	<i>N</i>	Iteration	<i>N</i>
1	2	1	0
2	4		0
3	3		0
4	5		0
5	4		0
6	3		0
7	2		23

Table 2 Faults left to be answered at the end of the complete development cycle

Novel framework			Std. framework		
Iteration	N	N_{solved}	Iteration	N	N_{solved}
1	2	0	1	0	0
2	4	2		0	0
3	3	4		0	0
4	5	3		0	0
5	4	5		0	0
6	3	4		0	0
7	2	3		23	0
Faults left to be answered		2	Faults left to be answered		23

less cost as compared to $23 \times$ time taken by the std. framework incurring higher cost and less quality.

Finally, the novel framework will take $2 \times$ time to produce a quality product with less cost as compared to $23 \times$ time taken by the std. framework incurring higher cost and less quality.

4 Results

As stated in Eq. 1, the MTBF for the std. model of development will take $23 \times$ time and while using the novel framework the time period reduces to $1/n$ of the original time schedule and reliability increases by n times.

Hence, the $MTBF_{new}$ is

$$\begin{aligned}
 MTBF_{new} &= 1/n(23 \times) \\
 &= 1/7(23 \times) \\
 &= 3.28 \times
 \end{aligned}$$

This result directly relates to the very fact that the MTBF has gone up by 3.28 times. This also means that the time period of finding the faults and its rectification and isolation has reduced up to 3.28 times and directly verifies the fact that novel framework produces the products which are reliable and follow better quality standards taking into consideration of user.

The results show a viable improvement over the standard models and provide a novel approach for the ideal workflow in the development domain.

5 Outcomes

As shown in the results, the novel framework is reliable, steady and has one feature of isolating the MUST and HARD [7] tasks from holding up the future work. It has more focus on developing independent modules with specific functionality and hence results in reusable modules. These reusable modules will lay down path for shorter development cycles of future projects and will be a long-term gain in terms of costing and resource availability.

The framework is able to answer the delays in terms of resources and planning; however, it does not take into consideration the human factors and also the unavailability of the resources which could be a major hindrance in the product development cycle.

6 Future Scope

A framework is developed from the proposed algorithm which can be used in various environments such as software project management, operating system, construction models, etc. The future action that may be taken into account is developing a framework which is specific to the environment by doing changes or converting this generalized framework so that it could become a real-time project management solution to the problem whose nature can be clearly identified.

References

1. Schmidt, R.: Software engineering: architecture-driven development. In: NDIA 15th Annual Systems Engineering Conference, October (2012)
2. McCall, J.A., Richards, P.K., Walters, G.F.: Factors in software quality. Nat'l Tech. Information Service, vols. 1, 2 and 3 (1977)
3. Boehm, B., Brown, J.R., Kaspar, H., Lipow, M., McLeod, G., Merritt, M.: Characteristics of software quality. North Holland (1978)
4. ISO, International Organization for Standardization, ISO 9000:2000, Quality management systems—Fundamentals and vocabulary (2000)
5. Curtis, B., Hefley, B., Miller, S.: People capability maturity model® (P-CMM®), Version 2.0. Software Engineering Institute, Carnegie Mellon University (2001)
6. Kapur, P.K., Pham, H., Gupta, A., Jha, P.C.: Software reliability assessment with OR applications. Springer Series in Reliability Engineering (2011)
7. Saini, G.S., Dubey, S.K., Bharti, S.K.: Fuzzy based algorithm for resource allocation. Advances in intelligent systems and computing, Springer. In: Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, FICTA 2016, vol. 1–515, pp. 69–78 (2016)
8. Kelly, D.: Scientific software development viewed as knowledge acquisition: towards understanding the development of risk-averse scientific software. *J. Syst. Softw.* **109**, 50–61 (2015). Elsevier

9. Chong, C.Y., Lee, S.P.: Analyzing maintainability and reliability of object oriented software using weighted complex network. *J. Syst. Softw.* **110**, 28–53 (2015). Elsevier
10. Tyagi, K., Sharma, A.: A rule-based approach for estimating the reliability of component-based systems. *Adv. Eng. Softw.* **54**, 24–29 (2012). Elsevier
11. Jifeng, H., Li, X., Liu, Z.: Component-based software engineering-the need to link methods and their theories. 973 project 2002CB312001 of the Ministry of Science and Technology of China
12. Boehm, B.: Value-based software engineering: overview and agenda. USC-CSE-2005-504 (2005)
13. Dybå, T., Kitchenham, B.A., Jørgensen, M.: Evidence-based software engineering for practitioners. *IEEE Software*, Published by the IEEE Computer Society. January–February (2005)

Android-Based Blind Learning Application



Abhishek Ranjan and T. M. Navamani

Abstract There are lot of people in this world who want to study but they cannot due to some or the other reasons. Blind people come under this category, but they did not build these reasons themselves. Some children are visually challenged by birth. To overcome these issues, we have designed an android-based application named Blind Learning APP (BL-APP). This application (App) will assist learning processes for blind with an easy-to-use interface and a number of inbuilt learning materials. Interface of the app is designed in such a way that anyone can figure where the buttons are lying on the screen and tap them easily. All the functionalities will have the text description associated with it, which will be conveyed to users using text-to-speech mechanism. This app also provides vibratory feedback to the user, with this they can visualise the shapes and learn about it in a better format.

Keywords Blind learning · Visually impaired · Text to speech · Vibratory feedback · Virtual assistant · XML styling

1 Introduction

Blind people are facing many visual challenges every day from studying an article to reading a label, they even find it difficult to figure out if they are at a correct place and many more. In order to mitigate these problems many tools have been developed which uses Image processing, computer vision and other sensors [1]. In the present era of touch screen, mobile applications can be accessed very easily. If we search on Play Store or App Store there are a number of apps to make sighted people to learn [2–4] but there is no application for teaching visually impaired people. Hence in this paper we have addressed this issue by proposing an android application specifically

A. Ranjan (✉) · T. M. Navamani
School of Computer Engineering, Vellore Institute of Technology, Vellore, Tamil Nadu, India
e-mail: abhishek.ranjan2690@gmail.com

T. M. Navamani
e-mail: navamani.tm@vit.ac.in

for visually impaired people for learning purpose. Although applications designed for sighted ones can be used by blind using inbuilt feature of Google talkback, but these applications do not provide the feature of learning.

We have developed a learning application for visually impaired persons so that they can learn thing easily without using braille script. This application is made up of clean and simple interface along with large buttons which adds to the ease of accessibility for visually impaired. It also gives a new way of experiencing different shapes using vibratory feedback.

Featured phones have small-raised dot on “number five” which gives the relative position of other keys but in case of touchscreen phone we lag this component. Evenness of screen makes it difficult for users to identify position of their fingers. To overcome this, we have focused on speech on touch rather than on speech on tap. Many people use Google assistant or Apple Siri to communicate with devices using voice commands. This text-to-speech converter sometimes mess up things and make it more complicated even for a sighted people then how can blind people be benefitted from it? The apps currently available on app stores focuses on guiding person through their home or workplace, reading messages, making phone calls and converting text of inbuilt applications into speech. Market apps provide blind people way to live their life but our proposed application will make their lives better, as it opens the horizon of learning and earning both. This system can be used on a large scale for teaching visually impaired people in an interactive manner.

The rest of the paper is organised as follows. Section 2 discusses about the related work done in this field followed by Sect. 3 which gives complete overview of system. Implementation details are shown in Sect. 4. The paper is summarised in Sect. 5 with conclusion and future scope.

2 Related Work

In [5], Thakurde developed an application in which people can detect obstacle on their way while walking from one room to another or one place to another, room to room walk is facilitated using camera and one location to other location navigation is done using camera along with Global positioning system (GPS) technology. We cannot trust GPS technology perfectly as blind people are unaware of surrounding and GPS does not give the surround view of the place.

Ghosalkar et al. [6] have developed an android-based application on examination using speech technology for blind. This application uses text-to-speech technology and makes it easier for the examinee to answer the competitive examination questions. It collects response from user via speech mechanism. But this application is limited only to examination not for learning purposes.

An inimitable android application developed by Pampatiwar [7] for blind to interact completely with mobile system using text-to-speech method. It provides an easier interface to configure settings of interaction. But nowadays these features are inbuilt in android talkback. Smartphone accessibility application for visually

impaired, which is discussed in [8] an innovative idea used by Baravkar et al. Users need to configure gesture settings in this App to control different things using those gesture. User here are bound to learn gestures which do not have any standards.

Android text messaging application for visually impaired people, developed by Thyagaraju et al. [9] makes user send and receive text messages, place and receive phone calls through voice commands. This application is just limited to only messages and phone calls. So, users need more functionalities beside messages and phone calls. iSee is an Android-based application proposed by Ghantous et al. [10] benefits visually impaired to perform their day-to-day activities easily. A single screen tap in iSee provides virtual eye for sense of seeing to the blind person by audibly communicating the object(s) names and description.

The non-availability of application in the learning domain of blind people justifies the need to develop an application which can make blind people learn and prosper in their life. BL-App does not require Internet connectivity which makes it anytime learning application.

3 Blind Learning Application (BL-APP)

3.1 Overview and Assumptions

The BL-APP is made up of simple user screen which shows topics available for learning. It notifies user through when they voice comments when they hover their fingers above the buttons. Voice output while hovering will give clear information of elements present on the screen.

The application interface contains large size buttons, so blind users can use it without messing with other things on the screen. Application database contains pre-defined text information about the topics and study material included. When user clicks play button, text will be spoken using Google text-to-speech feature. Back button placed on each page of application assist users to go back easily. When user selects certain topic, information regarding that topic will be spelled-out.

One of the key features of this application is to provide vibratory feedback when user touches the outline of any image, present on the app screen. Using this vibratory feedback, we can visualise the shapes like square and circle easily. This method makes visualisation of shapes easy for the people who are blind by birth and it can be used on large scale for teaching visually impaired people in the efficient and interactive manner. Users can learn things in better format as this will give proper interface for them.

BL-App is developed using android studio and limited to android operating system only. For the first-time users handling, functionalities and working of application needs to be conveyed in an unambiguous manner. Along with this users should have some basic knowledge of using smartphones.

3.2 Applications of Proposed System

- This system can be used on large scale for teaching visually impaired people in an efficient and interactive manner.
- Users can learn things in better manner as this provides proper interface to them.
- They can visualise the shapes with the help of vibratory feedback given to them while hovering over figures.
- Visually challenged people will come to know how the shapes are if they are blind by birth.
- Sighted people can also use this app for learning purpose.

3.3 System Architecture

This app is built on Android Studio version 2.3.3 with apk 19, i.e. it will support android devices running operating system (OS) greater than 4.4 (kitkat). We have used Java as a logical language and XML as styling or designing language, text-to-speech module for conveying predefined text message and bitmap get pixel for obtaining pixel value of images. The system architecture of the proposed system is shown in Fig. 1.

This diagram (Fig. 1) shows the components used for conversion of text to speech, ways of getting input and providing output to the users. Here we have used text-to-speech module (TTS) provided by android studio for conversion of text data to speech. OnTouch event listener [11] is used to detect movement of fingers over screen. Vibratory Sensors are responsible for shaking the device when touched. User in this figure is representing people who are going to interact with this system.



Fig. 1 System architecture of blind learning application

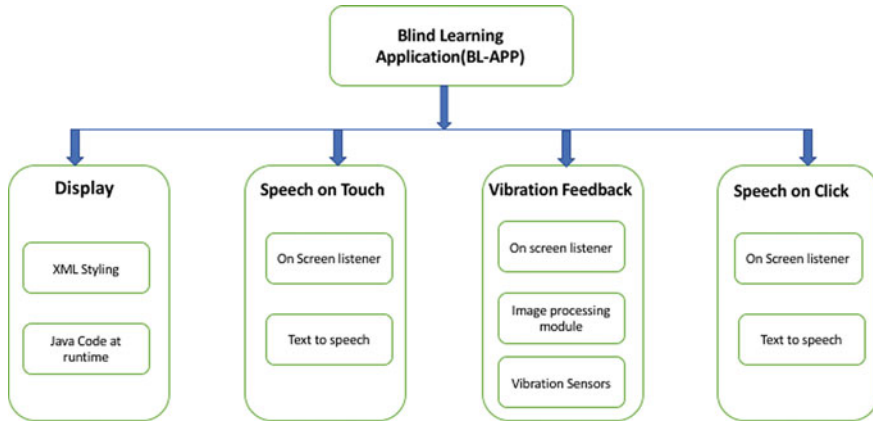


Fig. 2 Modular diagram of blind learning application

Modular Description

The functionalities and the modules of the application are discussed as follows (Fig. 2).

Blind learning application: This app helps blind people learn and recognise shapes using vibratory feedback. Applications have topics like learning numbers, shapes and alphabets.

Display: This module takes care of interaction with user and governs rules of display on screen. Although visually challenged people have nothing to do with the appearance of text or icons on the screen but they are concerned about the placement and size of the icons on screen.

Xml: Xml is used as the styling language in this app to design interface. XML enlists a set of rules for encoding documents that is both human-readable and machine-readable. Xml provide means for styling our icons and specifying the dimensions of buttons in this application.

Java: Java code is used to change content on the screen at runtime and interact with user according to their command. Be it is going back to previous page or jumping from one page to another, intents are created using Java.

Speech on Touch: This module guides user to know what is onscreen when person hover over screen. This functionality is implemented using On Screen Lister and google text-to-speech mechanism.

On Screen Listener: It keeps track of activity like touch on screen. In this case it keeps record of time at which it is touched on particular location. We used motion events to track movements in terms of an action code and a set of axis values. The action code provides the state change that occurred such as pointing down or up [11, 12].

Text to speech: This is the inbuilt feature of android phone which converts pre-defined text to speech for output [13]. We have used “TextToSpeech tts = new Text-

ToSpeech(this, this)” for creating new text-to-speech method and “tts.speak(“this is the topic page”, TextToSpeech.QUEUE_ADD, null);” for specifying our text to be spoken.

Vibratory module: This hardware module triggers vibration while touching the figure outline. getSystemService() method is used which returns a reference to a system service object. “context.VIBRATOR_OBJECT” is passed as parameter to get hold of vibration handle [14].

Image processing module: This module dissolves image according to pixel value and gives output for further processing. bitmap.getPixels() method is used which returns integer pixel colour value of the image outline. While moving our fingers on the screen, if pixel colour value exceeds threshold value (here 220) then device vibrates.

Speech on Click: This module gives speech output when clicked on button.

On screen listener: Keeps track of activity like touch on screen. In this case it will act on tapping the button. An event listener is method which is used to interface in the android view class which containing a single call back method. These methods are called by the Android framework when the View mapped to this is clicked.

4 Implementation Details

4.1 Snapshots of Working Application

This app is built on Android Studio version 2.3.3 with apk 19, i.e. it will support android devices running OS greater than 4.4 (kitkat). We have used java as a logical language and XML as styling or designing language, text-to-speech module for conveying predefined text message and bitmap get pixel for obtaining pixel value of images [15].

Figure 3a shows the application home page containing topics like alphabets, numbers, basic shapes and 3-D shapes. Figure 3b shows sub-topic of basic shapes whereas Fig. 3c, d show tetrahedron and circle respectively which can be felt with vibratory feedback when user touches perimeter of shape.

Home screen: On home screen we have created large buttons for the accessibility of the user. It shows the topics for learning such as alphabets, numbers, basic shapes and 3D shapes. We can further add to the list of topics. We have created home screen topic boxes using text view and have assigned width as that of parent view.

Topic page: On topic page this application provides list of sub topics which are available for example in case of basic shapes we have circle, square, rectangle and triangle. When the user clicks on any of the subtopics then they are taken to the page which contains detailed explanation of selected topic along with the shape. We have used **public boolean** on TouchEvent (MotionEvent event) function to detect touch or tap from the user and have used **switch** (action) to act as per user activity. If user touches any text view button the switch (action) intents to TTS module and play the

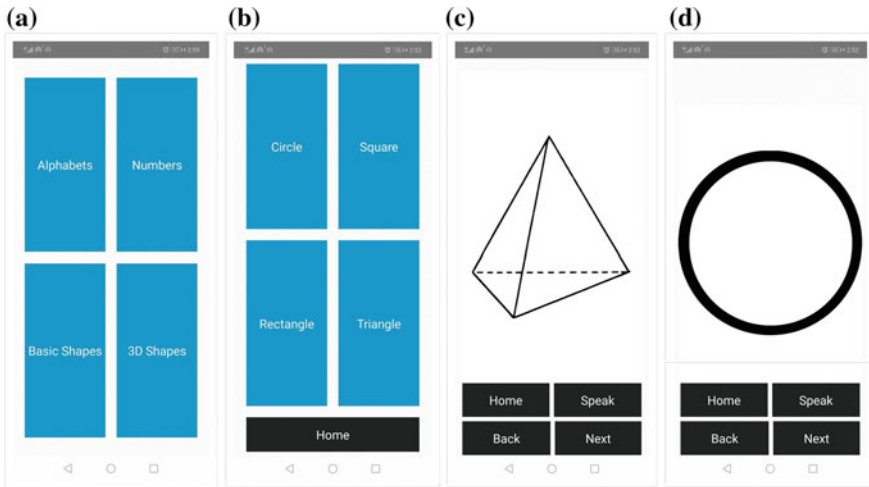


Fig. 3 Screenshots of working application

text assigned for that button and if they tap any button then switch (action) opens the page which is linked to the button.

Topic content page: This page provides the learning content for the topic selected by the user. When the user taps speak button then all the contents related to the topic is played using tts.speak function. We have kept the images of square, circle, rectangle in the drawable folder of application and it is linked with this page. We have also provided the back, next and home buttons on each page which eases the navigation from one page to another.

4.2 Analysis on Scope of Deployment

There are a number of blind people in the world. Many of them are completely blind from birth. To help these kind of people, and make them self-dependent our Blind Learning Application have great scope of deployment.

The pie chart shown in Fig. 4 represents that 253 million people are estimated to be visually impaired worldwide, 36 million people are blind and 217 people have low vision.

Along with these it is estimated that 19 million children are visually impaired. Out of 19 million children 12 million have a vision impairment due to refractive error. These statistics give us a clear view that there are large number of people who are visually impaired and we need another way of education for them.

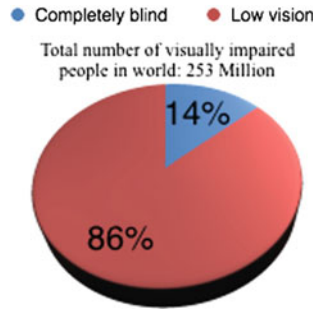


Fig. 4 Visually impaired people in world. *Source* World Health Organisation (WHO), October 2017

5 Conclusion and Future Work

In this paper, we have proposed a system, i.e. Blind Learning Application for the blind/visually impaired people. Earlier we had applications for assisting blind people for making phone calls, detecting obstacles in their path, appearing in examination using voice commands, our novel application for teaching visually challenged people addresses one more problem faced by them. We have also seen that text-to-speech feature in the applications designed for the sighted person creates confusion, which incited us to design an application specially for visually impaired persons using some basic concepts.

This application can be extended for taking a picture of text articles and converting it into voice message. One of the major problems which occurs with visually challenged people while capturing images for voice conversion is that they are unaware of their action whether they are capturing whole image or partial part of it. If they are missing some part of image, then they do not get complete information. So, functionality of notifying the users whether they are capturing whole image or not can be added to this application and it will be of great use.

References

1. Brady, E., Morris, M.R., Zhong, Y., White, S., Bigam, J.P.: Visual challenges in the everyday lives of blind people. Paris, France. Copyright 2013 ACM 978-1-4503-1899-0/13/04, 27 Apr–2 May 2013
2. Ramiah, S., Liong, T.Y.: Detecting text based image with optical character recognition for English translation and speech using Android. In: 2015 IEEE Student Conference on Research and Development (SCoReD). <https://doi.org/10.1109/scored.2015.7449339>
3. Blind Communicator App, Google Play (2016), Deaf-Blind Communicator App Google Play (2015). <https://play.google.com/store/apps/details?id=ar.com.lrusso.blindcommunicator>
4. Deaf-Blind Communicator App Google Play (2015), https://play.google.com/store/apps/details?id=appinventor.ai_lberis3.DeafBlindCommunicator

5. Tharkude, K.B.: Android application for blind people based on object detection. *4*(4) (2016). ISSN (Online): 2320-9801
6. Ghosalkar, S., Pandey, S., Padhra, S., Apte, T.: Smart, android application on examination using speech technology for blind people. *Int. J. Res. Comput. Commun. Technol.* *3*(3) (2014). ISSN (Online) 2278-5841
7. Pampattiwar, S.R.: Smartphone accessibility application for visually impaired. *Int. J. Res. Advent Technol.* *2*(4), 377–380 (2014). E-ISSN: 2321-9637
8. Baravkar, S.R., Borde, M.R.: Android text messaging application for visually impaired people. *IRACST—Eng. Sci. Technol. Int. J. (ESTIJ)* *3*(1), 58–61 (2013). ISSN: 2250-3498
9. Thyagaraju, G.S., Sowmya, N.K., Nithin N.: An Inimitable application for blind using android. *Int. J. Emerg. Trend Eng. Basic Sci. (IJEEBS)* *2*(1) (2015). ISSN (Online) 2349-6967
10. Ghantous, M., Nahas, M., Ghamloush, M., Rida, M.: *iSee: an android application for the assistance of the visually impaired*, Springer international publishing, Switzerland (2014)
11. Responding to Touch Events, Android Developers: <https://developer.android.com/training/graphics/opengl/touch.html>
12. Text to speech, Android Developers: <https://developer.android.com/reference/android/speech/tts/TextToSpeech.html>
13. Sharma, D., Kanwar, R.: Text to speech conversion with language translator under android environment. *Int. J. Emerg. Res. Manage. Technol.* *4*(6). ISSN: 2278-9359
14. How to make android device vibrate, Stack Overflow: <https://stackoverflow.com/questions/13950338/how-to-make-an-android-device-vibrate>
15. Blind Communicator App, Google Play (2016). <https://play.google.com/store/apps/details?id=ar.com.lrusso.blindcommunicator>

An Approach for Test Case Prioritization Using Harmony Search for Aspect-Oriented Software Systems



Abhishek Singhal, Abhay Bansal and Avadhesh Kumar

Abstract Regression testing is important part of testing during software maintenance. It ensures error-free software after modification during maintenance. Without any priority, execution of test cases is not cost-effective and time-consuming. Therefore, it is highly desirable to prioritize the test cases to achieve maximum fault coverage. In this study, a test case prioritization approach using harmony search for aspect-oriented software systems is proposed. In this paper, we have taken two objective functions such as minimum execution time and maximum fault coverage. Further, average percentage fault detection (APFD) metric was used to validate the results. Further, results are compared with random prioritization and no prioritization. The results indicate that proposed approach is performing well.

Keywords Regression testing · Software maintenance · Fault coverage · Aspect-oriented software systems · Execution time · Average percentage fault detection

1 Introduction

Regression testing is utmost important but time-consuming task under software maintenance. This task consumes a huge amount of time and cost of entire software development life cycle. The present era is very competitive and demands more reliable, cost-effective optimal time span. The ever-changing nature of customer requirements makes the software maintenance as well as regression testing more challenging [1].

A. Singhal (✉) · A. Bansal
Department of Computer Science & Engineering, ASET, Amity University Uttar Pradesh,
Noida, India
e-mail: abhi087@gmail.com

A. Kumar
School of Computer Science & Engineering, Galgotias University Uttar Pradesh,
Greater Noida, India

Regression testing is rerunning all the existing test cases of previously executable version to ensure the functionality of updated version.

Aspect-oriented programming (AOP) paradigm [2] is relatively new programming paradigm of software development, which overcomes the limitations of object-oriented software development. AOP implements cross-cutting concerns through aspect and classes. In AOP development, during regression testing, the changes are in either aspect or class or both and weaving of classes and aspects, and hence it is very costly and time-taking task to test the software again. Regression testing ensures error-free software after the changes due to technological changes and changes in customer requirements. Regression test case prioritization is an approach to decide the priority of available test cases based on some coverage criteria such as code, statement, branch, function coverage [3, 4], etc. It arranges the test cases in such a way that fault detection rate can be accelerated. Literature revealed that regression test case prioritization is an NP-hard optimization problem. Various researchers have applied various heuristic and metaheuristics approach such as genetic algorithm (GA) [5], ant colony optimization (ACO) [6], particle swarm optimization (PSO) [7], and artificial bee colony optimization (ABC) [4] to address the problem. But the scope of optimization still persists.

In this paper, we have proposed regression test case prioritization technique using harmony search for aspect-oriented software. The average percentage fault detection (APFD) metric approach is utilized to validate the performance of proposed algorithm on benchmarked test data. The rest of the paper is as follows: Section 2 discusses the related work done in the area of test case prioritization. Section 3 pertains to basics of approaches adapted for this study. Section 4 discusses the proposed algorithm for test case prioritization. Section 5 explains the results obtained after executing proposed approach. Finally, Sect. 6 concludes the proposed algorithm.

2 Related Work

In regression testing, test case selection and prioritization are most promising areas of research. The literature shows that test case prioritization is most important area to be explored.

Vedpal et al. [8] presented hierarchical test case prioritization approach for object-oriented software. They presented their approach on hierarchy at two different levels. First, all classes of given object-oriented program are prioritized, then, second, after obtaining prioritized classes, test cases of these classes are arranged in an order based on fault coverage. APFD metric was used to validate the results and further to calculate effectiveness of proposed approach.

Hla et al. [9] discussed a test case prioritization approach for object-oriented software by utilizing particle swarm optimization to prioritize the test cases to test modified software. They applied PSO to achieve maximum coverage such as statement, branch, function, and code coverage in test case prioritization.

Malhotra et al. [10] discussed a framework-based approach for test case prioritization. They validated their prioritized results with help of proposed metric modified average percentage of block coverage (mAPBC). With the application of this metric, rate of code coverage was computed.

Mahmood et al. [11] proposed a test case prioritization approach based on functionality coverage of a real-life software system. In this approach, maximum functionalities of the software are covered with the help of obtained prioritized test cases.

Gao et al. [12] utilized ant colony optimization metaheuristics for prioritization of test cases. Three objectives such as number of faults detected by test suite, execution time, and severity of fault are identified for prioritization of test cases. They validated their prioritized results with help of APFD metric.

Kaur et al. [13] utilized genetic algorithm for prioritization of test cases. They used a time-constrained execution environment.

Kaur et al. [14] proposed an approach for prioritization of test cases to test modified software using an ABC algorithm to obtain prioritized test cases. They validated their prioritized results with help of average percentage condition coverage metric (APCC).

Suri et al. [15, 16] presented a test case prioritization approach for modified software with the help of ACO algorithm to obtain prioritized test cases. They used objective function based on two factors such as fault coverage and execution time.

Fu et al. [17] proposed an approach for prioritization of test cases for modified based on relationship between program changes and method invocation to obtain prioritized test case. They validated their approach with help of APFD metric.

Luo [18] proposed a test case prioritization approach for object-oriented software using mutation analysis to obtain prioritized test cases on modified software. They validated their prioritized results with help of APFD metric.

Tiwari et al. [19] discussed a fault localization based on spectrum approach, which resolved the limitations of regression testing. Their proposed approach was used to identify the behavior between previous and modified version of the program.

Tiwari et al. [20] utilized a variant of particle swarm optimization with modified time varying acceleration coefficients for creating the new test cases to test modified version of software code due to any changes in requirement or due to any errors in original code. They further compared the proposed approach using benchmark functions to prove the superiority.

3 Harmony Search Technique

Harmony search metaheuristics [21] mimics the musician's improvisation process by adjusting pitches for creative musical process to achieve improved harmony. This optimization technique is a probability-based technique. Harmony search optimization technique utilizes size of harmony memory, consideration rate of harmony memory, the rate of adjustment of pitch, and bandwidth to perform intensification and diversification. The intensification and diversification balancing is the key of per-

formance of every optimization technique. Harmony search approach is successfully utilized in the area of software engineering [22, 23].

4 Proposed Prioritization Algorithm

The proposed prioritized algorithm, which utilizes harmony search technique, is as follows:

Step 1: Designing of objective function consisting main objective: (1) Fault covered by test case should be in descending order and time taken by test case should be in ascending order. These two objectives are considered in on objective function.

$$\text{Objective_Function}(\text{Test_Suite}) = w(\text{Objective1}) + (1-w)(\text{Objective2})$$

Minimize: Objective_Function (Test_Suite)

Here, $TS^i \in TS^1, TS^2, TS^3, \dots, TS^N$, TS contains a permutation of test cases.

Step 2: Generate the random initial harmony memory (HM) solutions of size $[HMS \times N]$ between ranges of $1-N$. Compute the objective function value (HF) for each solution value.

Step 3: Perform improvisation of the solution for each permuted test case in the test suite. For $i = 1$ to N .

- (a) Memory consideration: Select a random test suite from permuted test suite HM with a probability of HMCR.
- (b) Pitch adjustment: If ($\text{rand} < \text{PAR}$) then perform improvisation in selected test suite by selected test case TS' with a probability of PAR.
 - If ($\text{rand} < 0.5$), perform adaptive bandwidth selection on TS' and add to updated UTS (updated test suite)
 - else select a random test case which does not exist in updated test suite and add it to UTS.
- (c) If ($\text{rand} \geq \text{PAR}$), select a random test case, which does not exist in updated test suite and add it to UTS.

Step 4: The worst solution existing in harmony memory will be replaced by new solution in case its objective function value of new solution is better than worst solution.

Step 5: Repeat steps 2, 3, and 4 for updation of harmony memory and improvisation of existing solution until reach to the terminating criteria (Fig. 1).

5 Results and Validation

A benchmark banking problem [24] implemented in AspectJ [25] is considered for validation of approach. We have manually introduced eight faults in code and to

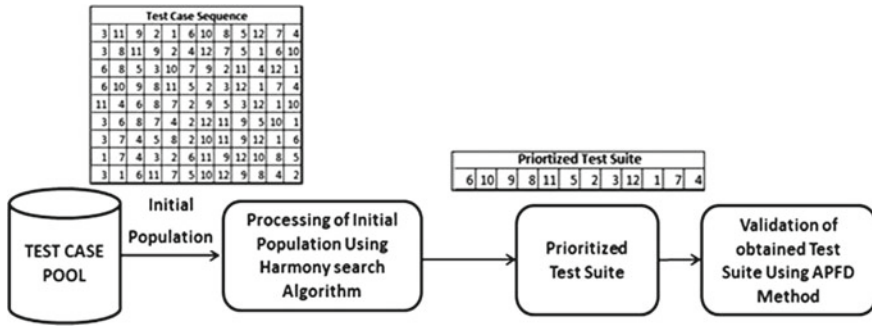


Fig. 1 Proposed test case prioritization approach using harmony search

Table 1 Fault matrix representing eight faults in banking application

Fault/test case	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12
1	0	0	1	0	1	1	0	0	1	0	0	0
2	1	0	0	1	1	1	0	1	0	0	0	0
3	0	0	0	0	0	0	1	0	0	0	0	0
4	1	1	0	0	0	0	0	0	0	1	0	0
5	0	0	1	1	1	0	1	0	0	0	0	0
6	0	0	0	1	0	1	0	1	1	0	1	0
7	1	0	1	0	1	0	1	0	0	1	0	1
8	0	0	0	0	0	0	0	0	0	0	1	1

detect these faults 12 test cases have been generated. The given Table 1 represents the fault matrix where eight faults are listed and each row represents that particular faults can be revealed or not by test cases between T1 to T12.

To analyze the performance as well as solution stability, we have executed the proposed prioritization approach 24 times for selected banking application. Results obtained for these 24 runs are shown in Table 2. Value of objective function represents the solution quality of the approach. Here, each row represents the permuted sequence of test cases in the particular run and its objective value to express the solution quality.

Further, average percentage fault detection (APFD) [26–28] metric is used to validate the proposed approach results. APFD is computed by taking weighted average of the % of faults which was detected during the application of test suite on the given application. Normally, it varies from 0 to 100. For achieving more effectiveness for the proposed approach, higher value of APFD is desired.

$$APFD = 1 - \left\{ \frac{(TP1 + TP2 + TP3 \dots + TPf)}{fc} \right\} + \frac{1}{2c}$$

Table 2 Prioritized test case sequence using proposed approach

Run no./Test case sequence													Value of objective function
1	3	11	9	2	1	6	10	8	5	12	7	4	2
2	3	8	11	9	2	4	12	7	5	1	6	10	2
3	6	8	5	3	10	7	9	2	11	4	12	1	2
4	6	10	9	8	11	5	2	3	12	1	7	4	2
5	11	4	6	8	7	2	9	5	3	12	1	10	2
6	3	6	8	7	4	2	12	11	9	5	10	1	2
7	3	7	4	5	8	2	10	11	9	12	1	6	2
8	1	7	4	3	2	6	11	9	12	10	8	5	3
9	3	1	6	11	7	5	10	12	9	8	4	2	2
10	10	4	5	3	1	6	11	7	2	12	8	9	2
11	6	10	8	12	7	4	3	11	5	2	1	9	2
12	6	11	2	12	1	9	3	8	4	5	10	7	3
13	1	8	5	2	6	7	3	10	11	9	12	4	3
14	3	1	6	10	5	2	9	8	11	4	12	7	3
15	1	11	7	6	8	5	3	12	4	9	10	2	3
16	10	8	11	5	2	3	7	4	9	12	1	6	2
17	6	11	4	5	2	12	8	9	3	7	1	10	2
18	3	12	11	7	9	5	1	8	6	10	4	2	2
19	1	11	9	6	3	10	8	7	4	2	12	5	2
20	3	10	5	6	7	9	2	12	1	8	11	4	2
21	3	12	6	8	9	7	2	1	11	4	10	5	3
22	6	8	7	3	10	9	5	2	4	12	1	11	3
23	10	8	5	7	9	2	3	1	6	11	4	12	3
24	3	8	11	7	5	4	9	12	6	1	10	2	2

Here, in this equation, TP1, TP2, ..., TPf are the position of test case, which detect the fault first in the given prioritized test sequence, f denoted number of faults, and c is representing the number of test cases in the test suite.

Comparison between APFD values applied for proposed approach, random prioritization, and no prioritization are given in Table 3.

Table 3 indicates that proposed harmony-based approach is providing highest APFD in comparison to other methods.

Table 3 Comparative results based on APFD

S. No.	Approach adopted	APFD
1	No prioritization [27]	0.65
2	Random prioritization [27]	0.73
3	Proposed approach	0.77

6 Conclusion

In this paper, we proposed harmony-search-based approach for prioritization of test cases. Proposed approach was compared against no prioritization and random prioritization approaches. The comparative result based on APFD parameter indicates the superiority of proposed prioritization approach. The proposed approach has proven the performance stability on various runs. The major advantage of the proposed approach is to find the optimal balance between test case time and fault coverage. In future, we wish to conduct the experiments under industrial environment with more complex software and large test suites to assess their effectiveness of proposed approach.

References

1. Beizer, B.: Software testing techniques, 2nd edn. Dreamtech Press (2003)
2. Kiczales, G., Hilsdale, E., Hugunin, J., Kersten, M., Palm, J., Griswold, W.: An overview of AspectJ. In: 15th European Conference on Object Oriented programming, Budapest, Hungary (2001)
3. Ahmed, B.S.: Test case minimization approach using fault detection and combinatorial optimization techniques for configuration-aware structural testing. *Int. J. Eng. Sci. Technol.* (2015)
4. Konsaard, P., Lachana, R.: Using artificial bee colony for code coverage based test suite prioritization. In: 2nd International Conference on Information Science and Security (ICISS), pp. 1–4, IEEE (2015)
5. Raju, S., Uma, G.V.: Factors oriented test case prioritization technique in regression testing using genetic algorithm. *Eur. J. Sci. Res.* **74**(3), 389–402 (2012)
6. Solanki, K., Singh Y., Dalal S.: Test case prioritization: an approach based on modified ant colony optimization (m-ACO). In: International Conference on Computer Communication and Control (IC4), pp. 1–6. IEEE (2015)
7. Mohemmed, A.W., Sahoo, N.C., Geok, T.K.: Solving shortest path problem using particle swarm optimization. *Appl. Soft Comput.* **8**(4), 1643–1653 (2008)
8. Vedpal, C.N., Kumar, H.: A hierarchical test case prioritization technique for object oriented software. In: International Conference on Contemporary Computing and Informatics (IC3I), pp. 249–254 (2014)
9. Hla, K.H.S.: Applying particle swarm optimization to prioritizing test cases for embedded real time software retesting. In: IEEE 8th International Conference on Computer and Information Technology Workshops, Sydney, Australia, pp. 528–532 (2008)
10. Malhotra, R., Tiwari, D.: Development of a framework for test case prioritization using genetic algorithm. *ACM SIGSOFT Softw. Eng. Notes* **38**(3), 1–6 (2013)

11. Mahmood, M.H., Hosain, M.S.: Improving test case prioritization based on practical priority factors. In: 8th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China (2017)
12. Gao, D., Guo, X., Zhao, L.: Test case prioritization for regression testing based on ant colony optimization. In: 6th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, China (2015)
13. Kaur, A., Goyal, S.: A genetic algorithm for regression test case prioritization using code coverage. *Int. J. Comput. Sci. Eng.* **3**(5), 1839–1847 (2011)
14. Kaur, A., Goyal, S.: A bee colony optimization algorithm for code coverage test suite prioritization. *Int. J. Eng. Sci. Technol.* **3**(4), 2786–2795 (2011)
15. Singh, Y., Kaur, A., Suri, B.: Test case prioritization using ant colony optimization. *ACM SIGSOFT Softw. Eng. Notes* **35**(4), 1–7 (2010)
16. Suri, B., Singhal, S.: Analyzing test case selection & prioritization using ACO. *ACM SIGSOFT Softw. Eng. Notes* **36**(6), 1–5 (2011)
17. Fu, W., Yu, H., Fan, G., Ji, X., Pei, X.: A regression test case prioritization algorithm based on program changes and method invocation relationship. In: 24th Asia-Pacific Software Engineering Conference (APSEC), Nanjing, China (2017)
18. Luo, Q., Moran, K., Zhang, L., Poshyvanyk, D.: How do static and dynamic test case prioritization techniques perform on modern software systems? An extensive study on GitHub projects. *IEEE Trans. Softw. Eng.* (2018)
19. Tiwari, S., Mishra, K.K., Kumar, A., Misra, A.K.: Spectrum-based fault localization in regression testing. In: 2011 Eighth International Conference on Information Technology: New Generations, Las Vegas, NV, pp. 191–195 (2011). <https://doi.org/10.1109/itng.2011.4>
20. Tiwari, S., Mishra, K.K., Misra, A.K.: Test case generation for modified code using a variant of particle swarm optimization (PSO) algorithm. In: 2013 10th International Conference on Information Technology: New Generations, Las Vegas, NV, pp. 363–368 (2013). <https://doi.org/10.1109/itng.2013.58>
21. Geem, Z.W., Kim, J.H., Loganathan, G.V.: A new heuristic optimization algorithm: harmony search. *Simulation*, Sage Publication **76**(2), 60–68 (2001)
22. Choudhary, A., Baghel, A.S., Sangwan, O.P.: An efficient parameter estimation of software reliability growth models using gravitational search algorithm. *Int. J. Syst. Assur. Eng. Manage.* **8**(1), 79–88 (2017)
23. Choudhary, A., Agrawal, A.P., Kaur, A.: An effective approach for regression test case selection using pareto based multi-objective harmony search. In: Proceedings of the 11th International Workshop on Search-Based Software Testing, pp. 13–20. ACM (2018)
24. Laddad, R.: AspectJ in action- enterprise AOP with spring applications, 2nd edn. Manning Publications, Greenwich (2009)
25. <https://www.eclipse.org/>
26. Elbaum, S., Malishevsky, A.G., Rothermel, G.: Test case prioritization: a family of empirical studies. *IEEE Trans. Softw. Eng.* **28**(2), 159–182 (2002)
27. Rothermel, G., Untch, R.H., Chu, C., Harrold, M.J.: Prioritizing test cases for regression testing. *IEEE Trans. Softw. Eng.* **27**(10), 929–948 (2001)
28. Do, H., Rothermel, G.: On the use of mutation faults in empirical assessments of test case prioritization techniques. *IEEE Trans. Softw. Eng.* **32**(9), 733–752 (2006)

Onboard Data Acquisition System to Monitor the Vehicle



Adesh Kumar Pandey and Sangeeta Arora

Abstract Road safety is very important in our day to day life. Every year thousands of accidents happen during the driving of vehicles. For many accidents, reasons are not identified. There is need of device for recording vehicle evidences when the person is on the way and safe the people from malicious activities i.e. accidents, kidnapping etc. The system having acquisition system on the board is designed to monitor and record the vehicle speed, video recording, acceleration, steering input etc. This device is robust due to its design and work in different temperatures also. This device helps in investigations of road accidents, vehicle theft and kidnappings etc. Initially this system is fabricated in car.

Keywords Raspberry · Data acquisition system · Arduino

1 Introduction

Now day's vehicles are increasing exponentially on the road which leads to the accidents on the roads. The safety is critical issue for which individual is responsible. The main reason behind the accidents is due to the lack of safety measures. If vehicle is equipped with safety measures then it will reduce the road accidents.

Kim et al. [1] discussed that with the analysis of accidents or road mishappening i.e. kidnapping cases has no physical evidence. Due to lack of evidences, sometimes reasons are unidentified and victims get no justice. Due to this, a data acquisition system scheme is proposed to enhance the security of vehicle. Prasad et al. [2] also proposed automobile black box system for the analysis of accidents. This system has twelve sensors used to record various driving data parameters.

A. K. Pandey · S. Arora (✉)
KIET Group of Institutions, Ghaziabad, India
e-mail: sangeeta.arora@kiet.edu

A. K. Pandey
e-mail: ak.pandey@kiet.edu

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_24

Wathanawisuth et al. [3] developed a compact data acquisition system with the components accelerometer, GPS and GSM module. Due to small size, it can be placed on motorcycle and bicycles also. Nugroho et al. [4] developed a system for flights which can convert data into binary form for the analysis. Kang et al. [5] proposed a black box system to record the video of vehicle when the person is driving. It also records the licence plate number and color of nearby vehicles. This system is also able to store the information on the server.

Fancello et al. [6] proposed methodology to determine the problematic sections of the road with the help of multicriteria approach. Dudziak et al. [7] discussed the behavior of vehicle on the roads and analysed to deal threats to deal.

This paper proposes the safety measure solutions which can be placed in vehicle. In this paper, a data acquisition system prototype is developed to place on the vehicles. This system is useful to track the location, review of accidental cases and help insurance companies to find the reason. In Sect. 2 need of this type of data acquisition system is discussed. The methodology of data acquisition system development is described in Sect. 3. In methodology, subject, research instruments are described. The Sect. 4 has the description of proposed system. In Sect. 5 results are discussed.

2 Need of Data Acquisition System

In India, vehicle is equipped with accessories but not necessities. These types of devices are not used in India. It may be lack of awareness or availability in the market. If vehicles are equipped with these types of devices, it not only helps for insurance claims but also in investigations and finds reasons in road accidents.

If the proposed system is used for vehicles in India, It may be of assistance to minimize casualties on roads, guard the women and keep law and order. This type of system facilitate transport department to equipped their vehicles with appropriate technology and maintain the vehicles.

The functions of the proposed Data Acquisition System are:

- Data Collection:
 - Visual Data: Data is recorded with the help of camera during driving from front and rear side both.
 - Driving Data: The vehicle status at the time of driving i.e. momentum, brake, seat belt wore seat belt or not and performance of steering.
 - Collision Data: When accident or any other activity happen, data like time, speed and shock power.
 - Positioning Data: The car positions are recorded in real time using the GPS.
- Storage of Data: This data is initially stored in RAM temporarily, further it is transferred in the Flash memory like SD card and hard disk. This information is analyzed in future in case of any accident, insurance settlements etc.

- **Data Transfer:** The data is recorded and transmitted to control center regularly when the vehicle is on the road. This is the great help at the time of any contingency. This data transfer may be triggered in any unusual situation i.e. steering performance, speed etc. This helps the control center for rescue operation also.

3 Methodology

Car is considered as subject in this proposed data acquisition system. The system has following research instruments:

Embedded System: Embedded Systems are systems which are the part of any electrical or mechanical system to perform a particular task having some real time constraints. An embedded system is a combination of three components i.e. hardware, software and Real Time Operating System (RTOS).

Hardware Resources: The proposed data acquisition system used following hardware components to perform the task of vehicle monitoring while vehicle is in running stage:

- (a) **Raspberry Pi:** The Raspberry Pi foundation evolves the computer which is on a single board in United Kingdom [8]. Raspberry Pi is well suited for physical setup as well for remote setup with existing computer. In physical setup, the requirement of power supplies or USB and SD cards is mandatory. For the remote setup, it can be connected with the help of wi fi or network switch and operated remotely by any computer in network. A protocol Virtual Network Computing can be used to control from one computer to another computer. It can be connected with Secure Shell (SSH) in Linux operating system or Chrome.
- (b) **Arduino:** It is one of the open source microcontroller which is used to read sensors and control things like turning on and off LED [9]. The programs can be uploaded in the Arduino board to interact with things in the real world. Arduino provides platform which is open-source and easily used with hardware as well as with software. An Arduino board accepts input in different ways i.e. light with the help of sensor, finger touch with button etc. Arduino board also provides its own Integrated Development Environment (IDE) for programming. The programming is one with the help of C++. The programs can be developed and run on computer system and afterward upload on the Arduino board with the help of Universal Serial Bus (USB).
- (c) **Accelerometer:** It is a device which is used for the measurement of acceleration [10]. This is an electromechanical device which generates the electrical signals. Accelerometers are used to measure static forces like gravity and dynamic forces i.e. any vibration or movement etc. The acceleration can be measured in different number of axes i.e. one, two or three. Accelerometer is made up of capacitive plates, which might be fixed or moving. The moving plates are fixed with tiny spring. These plates move as sensing the acceleration forces upon the sensor.

With the movement of plates, capacitance is changed from which acceleration can be resolved.

Accelerometer is used in car to measure whether car is climbing on a slope or any defect in the engine with help of vibration. In this system ADXL335 (along with AD0804 and LM358 IC) is used to measure any vibration or movement of car. It can be install on car's bonnet.

- (d) **Alcohol Sensor:** In this system, Alcohol Sensor is used to check whether drive has taken alcohol or not [11]. In this system MQ3 sensor is used. The analog resistive output is given by the sensor depends upon alcohol absorption. This is also work as alcohol alarm in domestic or industrial region. The portable alcohol sensor can be used to detect amount of alcohol taken by the person.
- (e) **GPS:** Global Positioning System (GPS) [12] is a navigation system which is based on satellite. To find the exact spot of the vehicle, GPS is used. The four GPS satellites are used to send information to receiver, which must be in the line of sight. GPS receiver fetches the information about the place and current time in regular periods. The information is radio signal with the speed of light, which is later interpreted by receiver. GPS is able to find the 2-D and 3-D location of the receiver with the help of trilateration process. Three satellites are used to find the 2-D location which tells latitude and longitude. Whereas for 3-D location provide altitude also using four satellites. In our system, GPS will be used for tracking place of vehicle.
- (f) **Camera:** In this system Qualcomm camera is used for video recording of the vehicle [13]. It made easy to know the accurate reason of collision. Qualcomm consists of custom-made chips which have features of Artificial Intelligence, Vision Intelligence etc. This chip can be used for security, 360 degree cameras etc. Due to Qualcomm, camera is light weight and has capability of full HD photo and video which can be controlled programmatically.

Software Resources: In proposed embedded system, data sent and received with the help of client server programming. For communication with the server, socket programming is done. Hardware resources provide the input to software resources.

In client server programming, the device request for services called a client and the device provide services to client is called server. Server provides the services such as data and resource sharing with several clients. The services provided by a server can give services to multiple clients or vice versa is possible. Server can be of any type i.e. database servers, web server, mail server etc., depends upon type of services used.

In socket programming, a connection is created using Transmission Control Protocol (TCP) between two systems, which enables the systems at both end to transmit the data. If multiple clients can be communicate with the same server with the help of separate TCP connection.

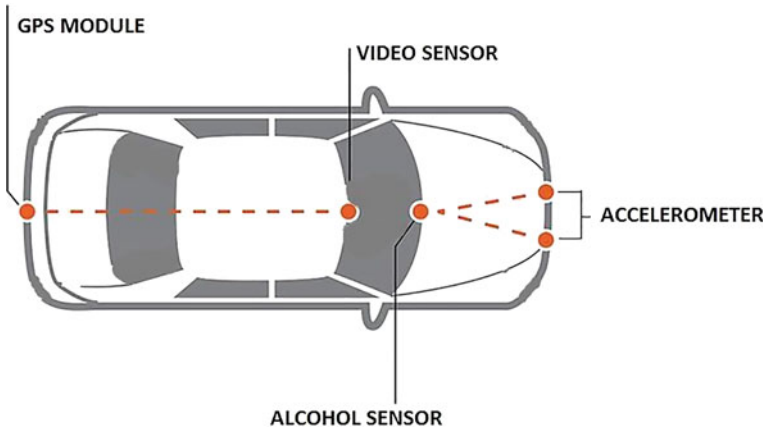


Fig. 1 Placement of sensors at desired location

4 Proposed System

In a proposed solution, a system is designed which acquire the data from the vehicle on the road. Initially a prototype is developed for the system; then system is tested by placing on vehicle along with the sensors placed at their desired locations. In Fig. 1, it is shown that sensor placed on the vehicle.

This data acquisition system is placed at the desired location on the vehicle and is tested on the car when it is moving position which discussed in next section. This data acquisition system is connected with the server which receives the data from it. The received data is stored in the server for further analysis if required. The tracking of position is done for safety issues. Video camera record video and sent to server time to time (Fig. 2).

5 Results and Discussion

The result evaluated with this developed system is stored in text format for both ends vehicle and server. The results are shown in Fig. 3.

The longitude and latitude is took from the file of particular interval and uploaded on Google map and results are displayed in the Fig. 2.

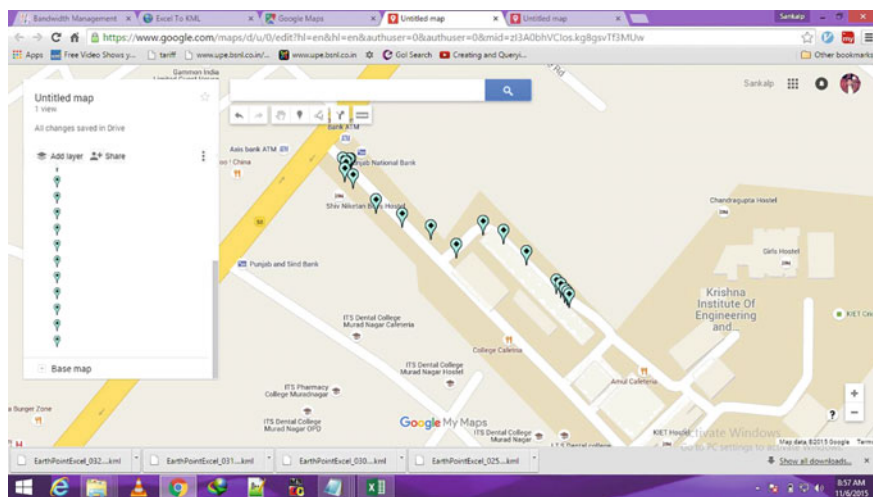


Fig. 2 Google map results

Date	KIDP	Latitude	Longitude	Fix Date	Time	Sex	Age	Course Name	Course Code	Course Credits	Course Sem	Course Status	Credits
11/12/2013	08	28.762743	77.430738	08:34:08	575	M	206.30	C. 06	7.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	447	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	744	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	951	M	206.30	C. 06	7.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	300	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	362	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	111	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	300	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	840	M	206.30	C. 06	3.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	871	M	206.30	C. 06	3.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	793	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	820	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	86	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	223	M	206.30	C. 06	6.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	374	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	843	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	818	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	397	M	206.30	C. 06	3.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	448	M	206.30	C. 06	2.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	740	M	206.30	C. 06	2.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	814	M	206.30	C. 06	2.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	78	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	210	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	373	M	206.30	C. 06	5.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	442	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	526	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	600	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	471	M	206.30	C. 06	4.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	741	M	206.30	C. 06	3.00	8	8	8	0
11/12/2013	08	28.762743	77.430738	08:34:08	832	M	206.30	C. 06	3.00	8	8	8	0

Fig. 3 Preprocessing results

6 Conclusions and Future Scope

In this paper, a prototype of crash proof Data acquisition system is developed to place on a vehicle. This data acquisition system record the information and sent it to server for use. This data acquisition system is tested on a moving vehicle i.e. car. This system is helpful for investigation of accidents, to track location, vehicle theft, drink and drive, kidnapping. This system will lead to help towards road safety.

References

1. Kim, M., Jeong, C.Y.: An efficient data integrity scheme for preventing falsification of car black box. In: Proceedings of International Conference on ICT Convergence (ICTC), IEEE, pp. 1020–1021 (2013)
2. Prasad, M.J., Arundathi, S., Anil, N., Harshika, Kariyappa, B.S.: Automobile black box system for analysis (2014)
3. Watthanawisuth, N., Lomas, T., Tuantranont, A.: Wireless black box using MEMS Accelerometer and GPS tracking for accidental monitoring of vehicles. In: International Conference on Biomedical and Health Informatics (BHI), 2012 IEEE-EMBS (2012)
4. Nugroho, S., Nasution, S.M., Azmi, F.: Analysis of cockpit voice recorder compression reliability for airplane on demand data acquisition system data transmission. In: International Conference on Control, Electronics, Renewable Energy and Communications (ICCREC) (2017)
5. Kang, C., Heo, S.W.: Intelligent safety information gathering system using a smart data acquisition system. In: IEEE International Conference on Consumer Electronics (ICCE) (2017)
6. Fancello, G., Carta, M., Fadda, P.: Road intersections ranking for road safety improvement: comparative analysis of multi-criteria decision making methods. Elsevier Transp. Policy (2018)
7. Dudziak, M., Lewandowski, A., Sledzinski, M.: Uncommon road safety hazards. *Procedia Eng.* (Elsevier) **177**, 375–380 (2017)
8. <https://projects.raspberrypi.org/en/projects/raspberry-pi-getting-started>
9. <https://www.arduino.cc/en/Guide/Introduction>
10. <http://www.analog.com/en/products/sensors-mems/accelerometers/adxl335.html> #product-overview
11. <https://www.sunrom.com/p/alcohol-sensor-module-mq3>
12. <https://www.gps.gov/systems/gps/>
13. <https://www.qualcomm.com/news/releases/2018/04/11/qualcomm-unveils-vision-intelligence-platform-purpose-built-iot-devices>

Part IV
Web and Informatics

Toward Adapting Metamodeling Approach for Legacy to Cloud Migration



Pooja Parnami, Aman Jain and Navneet Sharma

Abstract Migration of legacy application to Cloud is a fast-growing area of knowledge. Many IT-based organizations inclined toward empowering their legacy application with cloud computing capabilities. Many researchers, academicians, national, and international bodies are creating knowledge models to allow knowledge sharing and provide effective cloud migration model. This knowledge is scattered and huge, but lack of knowledge management. Our motive is to produce a metamodel, which could be able to generalize the cloud migration domain. Metamodel approach is an approach, to gather all domain concepts and their relationships. Using the metamodel, variety of domain solution models can be built. It can act as a language infrastructure which unifies describing the process model of moving legacy enterprise applications to the cloud environments. The benefits of the metamodel include simplifying the migration process, guidance, reuse specialized migration knowledge and support training and knowledge management activities. Furthermore, it reduces complexity and ambiguity in cloud migration domain.

Keywords Cloud migration · Legacy applications · Metamodeling approach

1 Introduction

Legacy system or software can be termed as outdated computer software, which is still in existence in the industries. Because legacy system holds valuable data and the organization have spent a huge amount of money and time, it is difficult to discard them. To preserve the existing system and take advantages of cloud computing, the industry and academia are eager to find solutions in legacy to cloud migration filed. Cloud computing is gaining popularity due to its properties like pay-as-you-go model,

P. Parnami (✉) · N. Sharma
IIS University, Jaipur, India
e-mail: pooja.parnami@yahoo.com

A. Jain
Maharishi Arvind Institute of Science and Management, Jaipur, India

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_25

on-demand self-service, scalability, multi-tenancy, etc. However, an integrated and generalized view of the cloud migration process could not be found in the literature. Software reengineering field relies on knowledge. The knowledge regarding cloud migration is scattered in the available literature, but there is a lack of management of this knowledge. Metamodeling provides a way of unifying common concepts and establish a relationship between the information derived from knowledge to form new concepts and ideas. This paper we proposed a new approach for cloud migration based on metamodeling. This could helps to reduce complexity and ambiguity in the field of cloud migration.

2 Legacy System

Different authors have provided different definitions for legacy application but to summarize, the legacy systems could be considered as large, monolith programs build on outdated programming language and technology which are difficult to manage, maintain and upgrade. The characteristics of legacy system cause various troubles for the organizations in running their business processes. The main problem with the legacy system is their poor and rigid architecture. Most of the monolith systems are big in size and build on mainframes. They are built on outdated technology. Mainly the hardware supporting the system are outdated are not getting support from the supplier and manufacturers. Due to lack of documentation and experts, the cost of daily operations and maintenance of is very high. It is difficult to add, update or modify the functionality of legacy systems.

3 Cloud Computing

According to the National Institute of Standards and Technology (NIST) [1] there are five essential characteristics of cloud computing:

- **On-demand self-service:** Without much interaction with provider the customer can avail services (such as storage, CPU time, extra nodes on network) instantaneously.
- **Broad network access:** The cloud services could be deployed on private or public network (cloud) and are available at any platform (such as mobile phone, laptops etc.).
- **Resource pooling:** The cloud service provider creates a pool of resources for multi-tenant use. These resources could be allocated and de-allocated as per need.
- **Rapid elasticity:** A client can purchase resource(s) at any time, in any quantity. The system can enable scaling on automatic or manual basis.
- **Measured service:** The cloud provider measure, audit and report the customer based on a metered system. A client is charged on the basis of quantity and level of services provided by the provider.

Based on virtualization of resources three service models of cloud computing [2–4]:

- **Infrastructure as a Service (IaaS):** In IaaS the service provider delivers IT infrastructure like hardware, storage, network, servers or other computing resources to the consumer. IaaS provider uses virtualization technique to meet on demand, shrinking and growing needs of consumer. Examples of IaaS are GoGrid, IBM SmartCloud, Amazon EC2 etc.
- **Platform as a Service (PaaS):** In PaaS model the service provider delivers comprehensive development environment including programming environment, tools, hosting environment etc. Examples of PaaS Windows Azure, AWS Elastic Beanstalk etc.
- **Software as a Service (SaaS):** In SaaS model software application is hosted by service provider delivered through the Internet to the user. The service provider also maintains the hardware, software and quality standards such as security, availability, performance, etc. Examples of SaaS are Google Apps, Dropbox, Salesforce, etc.

Four deployment models of cloud computing are [1] as follows:

1. Private cloud
2. Public cloud
3. Community cloud
4. Hybrid cloud.

4 Cloud Migration

There is no specific, clear, and unified definition of cloud migration. Cloud Migration could not be defined with a clear and unified definition, but it could be explained as a process of transition of traditional IT environment to cloud.

4.1 Migration Strategies

As per Gartner [5], there are five options for migrating legacy application to the cloud, which are given as follows:

1. **Rehost:** “Rehost” infers changing the hardware and/or infrastructure configuration of an application.
2. **Replace:** “Replace” means to discard an existing application with a commercial software which is delivered as a service.
3. **Refactor:** “Refactor” means to make changes in application code or configuration to connect old application with new infrastructure. Refactoring can use existing programming model and framework.

4. Rebuild: “Rebuild” means discarding the code of an existing application and rebuild solution on a provider’s application platform, application.
5. Revise: “Revise” cloud be defined as modifying or extending the existing code, then deploy it on the cloud with the help of rehost and refactor options.

Applications are usually build using three-layer architecture—Presentation, Business, and Data layer. One or more architectural layer(s) could be migrated to the cloud. Based on which layer(s) migrated on cloud following migration types has been identified [6].

1. Type I: In this type one or more component(s) can be replaced by Cloud services, which results in migrating data layer and/or business logic layer to the Cloud service.
2. Type II: In this type a partial migration of one or more application layers or set of architectural components or functionality takes place.
3. Type III: In this type of migration the application is migrated through encapsulating in Virtual Machine(s).
4. Type IV: Complete migration of the application takes place in this type of migration. The application functionality is implemented as a service on the Cloud.

In a white paper provided by Cisco [7], there are three application migration options: SaaS, PaaS, and IaaS.

1. Migration to IaaS: IaaS migration provides services to the user in the form of virtual machines. The first step is to build a virtual machine and load all software to be run in cloud. Then, this virtual machine is uploaded and deployed on IaaS provider’s hosting environment.
2. Migration to PaaS: PaaS migration is to move one software operating and deployment environment to another environment. The vendor provides the hardware and some application software like databases, middleware and deployment tools to the customer. At PaaS level, the customer need not manage their virtual machines.
3. Migration to SaaS: In SaaS model software and associated data are migrated to cloud and the user can access the application using a web browser.

SaaS migration can be further categorized into following sub-categories [8]:

1. Replacing by SaaS: There is no need for any kind of reengineering, only the local data is to be exported to cloud database.
2. Revising based on SaaS: Some legacy system’s functionality will be outsourced, and the integration of cloud and non-cloud services is done with the help of business processes.
3. Reengineering to SaaS: To reengineer, legacy systems it may require reverse engineering, structure redesign, service generation, etc.

In the paper [9], compares and analyzes and map the migration methods into five strategies, which are given as follows:

- Migrate to IaaS
- Migrate to PaaS

- Replace by SaaS
- Revise by SaaS
- Reengineer to SaaS.

In real-world scenario, the cloud migration strategy depends upon each organization's individual needs and properties of legacy system. Therefore, organizations must design and apply the most cogent strategy before the migration. Companies are inclined more towards IaaS strategy due to ease of implement and cost benefits. But IaaS migration not able to take full advantage of cloud platform. In PaaS migration, the existing system needs to adapt according to target platform, this results in some shortcomings, like missing capabilities, transition risk, and vendor lock-in.

There are two ways in SaaS strategy replacing or reengineer. Replacing a legacy system needs least migration effort but reengineering legacy system to cloud is very challenging and it may require lots of work like reverse engineering, structure redesign and service generation.

5 Issues and Challenges

Legacy application was build, without taking into account the characteristics of cloud environment, like scalability, elasticity, interoperability etc. Several existing surveys in the field of cloud migration identified the following key challenges of cloud migration:

- There is no universal method proposed and developed for legacy to cloud migration of applications [10].
- There is no abstract and structured technique to support multi-tenancy, elasticity, test and continuous integration [10].
- Existing cloud migration approaches are not tailorable i.e. developers should create a repository of reusable method fragments. These situation specific methods could be reused to fit in a given migration scenario [11].
- There is still very less tool support to automate and facilitate cloud migration tasks [11].

6 Metamodeling

For the purpose of design optimization, the best way is simulation, but in complex engineering the cost of simulation is very high. The surrogate option of simulation is metamodel, which is constructed in lieu of simulation models [12]. The metamodels are created by the combination of bottom-up and top-down analysis of existing methodologies and best practices.

For any given domain, the benefits of metamodel are [13]:

- Domain concepts are easier to apply for newcomers (concepts would be present in the single metamodel instead of having to look for them in a dispersed collection of extant ones).
- Increased portability of models across supportive modeling tools (they would refer to the same metamodel); better communication between researchers (they could use the same frame of reference, i.e., the unified metamodel).
- Research could focus on improving and/or realizing the unified metamodel instead of being spread across a number of existing metamodels.

A metamodeling process, i.e. developing a metamodel for a given domain, generally intends to build a collection of classes, relations, and constraints to express concepts in the domain, though they may not be expressed exactly in the same way [14]. Metamodeling is a complex process. It requires extensive knowledge of the diversified, distributed and poorly structured problem domain. A metamodel must be readable and understandable in terms of representing the problem domain [15].

There are several techniques suggested for metamodel creation. But each technique suggests an iterative way of finding the final version of a domain metamodel. The typical steps suggested by each technique could be defined in terms of the following steps:

1. Identifying domain concepts from sources.
2. Reconciliation and harmonization of collected concepts.
3. Defining the relationship between the concepts.
4. Designing initial metamodel.

This primary metamodel could be verified and refined to achieve expected quality.

7 Proposed Method

Migration of legacy software application to cloud is a collaborative work, which involves high level of complexity. The complete migration process requires the knowledge distributed across time, space and people. Our motive is to unify existing efforts done in the field of cloud migration and represent it in a reusable form. We are planning to follow metamodeling approach to create a generic process model for migrating existing software application stack to cloud environment.

The purpose of developing this metamodel is to get a broad understanding of the tasks carried out during legacy to cloud transition. The development of proposed metamodel is based on set of common and frequently used cloud migration concepts. The identification of these concepts is through exiting cloud migration literature. We need to identify cloud migration models, which includes processes, approaches, experience reports, etc. This collection provides a broad knowledge regarding cloud migration activities and operations. To achieve the final metamodel we need to follow an iterative path. The complete process is divided into the following seven steps (Fig. 1).

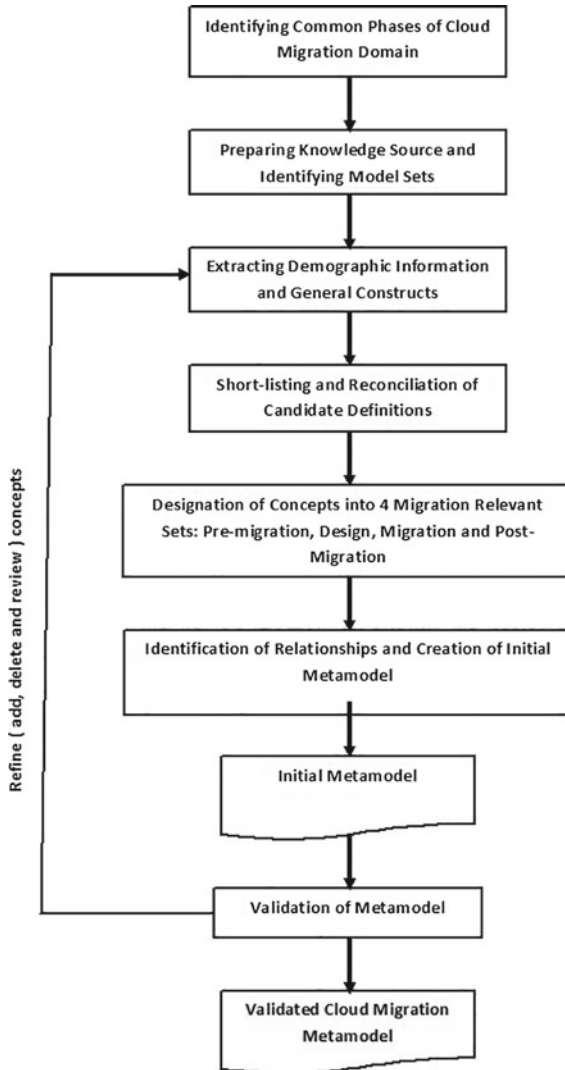


Fig. 1 The process of developing the cloud migration metamodel

7.1 Step 1: Identifying Common Phases of Cloud Migration Domain

After analyzing generic software development lifecycles proposed by Pressman [16] and Sommerville [17] 5 generic phases i.e. Analysis, Design, Implement, Test and Maintenance could be proposed in context of the current research. Additionally, with reference to existing reengineering lifecycles for example Butterfly [18], Renaissance

[19], Sneed's approach [20], and Architecture-Driven Modernization horseshoe [21], the cloud migration process could be categorized in following four generic phases:

- Pre-migration
- Design
- Migration
- Post-Migration.

7.2 Step 2: Preparing Knowledge Source and Identifying Model Sets

The main aim of this research is to create a metamodel and main knowledge source is literature available in cloud migration research. Essential feature of any research endeavor is rigorous review of relevant literature. Every year a considerable number of research papers are published in the field of cloud migration. The papers provide best practices, solutions, methodologies etc. to migrate a legacy application to cloud. Therefore, the first step is to collect available models and then divide them into three sets. SET 1 is used to initiate the metamodeling process, this will produce the initial metamodel. Another two sets, Set V1 and V2 are used for the purpose of validation of proposed metamodel. The sets are formed according to how broadly they cover the different phases of cloud migration

7.3 Step 3: Extracting Demographic Information and General Constructs

Through extensive study of models identified for SET 1, general constructs and their definitions need to be collected. While extracting the construct, the main criteria is that a construct must be sufficiently generic to a variety of cloud migration scenario regardless of a particular cloud technology and tool.

7.4 Step 4: Short-Listing and Reconciliation of Candidate Definitions

Since the literature selected for the study come from people with different background, the terminologies, phrases, and definitions of a given construct may differ. Therefore, it is essential to reconcile various definitions of assembled constructs. If two or more constructs had same name or two constructs with different names referring the same thing, then a harmonization process must be undertaken.

7.5 Step 5: Designation of Concepts into 4 Migration Relevant Phases

The reconciled and harmonized constructs designated into one of the cloud migration phases, identified in Step 1: Pre-Migration, Design, Migration, and Post-Migration.

7.6 Step 6: Identification of Relationships and Creation of Initial Metamodel

In this step, the major task was identification of relationship among the constructs and creation of initial metamodel with the help of UML notations.

7.7 Step 7: Validation of Metamodel

For the validation purpose following two validation techniques could be used—(a) Comparison to other metamodels. (b) Frequency-based selection.

8 Conclusion

Many IT-based companies wish to migrate their legacy applications to cloud platform. Several concerns like security, interoperability, vendor lock-in, etc. raise to incorporate cloud migration process. Proposed way of creating a metamodel could provide a generic cloud migration process model. The metamodel approach could help in getting a theoretical view in a domain which can be specialized and extended for a given context. The resultant metamodel could provide step by step process, with clear division of phases, activities, tasks and work products, during transition from legacy to cloud platform. The resultant metamodel will help in to fill the research gap in the domains where artefacts are not available. The current and future research work of introducing metamodeling will provide a foundation for creating a customized method to attune with a particular technology and help in to understand the process of technology shift.

References

1. Mell, P., Grance, T.: The NIST definition of cloud computing recommendations of the national institute of standards and technology. Natl. Inst. Stand. Technol. Inf. Technol. Lab. **145**, 7 (2011)

2. Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I.: Above the Clouds: A Berkeley View of Cloud Computing. Dept. Electrical Eng. and Comput. Sciences, Univ. California, Berkeley, Rep. UCB/EECS, **28**, 13 (2009)
3. Buyya, R., Yeo, C.S., Venugopal, S.: Market-oriented cloud computing: vision, hype, and reality for delivering IT services as computing utilities. In: Proceeding 10th IEEE International Conference High Performance Computing and Communications HPCC 2008, pp. 5–13, 2008
4. Low, C., Chen, Y., Wu, M.: Understanding the determinants of cloud computing adoption. *Ind. Manag. Data Syst.* **111**(7), 1006–1023 (2011)
5. Woods, J.: Five options for migrating applications to the cloud : rehost, refactor, revise, rebuild or replac (2011)
6. Abadi, D.J., et al.: Performance Antipatterns: detection and Evaluation of Their Effects in the Cloud. *Futur. Gener. Comput. Syst.* **29**(1), 758–765 (2013)
7. Bakshi, K.: Cisco cloud computing—data center strategy, architecture, and solutions point of view white paper. *Solutions*, 1–16 (2009)
8. Zhao, J.F., Zhou, J.T.: Strategies and methods for cloud migration. *Int. J. Autom. Comput.* **11**(2), 143–152 (2014)
9. Sabiri, K., Benabbou, F.: Methods migration from on-premise to cloud. *IOSR J. Comput. Eng. Ver. IV* **17**(2), 2278–2661 (2015)
10. Gholami, M.F., Daneshgar, F., Beydoun, G., Rabhi, F.: Key challenges during legacy software system migration to cloud computing platforms—an empirical study. *Inf. Syst.* **67**(June), 100–113 (2017)
11. Jamshidi, P., Ahmad, A., Pahl, C.: Cloud migration research: a systematic review **1**(2), 142–157 (2013)
12. Jin, R., Chen, W., Simpson, T.W.: Comparative studies of metamodelling techniques under multiple modelling criteria. *Struct. Multidiscip. Optim.* **23**(1), 1–13 (2001)
13. Beydoun, G., et al.: FAML: a generic metamodel for MAS development. *IEEE Trans. Softw. Eng.* **35**(6), 841–863 (2009)
14. Sowa, J.F.: *Conceptual Structures: Information Processing in mind and Machine*. Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA (1984)
15. Papers, R.S.: Bottom-up meta-modelling: an interactive approach. *MoDELS* **7590**, 3–19 (2013)
16. Pressman, R.S.: *Software Engineering: A Practitioner’s Approach, 7/e*. LLC, McGraw-Hill Global Education Holdings (1992)
17. Sommerville, I.: *Software Engineering*. Addison-Wesley Pub. Co.,—International computer science series (1992)
18. Wu, et al.: The butterfly methodology: a gateway-free approach for migrating legacy information systems. In: Proceedings of Third IEEE International Conference on Engineering of Complex Computer Systems, pp. 200–205 (1997)
19. Ian Warren, M.B., Avallone, D.: *The Renaissance of Legacy Systems*. Springer-Verlag (1999)
20. Sneed, H.M.: Planning the reengineering of legacy systems. *IEEE Softw.* **12**(1), 24–34 (1995)
21. Khusidman, V., Ulrich, W.: Architecture-driven modernization: transforming the enterprise. *Semin. Softw. Anal. Trasformation* pp. 1–7 (2007)

Application of Cloud Computing for Priority Job Scheduling by Multiple Robots Operating in a Co-operative Environment



Amitava Kar, Ajoy K. Dutta and Subir K. Debnath

Abstract In this paper, we describe the application of cloud computing in the field of robotics where one robot can locate the position of the other robot by consulting database residing in the cloud. The robots are assigned to pick up jobs lying on the floor according to the priority of the jobs and to keep them in the given spaces. In the co-operative environment, both the robots try to follow their route of most-priority-jobs-first till such situations arise that the job has been exhausted by the other robot. Then the robot searches for the next-priority-job. In case of keeping the jobs in spaces, the robots follow the nearest space where the jobs are to be kept till such situation arise that the space has been exhausted by the other robot. Then the robot searches for the next-nearest-space. Cloud computing minimizes the calculation overhead of the robots as the cloud calculates the path to the jobs and the spaces for both the robots and therefore the robots do not need heavy processor as they offload the calculations to the cloud. The robots use the resources residing in the cloud particularly memory and so they do not need to store all the data they need.

Keywords Automated robots · Cloud · Priority job scheduling · Co-operative environment

1 Introduction

A problem is said to be optimized when it obtains the best solution from all the feasible solutions, but the solution may be best with some parameters but it may not be the best with some other parameters. An optimization problem may yield optimal result under certain given conditions but when some additional constraints are applied to make one variable more optimal, the feasibility may be lost and the next most optimal solution has to be selected. The selection of the next most optimal solution to satisfy some additional constraint leads to reduction of the value of certain

A. Kar (✉)

Department of Computer Application, Asansol Engineering College, Asansol, India
e-mail: karamitava@gmail.com

A. K. Dutta · S. K. Debnath

Production Engineering Department, Jadavpur University, Kolkata, India

© Springer Nature Singapore Pte Ltd. 2019

Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_26

other variable against marginal increase of some variable. In this paper, the operation of multi-robot leads to the appearance of additional constraints, that one robot cannot pick up a certain job that has already been picked up by the other robot or a space, which has been utilized by one robot for keeping up a job, cannot be utilized by the other robot for keeping a different job. In this paper, the jobs are to be picked according to the priority. This leads to an increase of distance travelled by the two robots but at the same time, it reduces the time of work as both the robots operate simultaneously.

2 Related Work

The launching of the Internet in the 1990s led to the limited sharing of resources. Cloud Computing is an application of such a resource sharing mainly software rather than hardware. The hardware cannot be shared over the internet but load on the hardware can be reduced by offloading the complex calculations to the cloud. It relieves remote devices from the burden of carrying out extensive computations [1]. Cloud robotics uses the idea and includes the possibility of reducing the hardware requirement of a robot by storing the data in the cloud and getting them as and when required by querying them. It can also offload the complex calculations to the cloud and can query the result of the calculations as and when needed. Arumugam et al. [2] spoke about the Cloud Computing framework for service robots. Guizzo [3, 4] talked about the help that robot receives from the cloud. Robot's navigation involves the determination of its own position and planning to reach a desired location. Wen et al. [5] spoke about energy optimal execution policy for a cloud-assisted mobile application platform. Wen et al. [6] spoke about the energy optimality for cloud-assisted mobile platforms during their navigation. Durrant-Whyte et al. [7] talked about the path planning. Kar et al. [8–11] talked about the offloading of the complex calculations to the cloud and path planning by a single robot. Kar et al. [12] spoke about the data sharing using cloud by the two robots during their operation of carrying the jobs and to keep them in the given spaces but in an environment where any job can be kept at any spaces.

3 Proposed Approach

The robots are allotted the delivery of the jobs from being scattered on the floor to the given spaces. The location of the jobs and the spaces are there with the cloud. The cloud chalks out the route of each robot following the priority of each job and then to the minimum-distant space and again to the next prior job and then to the next space. This goes on until all the jobs or the spaces are exhausted.

The two robots have different routes via each job and each space and each of their movement is noted into the database housed in the cloud. So before each move, it checks with the database whether any targeted job has already been attempted for

pickup. The robot that is at minimum distance from the job goes for the most-prior job. The other robot goes for the next prior job. After pickup of a job, the robot moves to the minimum-distant space for delivering the job there.

The algorithm that the two robots follow is given as under:

Algorithm

Let
 $D_1 = \text{Distance travelled by } R1 = 0,$
 $D_2 = \text{Distance travelled by } R2 = 0$

(1) Input the jobs according to their priorities
 (2) Let us suppose that Job i has the highest priority, then Calculate
 $D_{1i} = R1J_i$ for all i from 1 to n
 $D_{2i} = R2J_i$ for all i from 1 to n
 if $D_{1i} < D_{2i}$ then,
 Job i is assigned to $R1$.
 Find the job that has the next highest priority (let us suppose that the job having the next priority is Job j).
 Job j is assigned to $R1$.
 $D_1 = D_1 + D_{1i}$
 $D_2 = D_2 + D_{2j}$
 Else
 Job i is assigned to $R2$.
 Find the job that has the next highest priority (let us suppose that the job having the next priority is Job j).
 Job j is assigned to $R1$.
 $D_2 = D_2 + D_{2i}$
 $D_1 = D_1 + D_{1j}$
 End if

(3) If $D_1 < D_2$ then,
 For $R1, D_1 = D_1 + \min (J_i S_k)$ for k and i from 1 to n
 For $R2, D_2 = D_2 + \min (J_j S_l)$ for j and l from 1 to $(n-1)$ and $l \neq k$
 Else
 For $R2, D_2 = D_2 + \min (J_j S_l)$ for j and l from 1 to n
 For $R1, D_1 = D_1 + \min (J_i S_k)$ for k and i from 1 to $(n-1)$ and $k \neq l$

(4) Let us suppose that Job m has the next priority followed by Job l .
 If $D_1 < D_2$ then
 For $R1, D_1 = D_1 + (S_i J_m)$ for i and m from 1 to $(n-2)$
 For $R2, D_2 = D_2 + (S_k J_l)$ for k and l from 1 to $(n-3)$
 Else
 For $R2, D_2 = D_2 + \min (S_k J_l)$ for j and l from 1 to $(n-2)$
 For $R1, D_1 = D_1 + \min (S_k J_m)$ for k and i from 1 to $(n-3)$
 Repeat Step 3 and 4 until all the jobs and spaces are respectively exhausted.

This algorithm is tested assuming that there are only four jobs and four spaces.

4 Experimental Setup and Results

Assumptions

Let us suppose that the names of the robots are R1 and R2.

1. Both the robots are assumed to be point robots.
2. The jobs may be picked up according to priority, given in the priority table.
3. Any job may be kept at any space.
4. R1 start at(0, 0) whereas R2 start at (120, 0).
5. Velocity of both the robot is same.
6. When one robot picks up a job then the location of the job becomes the location of the robot.
7. When one robot reaches a particular space and keeps the job there, then the location of the space becomes the location of the robot.

The details are given in the following diagram (Fig. 1).

The locations of the different jobs along with their priorities are given in Table 1.

The locations of the different spaces are given in Table 2.

The distance matrix for R1 is given Table 3.

For the algorithm to run successfully, the cloud uses two database tables:

- (i) Location (ObjName, ObjX, ObjY)
- (ii) Distance (FromObj, ToObj, Distance).

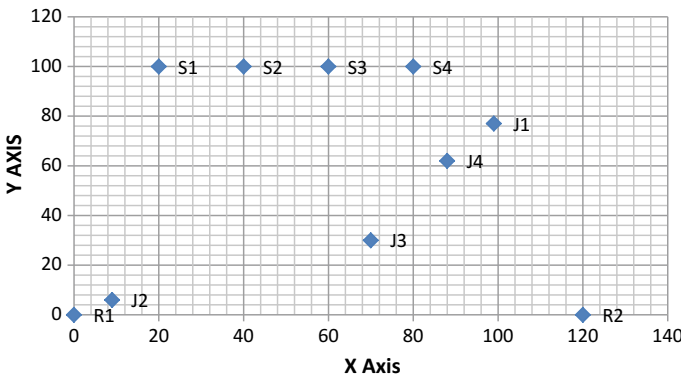


Fig. 1 Floor where the robots are operating

Table 1 Location of different jobs

Job	x	y	Priority
1	99	77	1
2	9	6	3
3	70	30	2
4	88	62	4

Table 2 Location of Spaces

Spaces	x	y
1	20	100
2	40	100
3	60	100
4	80	100

Table 3 Distance Matrix for R1 and R2

	J1	J2	J3	J4
R1	125.42	10.82	76.16	107.65
R2	79.81	111.16	58.31	69.77
S1	82.28	94.64	86.02	77.90
S2	63.32	98.98	76.16	61.22
S3	45.28	106.94	70.71	47.20
S4	29.83	117.80	70.71	38.83

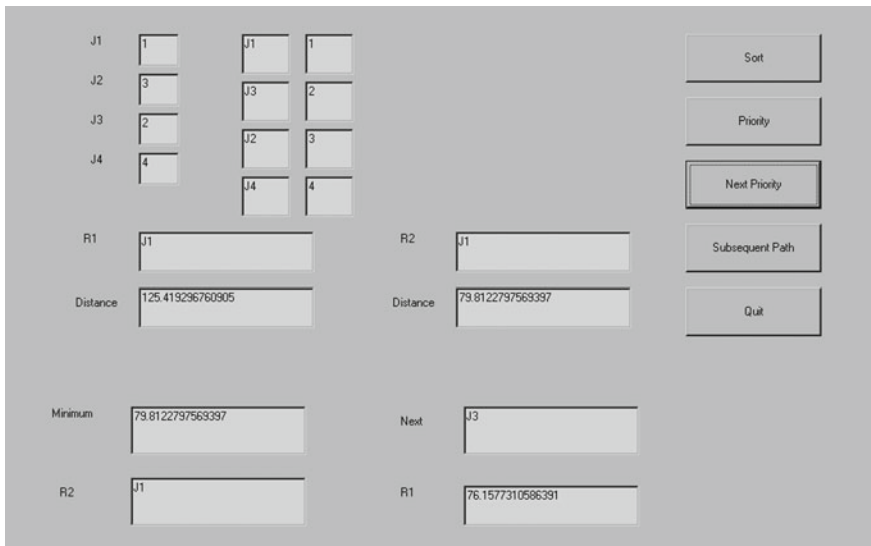


Fig. 2 Software where input is given in the form of priorities

The location table stores the location of different objects like the initial position of the two robots, the location of the different jobs and the spaces in their X and Y coordinates. The cloud calculates the distance between different objects and stores them in the distance table. The two robots query from the distance table and find out the objects that they have to reach and querying the location table they find out the location of the objects and reach the object (Figs. 2 and 3).

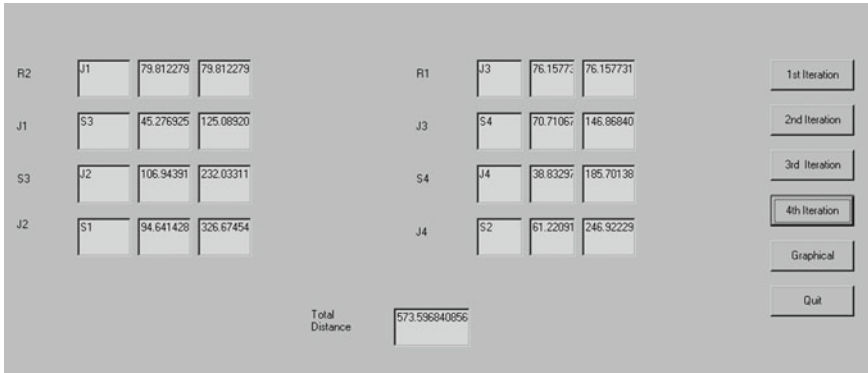


Fig. 3 Paths of the two robots R1 and R2

Table 4 Path of R1 and R2 along with stage-distance and cumulative distance

Object	Dist.	Cumu. Dist.	Object	Dist.	Cumu. Dist.
R2-J1	79.81	79.81	R1-J3	76.16	76.16
J1-S3	45.28	125.09	J3-S4	70.71	146.87
S3-J2	106.94	232.03	S4-J4	38.83	185.70
J2-S1	94.64	326.67	J4-S2	61.22	246.92

Table 5 Sequence of Jobs done by the two robots

Job	Space	Done by
J2	S1	R1
J3	S4	R2
J1	S3	R2
J4	S2	R1

The two robots are working simultaneously. All the jobs will certainly not be carried by one robot. To find out which robot does which jobs, the path of both the robots along with the distance of each stage and the cumulative distances are calculated by the cloud and are given Table 4.

Now, the velocities of the two robots are same and so one robot can analyze the position of the other robot. The most priority job is J1. The distance of R1 from J1 is more than the distance of J1 from R2. So R2 goes for the job J1 and R1 go for the next prior job J3. J3 is picked up by R1 earlier than R2’s picking up of J1 and so R1 will go for the space that is nearest to it. It goes to S4 to keep that job. The next nearest available space from R2 is S3. So it goes to S3 to keep the job J1. R2 reaches S3 earlier than R1’s reaching of S4 and so the picking up of next prior job is done by R2. This goes on until all the jobs and the spaces are exhausted.

The final table of the jobs kept at spaces is given Table 5.

The graphical view of the path of the two robots is given in Fig. 4.

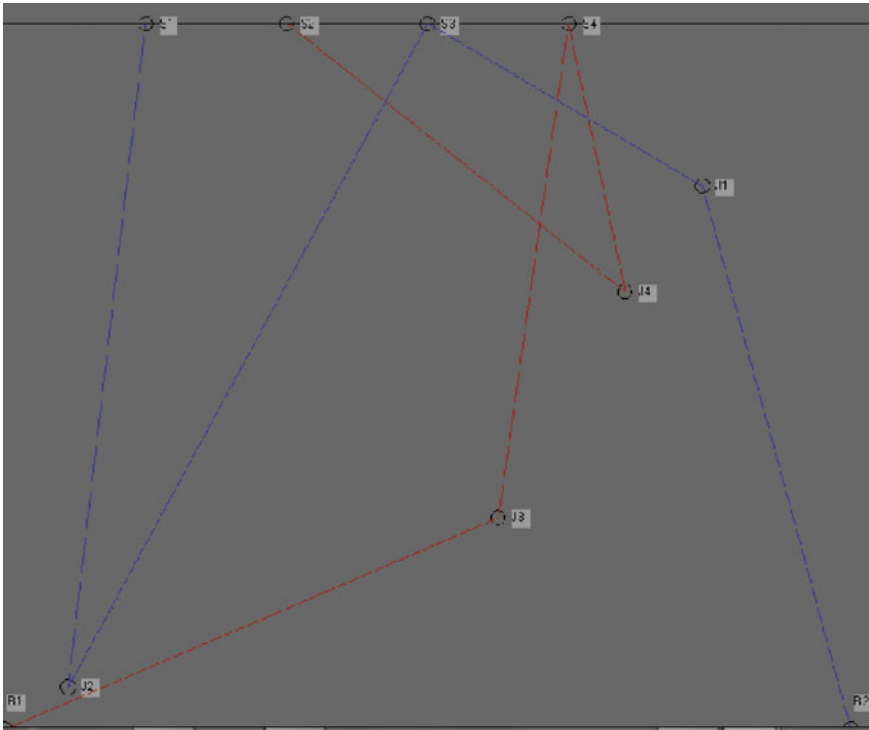


Fig. 4 Path of the two robots shown graphically

5 Collision Detection

From the diagrammatic view, there are possibilities of collision when R2 moves from J1 to S3 and R1 moves from J3 to S4 and again when R1 moves from J4 to S2 and R1 moves from S3 to J2.

Case—1:

J1(99, 77) to S3(60, 100). The equation of the line of movement J1S3 is $23x + 39y = 5280$.

J3(70, 30) to S4(80, 100). The equation of the line of movement J3S4 is $7x - y = 460$.

Solving these two equations, we get the point of intersection of the two lines $x = 78.45$ and $y = 89.15$.

The collision may occur at the point (78.45, 89.15).

The distance of the point (x, y) from J1 is $\sqrt{(99 - 78.45)^2 + (77 - 89.15)^2} = 23.87$.

The total distance travelled by the robot R2 to reach (x, y) is $79.81 + 23.87 = 103.68$. If the velocity of both the robots is the same (v say), then the total time taken by the robot R2 is $103.68/v$.

The distance of the point (x, y) from J3 is $\sqrt{(70 - 78.45)^2 + (30 - 89.15)^2} = 59.75$.

The total distance travelled by the robot R1 to reach (x, y) is $76.16 + 59.75 = 135.91$.

Then, the total time taken by the robot R2 to reach (x, y) is $135.91/v$.

But, $135.91/v > 103.68/v$. i.e. The robot R2 passes the point (x, y) much earlier than R1. So there is no collision.

Case—2:

S3(60, 100) to J2(9, 6). The equation of the line of movement S3J2 is $94x - 51y = 540$.

J4(88, 62) to S2(40, 100). The equation of the line of movement J4S2 is $19x + 24y = 3160$.

Solving these two equations, we get the point of intersection of the two lines $x = 53.99$ and $y = 88.92$.

The collision may occur at the point $(53.99, 88.92)$.

The distance of the point (x, y) from S3 is $\sqrt{(60 - 53.99)^2 + (100 - 88.92)^2} = 12.61$.

The total distance travelled by the robot R2 to reach (x, y) is $125.09 + 12.61 = 137.70$. If the velocity of both the robots is the same (v say), then the total time taken by the robot R2 is $103.68/v$.

The distance of the point (x, y) from J4 is $\sqrt{(88 - 53.99)^2 + (62 - 88.92)^2} = 43.37$.

The total distance travelled by the robot R1 to reach (x, y) is $185.70 + 43.37 = 229.07$.

Then the total time taken by the robot R1 is $229.07/v$.

But, $229.07/v > 137.70/v$. i.e. The robot R2 passes the point (x, y) much earlier than R1. So there is no collision.

6 Conclusions

R2 moves 326.67 units of distance and R1 moves 246.92 units. The total distance moved by the two robots is $(326.67 + 246.92) = 573.59$. The total time spent by the two robots is $\max(326.67/v, 246.92/v) = 326.67/v$ which is much less than the time taken by a single robot had the total work been done by it alone. The problem is of co-operation between the two robots in jointly conducting the jobs and cloud acts as a catalyst such that the jobs are performed.

This problem can be categorized as that of the path minimization problem in Operations Research with additional constraints that the robots cannot move from one job to another job or from one space to another space and the jobs are to be done according to the priority. In addition to it, the cloud acts as an information provider for each of the robots about the other robot.

This problem was formulated keeping in mind of the situation of a disaster, e.g. a flood or a draught with the jobs as the people who got stuck in the disaster and the

spaces as the safe zone where the people are to be transferred. In this paper, priority of the job implies the people who are to be saved with priority, i.e. saving a child is the maximum priority, followed by old people, followed by a woman and then comes the turn of a man.

References

1. Mell, P., Grance, T.: The NIST Definition of Cloud Computing, National Institute of Standards and Technology Special Publication, pp. 800–145 (2011)
2. Arumugam, R., Enti, V.R., Bingbing, L., Xiaojun, W., Baskaran, K., Kong, F.F., Kumar, A.S., Meng, K.D., Kit, G.W., Rakotondrabe, M., Ivan, I.: Davinci: a cloud computing framework for service robots. *Int. Conf. Robot. Autom.* 3084–3089 (2010)
3. Guizzo, E.: Robots with their heads in the clouds, *IEEE Spectrum* (2011)
4. Guizzo, E.: Cloud robotics: connected to the cloud robots get smarter. *IEEE Spectrum* (2011). 2012 Florida Conference on Recent Advances in Robotics 6 Boca Raton, Florida, May 10–11, 2012
5. Wen, Y., Zhang, W., Guan, K., Kilper, D., Luo, H.: Energy-optimal execution policy for a cloud-assisted mobile application platform. In: Rabaey, J.M., (ed.) *Digital Integrated Circuits*. Prentice Hall, 1996. NTU Technical Report, 2011
6. Wen, Y., Zhang, W., Guan, K., Kilper, D., Luo, H.: Energyoptimal execution policy for a cloud-assisted mobile application platform. NTU Technical Report (2011)
7. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part I. *IEEE Robot. Autom. Mag.* **13**, 99–110 (2006)
8. Kar, A., Dutta, A.K., Debnath, S.K.: Task management of robot using cloud computing. *IEEE Int. Conf. Comput. Electr. Commun. Eng.* (2016)
9. Kar, A., Dutta, A.K., Debnath, S.K.: Task management of robot using cloud computing. *IEEE Xplore* (2016)
10. Kar, A., Dutta, A.K., Debnath, S.K.: Task management of robot using priority job scheduling. *Int. J. Comput. Sci. Inf. Technol.* **8**(2) (2017a)
11. Kar, A., Dutta, A.K., Debnath, S.K.: Optimising task management of robot and deciding whether cloud computing is feasible. *Int. Organisation Sci. Res. J. Comput. Eng.* **19** version 4 (2017b)
12. Kar, A., Dutta, A.K., Debnath, S.K.: Application of cloud computing for optimization of tasks by multiple robots operating in a co-operative environment. *IEEE Int. Conf. Comput. Electr. Commun. Eng.* 2018 (2016)

A Framework for Security Management in Cloud Based on Quantum Cryptography



Priya Raina and Sakshi Kaushal

Abstract Cloud Computing (CC) and Quantum Computing, both have been interesting areas of research, individually. However, integrating the two can come with mutual benefits for both the fields. Cloud platforms can offer Quantum Computing as a service, but more importantly, they are capable of offering the flexibility, inherent in architecture, to accommodate new developments in Quantum Computing. On the other hand, Quantum Computing in general and Quantum Cryptography (QC), in particular, can help in alleviating the security concerns associated with CC, which have prevented users from migrating to cloud. The paper proposes a model framework contemplating the use of Blind Quantum Computation (BQC) between cloud servers involved in multiparty computations during online phase and using authenticated Quantum Key Distribution (QKD) for secure distribution of keys (used for encrypting the secret files) when going into offline phase. Finally, a proof of the model has been presented, using Universal Composability (UC) framework.

Keywords Cloud · Cloud security · Quantum computing · Quantum cryptography · Quantum key distribution · Blind quantum computing

1 Introduction

Cloud Computing (CC) [12] has proliferated in a short time span, owing to the flexibility, scalability, and economies of scale it offers. Useful as it may be, outsourcing critical data and processes to a third party, without adequate security guarantees, has deterred most users, from individuals to large organizations, from migrating to cloud. On the other hand, we have Quantum Computing [9], which holds the promise of

P. Raina (✉)
Chitkara University, Himachal Pradesh, India
e-mail: priyaraina1807@gmail.com

S. Kaushal
UIET, Panjab University, Chandigarh, India
e-mail: sakshi@pu.ac.in

immense computational and processing power, bringing a paradigm shift in the way we compute. However, the cost of the technology, as and when it becomes available, is likely to be too high, much like supercomputers. In such a situation, CC can make it accessible (Quantum Computation as a Service!). CC has much to gain from this union, as Quantum Cryptography (QC) [1, 3] and Information Theory can help overcome security concerns [17]. They can be useful in enhancing network and data security in cloud, particularly on the server side. Guarantees related to privacy, confidentiality, and authentication can be ensured by mechanisms, e.g., protocols for key management, bit commitment, MPC, Blind Quantum Computation, etc. [14].

In this paper, based on the literature survey (discussed briefly in Sect. 2), the authors propose a protocol using QC for secure key management for servers engaged in multiparty computation. In online phase, when servers are busy performing multiparty computations, the protocol suggests use of Blind Quantum Computation (BQC) between the cloud servers. When no computation is being performed, to save on cloud resources, the servers go to sleep, i.e., offline state. Before doing so, they need to encrypt their secret files and securely distribute them, for which a (sub) protocol P_{CKMQ} has been proposed that makes use of authenticated Quantum Key Distribution (QKD). Section 3 presents the formal model of P_{CKMQ} using the Universal Composibility (UC) [5] and Sect. 4 discusses the proof of security for P_{CKMQ} . Section 5 summarizes the paper with Conclusion and Future Work.

2 Related Work

According to existing body of literature, QC can be used for resolving security issues in CC in following ways [14]:

- (a) ***QKD integration with existing protocols***. QKD has been used for key generation in various protocols like PPP [8], IPsec [15], TLS/SSL [7].
- (b) ***Authentication using QKD***. Khan and Xu [11] have come up with an authentication scheme with unconditional security of QKD and flexibility of PKI for grid architecture with the communities connected via a quantum network like SECOQC. Their proposed architecture can be extrapolated to cloud. Khalid and Zulkarnain [10] have suggested the use of Tight Finite Key in addition to MQKD, i.e., Multi-QKD for distributing same secret key within a group in order to make multiparty authentication possible in cloud.
- (c) ***Confidentiality using Blind Quantum Computing(BQC)***. BQC protocols have been suggested for ensuring confidentiality of client's data and operations in a situation where the server provides Quantum Computation as a Service over the cloud. BQC based on measurement and teleportation was suggested in [4] and recently demonstrated in [2]. Another approach, called "ancilla-driven" model [16] uses actual quantum circuits rather than measurement.

3 The Proposed Model

Suppose there are n servers that want to compute a function of their “inputs”, without revealing them, a scenario typical of Multi-Party Computation (MPC) e.g., in awarding e-tenders the secret input would be the bid while the output will be finding the minimum bid. The input file is thus the server’s “secret file”. This paper proposes a model for maintaining privacy in a situation where the servers are “QC-capable”. In the model, it is assumed that the servers are not engaged in continuous computation, thus going into “sleep” mode to save on computing resources and the switch is either autonomous or semi-autonomous. The secret file needs to be secured in both the online as well as the sleep state.

In the online state, the servers can perform the computation using BQC [4], [16], suitably modified to work for a group of servers, thus computing the output function “blindly”, i.e., without actually divulging the secret files.

In the sleep state, the attacker can get the access to the secret file. If the server encrypts the file before going offline, but keeps the key with itself, it would defeat the purpose of encryption. If server S_i encrypts the file using a key K_i , and breaks it into n parts (where n is the total number of servers involved in BQC) such that $K_i = K_{i,1} || K_{i,2} \dots || K_{i,n}$. It keeps $K_{i,i}$ with itself, sending $K_{i,j}$ to j th server, the communication between S_i and S_j encrypted on the fly. The model suggests using QKD with one-time pad to achieve the same. Additionally, S_i needs to delete the clear text file and K_i . On “wake up”, it needs to collect the parts of K_i again using encrypted channels protected by session key, generate K_i and decrypt the secret file for use in the online phase. Figure 1 shows the interactions between three servers. K_1, K_2, K_3 are the secret keys the servers use for encrypting secret files SF_1, SF_2, SF_3 respectively. Functions E and D denote encryption and decryption operations respectively. $S_{i,j}$ is the shared secret key between servers i and j established using QKD to encrypt the channel while exchanging the parts of K_i . Note that $S_{i,j}$ at the initiation of “sleep” and “wake up” operations are established independently of each other and on the fly, and thus are distinct (although denoted using the same label). Arrows show the exchange while boxes show the servers’ local storage.

Thus, the protocol consists of two sub-protocols i.e, BQC [4] in the online phase and P_{CKMQ} (based on [6]) for transition between online and offline states.

3.1 The UC Framework

Security of a protocol depends upon not just the robustness of the algorithm, but also on the environment in which the protocol runs/ supposed to run. The difficulty in defining unpredictable environments like the Internet is why the notion of security is captured using the “stand-alone” approach. [5] proposed a solution by suggesting the idea of “Secure Composition”—there is no need to delve into the details of the specific environment as long as the protocol is secure when “composed” with other

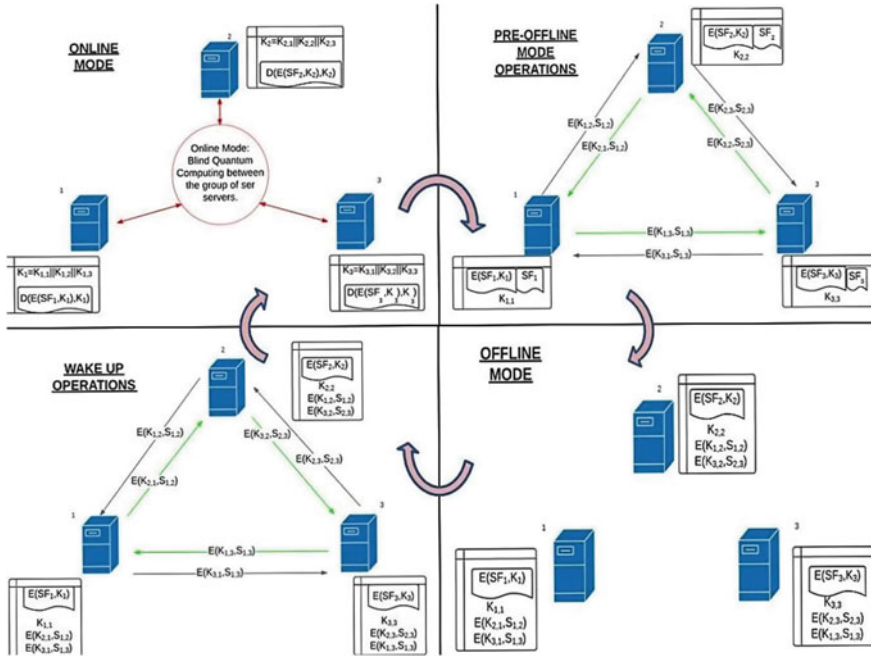


Fig. 1 Interactions between servers running the proposed protocol

protocols in the environment. The UC framework is a real/ideal world simulation paradigm i.e., a protocol is secure if the environment can't distinguish between ideal and real worlds. The entities in UC framework are modeled using the interactive Turing Machines (ITM model) [5]. Z denotes the environment, which is also the distinguisher. Π represents the real world protocol to be analysed i.e., the code run by each party. A is the real adversary whereas S is the simulator for Z , the parties and acts as the ideal adversary. $IDEAL - F$ represents the protocol emulation in the ideal world, realized using an ideal functionality F . The assumptions for the proposed model are as follows. Pair-wise asynchronous channels have been assumed to exist between the servers—quantum channel as well as public channel. The quantum channel is modeled by ideal functionality $F_{QUANTUM}$. Public channels connecting the servers are insecure but reliable, i.e., messages are eventually delivered. Servers are capable of erasing their states. This is modeled by requiring corrupted servers to reveal their current state to the adversary. Other aspects of security are modeled by following assumptions about the adversary:

- Static Adversary, i.e., parties to be corrupted in round r are a priori known to Z .
- Can corrupt any no. of channels but atmost $n - 1$ servers in each round.
- If a server is actively corrupted in, it remains so in all future phases and rounds.
- No cut off attack allowed.

Environment: Z represents everything that is external to the current protocol execution i.e., other protocols running in the system (BQC for this particular model), their adversaries and parties. It also represents interaction of the protocols with ongoing protocol execution. Z gives inputs to the parties and the adversary and collects all outputs. Finally, it outputs a bit denoting whether the interaction happened in the real/hybrid world or in the ideal world. The protocol model in UC framework has been presented in Figs. 2, 3 and 4. Figures 2 and 3 are associated with the ideal world. Figure 4 gives the real protocol in $F_{QUANTUM}$.

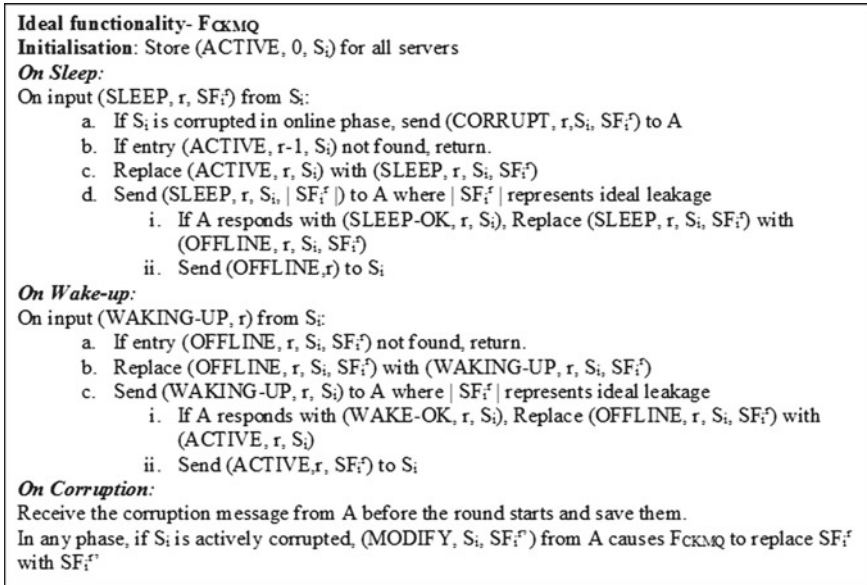


Fig. 2 Ideal Functionality

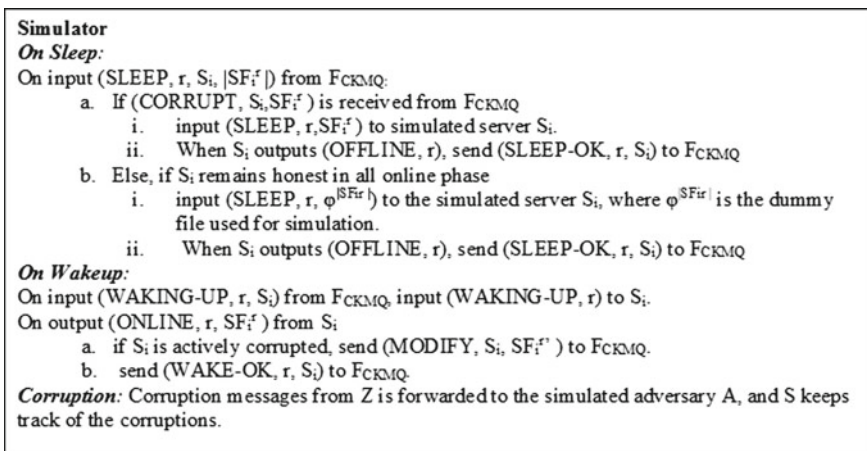


Fig. 3 Simulator

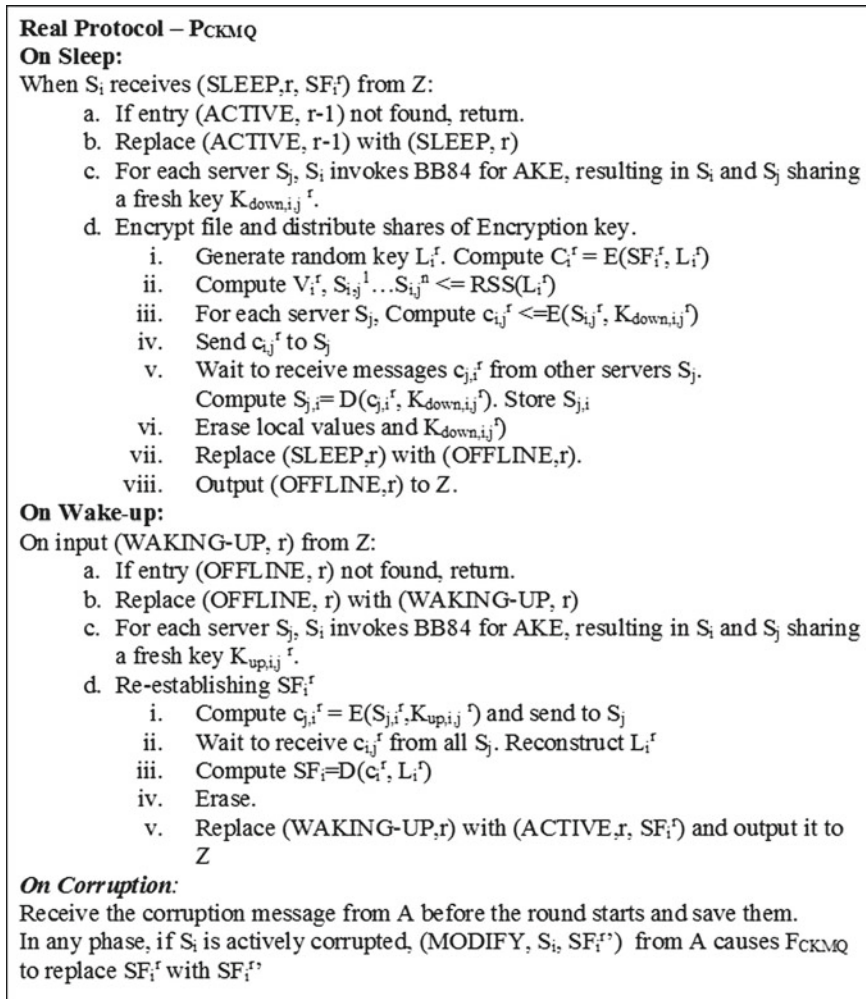


Fig. 4 Real Protocol

4 Proof of Security

The basic idea behind the proof is to use the transitivity of indistinguishability. The execution is broken into a sequence of intermediary protocol runs, the first one similar to the ideal world and the last one similar to hybrid world, w.r.t. Z . The aim is to prove that the environment cannot distinguish between two consecutive steps, which differ by single usage of one primitive, sequenced to prove indistinguishability.

Following the structure spelt out in [6], a parameterised version of P_{CKMQ} has been used in the proof. The protocol $P_{\text{CKMQ}}(a, b, c, d, e)$ is a modification of P_{CKMQ}

Table 1 Actions associated with bits in parameterised P_{CKMQ}

Bit	Action
$a = 1$	A new uniformly random key $L_i^{r'}$ replaces the original key L_i^r
$b = 1$	SF_i^r received from Z is replaced by the dummy file $0^{ SF_i^r }$ only if S_i is honest
$c = 1$ (only if $b = 1$)	Original file is copied to $SF_i^{r'}$ and stored throughout the protocol. It is not part of internal state of the server and thus not handed to adversary on corruption. For honest server, $SF_i^{r'}$ is returned to Z on wakeup
$d = 1$	S_i and S_j share random session keys that are unrelated to the keys generated via the Authenticated BB84 only if both are honest
$e = 1$	encrypted dummy shares $c_{i,j'} = Enc(r_{i,j}, K)$ are sent over the wires instead of the real encrypted shares $c_{i,j}$

and varies from it depending on which of the bits are turned on. Table 1 explains the actions. Since the protocol proceeds in a no. of rounds (which depends on A , Z and no. of random coins used by TMs), mathematical induction comes in handy throughout the proof.

Notation:

$$F_{CKMQ} = P_{CKMQ} \implies IDEAL_{F_{CKMQ}, S, Z} = EXEC_{P_{CKMQ}, A, Z}$$

$$P_{CKMQ}(a, b, c, d, e) = P_{CKMQ}(a', b', c', d', e') \implies \\ EXEC_{P_{CKMQ}(a, b, c, d, e), A, E} = EXEC_{P_{CKMQ}(a', b', c', d', e'), A, Z}$$

Given: $P_{CKMQ}(0, 0, 0, 0, 0) = P_{CKMQ}$

To Prove: Assuming that the encryption, MAC, robust secret sharing and authenticated BB84 primitives are secure, P_{CKMQ} UC-realizes the ideal functionality F_{CKMQ} with respect to the environment. That is, $F_{CKMQ} = P_{CKMQ}$.

Proof

$$\begin{aligned} F_{CKMQ} &= P_{CKMQ}(0, 1, 1, 0, 0) &= P_{CKMQ}(0, 1, 1, 1, 0) &= P_{CKMQ}(0, 1, 1, 1, 1) \\ &= P_{CKMQ}(1, 1, 1, 1, 1) &= P_{CKMQ}(1, 0, 1, 1, 1) &= P_{CKMQ}(0, 0, 1, 1, 1) \\ &= P_{CKMQ}(0, 0, 0, 1, 1) &= P_{CKMQ}(0, 0, 0, 1, 0) &= P_{CKMQ}(0, 0, 0, 0, 0) \\ &= P_{CKMQ} \end{aligned}$$

(1) $F_{CKMQ} = P_{CKMQ}(0, 1, 1, 0, 0)$

No corruption in online phase: S_i in $P_{CKMQ}(0, 1, 1, 0, 0)$ replaces SF_i^r with a dummy file 0^n (where $n = |SF_i^r|$) on sleep but returns the original file (stored as $SF_i^{r'}$) to Z on wakeup. Z cannot notice the initial and final replacement of files due to S in F_{CKMQ} .

Passive corruption in online phase: Indistinguishability due to static adversary.

Active corruption in online phase: F_{CKMQ} leaks the real secret file to S and there is no final replacement of the file at wakeup. Environment has the same view in both protocol executions, even if the adversary decides to modify the secret file.

(2) $\mathbf{P}_{CKMQ}(0, 1, 1, 0, 0) = \mathbf{P}_{CKMQ}(0, 1, 1, 1, 0)$ Let H_l , for $0 \leq l \leq R$, be the hybrid protocol execution that proceeds in the same way as $P_{CKMQ}(0, 1, 1, 0, 0)$ for all rounds $r > l$ and as $P_{CKMQ}(0, 1, 1, 1, 0)$ for all rounds $r \leq l$.

$H_0 = P_{CKMQ}(0, 1, 1, 0, 0)$ and $H_R = P_{CKMQ}(0, 1, 1, 1, 0)$.

Only difference between H_l and H_{l+1} is that all session keys generated by Authenticated BB84 in round $l + 1$ are used in H_l whereas they are immediately replaced by random session keys on both servers in H_{l+1} .

Let I_l be a hybrid protocol “half-way” between H_l and H_{l+1} , i.e., it behaves like H_{l+1} , except that only the session keys generated from Authenticated BB84 between honest servers in the sleep phase of round $l + 1$ are replaced by random keys. H_l differs only from I_l by the keys used in the sleep phase of round $l + 1$ and I_l differs from H_{l+1} only by the AKE keys used in the wakeup phase of round $l + 1$.

Let $P(r) : P_{CKMQ}(0, 1, 1, 0, 0) = H_r$ except with negligible probability.

$P(0)$ is true (since in H_0 , all rounds by definition use actual key).

$P(R)$ would imply that $H_R = P_{CKMQ}(0, 1, 1, 0, 0)$.

By definition, we know that $H_R = P_{CKMQ}(0, 1, 1, 1, 0)$.

$P(R)$ would imply $P_{CKMQ}(0, 1, 1, 0, 0) = P_{CKMQ}(0, 1, 1, 1, 0)$ from transitivity of indistinguishability.

Thus, proving $P(r)$ by using mathematical induction leads to the proof of this step.

For indistinguishability, we use game G_b (Fig. 5):

Let R^{G_b} simulate $EXEC_{H_r, Z, A}$ with following modifications:

- | |
|--|
| <ol style="list-style-type: none"> 1. Input : (KEY-GEN,i) for S_i
If message hasn't been input previously and no sessions have started, then <ul style="list-style-type: none"> • Compute and store (i, vk_i, sk_i) • Output vk_i • Mark S_i as honest 2. Input: (START-SESSION,i, j, sid) <ol style="list-style-type: none"> a. If key-gen has been invoked for both S_i and S_j and (i, j, sid) hasn't been stored
Store (i, j, sid) b. Output all messages between S_i and S_j to adversary. The adversary may even modify it or insert additional messages 3. Input: (CORRUPT,i)
If no session of S_i is test session and all sessions are completed <ol style="list-style-type: none"> a. mark S_i as corrupt b. Output sk_i 4. Input: (SK-QUERY, i, j, sid)
If all sessions are completed and sid is not test session <ol style="list-style-type: none"> a. Output $k_{i, j, sid}$ b. Mark the session with sid as exposed. 5. Input: (GET-TEST-SK,i, j, sid)
If the session is completed and unexposed and atmost T-1 sessions are marked as test-session <ol style="list-style-type: none"> a. If $b=0$: Output $k_{i, j, sid}$ b. If $b=1$: Output random key k c. Mark (i, j, sid) as test session |
|--|

Fig. 5 Game G_b for indistinguishability

1. In round r , instead of keygen in sleep phase for (sk, vk) R^{G_b} inputs $(KEY - GEN, i)$ to G_b and embeds vk received as vk_i^r in the protocol.
2. Everytime BB84 using digital signatures (with sk_i^r and vk_i^r) is initiated in wakeup phase of r (and shut down phase of $r + 1$), G_b is initiated with $sid = up$ and $sid = down$, respectively.
3. When S_i is corrupted, R^{G_b} receives sk_i^r , which is embedded in the protocol run.
4. When wakeup completes in round r of H_r , R^{G_b} sends $(SK - QUERY, i, j, up)$ to G_b and returned session key is embedded into the protocol. Z will not be able to distinguish because of this, since in H_r the resulting session key is already random and independent of Z .
5. When key exchange between S_i and S_j in sleep of round $r + 1$ is completed, R^{G_b} calls $(GET - TEST - SK, i, j, down)$ and embeds the test keys returned into protocol execution.

$R^{G_0} = H_r$ and $R^{G_1} = I_r \implies R^{G_0} = R^{G_1}$, by indistinguishability and also by the fact that in authenticated BB84, adversary cannot distinguish whether the key output to it is the actual key or any random key [13].

Similarly, it can be established that $I_r = H_{r+1}$ since wakeup keys in $r + 1$ can be replaced by random keys without E being able to distinguish.

Thus, $H_r = H_{r+1} \implies P_{CKMQ}(0, 1, 1, 0, 0) = H_{r+1} \implies P(r + 1)$.

(3) $P_{CKMQ}(0, 1, 1, 1, 0) = P_{CKMQ}(0, 1, 1, 1, 1)$ Dummy shares $c_{i,j'} = E(r_{i,j}, K)$ are sent instead of the real shares $c_{i,j} = Enc(s_{i,j}, K_{down/up})$ in $P_{CKMQ}(0, 1, 1, 1, 1)$. Due to Step (2), Z has no influence on which K is used.

Let L be the total number of session keys used by servers in $EXEC_{P_{CKMQ}, A, E}$, ordered according to $\pi(i, j, r, ?) \rightarrow 0, 1, \dots, L - 1$ such that the l th session key is used by S_i and S_j in the sleep/wakeup phase of round r for $(i, j, r, down/up) = \pi^{-1}(l)$.

Let H_l for $0 \leq l \leq L$ act as $P_{CKMQ}(0, 1, 1, 1, 0)$, except that in the sleep/wakeup phase of round r , servers S_i and S_j encrypt random shares as in $P_{CKMQ}(0, 1, 1, 1, 1)$ if $\pi(i, j, r, down/up) \leq l$, $P_{CKMQ}(0, 1, 1, 1, 0) = H_0$ and $P_{CKMQ}(0, 1, 1, 1, 1) = H_L$, by construction.

Let reduction with the modification that between the servers S_i and S_j in the round r and phase defined by the index l , neither the real nor dummy shares are encrypted. Instead $((s_{i,j}, r_{i,j}), (s_{j,i}, r_{j,i}))$ is input to G_{2b} (IND-CPA game) and the returned ciphertext $(c1, c2)$ embedded.

$R^{G_{2_0}} = H_l$ and $R^{G_{2_1}} = H_{l+1}$, by construction. Since $G_{2_0} = G_{2_1}$ and efficient transformations preserves indistinguishability, $R^{G_{2_0}} = R^{G_{2_1}}$.

And $H_l = H_{l+1}$ by transitivity of indistinguishability, which proves the proposition.

(4) $P_{CKMQ}(0, 1, 1, 1, 1) = P_{CKMQ}(1, 1, 1, 1, 1)$ The protocol $P_{CKMQ}(1, 1, 1, 1, 1)$ differs from $P_{CKMQ}(0, 1, 1, 1, 1)$ only in that a server S_i secret shares another key L_i^r instead of the actual key L_i^r used to encrypt the file SF_i^r . Privacy property of the secret sharing scheme can be used to prove indistinguishability. Let $\pi(i, r) \rightarrow 0, 1, \dots, L - 1$.

Let H_l be the hybrid protocol such that server S_i secret shares the same key L_i^r as used for encryption of SF_i^r as in $P_{CKMQ}(0, 1, 1, 1, 1)$ if $\pi(i, r) \leq l$ and samples and shares another key L_i^r in $P_{CKMQ}(1, 1, 1, 1, 1)$ if $\pi(i, r) > l$.

$H_0 = P_{CKMQ}(0, 1, 1, 1, 1)$ and $H_L = P_{CKMQ}(1, 1, 1, 1, 1)$. H_l and H_{l+1} differ only by the secret sharing of a key by one server S_i in one round r and it can be shown that for any $0 \leq l < L$ it holds that $H_l = H_{l+1}$ by using a reduction R^{G_b} (that simulates H_l with the modification that it accesses game G_b (Fig. 5) for generating shares of secret key and embeds the output in protocol run.

Since $G_0 = G_1$, $P_{CKMQ}(0, 1, 1, 1, 1) = P_{CKMQ}(1, 1, 1, 1, 1)$ by the fact that indistinguishability is transitive and preserved under efficient transformations.

(5) $\mathbf{P}_{CKMQ}(1, 1, 1, 1, 1) = \mathbf{P}_{CKMQ}(1, 0, 1, 1, 1)$ Similar to Step (3), except that the real file is encrypted and stored by S_i in $P_{CKMQ}(1, 0, 1, 1, 1)$ instead of a dummy.

(6) $\mathbf{P}_{CKMQ}(1, 0, 1, 1, 1) = \mathbf{P}_{CKMQ}(0, 0, 1, 1, 1)$ Similar to Step (4), except that the real file is used instead of dummy.

(7) $\mathbf{P}_{CKMQ}(0, 0, 1, 1, 1) = \mathbf{P}_{CKMQ}(0, 0, 0, 1, 1)$ In $P_{CKMQ}(0, 0, 1, 1, 1)$ an honest server always returns the correct file to Z on wakeup. In $P_{CKMQ}(0, 0, 0, 1, 1)$ the file returned to Z is that which is actually reconstructed in the protocol, so this step is essentially an argument for the correctness of the protocol: If no corruption occurs, correctness follows directly from the correctness of the encryption and secret sharing primitives. Further, it needs to be proved that Z cannot distinguish the file SF_i^r returned by honest S_i in round r from the correct file even if they have been modified. Consider a sequence of hybrid protocols H_l where the key rk_i^r output by reconstruction algorithm of the wakeup phase is replaced by the correct key L_i^r for $\pi(i, r) \leq l$. $H_0 = P_{CKMQ}(0, 0, 1, 1, 1)$ and $H_L = P_{CKMQ}(0, 0, 0, 1, 1)$.

The difference between H_l and H_{l+1} is only with regard to the secret sharing at S_i in round $r + 1$. Let E be the event that the reconstructed key is different from the key originally secret shared by S_i . We can prove that $P[E]$ is negligible in κ by constructing an adversary B that wins the robustness game with probability probability $P[E]$ (which is assumed to be non-negligible). B contradicts the robustness of secret sharing and hence the assumption that $P[E]$ is non-negligible stands falsified. Indistinguishability follows since $EXEC_{H_l, A, Z} | E = EXEC_{H_{l+1}, A, Z}$.

(8) $\mathbf{P}_{CKMQ}(0, 0, 0, 1, 1) = \mathbf{P}_{CKMQ}(0, 0, 0, 1, 0)$ Similar to Step (3), except that the real file is used instead of a dummy.

(9) $\mathbf{P}_{CKMQ}(0, 0, 0, 1, 0) = \mathbf{P}_{CKMQ}(0, 0, 0, 0, 0)$ Similar to Step (2), except that the real file is used instead of a dummy.

(10) $\mathbf{P}_{CKMQ}(0, 0, 0, 0, 0) = \mathbf{P}_{CKMQ}$ Follows from construction of $P_{CKMQ}(0, 0, 0, 0, 0)$.

Now, since the total number of reductions in the steps above is polynomial in the security parameter κ we get by transitivity of indistinguishability that $P_{CKMQ} = F_{CKMQ}$ or, using the standard UC notation, $EXEC_{P_{CKMQ}, A, Z} = IDEAL_{F_{CKMQ}, S, E}$.

5 Conclusions and Further Work

Integration of QC and CC has been suggested and based on the gaps in existing literature, a model was proposed for securing the servers in cloud by using QC techniques. Further, the modeling of the protocol was done using UC. However, only a partial analysis of the proposed model was done, i.e, the P_{CKMQ} that starts running when the system goes to “Sleep” mode and ends on “Wakeup”. This was possible because of the flexibility of the UC framework. In online mode, use of BQC after suitable modification has been proposed, the proof of which can be similarly done by means of modeling in UC framework. Finally, using modularity, security proof for the complete model can be conveniently achieved.

References

1. Alléaume, R., Branciard, C., Bouda, J., Debuisschert, T., Dianati, M., Gisin, N., Godfrey, M., Grangier, P., Länger, T., Lütkenhaus, N., et al.: Using quantum key distribution for cryptographic purposes: a survey. *Theor. Comput. Sci.* **560**, 62–81 (2014)
2. Barz, S., Kashefi, E., Broadbent, A., Fitzsimons, J.F., Zeilinger, A., Walther, P.: Demonstration of blind quantum computing. *Science* **335**(6066), 303–308 (2012)
3. Bennett, C.H., Brassard, G., Ekert, A.K.: Quantum cryptography. *Sci. Am.* **267**(4), 50–57 (1992)
4. Broadbent, A., Fitzsimons, J., Kashefi, E.: Measurement-based and universal blind quantum computation. In: *Formal Methods for Quantitative Aspects of Programming Languages*, pp. 43–86. Springer (2010)
5. Canetti, R.: Universally composable security: A new paradigm for cryptographic protocols. In: *Proceedings. 42nd IEEE Symposium on Foundations of Computer Science, 2001*, pp. 136–145. IEEE (2001)
6. Damgård, I., Jakobsen, T.P., Nielsen, J.B., Pagter, J.I.: Secure key management in the cloud. In: *IMA International Conference on Cryptography and Coding*, pp. 270–289. Springer (2013)
7. Elboukhari, M., Azizi, A., Azizi, M.: Implementation of secure key distribution based on quantum cryptography. In: *International Conference on Multimedia Computing and Systems, 2009. ICMCS'09*, pp. 361–365. IEEE (2009)
8. Ghernaouti-Hélie, S., Sfaxi, M.A.: Upgrading ppp security by quantum key distribution. In: *Network Control and Engineering for QoS, Security and Mobility, IV*, pp. 45–59. Springer (2007)
9. Gruska, J.: *Quantum Computing*, vol. 2005. McGraw-Hill London (1999)
10. Khalid, R., Zulkarnain, Z.A.: Enhanced tight finite key scheme for quantum key distribution (qkd) protocol to authenticate multi-party system in cloud infrastructure. In: *Applied Mechanics and Materials*, vol. 481, pp. 220–224. Trans Tech Publ (2014)
11. Khan, M.M., Xu, J.: Enhancing grid security using quantum key distribution. *Sersc. Org.* **6**(4), 67–76 (2012)
12. Mell, P., Grance, T., et al.: *The nist definition of cloud computing* (2011)
13. Mosca, M., Stebila, D., Ustaoglu, B.: Quantum key distribution in the classical authenticated key exchange framework. In: *International Workshop on Post-Quantum Cryptography*, pp. 136–154. Springer (2013)
14. Raina, P., Kaushal, S.: A study on quantum cryptography based security management in cloud. In: *Proceedings of the Sixth International Conference on Computer and Communication Technology 2015*, pp. 350–354. ACM (2015)

15. Sfaxi, M., Ghernaoui-Hélie, S., Ribordy, G., Gay, O.: Using quantum key distribution within ipsec to secure man communications. *Proceedings of metropolitan area networks (man2005)* (2005)
16. Sueki, T., Koshiha, T., Morimae, T.: Ancilla-driven universal blind quantum computation. *Physical Review A* **87**(6), 060,301 (2013)
17. Verma, A., Kaushal, S.: Cloud computing security issues and challenges: a survey. In: *International Conference on Advances in Computing and Communications*, pp. 445–454. Springer (2011)

Analyzing Student Performance Using Data Mining



Pankhurhi Mallik, Chandrima Roy, Ekansh Maheshwari, Manjusha Pandey and Siddharth Rautray

Abstract Analysis of student performance will help us understand the various factors that affect the overall of a student. Big Data Environment helps in analyzing the various concepts which are inbuilt for better strategies and the choices that are taken for an organization's overall development. Reduction in cost, time, the development of optimized and novice products, efficient and smart decision-making are some of the fields where it proves to be useful. Considering, the Higher Education System, which is inculcated in predicting the performance of students, this work will help various institutions in not only enhancing the quality of education, but also upgrading the overall accomplishments, identifying the pupil's at risk, and thereby refining the education resource management. This introspection will aid in identifying the patterns, where a comparative study between two distinct methods has been made in order to predict the student's success and a database has been generated.

Keywords Data mining · Student performance analysis · Clustering · K means · Mean shift

P. Mallik (✉) · C. Roy · E. Maheshwari · M. Pandey · S. Rautray
School of Computer Engineering, Kalinga Institute of Industrial Technology (KIIT)
Deemed to Be University, Bhubaneswar, Orissa, India
e-mail: pankhurhi.mallik@gmail.com

C. Roy
e-mail: chandrima.roy.1914@gmail.com

E. Maheshwari
e-mail: ekansh031998@gmail.com

M. Pandey
e-mail: manjushafcs@kiit.ac.in

S. Rautray
e-mail: siddharthfcs@kiit.ac.in

1 Introduction

Education is a key factor for accomplishing indelible economic progress. Student's academic performance centers around multiple aspects, making the analysis quite challenging. In later years, there has been an increase in the rate of interest and concern of people in the usage of data mining for educational and academic purposes. Data mining portrays rising and upcoming areas of researches in education, and it contains certain discrete requirements which other fields lack. In this paper, evaluation and analysis of the performance of students have been made. The goal is to do a comparative study of student's performance through different algorithms. For this, this research will be considering various parameters which affect a student's success [1].

All over the world, there is this rising consensus, a unison that the education system is degrading further day by day. On analyzing the elements, the organizations and other institutions can refine their models full of insight, creativity and above all an education system that works for their students and not against them. It will also help the mentors, the parents, and the whole education system to know about the areas of improvisation so that they work on it and uplift their academic progress and build a life full of experiences and knowledge. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. Two techniques have been used in this research. K means [2] is a method popular for cluster analysis in data mining [3]. It focuses on partitioning n results into k clusters. Basically, it is a kind of unsupervised learning which works on unlabeled data, i.e., data without defined groups. The other method used is the mean shift technique. It is a powerful nonparametric algorithm which is iterative and unsupervised in nature. Overall, it is a versatile algorithm which has found many practical applications.

Education, the process of promoting learning can also be defined as the acquisition of skills, knowledge, beliefs, and habits. Most of the education systems are designed in such a manner that it centers around a set of values or ideals which govern various educational decisions in that system. These choices encompass timetables/curriculum, facilities and functions, prototypes and model of the physical classrooms, interactions between the trainee and the trainer, techniques of evaluation, size of the class, educational activities, and more. The overall performance of any student might also be influenced by various psychological and social factors. Their performance/achievements vary in a strong way, but they differ from nation to nation and also from one individual to the other. Socioeconomic and personal aspects may also be taken into consideration. Socioeconomic Status (SES) is a term which consists of mainly two variables: the social status and economic status. Social status refers to the position in the society while economic status can be associated with the economy one owns or the monetary value possessed by an individual.

Attainment of Education is perceived as one of the main vehicles for spurring growth in the economic field and improving the living standards in developing coun-

tries. Data mining is used in this research to get a deeper probe into the education system and students' performance [4].

2 Background Details and Related Work

A lot of work has been done in recent years in order to provide a comprehensive review of data mining in education. Data mining represents promising areas of researches in the sector of education. It contains specific requirements which other fields lack. Different methods and techniques of data mining have been used to predict students' success, applying the data collected. Some of the work has been described below done by various researchers:

In the paper, "Data Mining Approach for predicting Student performance" published by E. Osmanbegovic et al., in the year 2012, they had applied three supervised machine learning algorithms to predict the success of a student in a course. They also predicted the performance of the learning methods where it was found that Naïve Bayes classifier had outperformed in prediction decision tree and neural network methods [5].

In the year 2016, Snehal Kekane et al. published a paper titled "Automatic Student Performance Analysis and Automatic Student Performance Analysis and Monitoring" in which they proposed a system which will display, in one single click, the results of student performance by the user which will not only initiate automation and also help in reducing manual efforts of the staff [6].

In the year 2015, titled "Predicting Academic Performance Of Students Using Data Mining Technique" by Indhu U Priya. et al. proposed system which focuses on the development of a tool for predicting the performance of a student for which preprocessing, classification, and clustering techniques were used. Indhu U Priya et al. published a paper titled "Predicting Academic Performance Of Students Using Data Mining Technique" in the year 2015 in which they proposed a system that focuses on the development of a tool for predicting the performance of a student [7].

In the paper titled "Student Performance Analysis System (SPAS)" published in 2015 by Chew Li Sa et al. few features were employed in the proposed framework during the design and implementation phase. Interface of the user and performance prediction were few of them which made sure that the objectives are achieved [8].

In the year 2013, Ramesh et al. published a paper titled "Predicting Student Performance- A Statistical and Data Mining Approach" where they investigated a survey cum experimental methodology which was adopted to generate a database and was constructed from a primary and a secondary source. It proved that Multilayer Perceptron (MLP) was considered to be the most appropriate classifier for prediction of student's performance. The other methods used were Naive Bayes, Multi-Layer Perception, SMO, J48, and REP Tree algorithms [9].

In the year 2007, a paper titled "Improving Student Performance in Public Primary Schools in Developing Countries: Evidence from Indonesia." by Daniel Suryadarm, investigated the correlation of student performance in mathematics and dictation tests

among school children in Indonesia. It finds a significant non-monotonic concave relationship between pupil–teacher ratio and student’s mathematics performance. Ordinary Least Squares (OLS) and Data Regression Analysis were the techniques used [10].

The paper titled “Academic self-efficacy and first year college student performance and adjustment.” published by Martin M. Chemers in the year 2001, investigated that self-efficacy can be considered an elementary feature in showing powerful relationships to academic progress and personal adjustment of the college students in the first year. A structural equation modeling (SEM) approach was used to test the adequacy of the hypothesized model [11].

In 2016, A. Bermejo Garcia published a paper titled “Student Performance Analysis” [12] where the analysis of the performance of a student was done with the help of techniques such as Machine Learning. Besides the above algorithms, the tools and techniques used were Logistic Regression along with Support Vector Machine.

3 Need for Proposal

The Higher Education department in the HR development ministry puts out an annual survey called the All India Survey On Higher Education (AISHE) [3]. It is supply-side and numerical. Based on the recent surveys, 65% of students who participate in an examination or the other secures a poor result despite taking enough time to appositely prepare for these tests. It, therefore, becomes extremely necessary to get to know the key responsible elements required to upgrade the performance of scholars.

All India Educational Survey (AIES) is conducted periodically. Its main purpose and objective is to collect, collate, and communicate information of the nation’s overall progress in the academics sector. There has been an improvement from 19.4% in 2010–11 to 25.2% in 2016–17, which is a significant achievement. However, AIES 2017 stated that the scenario is pretty staggering and a lot needs to be done. India, being a developing country has set an aggressive aim of attaining 30% GER in higher education by 2020.

Educational institutions will have to ramp up its efforts and get serious about what goes on in its organizations. For this, lot of analysis needs to be done, and that is why this research has come up with this introspection about the various factors that are responsible for the performance of a student.

4 Proposed Framework

A student’s performance can be evaluated on the basis of student’s learning outcomes. These outcomes can be in the form of various assessments. They give us essential information about what a pupil is learning and about the extent to which the teaching goals are being met. However, grading shall be the most optimum technique. Grading

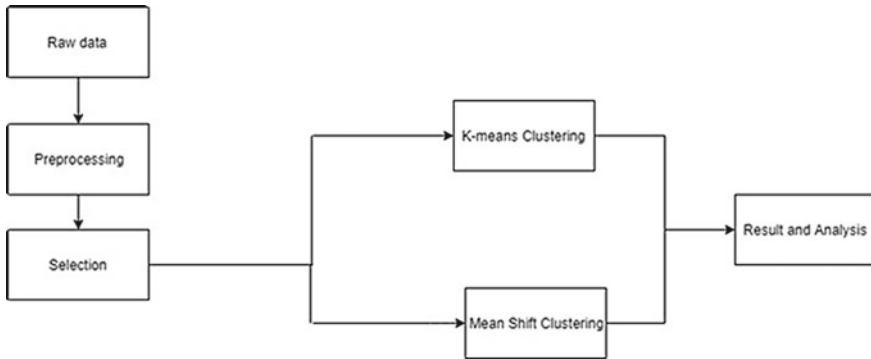


Fig. 1 Data Flow diagram for the proposed framework

is the application of the standardized measurements of varying levels of achievement. A grade is mainly classified as two types: Grade Point Average (GPA) and Cumulative Grade Point Average (CGPA). GPA is calculated by taking the number of grade points a student earned in a given period of time.

While there are various recognized correlating factors related to student’s academic progress, here in this investigation, this paper has scrutinized some basic key elements like age, sex, family size, parent’s cohabitation status, mother and fathers education, etc.

Based on the above key factors clustering techniques are used to analyze the performance of the student and check how they vary from one another depending on the data set variables. The data has been taken from the UCI Machine learning repository. It has been collected from the records of the two Portuguese schools using schools records and questionnaire. The dataset consists of attributes, namely age, sex, grade point for three different semesters, father’s job, mother’s job, extracurricular activities, romantic activities, travel time, number of absences of student, and alcohol consumption as mentioned above. The data consists of 394 students. The next step is preprocessing, in which any data with null values has been dropped. In the third step, after preprocessing, the data has been selected according to the requirements of the particular query. Then, the K Means and Mean shift clustering have been applied in python. Then, results so obtained have been discussed. A graph is plotted between GPA secured by a student and the above factors one by one. The clustering tool will help us to know what the factors are and is it directly or inversely proportional to the performance of the student (Fig. 1).

5 Experimental Setup

The hardware configurations of the desktop used are processor speed of 3.20 GHz, RAM of size 8.00 GB, and system type of 64-bit operating system with a ×64-

based processor. The software configurations used for this introspection include Python 3.6 version running on top of Windows 10. Python is an interpreted high-level programming language for general-purpose programming. It has a design philosophy that emphasizes code readability. scikit-learn, a python machine learning library was also used along with matplotlib, a plotting library. Package of Pandas has been used, which is a python package and provides fast, flexible data structures which manipulates numerical tables as well as time series. An active internet connection for data recovering over network was also required.

6 Result and Analysis

The analysis of student's performance helped us to know the exact factors affecting a student's performance. It gave a clear idea about which factors directly and indirectly lead to an increase in the overall accomplishments and what are the key elements that may affect adversely. This research compared two clustering techniques, namely k means and mean shift algorithm in our introspection to get a vivid conception.

This analysis gives answers to various questions, including the impact of father's and mother's job on a child's academic performance. Figure 2 shows the results obtained from mean shift clustering algorithm. Here, two clusters have been taken into consideration, namely high and low represented by red and blue colors, respectively.

However, when k means algorithm was applied to the same data, as shown in Fig. 3 It can be seen that three clusters are formed high, medium, and low represented by red blue and green colors, respectively. Thus, both the algorithms resulted in the output that parent's education is directly proportional to student's academic performance. Second, this research also analyzed how extracurricular activities and romantic relationships affect the progress of a student.

From Fig. 4, it can be implied that two clusters are there that are less active in extracurricular activities (blue) and more active students (red). It is observed that red clusters secure less grade point than the blue ones.

On applying K means algorithm, as shown in Fig. 5, it is observed that three clusters are formed. Blue denotes less involved in extracurricular activities, red denotes equal involvement while green denotes more inclination toward other activities. It can be implied from the two techniques that more involvement in extracurricular activities leads to less grade point. However, it cannot completely deny the fact that a little inclination toward these non-scholastic tasks help in the overall development too.

Other factors like health, age, sex, travel time of students to their respective institutions, family size, etc. are also some major key elements influencing the overall performance. It was observed that health status can be broadly categorized into five types—very poor being the lowest with poor, satisfactory, good being in the middle, and very good status implying no health issues. Students with less health issues secure more marks (GPAs) than those suffering with major problems.

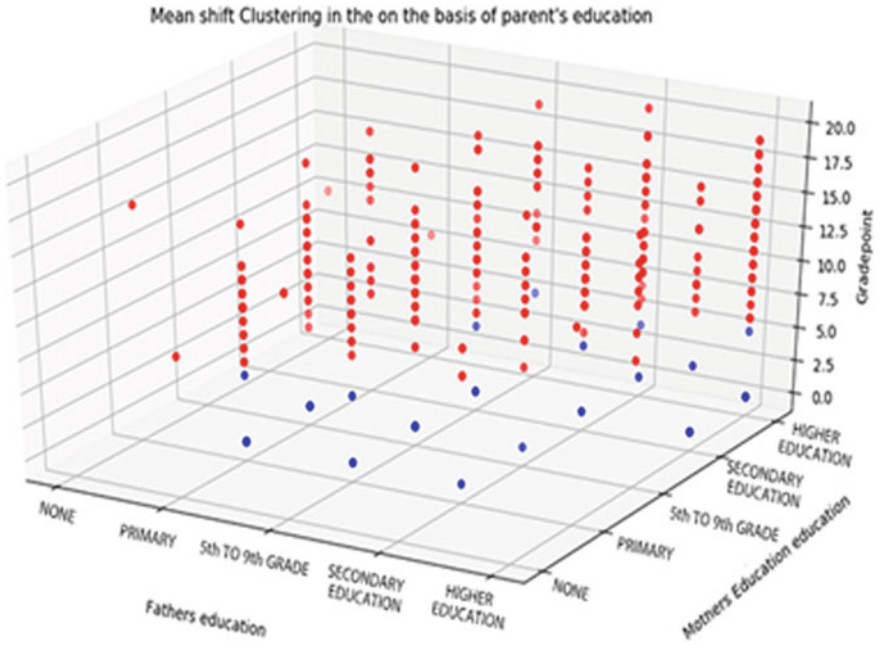


Fig. 2 Mean-Shift clustering on the basis of parent's education

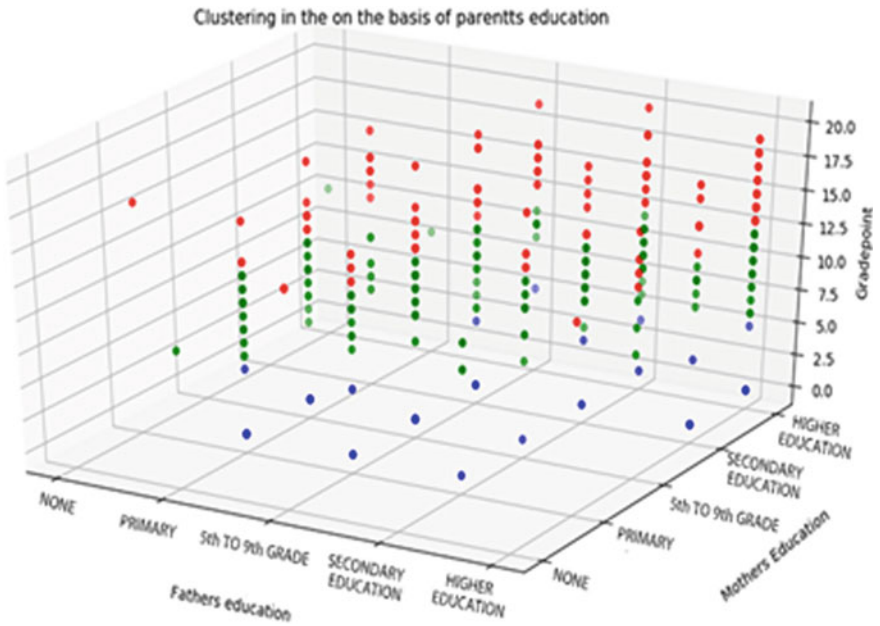


Fig. 3 K Means clustering on the basis of parent's education

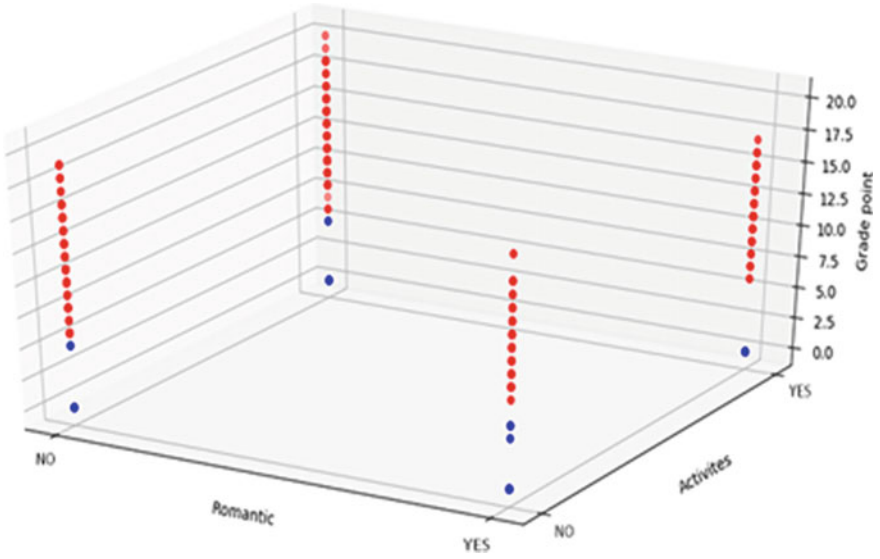


Fig. 4 Mean Shift on the basis of the impact of extracurricular activities

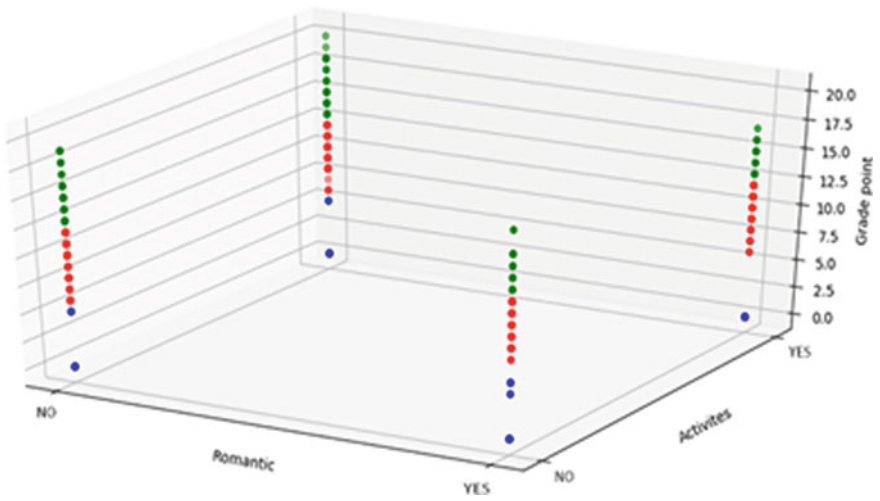


Fig. 5 K Means clustering on the impact of romantic and other activities

It was also observed that progress of those scholars was relatively poor for those who were absent from classes for a long period of time/more number of days, in comparison to those who regularly attended the classes. Thus, absence also had a major role to play. While health and number of absences are key elements, travel time also plays an important role. It was inferred from the investigations that students with travel time 15–30 min and less than 15 min could devote their time adequately to their studies and secured relatively better GPAs. While pupils with travel time more than an hour often performed less than the above. Similarly, both the clustering techniques came up with the output that those students who planned on going for higher education were better performers in terms of grade point average marks. While those scholars who were not inclined toward higher studies scored relatively less.

If travel time is being considered, an analysis regarding the study and the free time also needs to be done. Similarly, does rural and urban area have an impact on a student's performance?

Due to differences in social and organizational environments, there is a disparity of progress too. Better infrastructure, qualified teachers lead to quality learning and better academic achievement in urban areas. Thus, this gap needs to be bridged. While analyzing, one more question came across: Does the period have any impact on the marks of a student? k means clustering technique was used to find out that yes, it does have an impact. It could be implied that the percentage of students with first period grade had secured more marks in terms of grade point average (GPAs) than those who had second period.

Last, a graphical representation has also been done based on alcohol consumption and its impact on grade points. In today's generation large percentage of students consume copious amount of alcohol daily. This not only proves to be destructible for their health but also on their overall progress. Figure 6 reveals mean shift clustering where two clusters are formed: Low consumption of alcohol (Blue) and High consumption (Red). To perform a comparative study k means clustering was applied on the same data set. It can be implied from Fig. 7 that there are three levels of alcohol consumption, namely, less (blue), average (green), and high (red).

The workday (daily) and the weekend alcohol consumption can be categorized in numeric terms from 1.0 to 5.0, where 1.0 refers to very low consumption and very high consumption is denoted by 5.0. 2.0 represent low alcoholic consumption, 3.0 medium alcohol consumption and 4.0 high alcohol consumption.

The results were similar for the two techniques. It showed that consumption of alcohol was inversely proportional to their academic progress. It affected the GPA adversely. Students with level 1.0 were better performers than those with high level (>3.0).

Apart from the above results, from the techniques applied on the data set it can also be implied that students with large family size secure less grade points than those who belong to nuclear families. Two categories of family size had been taken into consideration: less than 3 (nuclear family) and greater than three. Similarly, in our analysis, it has been found out that the average marks for each failure groups and compared the results. Two groups have been made namely the ones with number of failures between one and three and the other ones with those greater than four.

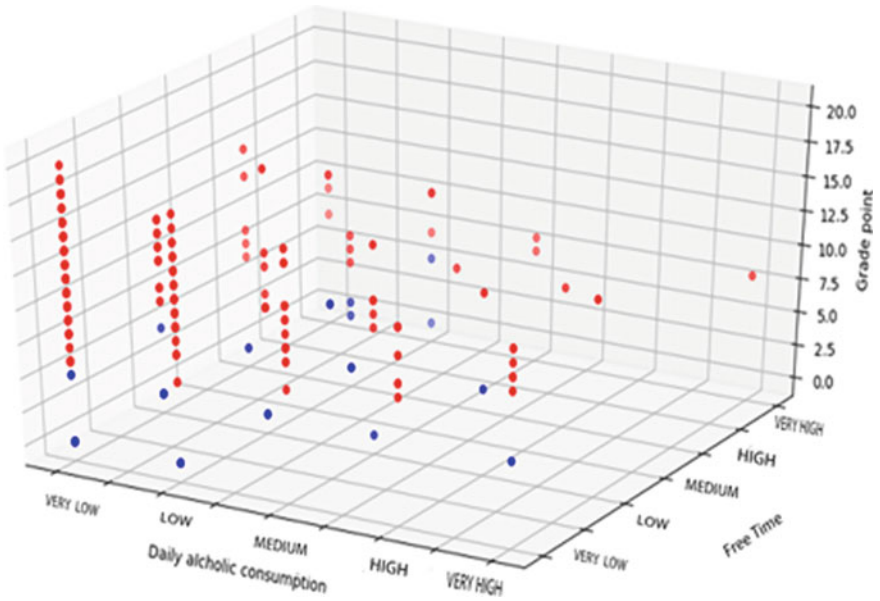


Fig. 6 Mean Shift clustering on the basis of daily alcohol consumption

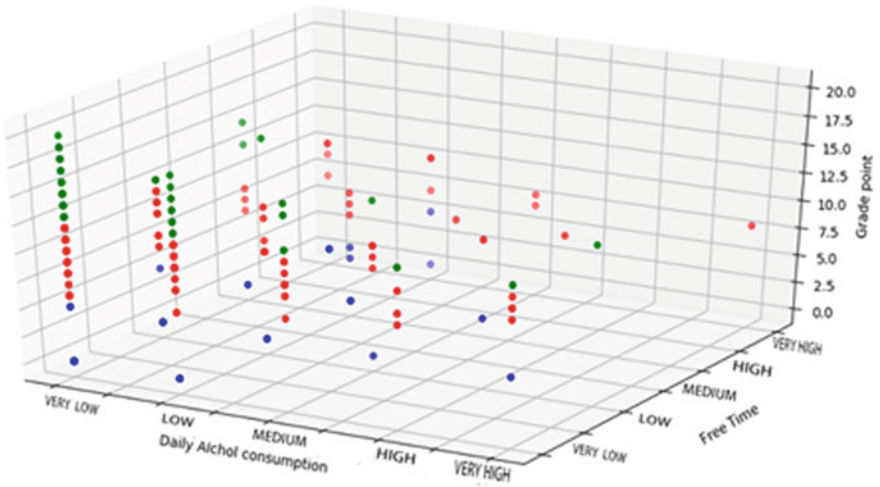


Fig. 7 K Means clustering on the basis of daily alcohol consumption

On plotting the graph, the students who have faced failure in the past were better performers than those who had not. These were the questions of some of the factors affecting the students' performance. This investigation was prone to find out the above results and do the analysis.

7 Conclusion

Today's youth is tomorrow's generation. The youth comprises of the students. The students play a vital role in society, they shape the future. It is, therefore, very important to see what is being shown today.

Results have shown that current and past health status of students affect their performance [13]. It also revealed that not only is the travel time of students from their homes to the institutions a major factor but also the very fact whether the student's parents are educated or not. Number of absent days along with one's involvement in extracurricular activities play a significant role. With increasing drinking culture among the students, not only is their health at risk but also their academic progress.

Big Data in today's world is of great use [14]. It is being used aggressively in the education sector also. As a result, a lot of work has already been done. However, a lot more is still left. Future is uncertain. Management of education needs to be one of the most coveted disciplines globally. It needs to be context specific and innovation should never end.

References

1. The journal for business education available from https://www.researchgate.net/publication/254344820_Academic_Performance_of_College_Students_Influence_of_Time_Spent_Studying_and_Working
2. Oyelade, O.J., Oladipupo, O.O., Obagbuwa, I.C.: Application of K means clustering algorithm for prediction of student's academic performance. *J.-Int. J. Comput. Sci. Inf. Secur.*
3. <http://aishe.nic.in/aishe/home>
4. Roy, C., Rautaray, S.S., Pandey, M.: Big data optimization techniques: a survey. *Int. J. Inf. Eng. Electron. Bus. (IJIEEB)* **10**(4):41–48 (2018). <https://doi.org/10.5815/ijieeb.2018.04.06>
5. Osmanbegovic, E., Suljić, M.: Data mining approach for predicting student performance. *Econom. Rev. J. Econom. Bus.* **X**(1) (2012)
6. Kekane, S., Khairnar, D., Patil, R., Vispute, S.R., Gawande, N.: Automatic student performance analysis and automatic student performance analysis and monitoring monitoring. *Int. J. Innovative Res. Comput. Commun. Eng.* (2017)
7. Priya, I.U.: Predicting academic performance of students using data mining technique
8. Sa, C.L.: Dayang Hanani bt. Abang Ibrahim, Emmy Dahlia Hossain, Mohammad bin Hossin "Student Performance Analysis System (SPAS)"
9. Ramesh, V., Parkavi, P., Ramar, K.: Published a paper titled "Predicting Student Performance- A Statistical and Data Mining Approach. *Int. J. Comput. Appl.* (2013)
10. Suryadarma, D., Suryahadi, A., Sumarto, S., Rogers, F.H.: Improving student performance in public primary schools in developing countries: evidence from Indonesia. Available from <http://www.tandfonline.com/loi/cede20>

11. Chemers, M.M., Hu, L., Garcia, B.F.: Academic self-efficacy and first-year college student performance and adjustment. *Am. Psychol. Assoc.*
12. García, Á.B.: Student Data Set
13. Das, N., et al.: *Big data analytics for medical applications* (2018)
14. Roy, C., Pandey, M., Rautaray, S.S.: A proposal for optimization of horizontal scaling in big data environment. *Advances in Data and Information Sciences*. Springer, Singapore, pp. 223–230 (2018)

Prediction of Employee Turnover Using Ensemble Learning



Shubham Karande and L. Shyamala

Abstract Employee turnover is now becoming a major problem in IT organizations, telecommunications, and many other industries. Why employees leave the organization is the question rising amongst many HR managers. Employees are the most important assets of an organization. Hiring new employees will always take more efforts and cost rather than retaining the old ones. This paper focuses on finding the key features of voluntary employee turnover and how they can be overcome well before time. The problem is to predict whether an employee will leave or stay based on some metrics. The proposed work will use the application of ensemble learning to solve the problem, rather than focusing on a single classifier algorithm. Each classification model will be assigned with some weight based on the individual predicted accuracy. The ensemble model will calculate the weightage average for the probabilities of the individual classification and based on this weightage average, an employee can be classified. Accurate prediction will help organizations take necessary steps toward controlling retention.

Keywords Employee turnover · Classification · Ensemble learning

1 Introduction

“You take away our top ten employees and we the Microsoft will become one of the mediocre companies”, this statement by Bill Gates is enough to prove the importance of the employee turnover. To retain employees is now become one of the major tasks of the Human Resource Managers (HR) of the company. One of the main goals of HRs is to retain their employees and make use of their knowledge for the growth

S. Karande · L. Shyamala (✉)
School of Computing Science and Engineering, VIT University,
Chennai Campus, Chennai 600 127, Tamil Nadu, India
e-mail: shyamalal@vit.ac.in

S. Karande
e-mail: Shubham.sadashiv2017@vitstudent.ac.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_29

of the company. The way companies can deal with this issue is predicting the risk of employee churn. Machine Learning Algorithms are frequently used in employee churn study. Implementation of these ideas in Employee Relationship Management (ERM) has now become new trend. Employee Turnover can be divided into two categories: voluntary turnover, where employee chooses to leave the company or retirement and involuntary turnover, where employer decides to let go the employee. Retirement is something which will not be needing prediction as it is legally enforced. We are focusing on voluntary turnover, therefore, involuntary turnover is out of the scope of this paper. The novel contribution of this paper is to explore the application of ensemble learning as an advancement of traditional algorithms. The objective behind the work is to provide an improved system to tackle the employee churn problem and give HRs of the companies a heads-up, so that they can plan some strategies to overcome the turnover [6]. The proposed work is comparing traditionally used classification algorithms with an ensemble learner which is combination of these same weak learners and by weighted average, we can give weights to these algorithms which will be the novelty if this paper.

The rest of the paper is organized as follows: we have discussed the related work in Sect. 2, Methodologies in Sect. 3 and Results in Sect. 4, and Conclusion in Sect. 5 which follows references at the end.

2 Literature Survey

The related work on the employee turnover is discussed here, at the first Rohit Punnoose et al. proposed a novel contribution of extreme gradient boosting in prediction of employee turnover, and comparison of XGBoost with six other historically used supervised classifiers [1]. The results showed that XGBoost gives higher accuracy, relatively low runtime, and efficient memory utilization than the other six. In the proposed work, the strongest predictors for voluntary turnover are job satisfaction, overtime, salary, distance from home, marital status, and employee's perception of fairness, an effective prediction model for predicting the employee turnover that have left the company has been introduced [3]. Decision tree is applied to predict the relevance of the attributes for turnover. This model can be used to decide employee will leave or not. A model is built using data from UCI repository to predict the status of employee turnover has been given. It uses three classification algorithms namely j48, bayesNet, and naive Bayes. The model was implemented using Weka and the best performing algorithm was j48 based on the accuracy [11]. An improved risk prediction clustering algorithm, which was multidimensional, was implemented to determine bad assets. In this work, primary and secondary levels of employee retention were used and association rule was integrated to avoid redundancy [4]. Two data mining models were developed for employee turnover to assist in decision. In this work, based on the accuracy obtained, regression model was found to outperform radial function model [10]. Three ensemble models were built and their performance in classifying the turnover as good risk group or bad risk group was

analyzed. The ensemble models were built using Adaboost, Bagging, Random Forest combined with three learning algorithms [7]. Ensemble machine learning algorithms were used to evaluate and decide the features which play a crucial role in predicting the risk involved in leaving the Company. Here, Tree-based classification was used and the algorithms were improved to favor the potential [2]. An improved ensemble algorithm based on automatic clustering and under sampling was proposed. In this method, clustering was done based on the weight of the samples and then a balanced distributed dataset was built which had a certain proportion of the majority class and all the minority class from each collection. By using Adaboost algorithm these datasets are used to build an ensemble classifier [5]. A methodology for improving the performance of the classification through ensemble learning was proposed. Here, classification was done using three different classifiers and the final classification was done by taking the majority voting from the classifiers [9]. Proposal for a customer churn problem in telecommunications industry, customer retention is given by Kim et al. [8]. Here, combination of SVM and PCA were used to not only get the higher accuracy but to boost the reliability of the model. All these references encourage us to build a model which will better the results, a model which is focusing on all the parts of architecture such as variable importance, algorithm selection, and performance matrices. Ensemble learner models can give the solution with weighted average at its tail. The approach used in the paper is ensemble learning model. Here, applied diverse machine learning algorithms on dataset to predict the employee turnover. Weighted average is calculated from the probabilities of individual classification, using which the final classification is done.

3 Methodologies

The various classifiers and techniques used in this paper are described in this section.

3.1 Support Vector Machines

Support Vector Machine is a supervised machine learning algorithm. SVM algorithm works by plotting the data points in n-dimensional feature space where n denotes the number of features. After plotting, depending on the number of dimensions, a line or a plane or a hyperplane is drawn separating the data points such that the data points in one side belong to one class and the data points on the other side belong another class making it as non-probabilistic binary linear classifier. The separating line is drawn in such a way that they are divided by a clear margin that is as wide as possible. New instances are then plotted into that same space and are classified as belonging to a class based on which side of the gap they fall.

3.2 *Random Forest Classifier*

Decision trees are of two types: classification and regression trees. A decision tree can be viewed as a flow chart like arrangement in which the internal node denotes test on a feature, each branch denotes the result of the test and each leaf node denotes the decision taken after calculating all features. However, the error rate is large for decision trees and they tend to overfit their training sets. A random forest is a meta-classifier that fits numerous decision trees classifiers on several subsamples on the dataset and use averaging or majority voting to increase the predicted accuracy and reduce overfitting.

3.3 *Logistic Regression*

Logistic Regression is another algorithm for Predictive Analysis borrowed from statistics. Despite the name Logistic regression, it is used from classification also. Unlike the other regression models, logistic regression does not try to give the value of numerical variable with given set of inputs instead it gives probability that the point belong to which class. Overfitting is very less for this model; therefore the model complexity becomes low. Equation (1) gives general logistic regression formula below,

$$g(E(y)) = \alpha + \beta x_1 + \gamma x_2 \quad (1)$$

The role of link function is to “join” the expectation of y to linear predictor.

3.4 *Variable Importance*

Classification trees analysis and Regression trees analysis can be collectively called as Classification and Regression Trees (CART) analysis. CART analysis produces a predictor ranking also known as variable importance on the basis of contribution predictors make to the building of the tree. Importance is decided by playing a role in the tree, either as a main splitter or as a surrogate. In this paper, random forest is used to calculate the variable importance. Instead of using all 34 features of classification, here selecting top 10 variables with better variable importance and use them for classification. This minimizes the time required to train the model.

3.5 Under Sampling

When training the model with an imbalanced dataset, the model tends to be biased toward the majority class. In this paper, the class attribute has 83.87% of one class and 16.23% of other class. Hence, by using under sampling we are making the proportion of the classes as 60 and 40%, respectively. By using this method, the biasing can be reduced while building the model and increase the sensitivity and specificity of the ensemble model.

3.6 Weighted Average Prediction

By making use of the accuracy from individual predictions of the classifiers, we are assigning weights to each classifier. In this paper, SVM was assigned 0.40, Logistic Regression was assigned 0.30 and Random Forest was assigned 0.30 as weights. By taking the product of probability that an instance will be assigned to a particular class for each classifier and the weights assigned to the respective classifier, we can predict the final classification. If the resultant of the product is greater than 0.5 it will be assigned to the class whose probability we took while taking the product otherwise it will be assigned to the other class. By using this approach we were able to increase the accuracy of the prediction for ensemble model. Example is given in Fig. 1.

3.7 Architecture Diagram

The proposed Ensemble Learning e Architectural is presented in Fig. 2 in this model, the dataset is taken and the features with variable importance are identified and collected. All these features are separately executed with different algorithms those are discussed earlier to get the individual score. Then to model the ensemble learning model, weighs are assigned for each model and given to the ensemble model. The classifier classifies and the accuracy is calculated.

	Model1	Model2	Model3	WeightAveragePrediction
Weight	0.4	0.3	0.3	
Prediction	45	40	60	48

Fig. 1 Weighted average prediction

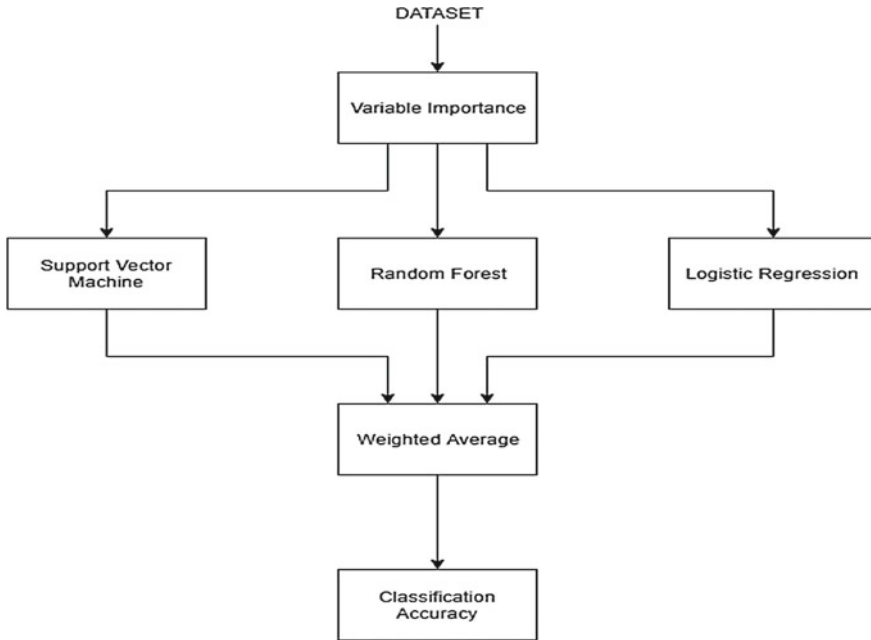


Fig. 2 Architecture diagram of ensemble model

4 Results

To validate the proposed method, the implementation is carried out on system with the configuration of Processor Intel Corei7, HDD 500 GB, RAM 8 GB, with Windows OS. The Tool used to implement the code is JetBrains Pycharm 2017.3.2.

4.1 Description of the Dataset

The dataset (<https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employeeattrition>) contains details of employee turnover from IBM Data, which is used for a case study of HR Analytics. The class attribute in the dataset, attrition represented as 0(employee did not leave) or 1(employee left). Exploratory data analysis is done on the dataset and it is revealed that the dataset is imbalanced in terms of target variable. 16.23% (employee will leave) of the total instances have the class variable “yes” and 83.87% of the total instances have class variable as “no” (employee will not leave). The results of the individual classifiers and the ensemble model are discussed in this section. Below given Table 1 represents the confusion matrix.

Table 1 Confusion matrix metrics and their definition

Actual class	Predicted class		
		No	Yes
	No	True positive	False positive
Yes	False negative	True negative	

4.2 Performance Metrics

The performance metrics used to validate the methodologies are given in Table 2 and these entries used to define and evaluate the performance of the classifiers discussed in this paper.

By making use of the confusion matrix, performance metrics such as accuracy, precision, sensitivity, and specificity for the classifiers and the ensemble model are calculated and their results are shown in Tables 3 and 4.

Table 2 Performance metrics and their definition

Metric	Equation	Definition
Accuracy	$(TP + TN)/(P + N)$	Ratio of the total number of predictions that are correct
Precision	$TP/(TP + FP)$	Ratio of the predicted positive cases that are correct
Sensitivity	$TP/(TP + FN)$	Ratio of the positive cases that are correctly identified
Specificity	$TN/(FP/TN)$	Ratio of negative cases that are correctly identified

Table 3 Confusion matrix for different methodologies used

Methodology	Reference	Prediction	
		No	Yes
Support vector machine	No	5954	1055
	Yes	956	1034
Logistic regression	No	6546	463
	Yes	1321	669
Random forest	No	6233	776
	Yes	839	1151
Ensemble model	No	6530	479
	Yes	961	1029

Table 4 Performance metrics

Methods	Accuracy (%)	Precision (%)	Sensitivity (%)	Specificity (%)
Support vector machine	77.65	84.95	86.16	49.50
Logistic regression	81.77	93.39	83.21	59.10
Random forest	82.64	88.93	88.14	59.73
Ensemble model	83.87	93.17	87.17	68.24

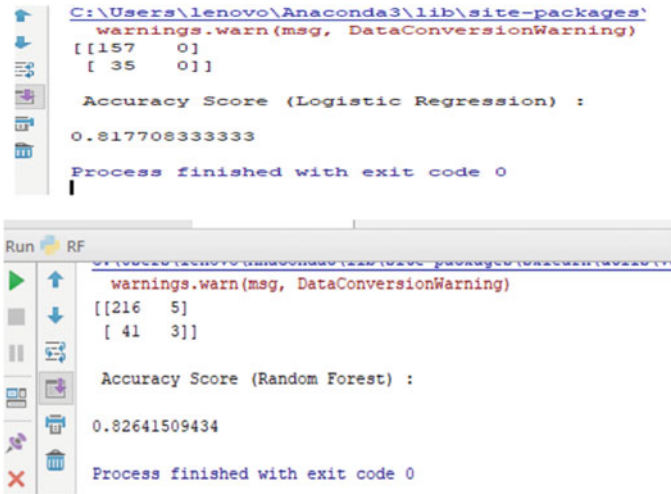


Fig. 3 Screenshots of implementation

From Tables 3 and 4, the result shows that the Ensemble model has more True positive and True negative compared to the other model. That means the prediction on the employees of their job interest to continue or not has been well predicted by this method than the other methods. Ensemble model outperforms in accuracy, precision, and specificity over other models. But its sensitivity of the data is slightly less than Random forest but good than other two techniques. From this, we conclude that the prediction accuracy of the ensemble model is better than other models. Figure 3 shows some screenshot of implementation.

5 Conclusion

The need of predicting employee turnover in companies and the use of machine learning algorithms in building these models were represented in this paper. The main challenge of building an Ensemble Learner Model which is a combination of Support Vector Machine, Logistic Regression, and Random Forest was highlighted. This model will be able to predict the employees turnover more precisely, based on

the accuracy obtained from the individual classifications weights were assigned, and calculated the weighted average. Based on the weighted average the final classification is done which gives an improved performance which is more superior to the results given by individual classifiers. In future, it can be fine-tuned more refrained feature to improve the sensitivity of the data.

References

1. Ajit, P.: Prediction of employee turnover in organizations using machine learning algorithms. *Algorithms* **4**(5), C5 (2016)
2. Alao, D., Adeyemo, A.B.: Analyzing employee attrition using decision tree algorithms. *Computing. Inf. Syst. Dev. Inf. Allied Res. J.* **4** (2013)
3. Cotton, J.L., Tuttle, J.M.: Employee turnover: a meta-analysis and review with implications for research. *Acad. Manag. Rev.* **11**(1), 55–70 (1986)
4. Holtom, B.C., Mitchell, T.R., Lee, T.W., Eberly, M.B.: 5 turnover and retention research: a glance at the past, a closer review of the present, and a venture into the future. *Acad. Manag. Ann.* **2**(1), 231–274 (2008)
5. King, G., Zeng, L.: Logistic regression in rare events data. *Political Anal.* **9**(2), 137–163 (2001)
6. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. *Emerg. Artif. Intell. Appl. Comput. Eng.* **160**, 3–24 (2007)
7. Kane-Sellers, M.L.: Predictive Models of Employee Voluntary Turnover in a North American Professional Sales Force Using Data-Mining Analysis. Texas A&M University (2007)
8. Kim, N., Lee, J., Jung, K.H., Kim, Y.S.: A new ensemble model for efficient churn prediction in mobile telecommunication. In: 2012 45th Hawaii International Conference on System Sciences, pp. 1023–1029. IEEE (2012)
9. Liaw, A., Weiner, W.: Classification and regression by random forest. *R News*, **2**(3), 18–22 (2017)
10. Peterson, S.L.: Toward a theoretical model of employee turnover: a human resource development perspective. *Hum. Resour. Dev. Rev.* **3**(3), 209–227 (2004)
11. Stovel, M., Bontis, N.: Voluntary turnover: knowledge management–friend or foe? *J. Intell. Capital* **3**(3), 303–322 (2002)

Automated Review Analyzing System Using Sentiment Analysis



A. C. Jishag, Vishnu Rakhesh, Suraj Mohan, N. Vinayak Varma,
Vaisakh Shabu, Lekshmi S. Nair and Maya Menon

Abstract An integral part of human behavior has always been to find out what others think about what would happen next. With the ongoing trend of e-commerce websites and personal blogs, people make active use of various technologies to understand and classify opinions. This paper introduces a new system to understand the emotions and feelings underlying the reviews provided by users on various e-commerce websites. This system holds an edge over the current rating system of star values by providing the users with a more precise and descriptive result. The main disadvantage of the star system is that it does not provide enough choice to the user. The methodology mentioned in this paper, named ARAS or Automated Review Analyzing System, overcomes this issue by using sentiment analysis which feeds upon each word in the review rather than a separate weighing system.

Keywords Sentiment analysis · Review system · Lexicon method · Web crawling

A. C. Jishag (✉) · V. Rakhesh · S. Mohan · N. Vinayak Varma · V. Shabu · L. S. Nair · M. Menon
Department of Computer Science, Amrita School of Engineering, Amritapuri
Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: jishagac@gmail.com

V. Rakhesh
e-mail: vishnurakhesh7@gmail.com

S. Mohan
e-mail: surajmohan44@gmail.com

N. Vinayak Varma
e-mail: vinayakvarma007@gmail.com

V. Shabu
e-mail: vaisakhshabu@gmail.com

L. S. Nair
e-mail: lekshmisn@am.amrita.edu

M. Menon
e-mail: mayamenon@am.amrita.edu

1 Introduction

Sentiment Analysis is a rapidly growing field of data analytics [1]. Sentiment is a thought, a judgment or an emotional response driven by a feeling. Sentiment Analysis, also known as opinion mining, deals with appraisal of peoples inclination towards an entity. An entity can be anything with a commercial value. Sentiment analysis has its foundation on the world wide web. From a users perspective, various contents related to their personal and public life can be hosted through social media platforms such as online social networking websites or e-commerce websites. From a technical perspective, most social media websites are empowered with built-in Application Programming Interfaces (APIs), prompting data collection and investigation by researchers and developers. Fundamentally, sentiment analysis affords a resourceful mechanism with the support of online data [2].

We propose Automated Review Analyzing System (ARAS), a new approach for assessment of customers market sentiments toward a product or service, though their transactions on of e-commerce websites. The prevailing system is the Star system, which analyzes a customers sentiment of a product expressed by a star value between 1 and 5 (Fig. 1). The basic flaw of this strategy is the lack of choice. Users are only given five options to rate the product. This can affect the realistic evaluation of customers experience that the users provide. There are categories of users who feel the five-star rating is insufficient to express their views. People who wish to rate a product less than 1 star or those who wish to rate a product between 1 star and 2 stars, etc., belong to this category. Some people feel that a particular product does not deserve even a single star or at other instances, a user feels that the product (or service) deserves more than 5 stars. For example, a user who had a horrible experience with the product is constrained to rate it 1 star whereas a user who had the best experience is unable to rate it over 5 stars. On the other hand, Sentiment analysis gives the user express more precise feelings about a product or service in the market. This is because, unlike the star system, sentiment analysis feeds off the reviews written by users. A user is free to use the complete English vocabulary to express his/her feelings.

Fig. 1 Review system at Amazon

Star Level	General Meaning
	I hate it.
	I don't like it.
	It's okay.
	I like it.
	I love it.

We propose an Automated Review Analyzing System (ARAS), a new approach for assessment of customers market sentiments toward a product or service, through their transactions on of e-commerce websites. The prevailing system is the Star system, which analyzes a customers sentiment of a product expressed by a star value between 1 and 5 (Fig. 1). The basic flaw of this strategy is the lack of choice. Users are only given five options to rate the product. This can affect the realistic evaluation of customers experience that the users provide. There are categories of users who feel the five-star rating is insufficient to express their views. People who wish to rate a product less than 1 star or those who wish to rate a product between 1 star and 2 stars, etc., belong to this category. Some people feel that a particular product does not deserve even a single star or at other instances, a user feels that the product (or service) deserves more than 5 stars. For example, a user who had a horrible experience with the product is constrained to rate it 1 star whereas a user who had the best experience is unable to rate it over 5 stars. On the other hand, Sentiment analysis allows the user express more precise feelings about a product or service in the market. This is because, unlike the star system, sentiment analysis feeds off the reviews written by users. A user is free to use the complete English vocabulary to express his/her feelings axed weight to every possible emoticon, indicating whether the opinion is positive, negative, or neutral.

2 Related Works

Marketing intelligence is the external data collected by a company about a specific market which it wishes to enter, to make decisions. It is the first set of data which the company analyses before making any investment decision. An important hurdle for Market Intelligence is the process of gathering consumer opinions of competing products and establishing thresholds as to benchmarks for products and services. Currently, no automated system is able to perform visual comparison of consumer opinions as proposed in this paper.

One of the main problems in sentiment analysis is the categorization of polarity [3]. Given a text span, it requires some intelligence or learning to categorize the text into a specific polarity. In [4], authors have defined three category document-level, sentence level, and entity aspect level The document-level is concerned about the overall sentiment of the document, While sentence-level and entity aspect-level analysis human opinion.

Singular value decomposition (SVD) is a means of decomposing a matrix into a product of three simpler matrices. Algorithms on machine learning and sentiment analysis normally use Singular Value Decomposition (SVD) based feature for sentiment prediction as it can capture the latent relation among the data. In [5], authors have suggested classification using SVM.

Hu and Liu [6] proposed several methods to analyze customer reviews from a corpus containing all types of reviews, such as negative, positive, and neutral. They identified the customer opinions and emotions and classified them as positive or

negative, ignoring the neutral ones. However, this technique is primarily based on unsupervised itemset mining, which is amenable for analysis when the user writes a complete, fully formatted review. This technique cannot be implemented for reviews which are very brief. Recent works [PSM14] [7] introduced a new model that allows neural nets (NN) to evaluate contextual sentiment: The proposed Global Belief-Recursive Neural Network (RecNN) represents the granular sentiment analysis. A backward step from upper tree nodes is introduced their methodology in order to faithfully capture contextual sentiment. A different approach is obtained by considering aspect-specific sentiment analysis, using hierarchical deep learning according to [LSM] [8]. Here, separate aspect sentiment (SAS) models or Joint Multi-Aspect Sentiment (JMAS) models train root node-level softmax classifiers, a commonly used classifier after SVM with a different loss function, with aspect and sentiment as classification outputs. However, a relabeling of product features is necessary for these models, mentioned above by Hu, M. and Liu, B. Their work should explore different recurrent neural networks (RNN) including bi-directional recurrent NN and Long-ShortTerm Memory (LSTM) RNNs that try to capture aspect-specific sentiment through context. In [9], authors have performed classification of text using various algorithms. Pang and Lee [10] suggested removal of objective sentences by extracting the subjective ones among them. They proposed a text-categorization technique that can be used in the identification of subjective content. A novel semi-supervised approach was proposed by Sanagar Gupta to construct polarity lexicon using iterative Latent Semantic Analysis technique [11] from unlabeled multiple source domains corpus. This technique was proven to have a considerable gain in the accuracy compared to the models until then. Hu and Liu [12] summarized a list of positive and negative words based on customer reviews. While the positive list contained 2006 words, the negative list had 47,883 words. The problem with this classification scheme was that both the lists contained many misspelled words, a common circumstance in the digital field. In [13], the authors have performed classification about 6799 tokens of Twitter data to identify sentiment score. Rafeek Remya in [14] discussed methods to detect polarity in sentiment analysis regression models. He considered product review dataset and the evaluation results obtained showed an improved classification accuracy.

3 Methodology

Numerous websites/companies resort to Sentiment Analysis as a process to appraise the current trends in customer satisfaction or product acceptance. On the other hand, users provided with the star rating system, discussed earlier, are restricted to provide an unfeigned assessment of the quality of any product or service. Companies rely on sentiment analysis rather than the star system, as the former offers their Customers unhindered options to express their experience of a product or service.

3.1 Web Scraping

Web scraping, also known as screen scraping, web data extraction, web harvesting, etc., is a technique used to extract large chunks of data from online websites. The retrieved data is then stored in a file in the local computer network, usually in the CSV format. Web scrapers perform the retrieval of data from websites. Web scraping (aka web crawling) can be divided into two steps, fetching the web pages, and extracting the web contents from the page. The contents of the web page may be searched, parsed, copied or reformatted into a spreadsheet. Our method would scrap reviews from a given Uniform Resource Locator (URL) and perform sentiment analysis on the retrieved data. The user needs to provide the URL of the website of interest. Web scraping is performed on the web pages directed by this URL and the retrieved data is stored in a local directory as a comma separated variable (csv) file.

3.2 Sentiment Analysis

Sentiment Analysis can be performed in two ways: Lexicon method and machine learning. In the Lexicon approach, sentiment value is obtained by comparing the dataset being reviewed with two pre-defined value sets, whereas in machine learning approach, the dataset is divided into test set and training set. Training set is used for the learning process, followed by processing the test set for prediction of the sentiment value. The proposed model uses the lexicon method rather than Machine learning to find the sentiment value of the reviews. This is because even though reviews are plentiful in any e-commerce website, a corpus generated from these reviews is not generally long enough to perform adequate supervised learning on. Initially, the whole review set is made into a corpus. This corpus would contain stop words, punctuation, and numbers, etc., which provide no impact on the sentiment value. Hence, these unwanted peripheral items are removed in the preprocessing step. Punctuation numbers are removed first followed by the removal of white spaces. All the words in the corpus now are transformed into lower case. Next, a predefined dictionary of stop words is now loaded and is compared with the corpus to eliminate stop words. The process of Stemming is now applied to the corpus and the final, preprocessed corpus is made into a term–document matrix. A document-term matrix represents frequency of words with respect to collection of documents in matrix form. Each term in the matrix is now compared with two dictionaries, positive lexicon set and negative lexicon set. Positive Lexicon set would contain words that provide a positive sentiment to the sentence negative lexicon set would contain words that provide a positive sentiment to the sentence. Each word is tagged with a weight with respect to the intensity of the sentiment it provides to the context.

4 Experimentation Result

This comparison results in obtaining the overall sentiment of each review, which in turn gives the overall sentiment of the product mentioned in the reviews. This obtained value, which would give us the percentage acceptance or rejectance of the product in the market is then used to plot different graphs. Graphs provide greater insight into the growth and decay of the products market performance. Four graphs are generated for this purpose. Figure 2 is a graph that shows the sentiment value obtained for the first 100 reviews. This graph will help the user understand the latest ratings of the hotel/product, i.e., if the hotel was receiving better reviews lately or were the sentiment value going down ever since. Every value in the star system has an upper bound and a lower bound. For example, 1 star can be taken in place of any value between 0 and 20%. This lower range and higher range is compared with the ARAS-generated sentiment value in Fig. 3, where its X-axis contains the name of each hotel while the Y-axis shows the percentage sentiment value. Figure 4 is an example of the results of a hotel, JW Marriot, shown as a bar graph portraying the percentage of positive and negative reviews. It acts as a final answer to the question of how good the hotel is?

This indicates the inefficiency of the star system in delivering a precise rating to the user. Also, at odd times, the ARAS ratings fluctuated marginally from the star reviews. While going through the reviews, it was found that this happened for those cases where the users had written more negative comments against the hotels and had rated it between 1 and 3. For example, for a one-star rating, its highest possible value is 20%, while the users who gave such a rating were normally those who completely disliked the hotel. Reviews from such users would contain maximum

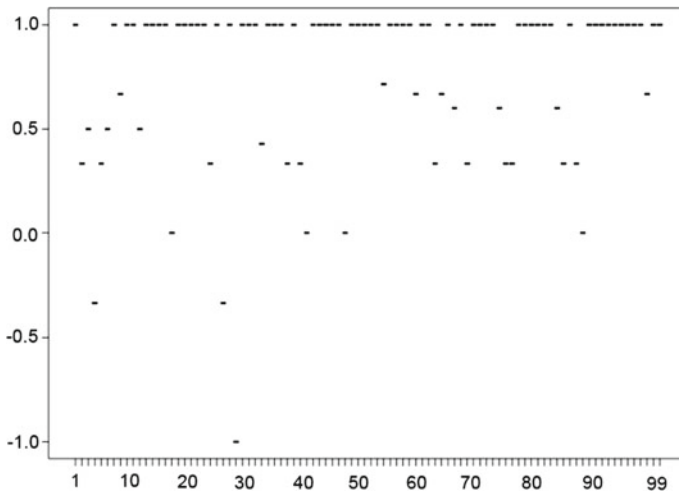


Fig. 2 Graph containing sentiment value for the first 100 reviews

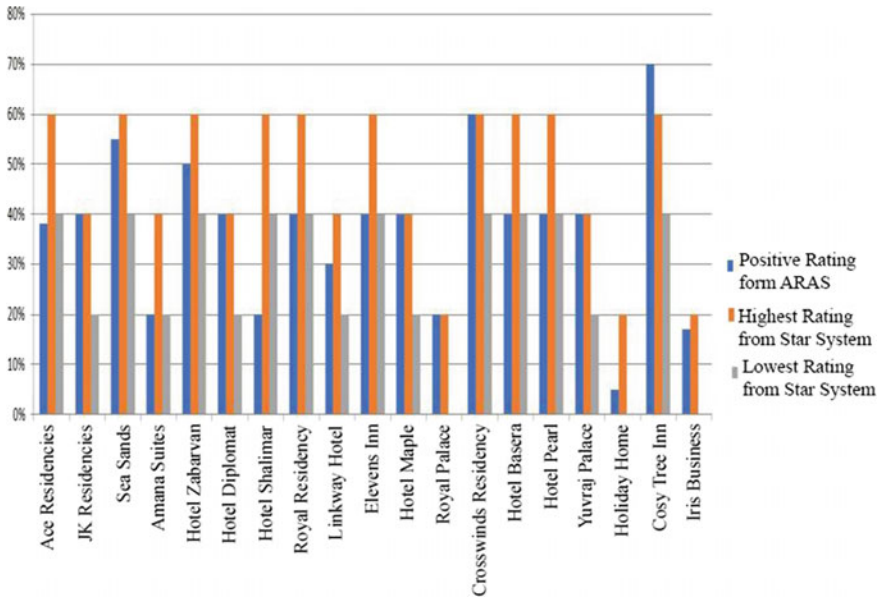


Fig. 3 Sentiment value generated for each hotel

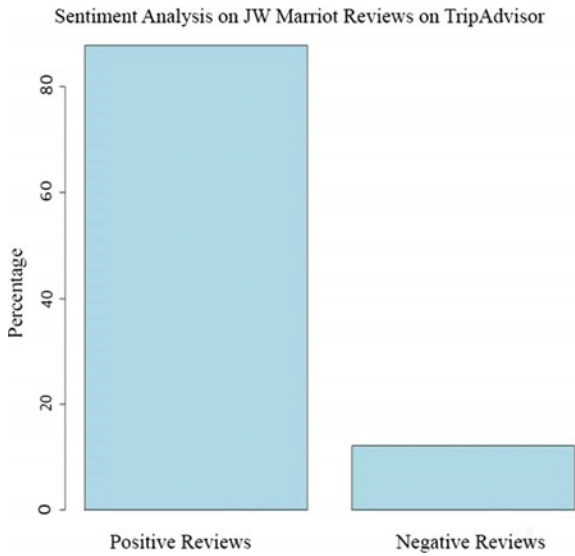


Fig. 4 Sentiment Analysis result generated for each hotel

number of negative words for the hotel than the positive words since the ARAS reviews are completely based upon the sentiment analysis on this review set, it showed the deviation.

ARAS-generated sentiment values were empirically verified by comparing it with the results of the star system for different hotels in TripAdvisor. The main disadvantage about the star system observed from this experiment was the lack of accuracy. Normally there are five-star ratings available for a product. Hence 1 star in the star ratings can be equated with any value between 0 and 20%, mathematically. On the other hand, ARAS provides a precise absolute sentiment value to each product. Hence for the comparison, we had to consider an upper limit, highest star value possible, and a lower limit, lowest star value possible, for each of the star rating.

As Fig. 3 indicates, result generated by ARAS mostly stood in between the highest and lowest values of the star rating. This indicates the inefficiency of the star system in delivering a precise rating to the user. Also, at odd times, the ARAS ratings fluctuated marginally from the star reviews. While going through the reviews it was found that this happened for those cases where the users had written more negative comments against the hotels and had rated it between 1 and 3. For example, for a one-star rating, its highest possible value is 20%, while the users who gave such a rating were normally those who completely disliked the hotel. Reviews from such users would contain maximum number of negative words for the hotel than the positive words since the ARAS reviews are completely based upon the sentiment analysis on this review set, it showed the deviation.

5 Conclusion and Further Improvements

We have done experimentation in a larger database considering reviews from 10 different websites and attained an accuracy of 94%. The future scope for this model on the followings.

5.1 *Semantics*

Our algorithm computes the overall sentiment of user reviews. Polarity plays a big role in sentiment analysis. For example, imagine the case of a review where the user praises the hygiene of the hotel but is disappointed with its locality. This is a positive review for a user, who gives priority to the hygiene but on the other hand a user who cares more about the locality would rate this as a negative review. This difficulty of classification was overcome in the suggested model as the classification depended on the query term this emphasizes on the semantics.

5.2 *Part of Speech (POS) Tagger*

The POS tagger takes a notably long time for training the system. This, in turn, led to a big spike on the execution time of the model. It did improve the computed accuracy, but we did not have enough time to conduct these tests.

5.3 *Support Vector Machines*

Pang and Lee [15] showed that SVM performed the best when classifying movie reviews as positive or negative. An important next step would be to further explore SVM parameters for classifying reviews.

5.4 *Dealing with Contrapositive Word*

The model suggested in this paper does not support contrapositive words.

5.5 *Machine Learning*

Use of machine learning instead of lexicon method can considerably improve the accuracy of the system. Though it also adds on the complexity of the system. However, machine learning will increase the execution times as the machine has to learn on its own rather than comparing lexicon sets.

Acknowledgements We are extremely thankful to Computer Science Engineering Department, at Amrita Vishwa Vidyapeetham, for providing all the resources and facilities for conducting the experimentation.

References

1. Kim, S.-M., Hovy, E.: Determining the sentiment of opinions In: Proceedings of the 20th International Conference on Computational Linguistics, p. 1367. Association for Computational Linguistics, Stroudsburg, PA, USA
2. Xing, F., Zhan, J.: Sentiment analysis using product review data, Department of Computer Science, North Carolina AT State University, Greensboro, USA (2015)
3. Chesley, P., Vincent, B., Xu, L., Srihari, R.K.: Using verbs and adjectives to automatically classify blog sentiment. *Training* **580**(263), 233 (2006)
4. Liu, B.: Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*. Morgan Claypool Publishers

5. Thara, S., Sidharth, S.: Aspect based sentiment classification: SVD features. International Conference on Advances in Computing, Communications and Informatics, ICACCI (2017)
6. Hu, M., Liu, B.: Mining and summarizing customer reviews. *KDD* **04**, 2004 (2004)
7. Paulus, R., Socher, R., Manning, C.D.: Global belief recursive neural networks. In: Advances in Neural Information Processing Systems 2014, pp. 2888–2896
8. Lakkaraju, H., Socher, R., Manning, C.: Aspect specific sentiment analysis using hierarchical deep learning. In: Advancesmensts and Applications of Deep Learning in computer Systems (2014)
9. Vijayan, V.K., Bindu, K.R., Parameswaran, L.A.: Comprehensive study of text classification algorithms. International Conference on Advances in Computing, Communications and Informatics, ICACCI (2017)
10. Pang, B., Lee, L.L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, ACL 04. Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
11. Sanagar, S., Gupta, D.: Adaptation of multi-domain corpus learned seeds and polarity lexicon for sentiment analysis. International Conference on Computing and Network Communications, CoCoNet (2015)
12. Hu, M., Liu, B.: Mining and summarizing customer reviews In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177. ACM, New York, NY, USA (2004)
13. Gann, W.-J.K., Day, J., Zhou, S.: Twitter analytics for insider trading fraud detection system In: Proceedings of the Sencond ASE International Conference on Big Data. ASE (2014)
14. Rafeek, R., Remya, R.: Detecting contextual word polarity using aspect based sentiment analysis and logistic regression. IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials, ICSTM 2017—Proceedings (2017)
15. Pang, B., Lee, L.: Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2(1–2):1135 (2008)

A Comprehensive Survey on Extractive and Abstractive Techniques for Text Summarization



Abhishek Mahajani, Vinay Pandya, Isaac Maria and Deepak Sharma

Abstract Over the years as the technology advanced, the amount of data generated during the simulations and processing has been constantly increasing. Techniques for creating synopses of this massively generated data have been in the forefront of the research in the recent times. Text Summarization was one such aspect of the research which focused on representing the idea of the context in a short representation. Efforts were put to create a system which was able to generate effective summaries providing an overview of all the ideas represented by the article. Text Summarization techniques can be broadly classified into Extractive and Abstractive Text Summarization techniques. The paper compares all the prevailing systems, their shortcomings, and a combination of technologies used to achieve improved results. The paper also draws attention towards the state-of-the-art standardized datasets used in developing the summarization systems. The paper also focuses on testing parameters and techniques used to test the efficiency of the summarizing systems.

Keywords Extractive text summarization · Abstractive text summarization

1 Introduction

As of late, Information is considered as an essential resource which can have diverse applications running from quantifiable purposes to learning portrayal purposes. With the measure of information gathered every day, there is plenty of data accessible and

A. Mahajani (✉) · V. Pandya · I. Maria · D. Sharma
Department of Computer Engineering, KJSCE, Mumbai, India
e-mail: a.mahajani@somaiya.edu

V. Pandya
e-mail: vinay.hp@somaiya.edu

I. Maria
e-mail: isaac.m@somaiya.edu

D. Sharma
e-mail: deepaksharma@somaiya.edu

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_31

available to us. In such conditions, it becomes imperative to separate all the vital data while neglecting the monotonous and less noteworthy data. Thus, the process of summarization turns into an extremely difficult assignment since it needs to catch all the critical data and aggregate it into a record which is semantically and etymologically right while mulling over the reasonability of the data mixed. It is astoundingly troublesome for individuals to comprehend large contexts, particularly when it is vast. A solution to this problem is having an efficient summarization technique capable of condensing the information in a concise manner while preserving the semantics. So, what does an effective summarization mean? The most accurate summary produced can be defined as the one which takes into account all the aspects of the content, prioritizes the most eminent information in the document, and creates a document which is significantly smaller than the original content. As mentioned above, summarization techniques are classified into Extractive and Abstractive techniques. Extractive Text Summarization as the name suggests extracts important information from the original content. It then combines this information to form a shorter representation of the original context. This involves identifying the important sentences in the article and combining them to form a summary. On the other hand, Abstractive Summarization means creating an abstract of the original content. The generation of summary is done by understanding the whole content and representing it in own terms. This is achieved using a Recurrent Neural Network consisting of Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM) cells. There are different approaches to implement both the techniques which will be discussed subsequently.

The paper discusses the different methods of summarization in both abstractive and extractive techniques. In Sect. 2, techniques of summarization used in past are discussed. Section 3 focuses on different extractive techniques used for summarization. Section 4 highlights the recent abstractive techniques used for text summarization. Section 5 provides information on the standardized datasets and testing methodologies used to evaluate the performance of the system. Finally, the Conclusion and Future Scope is discussed in Sect. 6.

2 Summarization Over the Years

Headways in Text Summarization rose in the early 50s. Text Summarization can be comprehensively described into two categories, viz., Extractive Text Summarization and Abstractive Text Summarization. Extractive Text Summarization is the method of extracting content from the document and combining it to form a text smaller in size. Abstractive Text Summarization takes summarization to a stride further. It is capable of depicting information by creating new sentences. Abstractive Summarization can be divided into Structured and Semantic approaches. Each of these classifications can be subdivided into subcategories based on various methods [1].

Structured approach fundamentally encodes the most indispensable information from the document(s) through mental blueprints like layouts, extraction principles, and elective structures like tree, ontology, rule, and graph structure.

- In tree-based approach [2], sentences from multiple documents are clustered according to the themes they represent. Second, these themes are re-ranked, selected, and ordered according to their significance. This is followed by the identification of common information using syntactic trees. The syntactic trees formed are subsequently merged using Fusion Lattice Computation to assimilate information from different themes. Linearization is carried out for the formation of sentences from the merged tree using Tree traversal.
- Ontology-based method [3] is based on the predefined knowledge base to create summaries belonging to a particular domain, i.e., it is domain specific. In this approach, the domain ontology for generating Chinese news articles is laid out by the area specialists. The archive preprocessing stage creates the important terms from the news corpus and furthermore the Chinese news lexicon. The important terms generated are classified using term classifier and the classified meaningful terms are passed to fuzzy inference mechanism to generate fuzzy ontology.
- Rule-based method [4] is based on random forest classification and feature scoring. The scoring is based on the constraints laid down by the user. The rules can be set in many ways such as: using verbs and nouns which are related to each other; keywords and syntactic constraints; domain constraints.
- Graph-Based Approach [5, 6] uses the graph data structure for language representation. Here, every word unit is represented by a node and the structure of the sentences is determined by directed edges. These edges represent the relationship between any two words. The underlying feature of this method is that it uses a shortest path algorithm to find the smallest sentences with considerable amount of information. The sentence formation is subjected to constraints such as it is mandatory to have a subject, verb, and predicate in it. Along with this, a compendium is used for Linguistic and Summary Generation purposes.

General notion in Extractive Text Summarization is to weight the sentences of a document as a function of high-frequency words, disregarding the very high-frequency common words. Extractive summarization framework later, in addition to the abovementioned method (i.e., recurrent dependent weights), also used the following approaches for deciding the sentence weight [7].

- Cue Method [8] depends on the hypothesis that the significance of a sentence is ascertained by the nearness or nonattendance of certain prompt (imply) words in the sign lexicon. This is like consideration instrument wherein the emphasis is on words which draw our consideration and are exceptionally impactful which helps in understanding the unique circumstance.
- Title based method [9] takes into consideration the heading and subheading in a document. For the most part, the heading labels speak to the entire thought in a document or a section in a couple of words. Following this approach will help in understanding an expansive perspective of the specific situation. Be that as it may, the data absorbed utilizing this technique is exceptionally constrained and can be deceiving.
- Location Method [10] exploits the idea of identifying important information in certain part of context. It is an apparent thought that the content at the beginning

or any record can be dealt with as the presentation and that it will give a general thought regarding what the report depends on. Moreover, the completion is considered as the finish of what is talked about in the record and can help in understanding the general result of the archive. This technique skirts the definite data specified in the body some portion of the archive. This is on account of the body of any setting comprises of point by point data expounding the given thoughts.

Other than this, sentence extraction should be possible utilizing Neural Network Architectures. One of these strategies is a [11] classifier which includes navigating the archive consecutively, and choosing whether to include the sentence into the rundown. Extractive techniques include picking sentences in an arbitrary way. Classifier architecture has been turned out to be superior to select design after different test investigations. Other technique [12] centers around the side information of the report. The pith of the substance lies in side information, for example, title, picture inscription, and twitter handle. In this manner, the given structure abridges an archive by making utilization of report encoder and consideration instrument over side data.

3 Summarization Using Extractive Techniques

Recently, Extractive text summarization is conducted using neural model. The advantage of using this method over the traditional pure mathematical and NLP techniques is to understand more contexts. The models improve the depiction of sentences by combining the important sentences to shorten the size and maintain the semantics at the same time.

A similar method was implemented [13] by (Narayan et al.) to achieve summarization using sentence extraction. The model was based on an encoder—decoder architecture consisting of RNNs and CNNs. The idea behind it was to acquire portrayals of sentences by applying single-layer convolutional neural systems over sequences of word embeddings and then depend on a recurrent neural network to make groupings out of sentences. CNNs were used to capture the important patterns amongst the sentences in the article. This CNN was used to extract the sequence of words in a sentence. Similarly, a document encoder was used to identify the sequence of sentences to get the document representation. Figure 1 shows the functioning of such a neural model.

Another method proposed by Erkan and Radev [14] is based on stochastic graph-based method. The method performs a multi-document summarization by considering three measures, viz., Centrality Degree, LexRank with threshold, and continuous LexRank. Centrality Degree identifies the correspondence of a sentence to other sentences. This helps in eliminating sentences with less similarity. To avoid important sentences getting avoided, a concept of page rank is used to calculate the centrality degree based on the centrality degree of its neighbors. The links between the sentences are weighted. This helps in strengthening the links having higher similarity

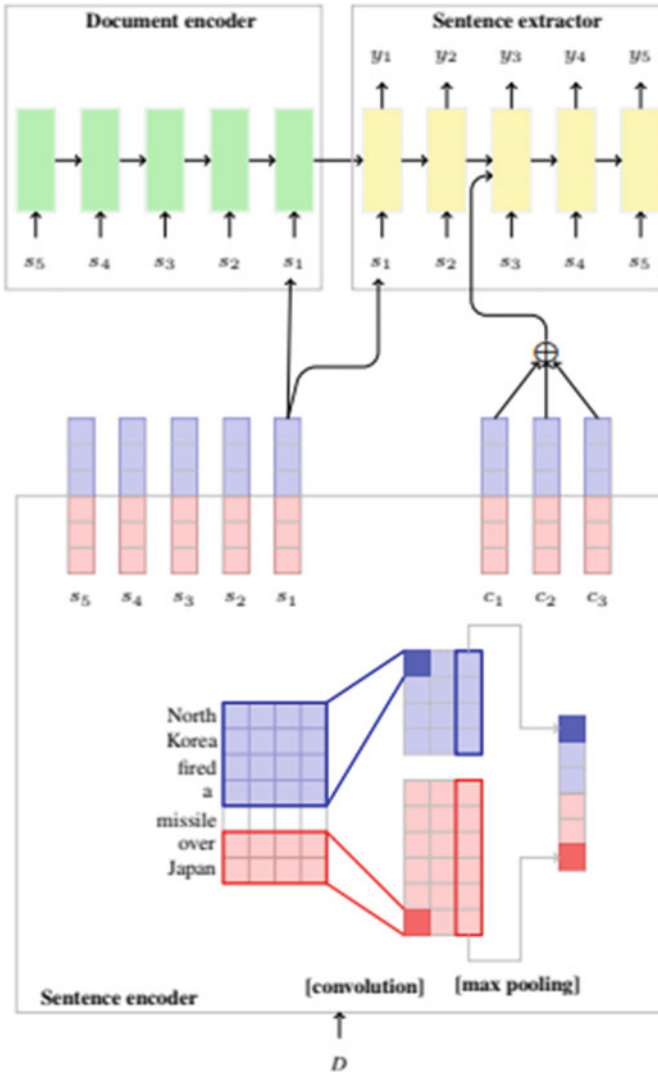


Fig. 1 Neural model for extraction of important sentences

instead of having a binary representation of the links. This creates a stochastic matrix for representing similarity.

Method proposed by Steinberger and Ježek [15] used singular value decomposition (SVD) in summarization. SVD is capable of capturing and re-modeling interrelationships for identifying similar sentences and terms. SVD of each term is computed using sentence matrix. Now, for each sentence vector in matrix V (its components are multiplied by corresponding singular values), we compute its length. The reason

of the multiplication is to favor the index values in the matrix V that correspond to the highest singular values (the most significant topics). The calculation is given by the following equation [15].

$$S_k = \sqrt{\sum_{i=1}^n v_{k,i}^2 \sigma_i^2}$$

where S_k is the length of the vector of k th sentence in the modified latent vector space. It is its significance score for summarization too. It is a number of dimensions of the new space.

This value is independent of the number of summary sentences (it is a parameter of the method). In our experiments, we chose the dimensions whose singular [15].

4 Summarization Using Abstractive Techniques

Lately, Abstractive Summarization has been achieved using a sequence to sequence encoder–decoder model. This model has its famous application in Neural Machine Translation. These language models are capable of taking an input of size N and give an output of size M . It is essentially used to preserve the dependencies LSTM cells are used. These cells are the most atomic unit of an encoder and decoder. Similar to LSTM, GRU cells can also be used at the expense of some accuracy.

There are ways in which Encoders have been designed over the few years. The most basic implementation is to use a bag of words. These encoders are capable of capturing the important words while ignoring the relationship between the neighboring words.

A [16] better method to address the problem of dependency among different words is the use of a convolution encoder or a sequence to sequence model. TDNN is used to achieve this along with max pooling layers. TDNN receives the output of the bottom layer to the input layer. This helps in preserving the dependency over long sentences.

Recently, Attention Models have been extensively used for various applications including Text Summarization. Attention Mechanism helps in laying emphasis on certain important words. Attention Mechanism can be implemented in two ways wherein either the same hidden units can be used for computing the attention weights or certain hidden units can be kept aside specifically for calculation of Attention Weights. The attention weights for a particular word are calculated by using the formula 1 [16].

$$a_{y'}(t) = \frac{\exp(h_{x_t}^T h_{y'})}{\sum_{t'} \exp(h_{x_{t'}}^T h_{y'})} \quad (1)$$

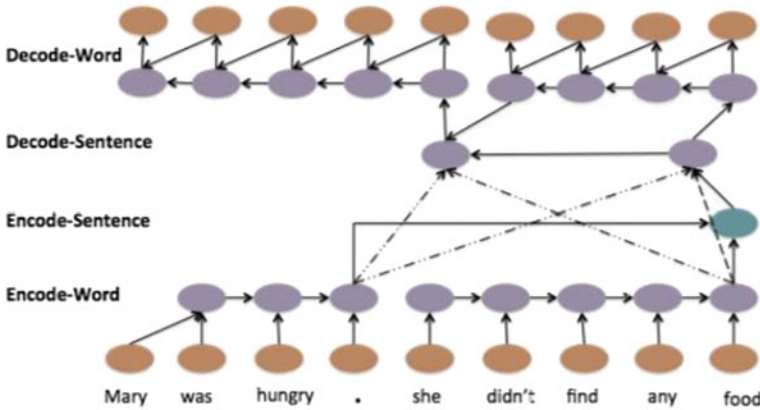


Fig. 2 Architecture of a hierarchical neural network [17]

Hierarchical Neural Models [17] have been introduced by (Li et al.). These models used a hierarchical LSTM model for capturing local compositionality. These models are capable of building encoding for paragraphs using the encodings of sentences which are created from the encoding of words and tokens. This architecture is used for paraphrasing documents and paragraphs. But on close observation of the model, this algorithm can be used for efficient summarization of long archives. Figure 2 represents such a model.

The method [18] proposed by (Dohare and Karnick) created an Abstract Meaning Representation of a story. The AMR is a graph representing the concepts and the relations between them. The paper detailed three methods for parsing such a graph to create a subgraph of the corresponding summary, viz., JAMR parsing (align based parsing), CAMR (grammar-based parsing), and Neural parsing (Using Seq 2Seq models). The algorithm tries to identify the important sentences and capture the relation between the important entities. Extracting the subgraph includes searching the position of the most referred entity and extracting the sub-tree of that entity from the graph. Sentences having similar meaning will have a similar AMR. An example of AMR can be viewed in Fig. 3.

A method [19] proposed by (Cheng et al.) suggested using data-driven approach for extracting sentence features. The architecture developed consisted primarily of hierarchical document encoder and attention-based extractor. This model focuses on sentence and word extraction. It helps in overcoming the difficulties caused by low-frequency words which are not present in the vocabulary generated from the dataset during the training. Architectures for both mechanisms have been illustrated below in Fig. 4 (Fig. 5).

The paper [20] proposed by (Nallapati et al.) sheds light on various algorithms to address the problems in summarization. Similar to Word Embeddings, embedding matrices were created for each feature such as parts of speech and cue words which we finally concatenated to form a single representation for a word. This method helps

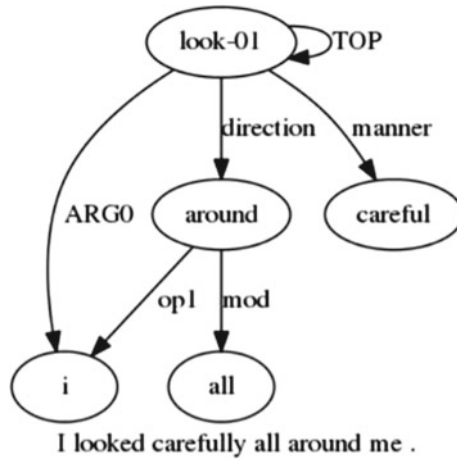


Fig. 3 AMR representation of a sentence [18]

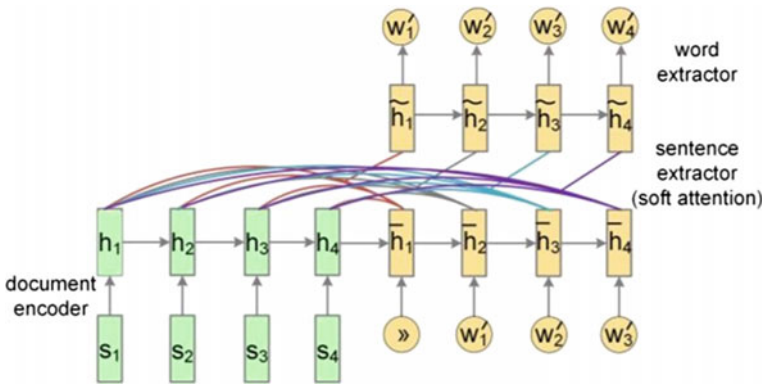


Fig. 4 Neural attention mechanism for word extraction [19]

in taking into consideration all the tags at the encoder level. On the decoder side, it can easily accept this type of word embedding for summary prediction. Such type of a model is shown in Fig. 6.

As mentioned before, the vocabulary generated during the training time contains limited number of words. Hence, there may be low-frequency words which are not present in the dictionary. In such cases, the system replaces that word with <UNK> but here, the system deals with it by providing a pointer to the location in the original article.

The pointer system might be more powerful in taking care of uncommon words since it utilizes the encoder's shrouded state portrayal of uncommon words to choose which word from the report to point to. Since the concealed state relies upon the content of the word, the model can precisely point to inconspicuous words in spite

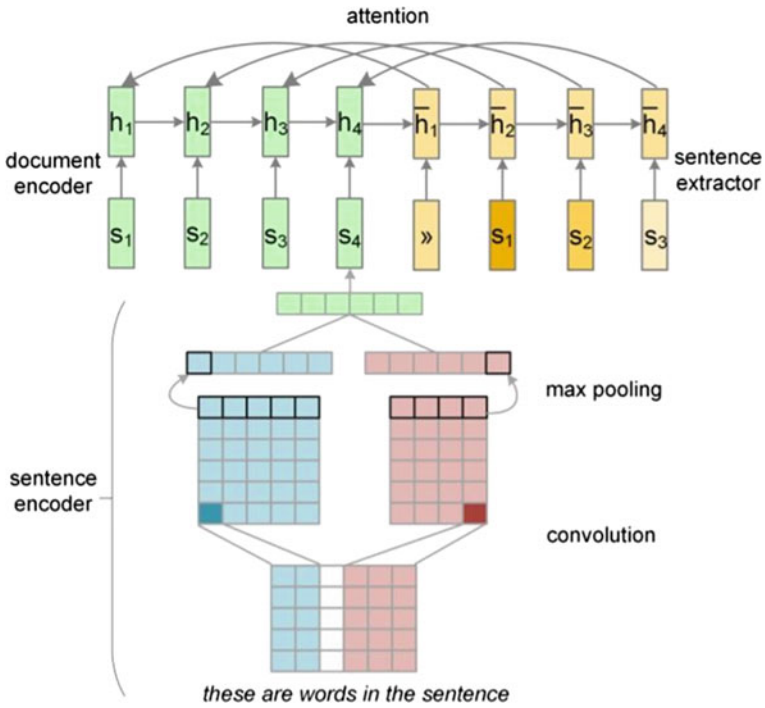


Fig. 5 R—CNN model for sentence extraction [19]

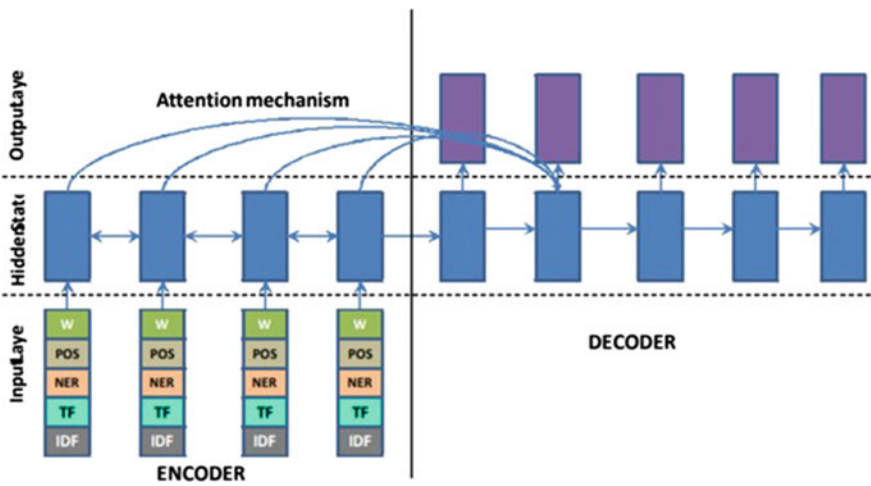


Fig. 6 Encoder creating an encoding matrix for POS, NER, and TF—IDF [20]

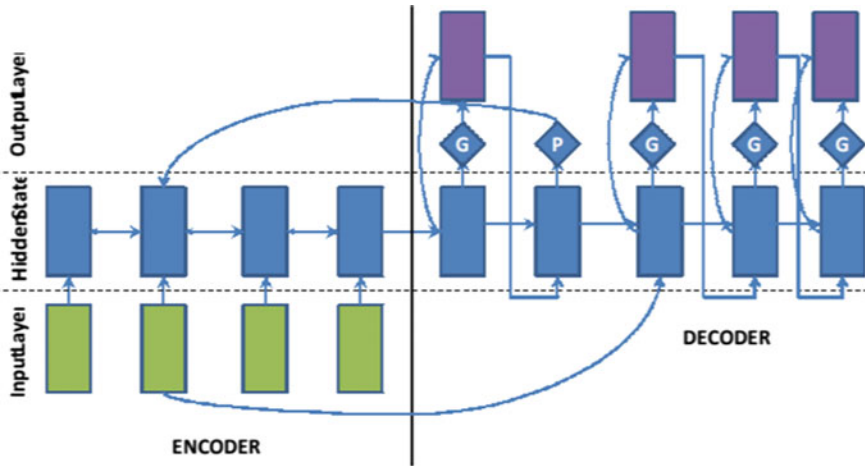


Fig. 7 Model showing the pointer system at the decoder level for handling low-frequency words [20]

of the fact that they do not show up in the objective vocabulary. When the pointer is activated, the embedding from the source is used as input to the next time step. Figure 7 shows this functionality.

Paulus et al. [21] proposed a reinforced model for summarization. The model attempted at generating summaries of longer articles while reducing the repeated and incoherent phrases. This was achieved by using an intra-temporal attention in the encoder that records previous attention weights for each of the input tokens while a sequential intra-attention model in the decoder takes into account which words have already been generated by the decoder.

The intra-temporal attention on the input sequence prevents the model from attending over the same parts of the input at different decoding steps. The intra-decoder attention incorporates more information from the previous decoding steps to avoid repetition and make better predictions. The decoder focuses on not generating the same trigram to avoid repetition during the beam searching. This attention model is depicted as follows in Fig. 8:

Along with these methods, different learning policies were implemented such as Supervised Learning with Teacher Forcing which minimizes the maximum likelihood loss at each step. The calculation is shown in 2 [21].

$$L_{ml} = - \sum_{t=1}^{n'} \log p(y_t^* | y_1^*, \dots, y_{t-1}^*, x) \tag{2}$$

Policy Learning introduces a reward function by comparing it to the ground truth to help maximize a specific discrete metrics. Equation 3 explains the loss function [21].

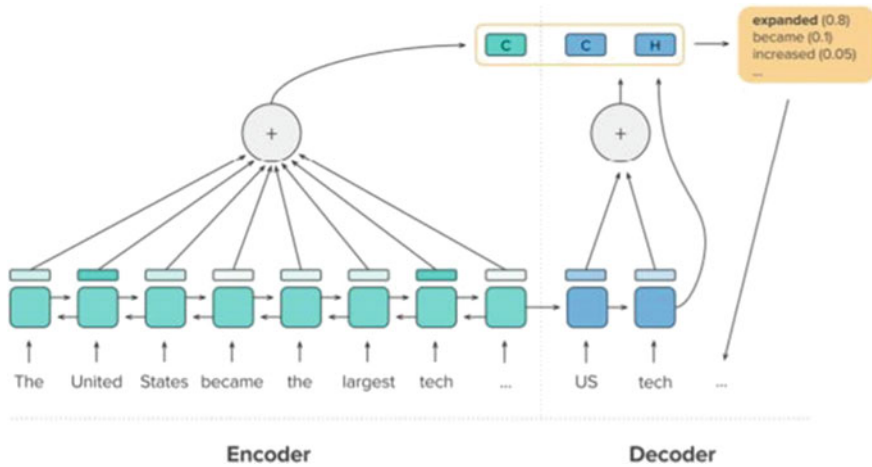


Fig. 8 Illustration of the attention-based encoders and decoders [21]

$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \dots, y_{t-1}^s, x) \tag{3}$$

The literature review provided extensive information about the prevailing systems. This helped in understanding the different technologies used and their effectiveness in generating summaries. The proposed system aimed at improving the 1—Gram metrics while generating summaries. The proposed system aimed at achieving this by using a combination of technologies focusing more on the word selection. 1—Gram approach emphasizes on whether the words selected for the summary are in conformance with the words from the reference summary. This was achieved using a combination of different technologies mentioned in forthcoming chapter.

This system aimed at improving the existing summarization system using computationally less expensive optimizations. Instead of using the traditional LSTM networks for maintaining the dependency, the proposed system is based on GRU cells which are computationally less expensive. Likewise, the system was developed on Tensorflow framework, which makes the deployment of system easier. AdaDelta hyper-parameter is used instead of AdaGrad and Adam for optimizing the learning rate for effective convergence. The most important aspect is the loss function, perplexity, which is calculated as 2^j in the traditional systems mentioned in the Literature Review. The proposed system used e^j to calculate the perplexity. This was done to adapt the cost function to the exponential functions used in the Neural Model.

5 Datasets and Testing Methods

The datasets play a crucial role in determining the outcome of the experiments. The most standardized dataset used by reputed institutions for any computational linguistics related problem is the Gigaword Dataset. It contains 10 Million tuples of items. These can be used for Summarization and Translation models. The vastness of the dataset helps in capturing various semantics and also increasing the dictionary. Other open source datasets included the BBC dataset and CNN Daily Mail dataset consisting of 400,000 documents. DUC datasets can be used for training as well as testing since they provide the reference summaries or translations which can be used during the ROUGE and BLUE testing.

The testing for such linguistic models is done by using the ROUGE and BLUE testing. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation and BLUE stands for Bilingual Evaluation Understudy. These testing paradigms focus on matching the N-Grams of the candidate and reference text. These tests give a score between 0 and 1 with 1 as highest benchmark.

6 Conclusion and Future Scope

In this paper, different techniques for text summarization were discussed. It can be concluded that the Extractive techniques can be easily scaled to provide summaries for articles pertaining to any domain. On the other hand, Extractive techniques lack the punch of creating anthropocentric summaries. Abstractive techniques handle this issue efficiently. They are able to replicate the summaries which are more appealing to read and comprehend. However, the downside of using Abstractive techniques is limitation imposed due to limited dictionary size and the constraints on the size of input (articles). After researching the existing systems system, it can be concluded that the most efficient system can use the advantages of both Extractive and Abstractive Techniques. Such systems can identify the relevant sentences from the article and perform Abstractive Summarization only on those sentences. Such a study still remains to be explored and overwhelming research is being conducted in condensing records consisting of unstructured and semi-structured data.

References

1. Moratanch, N., Chitrakala, S.: A survey on abstractive text summarization. In: 2016 International Conference on Circuit, power and computing technologies (ICCPCT) (pp. 1–7). IEEE (2016)
2. Barzilay, R., McKeown, K.R.: Sentence fusion for multidocument news summarization. *Comput. Linguist.* **31**(3), 297–328 (2005)
3. Lee, C.S., Jian, Z.W., Huang, L.K.: A fuzzy ontology and its application to news summarization. *IEEE Trans. Syst. Man, Cybern. Part B (Cybern.)* **35**(5), 859–880 (2005)

4. John, A., Wilsy, M.: Random forest classifier based multi-document summarization system. In: 2013 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 31–36. IEEE (2013)
5. Ganesan, K., Zhai, C., Han, J.: Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd International Conference on Computational Linguistics, pp. 340–348. Association for Computational Linguistics (2010)
6. Lloret, E., Palomar, M.: Analyzing the use of word graphs for abstractive text summarization. In: Proceedings of the First International Conference on Advances in Information Mining and Management, pp. 61–6. Barcelona, Spain (2011)
7. Gupta, V., Lehal, G.S.: A survey of text summarization extractive techniques. *J. Emerg. Technol. web Intell.* **2**(3), 258–268 (2010)
8. Zhang, J., Sun, L., Zhou, Q.: A cue-based hub-authority approach for multidocument text summarization. In: Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05, pp. 642–645. IEEE (2005)
9. Ferreira, R., Freitas, F., de Souza Cabral, L., Lins, R.D., Lima, R., França, G., Favaro, L.: A context based text summarization system. In: 2014 11th IAPR International Workshop on Document Analysis Systems (DAS), pp. 66–70. IEEE (2014)
10. Zhang, P.Y., & Li, C.H.: Automatic text summarization based on sentences clustering and extraction. In: 2nd IEEE International Conference on Computer Science and Information Technology, 2009. ICCSIT 2009, pp. 167–170. IEEE (2009)
11. Nallapati, R., Zhou, B., Ma, M.: Classify or select: neural architectures for extractive document summarization. arXiv preprint [arXiv:1611.04244](https://arxiv.org/abs/1611.04244) (2016)
12. Narayan, S., Papasrantopoulos, N., Lapata, M., Cohen, S.B.: Neural extractive summarization with side information. arXiv preprint [arXiv:1704.04530](https://arxiv.org/abs/1704.04530) (2017)
13. Narayan, S., Papasrantopoulos, N., Cohen, S.B., Lapata, M.: Neural extractive summarization with side information. arXiv preprint [arXiv:1704.04530](https://arxiv.org/abs/1704.04530) (2017)
14. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res.* **22**:457–479 (2004)
15. Steinberger, J., Ježek, K.: Text summarization and singular value decomposition. In: International Conference on Advances in Information Systems, pp. 245–254. Springer, Berlin, Heidelberg (2004)
16. Peddinti, V., Povey, D., Khudanpur, S.: A time delay neural network architecture for efficient modeling of long temporal contexts. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
17. Li, J., Luong, M.T., Jurafsky, D.: A hierarchical neural autoencoder for paragraphs and documents. arXiv preprint [arXiv:1506.01057](https://arxiv.org/abs/1506.01057) (2015)
18. Dohare, S., Karnick, H.: Text summarization using abstract meaning representation. arXiv preprint [arXiv:1706.01678](https://arxiv.org/abs/1706.01678) (2017)
19. Cheng, J., Lapata, M.: Neural summarization by extracting sentences and words. arXiv preprint [arXiv:1603.07252](https://arxiv.org/abs/1603.07252) (2016)
20. Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B.: Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint [arXiv:1602.06023](https://arxiv.org/abs/1602.06023) (2016)
21. Paulus, R., Xiong, C., Socher, R.: A deep reinforced model for abstractive summarization. arXiv preprint [arXiv:1705.04304](https://arxiv.org/abs/1705.04304) (2017)

Cloud Storage–Optimization of Initial Phase for Privacy-Preserving Public Auditing



Deepak Kumar Verma, Purnima Gupta and Rajesh Kumar Tyagi

Abstract The integrity of data confirms that an unapproved user cannot temper the outsourced data in cloud storage. The cloud storage needs to be secure from unapproved tempering. For integrity verification process, data owner generates the signature which is sent to the third-party auditor. It is a one-time process, but it increases the overhead of the data owner. We have analyzed the existing data integrity auditing schemes along with their distinctions. The proposed system supports privacy-preserving integrity auditing by using homomorphic linear authentication and employing Boneh–Lynn–Shacham-based signature technique. We extend our research to enable the data owner to speed up the initial phase through detailed experiments and comparisons between single-thread and multi-thread models using different core of CPUs. The proposed scheme demonstrated by using multithreading architecture on multi-core CPU for getting better performance.

Keywords Cloud service provider · Third-party auditor (TPA) · Data integrity · Provable data possession (PDP) · Proof of retrievability (POR)

1 Introduction

Cloud computing provides on-demand access to a shared network resource like storage servers, applications, platforms and many other shared or individual services [10]. It can be speedily administered with minimum effort or the service provider interaction. Lin et al. proposed a method for selecting cloud services with the best

D. K. Verma (✉) · P. Gupta
IEC College of Engineering and Technology, Greater Noida, Uttar Pradesh, India
e-mail: deepak.verma1980@gmail.com

P. Gupta
e-mail: purnimaa018@gmail.com

R. K. Tyagi
Krishna Institute of Engineering and Technology, Ghaziabad, Uttar Pradesh, India
e-mail: profrajeshkumartyagi@gmail.com

security and privacy features [8]. In cloud computing, the workload of users can be managed efficiently and economically using virtualization [7]. Data outsourcing reduces the burden of local data management and maintenance. Data owner loses the physical control over this data after uploading it on the cloud server. Hacigumus, Iyer and Mehrotra were first to introduce the concept of data outsourcing [5].

Due to eminent mobility, coherent storage and retrieval of data, users are being attracted towards accessing cloud services [6]. Apart from the benefits of cloud computing, there are three main pillars of security are CIA (Confidentiality, Integrity, and Availability) [22]. There are many reasons for cloud service providers to be dishonest. Dishonest CSP may produce the wrong status of outsourced data. Outsourced data, which has been accessed rarely, can be discarded by CSP because of budgetary reason or CSP may hide the data loss information for his reputation [1, 23].

We are focusing on integrity auditing of the outsourced data in cloud storage. There are two categories of auditing schemes, private auditing and public auditing. In private auditing scheme, all computation that is needed for checking integrity is directly performed between data owner and cloud service provider. Second is public auditing, in which the integrity verification process is done by TPA (Third-Party Auditor) [21, 22]. This scheme reduces the computation overhead of users because all computations are done through TPA, and integrity verification results produced by TPAs are commonly accepted by both data owner and CSP [23].

The main contribution of our scheme is as follows:

1. We have analyzed the existing data integrity auditing methods along with their distinctions and proposed an integrity auditing system that supports privacy-preserving integrity auditing using homomorphic linear authentication and employing Boneh–Lynn–Shacham (BLS) based signature technique.
2. We have extended our research to speedup the initial phase of integrity auditing by implementing multithreading architecture on multi-core CPU for better performance. Finally, we justified our proposed approach with the existing methods.

2 Integrity Auditing System

As shown in Fig. 1, the auditing model consists of three main entities like user/data owner (DO), third-party auditor (TPA) and cloud service provider (CSP). DO sets up their data and uploads it to the cloud server. The CSP stores the data in the cloud and allows accessing the data from anywhere and at any time. When the DO sends a request to TPA for checking the integrity of data, the TPA sends a challenge to CSP, and as the answer of that challenge, the CSP sends the proof to the TPA. In this way the TPA ensures the integrity of outsourced data.

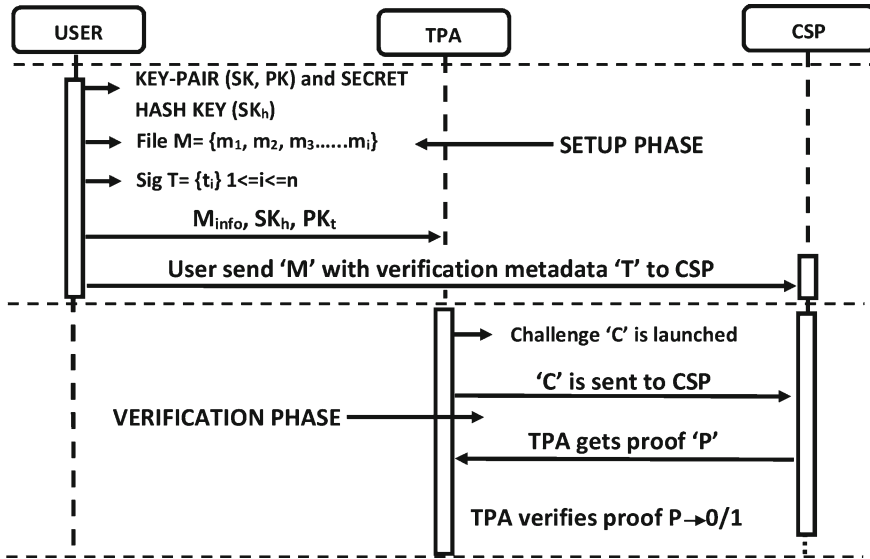


Fig. 1 Auditing model

3 Related Work

Fu et al. proposed a new privacy-aware public auditing system for pooled cloud data by constructing a homomorphic verifiable group signature [4]. Their system ensured that group users can trace data changes through the designated binary tree and can recover the latest correct data block when the current data block is damaged. Public auditing can be achieved through two basic concepts like MAC-Based and HLA-based. Many schemes have been proposed for dynamic data auditing by [2, 11, 15, 16, 18, 21, 25, 26]. These schemes have achieved dynamic auditing through implementing techniques like the indexed hash table [23], Merkle hash tree [19] and dynamic hash table [21]. Zhang and Tu [24] offered a new approach, based on batch-leaves-authenticated Merkle Hash Tree (MHT), to batch-verify multiple leaf nodes and their own indexes altogether, which is more suitable for the dynamic outsourced auditing system than traditional MHT-based dynamism approach that can only verify many leaf nodes one by one. Li et al. [9] concentrated on the multifarious key management in cloud data integrity checking by introducing fuzzy identity-based auditing. They verified the security of the proposed protocol based on the computational Diffie-Hellman postulation and the discrete logarithm hypothesis in the selective-ID security model. Table 1 gives the comparison of the existing integrity auditing schemes.

Table 1 Comparison of existing data integrity auditing schemes

Data integrity auditing scheme	Technique used	Proposed By	Year	Strength	Weakness
PDP First privacy preserving	Integrating Homomorphic authenticator with random masking	Wang et al.	2010	Supports public auditing privacy preserving	Does not support data dynamics
Fully dynamic PDP	Combined BLS based HLA with MHT	Wang et al.	2011	Supports dynamic auditing	No Privacy preserving
CPDP (corporate provable possession)	Hash index hierarchy	Zhu et al.	2012	Supports public auditing Privacy-preserving Batch auditing in multi-cloud	It does not support the dynamic audit. Does not support auditing for multiuser
IHT-PA (Index hash table-public audit)	Index hash table	Zhu et al.	2013	Supports public auditing Privacy-preserving Supports dynamic auditing	Batch auditing is not mentioned
Privacy-preserving public auditing and multi-owner authentication	Homomorphic linear authenticator and random masking	Nandini et al.	2014	Eliminated the trouble of cloud user from the dull and possibly high-priced auditing task	Does not support data dynamics operations
DPDP (Dynamic PDP)	Using ranked based authenticated skip list	Erway et al.	2015	Dynamic data auditing No demand for privacy-preserving	No public auditing Does not support Batch auditing

(continued)

Table 1 (continued)

Data integrity auditing scheme	Technique used	Proposed By	Year	Strength	Weakness
DHT-PA (Dynamic hash table-public audit)	Dynamic hash table	Tian et al.	2015	Supports public auditing Supports dynamic auditing Supports batch auditing in multi-cloud	Communication cost is greater than DAP and IHT-PA
Integrity checking for outsourced data	Suggested various methods	Kaustubh and Jog	2016	Only be efficient when the verifier of the service maintains a copy of same outsourced data	The original file is the necessity for integrity examine
Secure preserving public auditing	Homomorphic linear authenticator	Poornima and Ponmagal	2016	Supports batch auditing Privacy preserving	Does not support data dynamics
NPP: a new privacy-aware public auditing Scheme	Homomorphic verifiable group signature	Fu et al.	2017	Eliminates misuse of single-authority control Ensures non-frameability	Group users can trace the data changes through the binary tree
Dynamic data operations with deduplication	Markle hash tree (MHT)	Wu et al.	2017	Support dynamic data operations Privacy preserving	Does not support batch auditing

4 Problem Statement

As discussed in literature review, authors have compared existing data integrity verification schemes for cloud storage and noticed that many of them require dividing the data into blocks. All blocks are then required to generate authenticators so that it can be sent to TPA for verifying the integrity of user’s data. These authenticators/signatures are highly required to accumulate confidentiality and privacy preser-

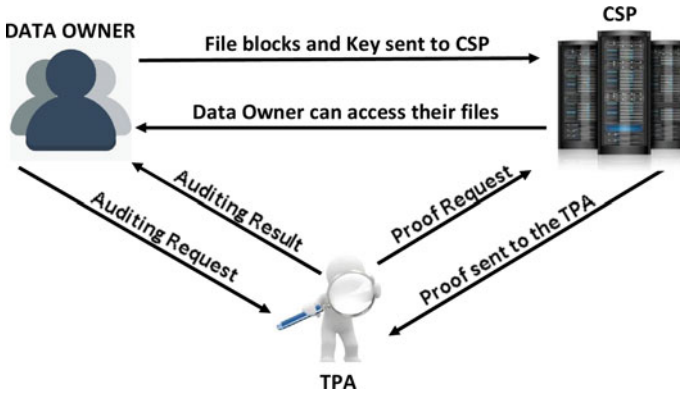


Fig. 2 System model

vation of cloud user’s data from TPA. Authenticator/signature generation for the large set of blocks is quite tedious and time-consuming task for the DO. Nandini et al. [12] and [13, 17, 20] have indicated that the data setup happens only once, traditionally it was generated sequentially and expanded the computational overhead at the user side. Initial phase can speed up by implementing it to multi-threading model on latest multi-core CPU system [3, 14, 20].

DO, CSP and TPA are three main entities in the proposed system model shown in Fig. 2. It means that there should not be any discrepancy while retrieving the data by the users. The TPA verifies the data integrity of the outsourced data for the data owner. To verify the outsourced data, the data owner provides the metadata instead of complete data because it is in encrypted form for data privacy. When data owner sends the request to TPA for checking the integrity of data, then TPA challenges the cloud service provider, and then CSP sends the algebraic signature to the TPA for proofing.

5 Proposed System Model

5.1 The Proposed Solution by the Authors for Integrity Auditing Initial Phase

In this paper, HLA (Homomorphic Linear Authenticator) technique has been used to achieve public auditing without downloading heavy count (complete data) of data blocks. To achieve a privacy-preserving public auditing, File “ M ” is divided into fixed size blocks and then HLA $\phi = \{\sigma_i\} 1 \leq i \leq n$ is aggregated for dataset $M = \{m_i\}$ where “ n ” is a total number of blocks. This process of generating authenticators is very tedious and time-consuming operation for the data owner. In the

proposed system, data is first divided into the fixed size of blocks $M = \{m_i\}$. These blocks are further encrypted with AES-128 encryption technique. After encryption of each block, dataset $M = \{m_i\}$ is ready to compute authenticators for each “ p ”. A short signature scheme has been used based on the computational Diffie-Hellman assumptions, i.e., certain elliptic and hyperelliptic curves. Traditional single-threaded models for the generation of authenticators on multi-core CPU systems may be approximate “ $p-1$ ” times slower than our proposed model, which is a multi-threaded model with multi-core CPU systems where “ p ” is CPU cores available on the system.

In the proposed system, authors used a single dimensional array of encrypted data blocks $\{m_i\}$ to generate signature $\{\sigma_i\}$. The single-threaded model may take several minutes to complete its task if the array of data blocks is very large. And that’s why authors have implemented a multi-threaded model, which breaks the array up to the size of the threshold. The authors have taken threshold size 50.

5.2 Experimental Setup

We applied Java pairing-based cryptography (JPBC) API with library version 2.0.0, the Fork-Join-Pool framework of JDK 1.8 and “J-Free-Chart” API for generating charts. All these supporting libraries are integrated with Net-Beans 8.0 IDE to make application development easy and bug-free. The Boneh-Lynn-Shacham (BLS) with the short signature scheme is used to generate signatures for blocks. This scheme used a verification pairing function and signatures under some elliptic curves. Authenticator generation on each block is independent, and it is the one-time operation.

5.3 Preliminaries

Bilinear Map: Let $G_1, G_2,$ and G_T be cyclic groups of prime order r . Let g_1 be a generator of G_1 and g_2 is a generator of G_2 . A bilinear pairing or bilinear map e is an efficiently computable function $e: G_1 \times G_2 \rightarrow G_T$ such that:

Bilinearity: For all a, b in Z_r (the ring of integers modulo r) it holds that $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab}$.

Non-degeneracy: $e(g_1, g_2) \neq 1$. Our creation of multi-threaded data setup model includes split, encrypt, setup, keyGen, sigGen, upload, delete, decrypt and join operations as follows:

- (i) $split(M) = \{m_1, m_2, m_3 \dots m_i\}$
- (ii) Encrypt (key, $m_i, m_{i\ enc}$)
- (iii) Setup
- (iv) Private Key (x) and public key (g^x) generates from selecting a random integer $x \leftarrow Z_r$ and a random “ g ” $\leftarrow G_2$.
- (v) Signature $sig = h^x$.

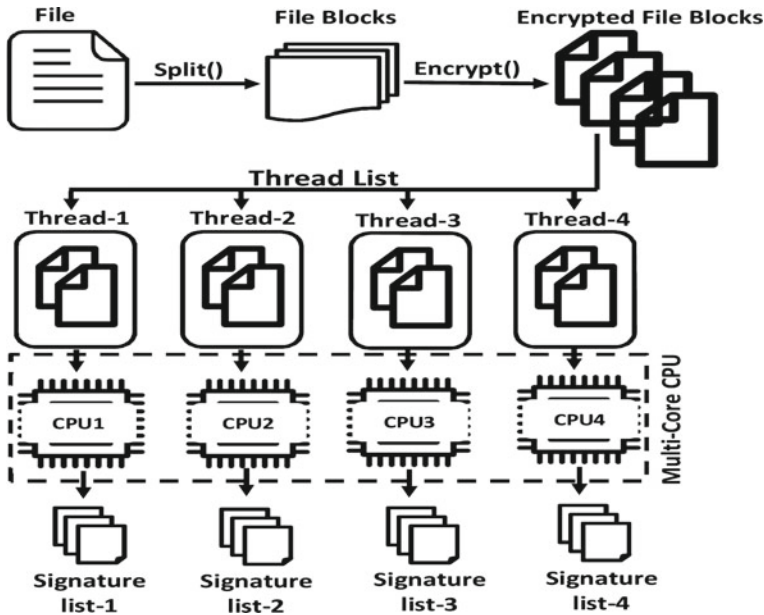


Fig. 3 Multithreading model

Figure 3 represents the architecture of our proposed multithreading model.

6 Result and Discussion

The results of the experiments are shown below in Tables 2, 3 and 4 respectively.

Each file M is splitting into “ n ” data blocks as $M = (m_1, m_2, m_3 \dots m_i)$, where $1 \leq i \leq n$. Due to computation overhead, 50 KB data block size has been taken as in [18]. However, if we get the less size of data blocks, then more blocks will be generated and for each block signature will be produced by the system.

On the basis of the block size (50 KB), 8 MB data component has been divided into 164 blocks which will generate 164 signatures. Data component splitting time and block encryption time have been reduced up to 50% by implementing the proposed system with the multi-threaded model on multi-core CPU system.

The comparison has been made on different file which has different size with different count of blocks like 8 MB (164 blocks), 16 MB (328 blocks), 32 MB (656 blocks), and 64 MB (1312 blocks).

Figure 4 shows the performance comparison graph of Table 3. As our paper focuses on the initial phase setup at DO side, it has been observed that up to 50% computation time reduced by using multithreading model as compared to the single-threaded model.

Table 2 List of symbols

Symbol	Meaning	Symbol	Meaning
x	Private key	i	Block/signature sequence number
g^x	Public key	n	Number of blocks
t	Threshold size	p	Number of available CPU cores
M	Data set	σ_i	Signature of a block
Φ	Set of signatures	h	Mapping element of the signature
e	Bilinear map	f	The first index of the block array
m_{i_enc}	Encrypted data block	l	Last index of the block array
m_i	Data block		

Table 3 Computation time comparison on different CPUs (50 KB block size)

		Split time (s)		Encryption time (s)	
File size (MB)	Number of blocks	2 CPU core	4 CPU core	2 CPU core	4 CPU core
8	164	0.56	0.234	3.883	1.645
16	328	0.873	0.506	3.52	2.568
32	656	1.434	0.789	4.49	3.358
64	1312	2.484	1.89	8.215	4.984

Table 4 Computation time on the single-threaded model

File size (MB)	Computation time on 2 CPU cores at 25 KB block size (s)	Computation time on 4 CPU cores at 25 KB block size (s)	Computation time on 2 CPU cores at 50 KB block size (s)	Computation time on 4 CPU cores at 50 KB block size (s)
8	101	23.07	28.46	25.4
16	160	41.2	54.5	47.6
32	286	103.86	104.05	102.1
64	582	234.36	207.8	198.5

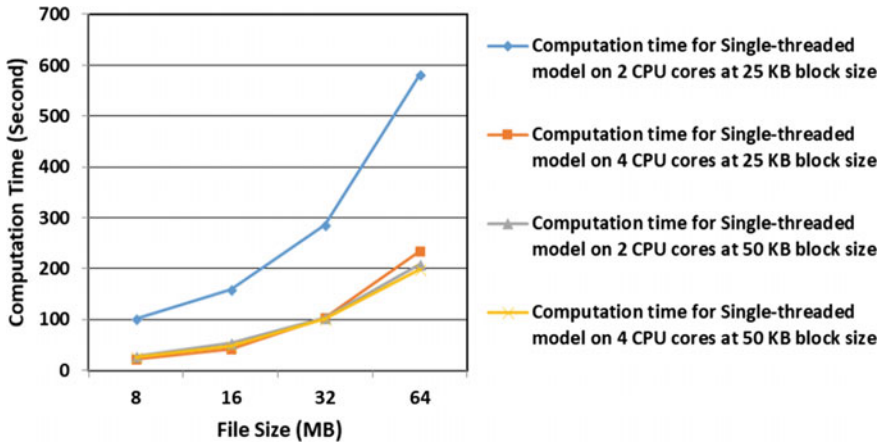


Fig. 4 Computation time on the single-threaded model

Table 5 Computation time on multi-threaded model

File size (MB)	Computation time on 2 CPU cores at 25 KB block size (s)	Computation time on 4 CPU cores at 25 KB block size (s)	Computation time on 2 CPU cores at 50 KB block size (s)	Computation time on 4 CPU cores at 50 KB block size (s)
8	62	10.72	17.71	12.5
16	119	19.52	34.07	23.2
32	246	43.49	69.29	50.5
64	477	81.59	136.43	96.3

Table 5 shows the performance on the multi-threaded model for both 25 KB and 50 KB data block sizes on two parallel CPU cores and four parallel CPU cores.

The performance differences between single-threaded and multi-threaded model have shown in Fig. 5.

Figure 6 shows the difference between throughputs on dual core and core i3 (4 CPU cores) systems with different block sizes.

We achieved the better performance on multi-core CPU system by generating the limited number of threads depends on available CPU cores. Figure 7 represents the throughput analysis (in KB/second) of our proposed system based on different CPU types (multi-core CPUs).

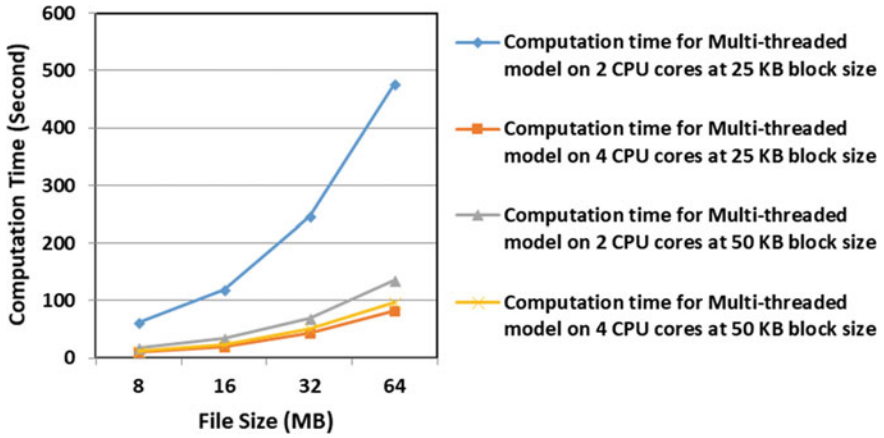


Fig. 5 Computation time on multithreading model

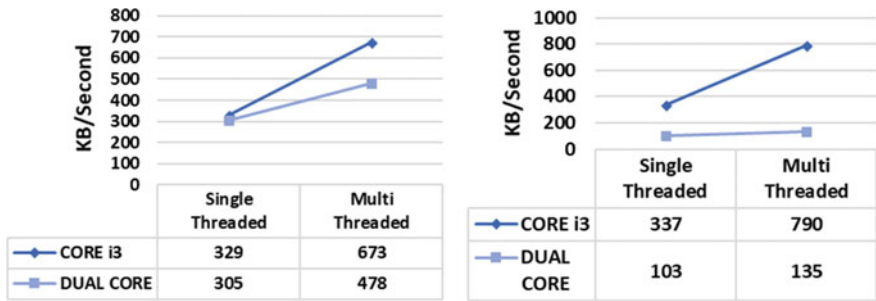


Fig. 6 Throughput analysis based on different block sizes

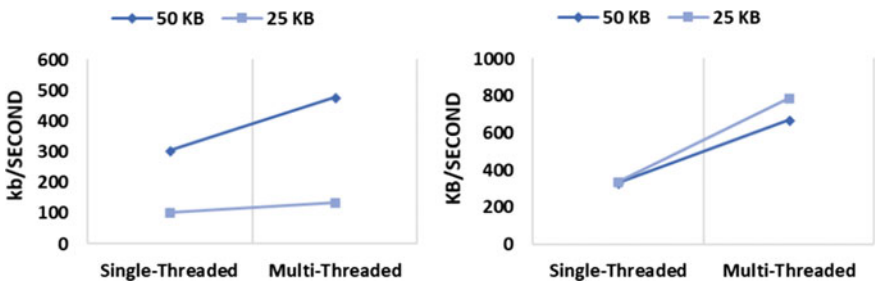


Fig. 7 Throughput analysis based on different CPU types

7 Conclusion

In this paper, a detailed comparative analysis of different types of auditing schemes has been presented. Authors proposed an enhanced scheme for the initial phase setup for data integrity auditing to reduce the computation cost at user side. We used multi-threading parallel execution for initial phase by utilizing all available CPU cores through concurrent execution of tasks. The results show the significant reduction in the computation cost at user side. In future, the proposed scheme can be extended by utilizing more processing units with the help of graphical processing units (GPUs).

References

1. Ateniese, G., Burns, R., Curtmola, R., Herring, J., Kissner, L., Peterson, Z., Song, D.: Provable data possession at untrusted stores. In: Proceedings of the 14th ACM conference on Computer and communications security, pp. 598–609 (2007)
2. Ateniese, G., Di Pietro, R., Mancini, L.V., Tsudik, G.: Scalable and efficient provable data possession. In: Proceedings of the 4th International Conference on Security and Privacy in Communication Networks, p. 9 (2008)
3. Chen, K.Y., Chang, J.M., Hou, T.W.: Multithreading in Java: performance and scalability on multicore systems. *IEEE Trans. Comput.* **60**(11), 1521–1534 (2011)
4. Fu, A., Yu, S., Zhang, Y., Wang, H., Huang, C.: NPP: a new privacy-aware public auditing scheme for cloud data sharing with group users. *IEEE Trans. Big Data* (2017)
5. Hacigumus, H., Iyer, B., Mehrotra, S.: Providing database as a service. In: Proceedings of 18th International Conference on Data Engineering, pp. 29–38 IEEE (2002)
6. Kaustubh, S., Jog, V.V.: Article: A Survey on Integrity Checking for Outsourced Data in Cloud using TPA. *IJCA Proceedings on International Conference on Internet of Things, Next Generation Networks and Cloud Computing ICINC 2016*(1), 6–9 (2016)
7. Kumar, N.S., Lakshmi, G.R., Balamurugan, B.: Enhanced attribute based encryption for cloud computing. *Procedia Comput. Sci.* **46**, 689–696 (2015)
8. Lin, L., Liu, T., Hu, J., Ni, J.: PQsel: combining privacy with quality of service in cloud service selection. *Int. J. Big Data Intell.* **3**(3), 202–214 (2016)
9. Li, Y., Yu, Y., Min, G., Susilo, W., Ni, J., Choo, K.K.R.: Fuzzy identity-based data integrity auditing for reliable cloud storage systems. *IEEE Trans. Dependable Secure Comput.* (2017)
10. Mell, P., Grance, T.: The NIST definition of cloud computing. *Nat. Inst. Stand. Technol.* **53**(6), 50 (2009)
11. Mutyalanna, C., Srinivasulu, P., Kiran, M.: Dynamic audit service outsourcing for data integrity in clouds. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCET)* **1**(2), 2482–2486 (2013)
12. Nandini, J., Sugapriya, N.P., Vinmathi, M.S.: Secure multi-owner data storage with enhanced TPA auditing scheme in cloud computing. *Int. J. Adv. Comput. Sci. Cloud Comput.* **2**, 2321–4058 (2014)
13. Poornima, S.N., Ponmagal, R.S.: Secure preserving public auditing for regenerating code based on cloud storage. *Networking Commun. Eng.* **8**(5), 200–204 (2016)
14. Rinku, D.R., Rani, M.A.: Analysis of multi-threading time metric on single and multi-core CPUs with Matrix Multiplication. In: 2017 Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), pp. 152–155. IEEE (2017)
15. Rizvi, S., Razaque, A., Cover, K.: Cloud data integrity using a designated public verifier. In: High Performance Computing and Communications (HPCC). 2015 IEEE 7th International

- Symposium on Cyberspace Safety and Security (CSS), 2015 IEEE 12th International Conference on Embedded Software and Systems (ICCESS), 2015 IEEE 17th International Conference on, pp. 1361–1366. IEEE (2015)
16. Ryoo, J., Rizvi, S., Aiken, W., Kissell, J.: Cloud security auditing: challenges and emerging approaches. *IEEE Secur. Priv.* **12**(6), 68–74 (2014)
 17. Shirahatti, A.P., Khanagoudar, P.S.: Preserving integrity of data and public auditing for data storage security in cloud computing. *Int. Mag. Adv. Comput. Sci. Telecommun.* **3**(3), 161 (2012)
 18. Tian, H., Chen, Y., Chang, C.C., Jiang, H., Huang, Y., Chen, Y., Liu, J.: Dynamic-hash-table based public auditing for secure cloud storage. *IEEE Trans. Serv. Comput.* (2015)
 19. Wang, C., Chow, S.S., Wang, Q., Ren, K., Lou, W.: Privacy-preserving public auditing for secure cloud storage. *IEEE Trans. Comput.* **62**(2), 362–375 (2013)
 20. Wang, C., Wang, Q., Ren, K., Lou, W.: Privacy-preserving public auditing for data storage security in cloud computing. In: *Infocom, 2010 Proceedings IEEE*, pp. 1–9. IEEE (2010)
 21. Wang, Q., Wang, C., Ren, K., Lou, W., Li, J.: Enabling public auditability and data dynamics for storage security in cloud computing. *IEEE Trans. Parallel Distrib. Syst.* **22**(5), 847–859 (2011)
 22. Wu, Y., Jiang, Z.L., Wang, X., Yiu, S.M., Zhang, P.: Dynamic data operations with deduplication in privacy-preserving public auditing for secure cloud storage. In: *2017 IEEE International Conference on Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC)*, vol. 1, pp. 562–567 (2017)
 23. Yang, K., Jia, X.: An efficient and secure dynamic auditing protocol for data storage in cloud computing. *IEEE Trans. Parallel Distrib. Syst.* **24**(9), 1717–1726 (2013)
 24. Zhang, H., Tu, T.: Dynamic Outsourced Auditing Services for Cloud Storage Based on Batch-Leaves-Authenticated Merkle Hash Tree. *IEEE Trans. Serv. Comput.* (2017)
 25. Zhu, Y., Ahn, G.J., Hu, H., Yau, S.S., An, H.G., Hu, C.J.: Dynamic audit services for outsourced storages in clouds. *IEEE Trans. Serv. Comput.* **6**(2), 227–238 (2013)
 26. Zhu, Y., Hu, H., Ahn, G.J., Yu, M.: Cooperative provable data possession for integrity verification in multicloud storage. *IEEE Trans. Parallel Distrib. Syst.* **23**(12), 2231–2244 (2012)

A New Approach for Cloud Security Using Hybrid Querying System Over Cloud Scenario



Priya Sen, Ritu Prasad and Praneet Saurabh

Abstract Cloud computing is one of the most important and remarkable innovations of recent times. Since all the data is stored in servers located at different locations operating on different hardware so data consistency, availability, and security while accessing becomes a challenging task. Cloud introduces different data access management techniques that make sure of information security within the cloud even in the scenarios of outsourcing and untrusted cloud servers. Information access management in the cloud is a difficult issue due to the distributed architecture of cloud storage systems. Low trust of such a storage paradigm leads to limited proliferation and acceptability. These challenges further soar while addressing the issues of security, since it remains a massive risk. In this scenario, hybrid information prove a meaningful solution as it assists and stores all these varied information depending on privacy, needs of information and initial supply and offered capability. In this context, this paper proposes a cloud security hybrid querying system (CSHQS) that help to overcome data storage complexity and its subsequent access. The query input mechanism help in data computation, its processing, and deriving output with input test. The experimental results show the efficiency of the proposed CSHQS as it outperforms the existing state of art.

Keywords Cloud computing · Indexing analysis · Data storage · Data outsourcing · Big data analytics

P. Sen (✉) · R. Prasad
Technocrats Institute of Technology (Advance), Bhopal 462021, Madhya Pradesh, India
e-mail: psen8581@gmail.com

R. Prasad
e-mail: rit7ndm@gmail.com

P. Saurabh (✉)
Technocrats Institute of Technology, Bhopal 462021, Madhya Pradesh, India
e-mail: praneetsaurabh@gmail.com

1 Introduction

Cloud computing is a new advancement that offers more flexibility to users in terms of storage, hardware, and services [1]. Cloud services are designed, developed, and hosted on cloud illustrated in Fig. 1. Afterward, these services are customized and offered to the users according to their need and preferences. The cloud computing model works in pay-per-use, multi-tenancy, quantifiability, self-operability, on-demand, and value-effective manner [2, 3]. Due to this, much of flexibility cloud computing has become a synonym in the digital world due to the abovementioned services offered to the users [4, 5]. AWS (Amazon net Services), Sales force’s CRM system, Microsoft Azure, and Google Cloud Platform all exemplify this widespread notion of cloud computing [6, 7]. Design of cloud computing explains about the different features including application, service, cloud runtime, storage, and infrastructure [8]. It additionally explains regarding the side and also the face of a cloud computing run their operations [9]. The key advantage that cloud computing brings is agility, the flexibility to use abstracted calculate, storage, and associate degree network resources to workloads PRN and faucet into an abundance of pre-built services, however, in several conceptions from the cloud, it faces problems in knowledge security particularly in medical attention knowledge outsourcing. But all these flexibilities come with some limitations in terms of deployment and security.

This paper proposes a cloud security hybrid querying system (CSHQs) method of with the economical knowledge security in hybrid algorithmic program structure and subquery mechanism handling with totally different elements. This paper is organized in the following manner, Sect. 2 put forwards the related work, Sect. 3 presents the proposed work, Sect. 4 presents results and analysis, and Sect. 5 concludes the paper.

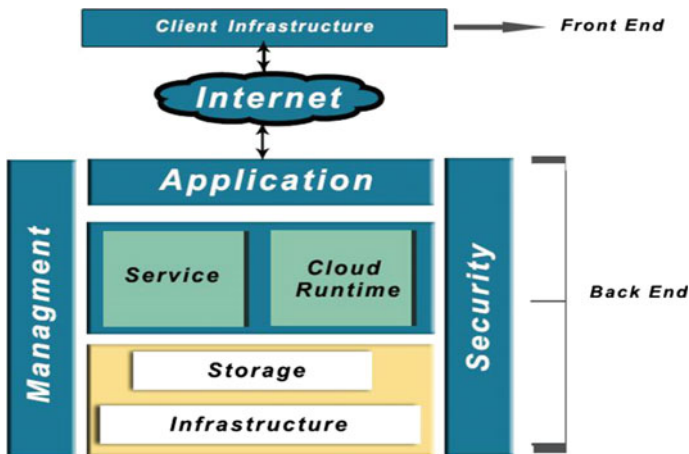


Fig. 1 Architecture of cloud computing

2 Related Work

Cloud computing is a common taxonomy used for storing and hosting data, services over the web. Cloud and its various services use virtualization and utility computing to accomplish different application for various domains [10]. Cloud computing can also be treated as an on-demand delivery model of the associate application to consumers [11]. Cloud computing architecture and servers are distributed but centralizes the operations to maintain consistency to provide demand CSPs and Internet Service Providers (ISPs) [12]. SAP HANA is accessible and it offers hybrid deployments to relinquish real choices to customers. It is a competitive advantage, whereas storing information in memory has performance benefits. Apart from the benefits, cloud computing also presents significant challenges and an organization needs to train the manpower that can specialize in high-value tasks [9]. Cloud computing is actually a mix of the existing technologies that are around since the primary 1990s [11]. Over the years, different interesting computational intelligence concepts have been introduced and implemented for different complex problems [13–15]. These new ideas can be used to find the desired solution [16, 17]. The next section introduces cloud security hybrid querying system (CSHQS) as it is important to maintain the security and efficient query. All the inputs and the data center set configuration must be accurate.

3 Proposed Method

This section put forwards the proposed mechanism cloud security hybrid querying system (CSHQS). CSHQS process explains about the cloud security hybrid querying system which states about collecting and storage of large amount of data. CSHQS help in providing data security along with relatively efficient processing of that data. The proposed algorithm makes hybrid cloud take advantage of query, processing with security and efficient query outcome processing system, and finally providing efficient results on requirement data usage.

3.1 Execution Process Steps

The different steps of CHSQS are given below that specifically revolve around the whole idea of data storage. CHSQS takes input datasets the files over cloud computing scenario which is given by the user.

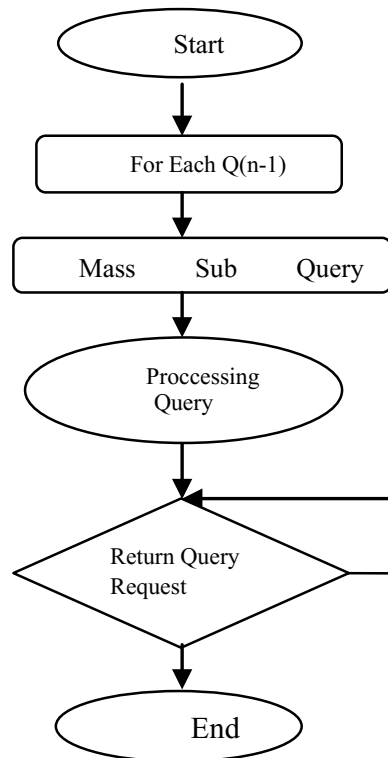
- Data conversion into the binary format for further processing in the next phase.
- Next step gives the security approach which is symmetric encryption approach helps in data security.

- An advance hashing approach which helps in hash code of data, duplicate records in large files, etc.
- Finally, a data query system which helps in querying the data is driven.

3.2 Flowchart of CSHQS Execution Process

The below flowchart in Fig. 2 explains about the working of cloud security hybrid querying system in which processing has been done on the given set of data, that is, for each $q (n - 1)$, the process further goes to the mass subquery accessing () which will process query either return query request if the query request does not proceed and the process end if all goes right. Flexibility of cloud offers on-demand approach to end users so that they can pay and use.

Fig. 2 Flowchart of cloud security hybrid querying system



3.3 Working of Cloud Security Hybrid Querying System (CSHQS)

The algorithm pseudocode is presented and described in below section, which gives the pseudocode. Cloud computing is a new advancement designed, developed, and hosted on the cloud that offers more flexibility to the users in terms of storage, hardware, and services.

Input:

Output:

Step 1: [CloudSim API extraction and loading]

Data configuration, setup configuration such as data center, virtual machine, input, and output processing system.

Taking input from the dataset which is given by the user browsed input data file and taking by simulation interface.

Input: Sensor dataset, Symmetric key encryption setup, Hashing system

Initialize the components;

Vm (i-n) initialization ();

DS (i-n) initialization ();

Step 2: [Setup starts and processing input] an availability of resources and processing it with the given component which process the input sensor data.

Data conversion into the binary format and further is processing with the next phase.

For each (input set information ds) {

Processing request ();

Processing symmetric key setup ();

XOR (ds); S-Phases evaluation (ds);

Step 3: [Binary conversion] the complete input data is obtained and conversion into binary is performed which helps in computing binary data dictionary.

Third phase gives the security approach which is Symmetric encryption approach help in data security.

Data security computation ();

Step 4: [Symmetric encryption with data] Simulation data encryption is performed before storage, thus the symmetric encryption mechanism is used for data encryption.

An advance hashing approach which helps in data storage is performed.

Performing hashing computation ();

Step 5: [Storage in hashing manner] Storage of encrypted data in hashing manner is get performed which help in hashing manner storage of data; it builds the data in hash values.

Finally, a data query system which helps in querying the data is driven.

Return storage updated information logs;

Step 6: [Accessing query system] A hybrid query is performed over the secure data which helps in accessing the system and query on provided data.

An effective computation, comparison using the provided comparison parameter is given.

Output: Data security, storage updating, hashing indexing generation.

Data security computation ();

[End]

The above algorithm explains the steps that are involved in the CloudSim for extraction and loading data in the cloud server, which took several inputs as $(i - n)$. The system gets initialized and further processes the different inputs that also include data conversion into binary format. In another step, symmetric encryption with data is performed and after that, it moves forward to storage in hashing manner and in the end access based on query system is achieved.

4 Experimental Result and Analysis

In this section, the proposed and developed CHSQS is tested under different conditions, and thereafter comparisons will be carried with the existing state of the art. Experiments are done to calculate the cost of file upload, data accessibility, and subsequent time of cloud computing scenario. A comparison analysis between the previously obtained technique which is compressive sensing and data processing approach is implemented. Further, the proposed approach in two phases which is data storage, and then query processing is given by the output interface. These comparisons are explained in tabular form by taking parameters such as Sensor dataset (Tset1) obtained from Symmetric key (Tset2) obtained from, Encryption setup (Tset3) obtained from, Hashing system (Tset4).

4.1 Execution Process Steps

- (i) **Performance improvement:** This parameter calculates the improvement between the proposed CSHQS and existing method.

$$\text{Performance Improvement \%} = \frac{\text{Proposed algorithm} - \text{Existing algorithm}}{100} * 100 \quad (1)$$

- (ii) **Time:** It is used to measure time to upload and then access data over servers between the proposed CSHQS and existing method.

$$\text{Final time (ms)} = \text{End time (ms)} - \text{Start time (ms)} \tag{2}$$

(iii) **Cost:** this metric determines the cost while uploading and accessing between the proposed CSHQS and existing method.

$$\text{Cost} = \text{File size}/\text{Final time} \tag{3}$$

4.2 Experimentation and Result Analysis

4.2.1 Comparison Among Algorithm on the Basis of Time

This experiment is performed to find the total time required to upload and subsequently access the data from the servers. The experiments are carried on dataset symmetric and encryption.

Table 1 shows the comparison among different existing algorithm (Indexing security algorithm) with the proposed CHSQS with respect to completion time. Results in Table 1 and subsequent Fig. 3 shows that the proposed CHSQS takes lower time reports and less time with respect to different number of tweets up to 1 crore in cloud analyst simulation environment.

Table 1 Time comparison among algorithms

Different file processing test	Existing algorithm (ms)	Proposed algorithm (ms)	Improvement (%)
Tset 1	53	44	0.09
Tset 2	330	112	2.18
Tset 3	432	321	1.11
Tset 4	643	490	1.53

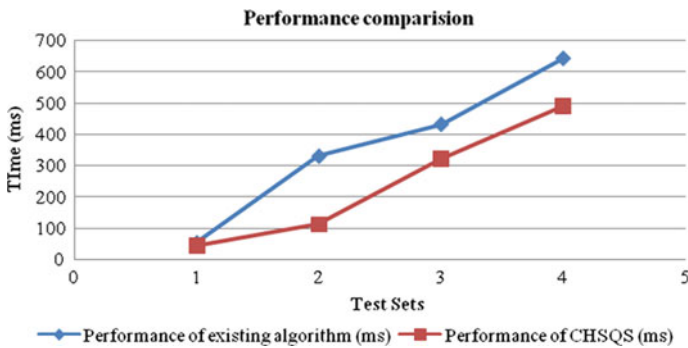


Fig. 3 Time-based comparison among query processing algorithms

Experimental findings presented in Table 1 and subsequent Fig. 3 very clearly explain that the CHSQS takes less time (according to the improvement %) under different test sets as compared to the existing method.

4.2.2 Comparison Among Algorithm on the Basis of Cost

This experiment is done with the intent to find the total cost based on file size. The experiments involved dataset with same hashing values and encryption.

Table 2 and Fig. 4 shows the comparative analysis between the computations cost among proposed algorithm and the existing system under different test sets. Experimental results illustrates that the proposed CHSQS performs better under all the test conditions as compared to the existing system. This result also establishes the fact that the newer incorporation of the query input mechanism help in data computation, its processing, and deriving output.

Table 2 Cost comparison among algorithms

Different file processing test	Existing algorithm (per unit)	CHSQS (per unit)	Improvement (%)
Tset 1	10	8	0.09
Tset 2	21	16	2.18
Tset 3	41	38	1.11
Tset 4	156	112	1.53

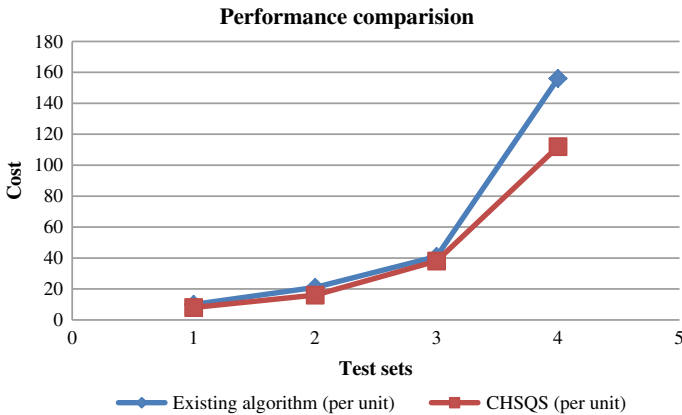


Fig. 4 Cost-based comparison among query processing algorithms

5 Conclusion

Cloud computing is an important mechanism today in any segment. It helps in data storage, data computation and data accessing mechanism with different input and output. The proposed cloud security hybrid querying system (CSHQS) helps in solving the data storage complexity, as well as its accessing issues. The query input mechanism helps in data computation, its processing, and deriving output with input test. CSHQS overcomes the limitations of the previous approaches, that is, compressive sensing and using of devices data in inappropriate format. Experimental results show that CSHQS using the clouds library with the web framework for data processing performed better. Evaluation result shows the efficiency of CHSQS approach over existing technique.

References

1. Pancholi, V.R., Patel, P.B.: Enhancement of cloud computing security. *Int. J. Innov. Res. Sci. Technol.* **2**(09), 18–21 (2016). ISSN (online) 2349-6010
2. Rashmi, S.G., Mehruz, S.: Securing software as a service model of cloud computing: issues and solutions. *Int. J. Cloud Comput. Serv. Architect. (IJCCSA)* **3**(4), 1–11 (2013)
3. Joshi, P., Saurabh, P., Prasad, R., Mewada, P.: A new neural network based IDS for cloud computing, progress in computing, analytics and networking. In: *Advances in Intelligent Systems and Computing*, vol. 710, pp. 161–170. Springer (2018)
4. Kumar, S., Goudar, H.R.: Cloud computing—research issues, challenges, architecture, platforms and applications: a survey. *Int. J. Future Comput. Commun.* **1**(4), 356–360 (2012)
5. Padhy, R.P., Patra, R.M., Satapathy, S.C.: Cloud computing: security issues and research challenges. *IRACST Int. J. Comput. Sci. Inf. Technol. Secur. (IJCSITS)* **1**(2), 136–146 (2011)
6. Vyas, I., Saurabh, P.: Danger theory based load balancing (DTLB) algorithm for cloud computing. *IJCSI* **12**(2), 301–306 (2015)
7. Batham, S., Prasad, R., Saurabh, P., Verma, B.: A new approach for data security using deduplication over cloud data storage. In: *3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, MNIT Jaipur, SSRN-Elsevier, pp. 961–966 (2018). ISSN 1556-5068
8. Han, J., Susilo, W., Mu, Y.: Identity-based data storage in cloud computing. *Future Gener. Comput. Syst.* **29**, 673–681 (2013)
9. Durairaj, M., Kannan, P.: A study on virtualization techniques and challenges in cloud computing. In: *J. Sci. Technol. Res.* **3**(11), 147–151 (2014). ISSN 2277-8616
10. Casola, V., Cuomo, A., Rak, M., Villano, U.: The cloud grid approach—security analysis and performance evaluation. *Future Gener. Comput. Syst.* **29**, 387–401 (2013)
11. Arshad, J., Townsend, P., Xu, J.: A novel intrusion severity analysis approach for Clouds. *Future Gener. Comput. Syst.* **29**, 416–428 (2013)
12. Gonzalez, N., Miers, C., Redfólo, F., Simplicio, M., Carvalho, T., Näslund, M., Pourzandi, M.: A quantitative analysis of current security concerns and solutions for cloud computing. *J. Cloud Comput. Adv. Syst. Appl.* **1**, 11 (2012)
13. Saurabh, P., Verma, B.: An efficient proactive artificial immune system based anomaly detection and prevention system. *Expert Syst. Appl.* **60**, 311–320 (2016). Elsevier
14. Saurabh, P., Verma, B.: Immunity inspired cooperative agent based security system. *Int. Arab J. Inf. Technol.* **15**(2), 289–295 (2018)
15. Saurabh, P., Verma, B.: Cooperative negative selection algorithm. *Int. J. Comput. Appl.* **95**(17), 27–32 (2014). 0975-8887

16. Saurabh, P., Verma, B., Sharma, S.: An immunity inspired anomaly detection system: a general framework. In: 7th BioInspired Computing: Theories and Applications, pp. 417–428. Springer (2012)
17. Saurabh, P., Verma, B., Sharma, S.: Biologically inspired computer security system: the way ahead. In: CICS, vol 335, pp. 474–484. Springer (2011)

A Novel Approach for Meta-Search Engine Optimization



S. Siji Rani and S. Goutham

Abstract Search engines are turning out to be the greatest tools for gaining valuable data from the internet. Search engines return the search result to the user query which can be an important result or non-important result. Because, the users naturally look only at the first few pages of search results, and search engine ranking can introduce significant bias to their understanding of the internet and their information gain. When a search query is delivered to several search engines, each individual returns a list of pages based on the ranking. Scientists have confirmed that merging search results in a meta-search engine makes a substantial progress in a search result. Current meta-search engines use several search engines for fetching the results but do not emphasize on the semantic relation of the query for finding the best result. In order to overcome this limitation, a new approach is proposed. The proposed approach can optimize meta-search results using the combination of linear search and semantic search.

Keywords Meta-search engine · Linear search · Optimization · Semantic

1 Introduction

The rapid growth of the web, and searching for data on the web is becoming a difficult task. Web searching is the main activity on the World Wide Web and the main web search engine such as DuckDuckGo, Yahoo, Bing, Google, etc., are commonly used search engines for finding exact results in the massive area of the web. Search engine algorithms are extremely surreptitious. Nobody knows precisely what each search engine weighs and what importance it attaches to each factor in the formula.

S. Siji Rani (✉) · S. Goutham
Department of Computer Science and Engineering,
Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India
e-mail: sijiranis@am.amrita.edu

S. Goutham
e-mail: gouthamsajeev1996@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_34

Each search engine employs its own filters to eliminate spam. For getting data or information, a user enters the query on the search engine and expects the top possible result. Search engine returns a list of web pages which contains the anticipated information. But when more than one search engines are used to performing the similar task, the diversity of results increases which can enhance the satisfaction level of the users. Due to this diversity of web, meta-search engine is getting more popularity. Scientists have confirmed that merging search results in a meta-search engine makes a substantial progress in a search result. Combining the results is an important phase in meta-search engines. When a person enters a query to the meta-search engine, then it selects the suitable set of web search engines based on the search query. Execute a search query in different search engines and extract the data. Then combine these extracted searched data into a single search result list. Combining the search results into a particular result list is a difficult action because of the diversity of results. The main contribution of this paper is to introduce a new approach to optimize meta-search results using the combination of linear search and semantic search.

Some related research work connected to this topic is discussed in Sect. 2. The explanation of solution is discussed in Sect. 3. The algorithm and architecture design are discussed in Sect. 4. In Sect. 5, experimental result are shown. In Sect. 6, conclusion are presented.

2 Related Work

The study in [1] shows that not even one research used mathematical optimization theory for optimizing the search results. Linear programming is a margining technique, which is used to combine the search result obtained from meta-search engines. In this proposed approach, pages are ranked based on the position of each document from member search engines, and then linear programming is used for combining the result for find best page with the highest rank in the result. This approach is one of the best existing techniques for optimizing the meta-search engine result.

A novel search result integration algorithm for meta-search engine is introduced in [2] and the efficiency of the merging algorithm is evaluated. The algorithm contains position merge algorithm and the titles and snippets merging process. The ideas of position merging technique are to take advantage of the real position of the document from individual search engines. If the same search query is passed to other search engines, there is a change that pages which will occur in several result lists of not the same search engines, and the position of the document in the result cannot be equal. By the use of titles and snippets technique, the similarity among the query and documents is calculated. It concludes that the improved merging algorithm could increase the excellence of searching.

The paper [3] describes the Meta search engine ranking method based on user vote system. The user model proposed based on personal behavior is contrasted, which helps the user to vote to search results obtained from meta-search engine, and those updates repeatedly change the result list. Based on the user model, a search result list ranking of the meta-search engine is calculated. The paper [3] discuss a user model based search engine query result ranking method. It constructs a user model, based on user preferences. In addition, it allows users to take an ingenuity vote to the search results. Moreover, it updates dynamically and automatically.

The study in [4] shows no single search engine can index the whole Internet, proposes a new method for vertical search engines with page rank algorithm and latent semantic indexing. The system takes the web pages with same ranks and those are semantically related to a relevant query for the crawling process. Sajeev and Ramya [5] proposes a new system that accepts time information, logical information, or semantic information combining the navigational mode, and sorting out users according to their interest and nature. In this paper, [6] describes a dictionary-based synonyms technique, which helps to find a semantically similar word. Here, the dictionary consists of the most commonly used English word. When a word appears, the learning model will generate synonym for that word.

In this paper, [7] a comparative study of various search engine techniques has been done. For that, they used different factors like n-gram indices, keyword, form, and flash optimization. This tells crawling and fetching are the best method for faster and efficient data retrieval. This paper [8] also describes the different search engine optimization techniques. Rather than giving more importance to search engine ranking, it enables the website owner. For each website owner, it results better user interaction to a website which has owned by them.

3 Proposed Solution

In this section, we give a brief explanation of linear search and semantic search in general. The linear search method is used to re-rank and optimizes meta-search engine result. In the proposed method, initially the user input the query to search. After that, the meta-search engine repeats the query and passes the search query to several other existing search engines. The different results will be obtained from different search engines. For each document or page, it has a dissimilar rank in different search engines so the linear method is used to merge the rank from several existing search engines and re-rank those documents.

For example, for a meta-search engine if there is a use of four separate search engines, then let us have an assumption that the first page of each search engine has 10 documents consider D1, D2, D3, D4, D5, ..., D10 in which some of the documents may be common. For each document, assign some integer value to the search engine result in a particular position. Search engine result contains D1 (Document 1) is on

the first position. Then assign a rank 10 to it. If it is on the last position, then assign a rank 1 to it, and if it is on the second position, then assign a rank 9 to it. Repeat the same procedure for all documents. There may be a chance that document D1 can appear on different search engines on a different position of a search result. As per this example, rank of D1 calculated on the meta-search engine will be the sum of result position on all four separate search engines, i.e., $(10 + 1 + 9 + 5 = 25)$. The similar way for document D2 on the first search engine on the first position, D2 on the second search engine on the second position, D2 on the third one on the third position, and then if D2 is not present on the fourth search engine, its rank will be $(10 + 9 + 8 + 0 = 27)$.

Once the result is obtained from the linear search, those search result rank will be recalculated by semantic search. For semantic search WordNet used, WordNet is one of the best lexical databases that covers all possible synonyms, and assemble words into the synonyms which will return the semantic meaning between them. In this method, a semantic search performed on each document is obtained from the linear search. Then, if any similar word present on the document, its rank will get updated.

Linear search rank of each document is calculated by Eq. (1).

$$LP_{W_i} = \sum_{i=1}^n \text{Pos} * \text{Rank}(P_i) \quad (1)$$

where LP_{W_i} denotes linear rank of each document, Pos denotes position of document in a search result, and n denotes number of search engines. Based on Eq. (2), the final rank of each document is calculated.

$$P_{W_i} = LP_{W_i} + SP_{W_i} \quad (2)$$

where SP_{W_i} denotes semantic rank of each document and P_{W_i} denotes combined rank that contains the best rank for each document.

4 Architecture of Proposed System

4.1 Architecture of Proposed System

In this section, we present a meta-search engine architecture design based on a new result ranking approach shown in Fig. 1. The main objective of our system is to help any user for receiving more relevant search results from the Web. This is accomplished by querying multiple search engines at the same time, then retrieving the results, combine those result using the proposed ranking method and presenting them to the user in a single ranked list.

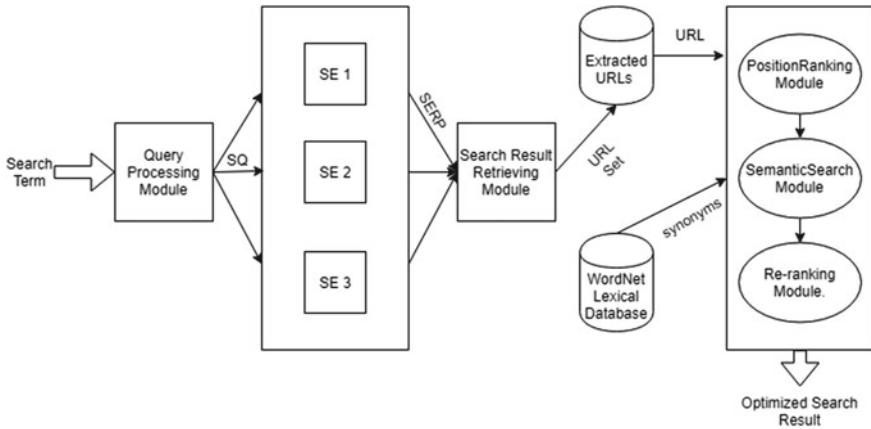


Fig. 1 Architecture of proposed system

The proposed system has the main six important modules, Interface Module, Query Processing Module, Search Result Retrieving Module, Position Ranking Module, Semantic Search Module, and Re-ranking Module. To better understand the proposed system, we will describe the functioning of each module.

Interface module lets the users to type their queries in a simple and natural way using their own keywords or search term without any particular search engine representation or restriction. It receives the query as input, then it sends to query processing module. Query processing module receives the query from a user, then it converts it into a query specific to the search engine and it encapsulates it in the search engine URL. After executing each URL one by one, the search result-retrieving module receives the search result from different search engines as input. Then it extracts the top specified number of URL from each search engines. Search result retrieving module receive the search result from different search engines as input, then it extracts the top specified number of URL from each search engine. Then, each URL sets sends to the next module. Position ranking module receives the URL sets as input. It will perform the proposed linear search algorithm, and then calculate the rank for individual URL. Semantic search module receives the document as input. It will find the synonym for the keyword entered by the user. Then, it will search for that synonym in the document, calculate the rank for each document. Finally, the re-ranking module calculates a new rank for each document based on the result obtained from linear search and semantic search.

4.2 Meta-Search Engine Optimization Algorithm

Algorithm 1: Meta-Search Engine Optimization Algorithm

Input :Search Term

Output :Optimized search results

$i \leftarrow 1, q \leftarrow query;$

while $i < SE_n$ **do**

q is passed to SE_i ;

$D_i \leftarrow$ result of search engine SE_i ;

end

foreach D_i **do**

foreach P_i in D_i **do**

$P_{W_i} \leftarrow P_{W_{D_i}} + P_{W_i}$;

end

end

$list \leftarrow keyword.getSynonyms$;

foreach P_i **do**

if $P_{words} = list$ **then**

$P_{W_i} \leftarrow P_{W_i} + 1$;

end

end

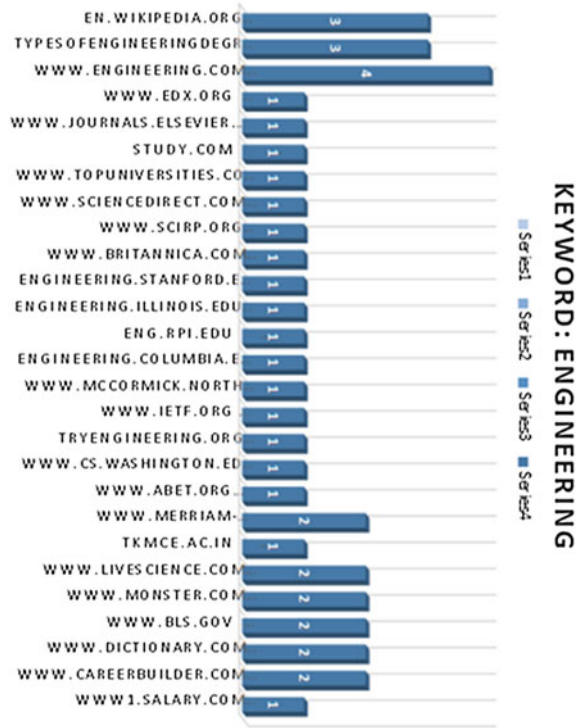
Assume n search engines symbolized by SE_1, SE_2, \dots, SE_n and D_1, D_2, \dots, D_n denotes search result from i th search engine. Search query q containing some keywords is sent to the search engines, and individual search engine returns the initial m result. Each m set of URLs from SE_i search engines is stored in D_i . P_i denotes i th web page presented in the D_i search engine result where P_{W_i} denotes position rank of each web page. $list$ contain set of synonyms for keyword contained in the search query.

This proposed algorithm can be able to combine the results of meta-search engine, finding best list involving of m documents between search result obtained from n search engines for a given search query q .

5 Experimental Result

For testing efficiency of our new meta-search result ranking algorithm, we implemented a prototype of the proposed system. Our meta-search engine supports querying several search engines, and we selected in the first step to fix the number of search engines to 4. And, we selected most of the famous and most used search engines such as Google, DuckDuckGo, Bing, and Yahoo. Instead of the testing system on a document collection, we decided to test it directly on the World Wide Web. The following graph made in the first 10 results of 4 search engines for a keyword engineering. Numbers represent the presence of a particular page, for example, if the

Fig. 2 Meta-search result for the keyword engineering



number is 3 then that particular page is presented in three search engines. Similarly, if the number is 1, then that page is presented only in one search engine. Every search engine may return the same page but with different positions. From the following graph, we can understand the diversity of results. Because, we only extracted the top 10 pages from each search engines. From that search engines, we obtained 27 pages as result. It showed in Fig. 2.

The proposed system shown in Fig. 1 can be used for optimizing the results of the meta-search engine. For this purpose, the table is given below, consisting of four search engines and five documents from each search engine (Fig. 3; Table 1).

Performing linear search on meta-search result gives the best document out of all other documents or search results. From those best documents, semantic search gives the document that more relevant information. Table 2 shows that final optimized result almost similar to Google search result also final result contain document from other search engines DuckDuckGo and Yahoo.

Table 3 shows the similarity score of each search engine with meta- search engine result. Search engines search result similarity increase if the search query is more popular for example above table shows some sample keyword and matching of each search engine result with the final optimized result. For the keywords in the IT technology, all four search engines results are almost the same.

Fig. 3 Search engines participation in the top 10 of the search result

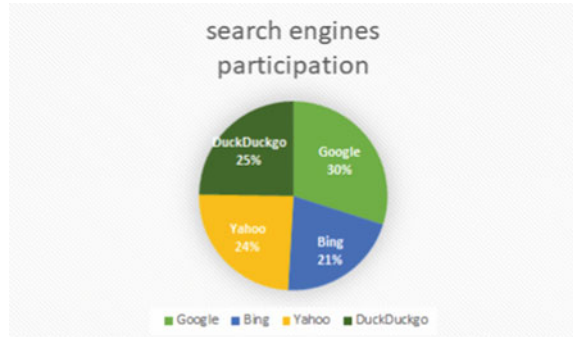


Table 1 Search result for the keyword engineering

Search engines	Document position				
	1	2	3	4	5
Google	D1	D2	D3	D4	D5
Bing	D6	D3	D7	D8	D9
Yahoo	D3	D1	D10	D11	D12
DuckDuckGo	D3	D1	D11	D10	D13

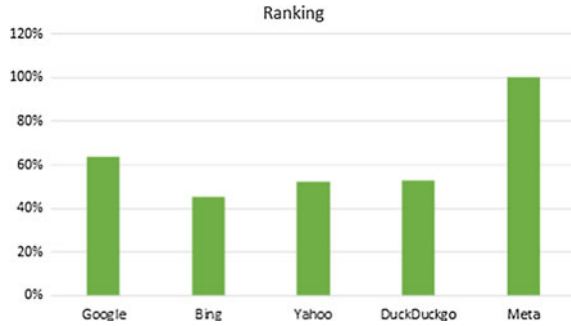
Table 2 Document position for keyword engineering of proposed approach

Approaches	Document position				
	1	2	3	4	5
Linear	D3	D1	D10	D6	D2
Semantic	D1	D2	D3	D10	D6
Combined	D1	D2	D3	D10	D6

Table 3 Result obtained for different keywords

Similarity score table					
Category	Keywords	Google (%)	Bing (%)	Yahoo (%)	DuckDuckGo (%)
Sports	Bat football worldcup	58	40	50	51
Entertainment	Game star gossip	60	35	40	40
IT technology	Iphone samsung	72	60	68	68

Fig. 4 Comparison of the ranked list results



As we can see in Fig. 4, Google has the highest participation in search engine result. Because, final meta-search engine result contains 30 percentage of Google search result. However, DuckDuckGo and Yahoo has almost equal proportions of participation. Bing has the lowest participation among all other, and the reason behind that Bing search result does not match with any other search engine result. Figure 4, shows the ranking of each search engine, Google web page ranking order of 60 percentage similar to final meta-search engine result followed by DuckDuckGo and Yahoo. From this experiment, we can rank these search engines (first Google, second DuckDuckGo, and third Yahoo).

The experimental results shown in this paper are average of testing different search query. It is highly possible that the most significant document with respect to the user query is located in the top of meta-search engine result.

6 Conclusion

In this paper, we presented a novel approach to meta-search engine result optimization algorithm that combines two techniques, linear search, and semantic search. According to the experimental results, we can see that the proposed system produces a well-optimized search result list. In future, we need to integrate multi-threading technique to our system, that will speed up the processing and save time. Multi-threading technique can put the programs that need time for processing background. So, we can do the semantic search on multiple web pages at the same time, and it will help the system to run so faster and deliver the best search result.

References

1. Amin, G.R., Emrouznejad, A.: Optimizing search engines results using linear programming. *Expert Syst. Appl.* **38**(2011), 11534–11537 (2011)
2. Fu-yong, Y., Jin-dong, W.: An implemented rank merging algorithm for meta search engine. In: *International Conference on Research Challenges in Computer Science* (2009)

3. Yan, L.U., Meng, X.U., Yuanyi, L.I., Weihui, H.U.: A user model based ranking method of query results of meta-search engines. In: *International Conference on Network and Information Systems for Computers* (2015)
4. Pavani, K., Sajeev, G.P.: A novel web crawling method for vertical search engines. In: *2017 International Conference Advances in Computing, Communications and Informatics (ICACCI)* (2017)
5. Sajeev, G.P., Ramya, P.T.: Effective web personalization system based on time and semantic relatedness, *Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sept. 21–24, 2016 (2016)
6. Rani, S.S., Sreejith, K., Sanker, A.: A hybrid approach for automatic document summarization. In: *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*
7. Chhabra, S., Mittal, R., Sarkar, D.: Inducing factors for search engine optimization techniques: a comparative analysis. In: *2016 1st India International Conference on Information Processing (IICIP)*, Delhi, 2016, pp. 1–4
8. Lemos, J.Y., Joshi, A.R.: Search engine optimization to enhance user interaction. In: *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, 2017, pp. 398–402

Why Adopting Cloud Is Still a Challenge?—A Review on Issues and Challenges for Cloud Migration in Organizations



Mohammed Shuaib, Abdus Samad, Shadab Alam
and Shams Tabrez Siddiqui

Abstract Adoption of new technical innovation contributes not only by increasing the business goals, but also facilitates the growth and competitiveness of the organization. Organizations including SME are reluctant to migrate their existing system to cloud platform because of various cloud adoption challenges. Various technical and nontechnical factors are responsible for cloud adoption and migration. Unavailability of a well-defined cloud strategy development model, less prior expertise of the cloud domain and unsurely about how and when to initiate cloud adoption or migration are the key challenges while moving to cloud platform. Most of the organizations are now thinking to migrate to a cloud platform, it is imperative for organizations to critically explore the challenges related to their business. Thus, there is the need of defining challenges associated during cloud adoption and a well-defined cloud strategy model for a successful migration to the cloud domain. This paper aims to investigate the key determinants affecting cloud adoption most in the organizations. Further, the paper identifies existing cloud frameworks and critically evaluates them based on their effectiveness and drawbacks. Factors affecting the cloud adoption process are identified, and a hypothetical framework is proposed based on identified variables. Results suggest that technical factors, organizational factors, and some external factors have a positive impact on cloud adoption.

Keywords SLA (Service-Level Agreement) · CSP (Cloud Service Provider) · TOE (Technological, Organizational, and Environmental) · DOI (Diffusion of

M. Shuaib · S. Alam · S. T. Siddiqui (✉)
Jazan University, Jizan, Saudi Arabia
e-mail: stabrezsiddiqui@gmail.com

M. Shuaib
e-mail: talkshuaib@gmail.com

S. Alam
e-mail: s4shadab@gmail.com

A. Samad
Aligarh Muslim University, Aligarh, India
e-mail: abdussamadamu@gmail.com

Innovation) · TAM (Technology Acceptance Model) · SME (Small and Medium Enterprises)

1 Introduction

Organizations are reshaping their business model by restructuring the existing IT model for organizational growth and competing with their existing peers. Organizations are in search of dynamic, elastic, and scalable IT infrastructure for a fast and dynamic work culture within the organization. Cloud computing is capable of delivering all such features like a low-cost IT solution for the organizations [1]. Lower cost IT solution, dynamic scaling, scalability, higher availability, QoS, pay-per use, and ease of use are some major benefits offered by cloud computing and can benefit any business organizations. Larger organization have higher rate of cloud adoption because of resource availability, on the other hand, small and medium organizations have a very low rate of cloud adoption.

To survive in this cut through an environment of marketplace and for having an extra edge in business, organization must think of adopting latest technologies. Cloud computing provides a more flexible and scalable computational infrastructure to the organization. Migrating or adopting the cloud technology does not have a single dimension. The migration to the cloud platform includes the migration of complete or partial part of organization data, application, or services. This migration of data, application, and services can lead to several challenges and vulnerabilities to the organization. Security threats, privacy and trust, service availability, SLA, interoperability, network bandwidth, and virtualization threat are some major risks associated with cloud migration/adoption process [2, 3].

Thus, a more focused research is required to develop a cloud strategy model for the cloud migration process. The paper aims to focus the key challenges having maximum impact on cloud migration in organizations. The research will identify the factors and challenges, which will slow down the growth of cloud adoption.

2 Background Study

2.1 Cloud Computing

Cloud computing paradigm has completely changed the perception of the organization about infrastructure, computing power, and services. The main concept of cloud computing is to offer various services and infrastructure in an on-demand basis with pay as you go model from anywhere and any device. In simple words, cloud computing is a dynamic service provisioning of resources over the Internet. Cloud computing ensures independence to its end user, to select, configure, and personal-

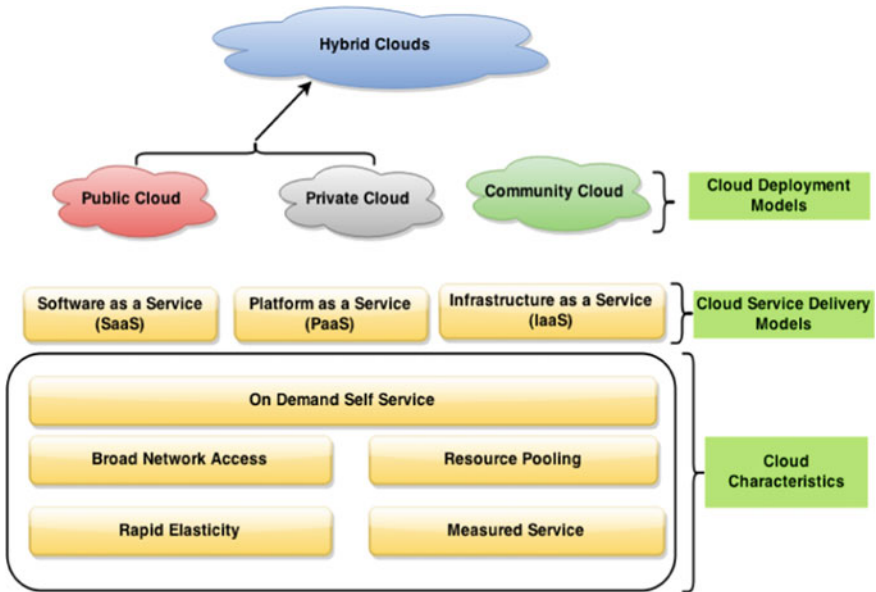


Fig. 1 [4] NIST-based cloud computing model

ize the cloud services based on the requirements. Virtualization, multi-tenancy, and network environment are the backbones of cloud computing [4]. These technologies play a vital role in delivering services and computing infrastructures dynamically. A virtualization technology uses virtual machines to offer the abstraction of real hardware, software, or network resources. Multi-tenancy is used for creating multiple instances of any application to fulfill the demand of various users at the same time [4].

Most of the available definitions of cloud computing in the literature are based on five key characteristics includes “*multi-tenancy, scalability, elasticity, pay-per-use, and self provisioning*” to define cloud system [4]. Multi-tenancy is used for creating multiple instances of the real resources to its user using one single implementation. Scalability serves as scaling of the resources, whereas the elasticity can be used for dynamic service provisioning of resources available based on the requirement of the user. In self-provisioning, cloud computing offers authorization to end user for personalizing the resources, while paying for those computing resources in pay as you go model. A NIST-based definition of the cloud computing model is represented in Fig. 1 [4].

3 Research Process

A literature review analysis methodology is used for the identification of key challenges and factors associated with the adoption of cloud computing. The major reason to use literature review methodology is to collect the pre-information available in the literature about the research domain. A step-by-step process is followed to conduct the research literature review. Searching, analyzing, reading, evaluating, and reviewing are the steps followed to conduct critical literature analysis.

The searching step comprised of systematic scanning of major research journals and conferences including (ACM, Springer, and Elsevier), Cloud Security Alliances report, McAfee security reports, and research articles containing relevant research of the area. Once all the desired articles, papers, and reports were collected those are read and analyzed.

Finally, all the articles related the research areas are critically reviewed. Several information like cloud computing background, benefits, cloud adoption factors, and challenges are assessed from available literature. This research has reviewed the available articles on the two perspectives: first technical perspective and second organizational perspective. Based on both the technical and organizational perspective, the factors affecting cloud migration most are identified. The domain research articles are critically analyzed and summarized in Table 1.

4 Literature Review

The research aims to identify the various technical and nontechnical challenges that affect the cloud computing the most. In this section, research explains the findings of the literature review methodology.

4.1 Literature Review Analysis

Previous research suggests that migrating to cloud platform has greater benefits and advantages to business organizations. Features such as dynamic service provisioning, on-demand service, remote access, and scalability offer greater flexibility to organizations; for establishing their new IT infrastructure or migrating existing IT platform to cloud platform. The research process implicates that the literature review has been performed based on the two major perspectives organizational and technical. TOE (Technological, Organizational, and Environmental) and DOI (Diffusion of Innovation) are the most widely used theories used in previous research studies.

TOE framework is useful for the identification of components that influence any organization while adopting innovative technologies in three different contexts including, Technological, Organizational, and Environmental [14]. The first context

Table 1 Analysis of previous research articles

Research article	Cloud computing background study	Cloud adoption/migration challenges	Cloud adoption theory	Study perspective
Low et al. [5]	✓	✓	Combined TOE and DOI theory	Organizational and technical
Khajeh-Hosseini et al. [6]	✓	✓	Cloud migration framework	Technical
Lin and Chen [7]	✓	–	Semi structured study	Organizational
Rafee et al. [8]	✓	✓	–	Technical
Jlelaty and Monzer [9]	✓	✓	Theoretical study	Technical
Morgan and Conboy [10]	✓	✓	Empirical analysis	Organizational
Abdollahzadegan et al. [11]	✓	✓	TOE	Organizational and technical
Oliveira et al. [12]	✓	✓	TOE and DOI	Organizational and technical
Gholami et al. [13]	✓	✓	Generalized framework	Organizational Technical

refers to technological components in the organization such as technical advantage, compatibility with existing system and complexity. The second context refers to organizational components in organization such as organization structure, support, organizational competitiveness, and prior domain experience. The third context refers to environmental components in organization including business competition, and peer pressure [15].

Several key studies found in the literature having adopted TOE as the base theory framework. In [15], the authors have proposed a framework to identify the key factors affecting cloud adoption. The proposed framework is validated using the data of service based organization [15]. Research concluded that “*reliability, security, resource availability, policy compliance, and adaptability to existing platform*” are the key factors associated with cloud adoption. However, the study lacks in providing a clear recommendation based on the framework. In [11], a conceptual framework is proposed to analyze the factors affecting cloud adoption in small and medium enterprises focusing on organizational context. The research concluded that firm size, organizational management support, technology readiness to adopt, and relative advantage are the key determinants of cloud adoption.

In [16], a more concentric study with risk analysis focusing on SME, to find barriers to cloud adoption has been performed by authors. The research concluded that factors including business concerns, IT capabilities, and perceived benefits are the key determinants of cloud adoption, while business concern is the most dominant determinant that influences the choice of deployment model for the organization [16]. However, research lacks in proposing technical challenges faced during cloud migration. In [17], the authors have proposed a framework using TOE model for the analysis of challenges associated with cloud adoption. The research concluded that relative advantage, complexity, compatibility, management support, firm size, and peer pressure are the most dominant determinants of cloud adoption. The proposed framework is validated using quantitative research using the data of 24 global organizations across the world. However, research does not focus on technical perspective during cloud migration and adoption.

DOI is another widely accepted theory used for identification of factors that hinder information system adoption. Theory concludes that the adoption of any technology is directly dependent on five key components “*relative advantage, compatibility, complexity, observability, and trialability*” [12]. These components are helpful in developing a better understanding from an organizational perspective. In [7], DOI theory is adopted to analyze the factors affecting cloud adoption in organization. Research concludes that relative advantage and compatibility are the key determinants of [7]. A semi-structured study with the qualitative analysis is used to validate the research framework.

In some latest research, a combination of TOE and DOI theory is used for cloud adoption. In [18], the authors have proposed a research framework for cloud adoption. The proposed framework has been validated using the data of a specific country and domain. Research concluded that “*organization management support, organization size, relative advantage, complexity, and technology readiness*” are the dependent variables of cloud adoption [18]. In [19], an integrated TOE, DOI framework is used for the realization of cloud adoption factors in Saudi Arabia. Research concluded that “*organization size, organization management support, business peer pressure, technological readiness*” are the determinants of cloud adoption. Except these adoption theories, “*TAM (Technology Acceptance Model) and UTAUT (Unified theory of acceptance and use of technology)*” are also used as adoption theory for cloud adoption. In [20], a more generic cloud migration model is proposed. The proposed model consists of three different phases including plan, design, and enable. However, the proposed model is a conceptual model based on a domain study. The major limitation of the available research can be summarized as, no study is available that combines both IT challenges and vulnerabilities, with organizational challenges to identify the major challenges during cloud migration and adoption.

4.1.1 Analysis of Cloud Adoption Frameworks

Previous studies can broadly be categorized into two sections first dealt with the concept of adoption theories and second is based on cloud adoption framework based

on IT service and infrastructure. Based on cloud deployment and service delivery models there exist several frameworks. Some most common cloud adoption frameworks are: “*CBMF (Cloud Business Model Framework)*, *LCCF (Linthicum Cloud Computing Framework)*, *ROI (Return on Investment)*, *Performance Metrics Framework*, *OCCCSF (Oracle Consulting Cloud Computing Services Framework)*, *IBM (IBM framework for cloud adoption)*, *ClouSim*, *BCF (BlueSky Cloud Framework for e-Learning)*, *Hybrid ITIL V3 for Cloud*, *CCBF (Cloud Computing Business Framework)*” are some well-known cloud frameworks [21, 22]. The detailed comparison of these entire frameworks is presented in Table 2. Each cloud adoption framework has its own pros and cons. The challenges associated with most of the frameworks need to be considered thoroughly and a more practical framework is required, which is free from all these ambiguities.

5 Cloud Adoption Factors

Literature review analysis is used as a research process to investigate the factors affecting the cloud migration process the most. On the other hand, cloud adoption can have different perspectives ranging from individual users to business organizations. Thus, in order to define the scope of this paper, research is limited to cloud adoption within the organizations. Most widely used cloud adoption theories and cloud adoption framework is studied and analyzed. The approach used in the research is also supported by several researches, as the approach used a fixed set of criterion to review the literature. Based on the detailed analysis of the current existing studies from organizational and technical perspective, cloud adoption factors are identified. The section consists of the factors that must be taken into account while migrating to cloud platform. These factors are also used for proposing a hypothetical framework for cloud adoption/migration.

5.1 Virtualization

Virtualization is the base technology used in cloud computing to offer resource sharing. Virtualization management ensures the administration of the virtualization process that includes virtual machine creation, monitoring, and service delivery. The virtualization management can add much vulnerability to the cloud adoption process. Virtual machine security and attacks, server consolidation, and resource sharing are some major challenges associated with virtualization [23, 24]. Virtualization management has never been the part of previous cloud adoption theories and adoption frameworks. Thus, a more positive insight needs to be provided in virtualization management during cloud adoption.

Table 2 A comparative analysis of cloud adoption framework

IT Framework	Contribution	Limitation
CBMF (cloud business model framework)	<ul style="list-style-type: none"> • The framework consists of four different layers as Infrastructure, application, platform, and business model • All layers presented in the model has specific roles and responsibilities and have a direct connection with the main business model layer 	<ul style="list-style-type: none"> • Recommendations are not provided that how an organization can incorporate this framework for cloud migration/adoption • Framework is not validated using a case study
LCCF (Linthicum cloud computing framework)	<ul style="list-style-type: none"> • A self-realized framework is provided 	<ul style="list-style-type: none"> • The framework is not validated • No recommendation is provided for the organizations
ROI (return on investment)	<ul style="list-style-type: none"> • A cost analysis based framework is provided for cloud adoption 	<ul style="list-style-type: none"> • The framework is proposed with a specific case study. However, no universal mathematical foundation is proposed for computing ROI • Each organization can't use mathematical foundation for cost analysis
IBM (IBM framework for cloud adoption)	<ul style="list-style-type: none"> • Most feasible and practical solution is proposed that covers most of the organizations with all deployment model 	<ul style="list-style-type: none"> • Most practical solution but no case study is provided that how an organization can incorporate it
CCBF (cloud computing business framework)	<ul style="list-style-type: none"> • Based on the literature reviews, insights, methodologies a conceptual framework is proposed • The framework is validated with detailed analysis and case study 	<ul style="list-style-type: none"> • The framework does not explain how it dealt with cloud adoption challenges

5.2 Security and Privacy

Security and privacy are a major hindrance in cloud adoption. Almost all previous research and security reports had highlighted security and privacy as the major threat to cloud adoption. Threats such as data theft, data loss/leakage, organization privacy, DDoS attacks, malicious attacks, and identity and access management are the major loopholes in cloud computing [24, 25]. However, complete assurance to security and privacy cannot be guaranteed but needs to improvise so that the probability of cloud adoption among organizations can be improved [8].

5.3 Interoperability

Interoperability is the major challenge in cloud migration [1]. Interoperability provides a mechanism to migrate from legacy system to the cloud platform. This integration of organization data, services, and application between these platforms is a daunting task. This challenge must need to address to improve cloud adoption ration among organizations. Compatibility is also another challenge as different cloud service providers have a different set of rules and regulation for migration. Thus, if an organization wants to switch between different CSPs then it proves to be a challenging task for the organization. This creates a situation like vendor lock-in during cloud migration [3].

5.4 Service-Level Agreement

SLA (service-level agreement) works as a bond between the service provider and organization. SLA consists of all the services and offerings such as data backup management, billing, legal documents, performance metrics, and others. However, there is no standard format for SLA and completely depends upon the service provider. Thus organizations are unsure about what exact services and offerings are provided by the service provider. Very few researchers have included SLA as the major determinant for cloud adoption in their research [2, 3].

5.5 Organizational Context

At the organizational level there exist several challenges during cloud adoption. Challenges are like organizations intention to adopt the new environment, the existing business environment, people's intention to adopt the technology, organization support, and size [26, 27]. The intent of cloud adoption in any organization is directly

proportional to the prior expertise of the organization in the domain and management willingness. Management willingness plays a vital role as they better know how to manage the resources and utilize them when adopting cloud computing.

However, except these lack of a defined cloud strategy model also creates the problem in the successful migration to the cloud platform that is “*rooted in cloud service delivery model*” [28]. Unsurely in the organizations about when and how to initiate cloud adoption, and limited understanding of the cloud domain are also responsible for cloud migration failure [28]. While migrating to cloud organization should have a clear insight into which service and deployment model they required and how these services change their legacy system. Lack of prior expertise to cloud computing, unavailability of proper cloud strategy and frameworks that guide the organization in adopting cloud computing makes it difficult for successful implementation of cloud.

6 Cloud Adoption Framework

TOE and DOI are the two most popular theories for Information System adoption. TOE is the most suitable adoption theory as it proposes three most concrete contexts regards to information system adoption.

However, TOE lacks in few organization-level contexts, which are found in DOI theory. Thus, it is more useful to propose a framework based on integrating both the adoption theories. This study is also supported by previous studies. This research concentrates on the factors affecting cloud computing the most focusing on organizational and technical context. The proposed adoption framework focusing on four different contexts with a combination of two IS adoption theories. The four different contexts are organizational, technological, economical, and eternal.

Based on the literature review analysis the most influential cloud adoption factors are identified. The constructs for cloud adoption are virtualization management, security, privacy and trust, QoS, availability, backup and recovery, compatibility, interoperability, scalability, return on investment, cost estimation, economical risk, relative advantage, management support, organization size, technological readiness, organizational competitiveness, SLA, and regional regulations. The proposed hypothesis argues that all these constructs have a positive impact on cloud adoption. The constructs are supposed to be the independent variables and cloud adoption as the dependent variable. The relation of the construct and hypothesis is shown in Fig. 2.

The proposed framework will be validated through the data collection from various organizations having adopted cloud computing or thinking to migrate on the cloud platform.

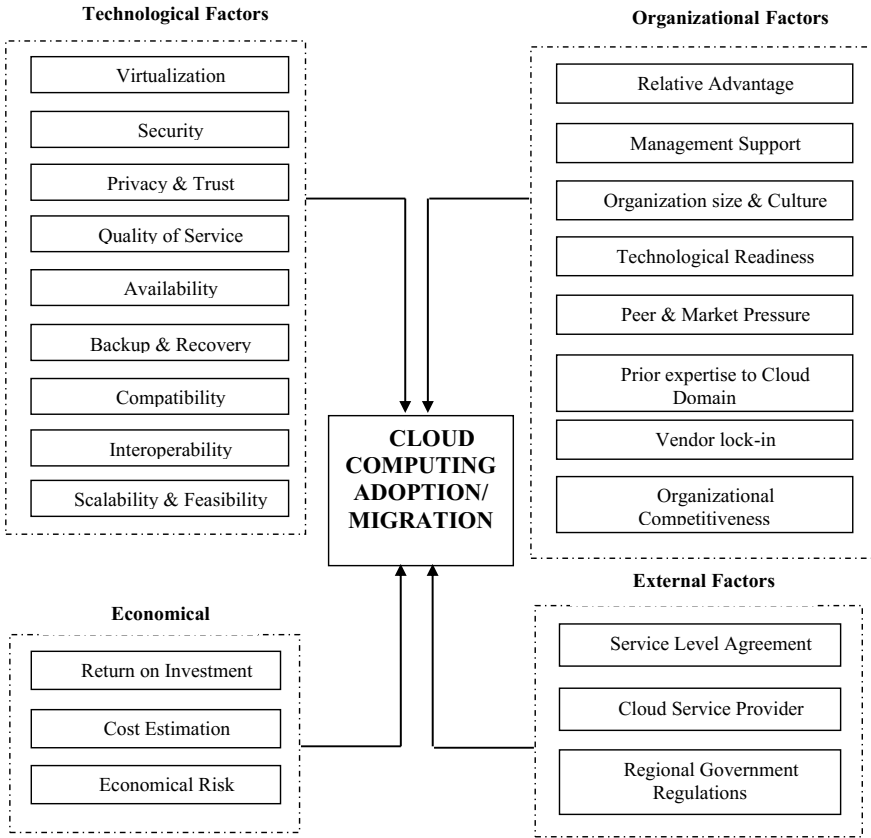


Fig. 2 Research framework for cloud adoption

7 Conclusion and Future Prospective

This research extends the cloud migration in more detailed analysis using both qualitative and quantitative analysis using both cloud domain expertise and organizations that have migrated and adopted cloud technologies. The research identified key technical and nontechnical factors affecting cloud migration. These challenges include organizational context, technical context, and external factors. Research has found that virtualization management, security, privacy and trust, backup and recovery, compatibility, interoperability, scalability, return on investment, cost estimation, economical risk, relative advantage, management support, organization size, and technological readiness are some major barrier to cloud adoption. Technical challenges like virtualization management, SLA, and government regulation, compliance are studied separately in the study and analyzed in more detail which assumes to be the most predominant factors in cloud migration.

This analysis is very critical for the organization perspective to answer the key questions like what are the key challenges an organization needs to address while migrating to cloud platform. The research will also be helpful to have an insight into the organization ready to migrate to a cloud platform? If yes then how they manage the challenges associated with the cloud migration process. The research framework is a must requirement for the successful migration to the cloud platform. Research has proposed a hypothetical cloud adoption framework based on the factors identified in this research. The proposed research framework will work as a tool for organizations to analyze the challenges while migrating to the cloud platform, as well as provide a better understanding of the drawbacks, benefits of other frameworks.

References

1. Avram, M.-G.: Advantages and challenges of adopting cloud computing from an enterprise perspective. *Proc. Technol.* **12**, 529–534 (2014)
2. Rad, B.B., Diaby, T., Rana, M.E.: Cloud computing adoption: a short review of issues and challenges. In: *Proceedings of the 2017 International Conference on E-commerce, E-Business and E-Government*, pp. 51–55 (2017)
3. Alharthi, A., Alassafi, M.O., Walters, R.J., Wills, G.B.: Towards a framework to enable the migration process to educational clouds in Saudi higher education. In: *2016 International Conference on Information Society (i-Society)*, pp. 73–76 (2016)
4. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. *J. Internet Serv. Appl.* **1**(1), 7–18 (2010)
5. Low, C., Chen, Y., Wu, M.: Understanding the determinants of cloud computing adoption. *Ind. Manag. Data Syst.* **111**(7), 1006–1023 (2011)
6. Khajeh-Hosseini, A., Sommerville, I., Bogaerts, J., Teregowda, P.: Decision support tools for cloud migration in the enterprise. *arXiv preprint arXiv:1105.0149* (2011)
7. Lin, A., Chen, N.-C.: Cloud computing as an innovation: perception, attitude, and adoption. *Int. J. Inf. Manage.* **32**, 533–540 (2012)
8. Ahmed, M., Chowdhury, A.S.M.R., Rafee, M.M.H.: An advanced survey on cloud computing and state-of-the-art research issues. *IJCSI Int. J. Comput. Sci. Issues* **9**(1), 1694–0814 (2012)
9. Jelaty, M., Monzer, Y.: Factors in cloud computing adoption (2012)
10. Morgan, L., Conboy, K.: Factors affecting the adoption of cloud computing: an exploratory study (2013)
11. Abdollahzadegan, A., Hussin, C., Razak, A., Moshfeqh Gohary, M., Amini, M.: The organizational critical success factors for adopting cloud computing in SMEs (2013)
12. Rogers, E.M.: *Diffusion of Innovations*. Simon and Schuster, New York (2010)
13. Gholami, M.F., Daneshgar, F., Beydoun, G., Rabhi, F.: Challenges in migrating legacy software systems to the cloud—an empirical study. *Inform. Syst.* **67**, 100–113 (2017)
14. Baker, J.: The technology–organization–environment framework. In: *Information Systems Theory*, pp. 231–245. Springer (2012)
15. Nkhoma, M.Z., Dang, D.P., De Souza-Daw, A.: Contributing factors of cloud computing adoption: a technology-organisation-environment framework approach. In: *Proceedings of the European Conference on Information Management & Evaluation*, pp. 180–189 (2013)
16. Hsu, P.-F., Ray, S., Li-Hsieh, Y.-Y.: Examining cloud computing adoption intention, pricing mechanism, and deployment model. *Int. J. Inf. Manage.* **34**, 474–488 (2014)
17. Borgman, H.P., Bahli, B., Heier, H., Schewski, F.: Cloudrise: exploring cloud computing adoption and governance with the TOE framework. In: *2013 46th Hawaii International Conference on System Sciences (HICSS)*, pp. 4425–4435 (2013)

18. Oliveira, T., Thomas, M., Espadanal, M.: Assessing the determinants of cloud computing adoption: an analysis of the manufacturing and services sectors. *Inf. Manage.* **51**, 497–510 (2014)
19. AlBar, A.M., Hoque, M.R.: Determinants of cloud ERP adoption in Saudi Arabia: an empirical study, In: 2015 International Conference on Cloud Computing (ICCC), pp. 1–4 (2015)
20. Gholami, M.F., Daneshgar, F., Low, G., Beydoun, G.: Cloud migration process—a survey, evaluation framework, and open challenges. *J. Syst. Softw.* **120**, 31–69 (2016)
21. Chang, V., Walters, R.J., Wills, G.: The development that leads to the cloud computing business framework. *Int. J. Inf. Manage.* **33**, 524–538 (2013)
22. Chang, V., Walters, R.J., Wills, G.: Review of cloud computing and existing frameworks for cloud adoption (2014)
23. Zhang, Q., Cheng, L., Boutaba, R.: Cloud computing: state-of-the-art and research challenges. *J. Internet Serv. Appl.* **1**, 7–18 (2010)
24. Ali, M., Khan, S.U., Vasilakos, A.V.: Security in cloud computing: opportunities and challenges. *Inf. Sci.* **305**, 357–383 (2015)
25. Shuaib, M., Samad, A., Siddiqui, S.T.: Multi-layer security analysis for hybrid cloud. In: 6th International Conference on System Modeling and Advancement in Research Trends, SMART-2017, pp. 526–531 IEEE (2017)
26. Schneider, S., Sunyaev, A.: Determinant factors of cloud-sourcing decisions: reflecting on the IT outsourcing literature in the era of cloud computing. *J. Inf. Technol.* **31**, 1–31 (2016)
27. Alkhater, N., Wills, G., Walters, R.: Factors influencing an organisation’s intention to adopt cloud computing in Saudi Arabia. In: 2014 IEEE 6th International Conference on Cloud Computing Technology and Science (CloudCom), pp. 1040–1044 (2014)
28. Smith, D.: Cloud computing deployments should begin with service definition. Gartner Report (2016)

Analysis of Block-Level Data Deduplication on Cloud Storage



L. Suresh and M. A. Bharathi

Abstract Data storage in the cloud is of prime importance for organizations, individuals, business as the storage of data is increasing exponentially. In the modern era, multiple data blocks of files across the sectors of storage disk are duplicated. Deduplication is a common technique for reducing a large amount of duplicate data. Deduplication storage optimization technique is an efficient method of reducing the chunks with pointer address mapping single instance across the storage data disk. Network transfer protocol for deduplication is major impact factor for performance. Network bandwidth, disk storage optimizations are parameters for real time data transfer in the deduplication protocol. In this analysis, we are examining the secure deduplication on cloud storage data on fixed block level and file level deduplication with new Protocol for transferring the data to cloud storage. Client–Server data transfer model architecture for deduplication technique will be analyzed for the new protocol data transfer.

Keywords Deduplication · Cloud storage · Block level · Dedupe

1 Introduction

Cloud storage [1] is an information storage model in which users application data is backed up in the server remotely over the network (internet), which can be maintained, managed for providing the service to users. Data in the cloud storage requires the elimination of repeated data across the file system, only a single instance of data stored in the storage. Data deduplication is a general way of removing the duplicate repeated blocks of data in the storage. Repeated blocks of data in the files are replaced with an address of data block pointer, which holds the single data segment across the

L. Suresh (✉) · M. A. Bharathi
Department of Computer Science, BMSIT & M, Bangalore, India
e-mail: Edu2mtech@gmail.com

M. A. Bharathi
e-mail: bharathi_m@bmsit.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_36

storage area. Data Storage area contains repeated copies of duplicate files. Storage disk is imaged or backed up, all the duplicate copies of files are saved increasing the IO operations for writing the duplicate files for backup. With the data segment deduplication, only single file is stored in the backup, duplicate multiple files are referenced back to the stored single copy of file [2].

Data deduplication is a single instance file system storage which reduces the redundant chunks of data and storage. It ensures the significant data benefits for archiving in the file storage media, disk storage or tape. Duplicate blocks of data across the storage are removed with the address of data block pointer which contains data block which references all the duplicate blocks. Deduplication eliminates the redundant data segments from the cloud storage by reducing the size of stored data. Deduplication in the cloud storage improves the faster transfer of data, bandwidth over the network, resulting in the efficient data protection.

Data Deduplication ensures the reduction of storage of files, data blocks and replication data backups which benefits the storage disk cost efficiently. Enterprise, application, end users can backup data with less cost of disk storage. The bandwidth of the network is increased efficiency by reduction of repeated data blocks of chunks transmitted over the network. Environmental benefits is attained by less power in terms of electricity, reduction of storage space is for data in both primary and remote locations of data storage servers [3].

2 Data Deduplication Methods

Deduplication techniques method has file level, block level for removing the redundant data across the storage area. Finding similar files, blocks of data is key success for deduplication. A major concern for deduplication is the maintenance of reference pointer for duplicate data segments, files. It aims at faster access of data across the storage.

File-level data deduplication is a method of reducing the duplicate data files and block-level deduplication works by unique block across the files in the storage media or server. Block segments of data in dedupe comparison fixed by block or variably sized block by eliminating the duplicate blocks.

File-level deduplication: File-level deduplication uses minimum resource for data storage over larger physical storage of data. File-level deduplication cannot remove duplicate data checks across the file. File level is less reliable and consistent compared to block level deduplication. A small difference in the file data affects the entire file need to be back up.

Block-level deduplication: The major advantage of block-level deduplication is that it will remove the data chunks across the file system of storage area. The disadvantage of block level is it cannot implement over large amount of physical storage or networks of server [4].

Reduction ratios of data only in the 5:1 or less range in file-level deduplication. Block level have redundant data segments in the ratio of 20:1 to 50:1. Block level

deduplication assembles data blocks depends on the hash checksum that targets the single copy of the data segment. File-based data deduplication store single copy of files and duplicate file points to the existing single file in the storage that is less reassemble.

3 Literature Review

Cloud storage is a new technical modern jargon shift in Internet information technology. Data deduplication saves storage space and reduces the amount of bandwidth for data transfer from client to server cloud storage. Several methodologies have been applied to data deduplication based on the file and block level with fingerprint or hash techniques. The objective of cloud storage is the reduction of cost and continuous service to business needs. Effective maintains of cloud storage in terms block level/file level deduplication is the major goal in days cloud storage competition.

International papers, Journals from the various publications are reviewed for deduplication on cloud storage; summary of the technicality is shown in Table 1.

All the above method/techniques of deduplication are based on the file/block level which is commonly used in the cloud storage. Comparison of the method which is mentioned in the reviews paper is explained in the below section.

4 Comparison of Various Deduplication Techniques

Deduplication method has many methods for reducing the redundant data, mainly there are two types file level and block level method. File-level method based the complete file content comparison versus cloud storage stored files for removing the duplicate data. Block level is a chunk of data comparison versus multiple unique data blocks stored across the files in the cloud storage. The below methods are compared from the various review paper for file/block level redundant data deduplication.

Redundancy Removal Multimedia File System (RRMFS) method is file level technique used in cloud storage for finding the dedupe data, where robin fingerprinting is based on block-level dedupe method. Block level method is more efficient method for removing redundant data across the storage media. Block level has many complexes in archiving the checksum comparison of the stored checksum for finding dedupe block data. FP tree based on the hash method for file dedupe, Signcryption method for encrypted file dedupe. All the methods in the above bases on single instance server centralized cloud storage, whereas DAC and Heterogeneous scheme of management is based on the multiple cloud storage dedupe. Chord algorithm is a two-side client-server-stored dedupe technique.

Table 1 Literature review of data deduplication on storage

Title of paper and publisher's	Methods/techniques	Merits	Demerits
A storage solution for multimedia files to support data deduplication, "Wang et al. [10]"	<ul style="list-style-type: none"> • Redundancy removal multimedia file system (RRMFS) for implementation of data deduplication • Distributed tree directory to remove duplicate multimedia files 	<ul style="list-style-type: none"> • High data redundancy to store multimedia files efficiently • Easy to find the directories across the data center using the distributed directory tree • HDFS compares the redundant multimedia files while saving to data center storage for deduplication 	<ul style="list-style-type: none"> • Finding the duplicate multimedia files in terms of speed and accuracy • Only single-format multimedia files is considered for redundancy • Identifying different formats of multimedia files for deduplication in terms of consistency accuracy and speed
An efficient and secure deduplication scheme based on Rabin fingerprinting in cloud storage, "Su et al. [11]"	<ul style="list-style-type: none"> • Rabin fingerprinting for chunking block level deduplication 	<ul style="list-style-type: none"> • Perform deduplication based on fingerprints of data blocks before these blocks are encrypted by users • Trusted third-party server is introduced to randomize the convergent keys and manage them • Dictionary brute force attacks are avoided by proposed method scheme which is more secure 	<ul style="list-style-type: none"> • Deduplication on data before encryption • Third party server for key randomization which is not trusted by the enterprises
A data deduplication method in the cloud storage based on FP-tree, "Haoran et al. [12]"	<ul style="list-style-type: none"> • FP tree based removal of redundant data • Consistent hashing method using the Data distribution algorithm 	<ul style="list-style-type: none"> • Improved FP tree algorithm for data deduplication • Decreased memory consumption while searching the redundant data 	<ul style="list-style-type: none"> • Not suitable for large data files in the cloud storage • Searching is time-consuming for large data files

(continued)

Table 1 (continued)

Title of paper and publisher's	Methods/techniques	Merits	Demerits
Big data cloud deduplication based on verifiable hash convergent group signcryption, "Cho et al. [13]"	<ul style="list-style-type: none"> • Signcryption method for redundant data removal • Verifiable hash convergent group signcryption (VHCGS) 	<ul style="list-style-type: none"> • Eliminate redundant encrypted data owned by different users 	<ul style="list-style-type: none"> • Lack of inter-domain user data deduplication on cloud storage
DAC: improving storage availability with deduplication-assisted cloud-of-clouds, "Wua et al. [14]"	<ul style="list-style-type: none"> • Deduplication-Assisted primary storage system in Cloud-of-Clouds (short for DAC) 	<ul style="list-style-type: none"> • Data reference among multiple cloud server providers for deduplication • Solves the vendor lock-in problem in the cloud computing 	<ul style="list-style-type: none"> • Security loophole in terms of encryption for data reference • Multiple data replica needed for reliability
Public auditing for encrypted data with client-side deduplication in cloud storage, "Kai et al. [15]"	<ul style="list-style-type: none"> • Client-side deduplication • Proxy re-encryption for public auditing of encrypted data 	<ul style="list-style-type: none"> • Data auditing of encrypted data TPA • Secure and efficient deduplication scheme for auditing 	<ul style="list-style-type: none"> • Encrypted Cipher text is altered by unknown third parties, which allows to decrypt by others
Smart data deduplication for telehealth systems in heterogeneous cloud computing, "Keke et al. [16]"	<ul style="list-style-type: none"> • Deduplication based on SD2M (smart data deduplication model) on telehealth systems 	<ul style="list-style-type: none"> • Performance-oriented implementation for current cloud-based telehealth systems to achieve real-time services when constrained cloud storage is applied • Improved performance in terms of hash collision and execution 	<ul style="list-style-type: none"> • Block-level data deduplication is not achieved • General system of cloud storage is not considered, referenced for telehealth systems

(continued)

Table 1 (continued)

Title of paper and publisher's	Methods/techniques	Merits	Demerits
Heterogeneous data storage management with deduplication in cloud computing, "Yan et al. [17]"	<ul style="list-style-type: none"> • Heterogeneous scheme of management to manage encryption data deduplication 	<ul style="list-style-type: none"> • Easy Management of dedupe and control access for Multiple cloud service providers • Cost-effective management of big data storage for multiple cloud service providers 	<ul style="list-style-type: none"> • Practical deployment • The Proposed scheme of demerits are rationality and security
Two-side data deduplication mechanism for non-center cloud storage systems, "Xu et al. [18]"	<ul style="list-style-type: none"> • Chord algorithm for fingerprint block of data detection 	<ul style="list-style-type: none"> • Enhanced Data deduplication for data storage in client and server data storage transfer • An improved version of load balance in the cloud system, searching time for fingerprint and response time 	<ul style="list-style-type: none"> • Non-cloud Server deduplication • Client-Side fingerprint checks delay Data transmission acknowledgement
DedupeSwift: object-oriented Storage System based on Data Deduplication, "Ma et al. [19]"	<ul style="list-style-type: none"> • Lazy method in deduplication to reduce to file I/O bottleneck 	<ul style="list-style-type: none"> • Compression, caching reduces the storage and CPU resources • Reduced Storage footprint in the cloud storage 	<ul style="list-style-type: none"> • Read and Write Performance reduction compared to Open Swift • Read overhead in data fragmentation

4.1 Evaluation Analysis of Deduplication

From the above methods, File level method deduplication saves only 10–20% of data in the cloud storage. File level data is duplicate in multimedia cloud storage like YouTube, WhatsApp, Instagram application-based cloud storage. Block-level data storage reduces the data around the 30–40% in the centralized cloud storage.

Block-level data storage has a huge impact of storage space reduction in the Finance, mobile, multimedia, text-based cloud storage data. Hash-based calculation for file level dedupe requires the entire content hash and time complexity for calculation. Block-level hash generation for content is faster than file-level dedupe [5].

Hash collision issues are more in block level deduplication compared to file-level deduplication. Compression and encryption data transfer on the wire consumes more bandwidth compared to non-compressed/encrypted data.

5 Issues in the Cloud Storage Data Deduplication

- Cloud storage offers data deduplication as to make the free storage available, Deduplication creates the chunk key or hash fingerprint for file or block level data. Compares the hash tags with file hash or block-level data hash before transferring the data from client to server cloud storage. Efficient generation of hash without collision and comparison is a major issue [6].
- Data transfer protocol which is TCP or UDP based methodology, implementation of the efficient protocol for data deduplication transfer to the centralized cloud server is a major concern in the cloud computing environment.
- Most Deduplication based data transfer protocol relies on the inline deduplication, post-process deduplication for data process for hash, comparison of hash keys [7].
- Security, quality of service, reliability, compression of data is the major threats to cloud storage data deduplication [8].
- Incremental backup of files, data blocks needed to be achieved for data deduplication. Cloud storage has the version of files. An efficient method for merging the chunks, files in cloud storage version need to be maintained.
- Source-side deduplication, server-side data deduplication are a major concern while deciding the data deduplication for process data. Faster transfer of data needs to be achieved in terms of time.
- Restoring the data blocks from the cloud storage to clients space without data block mismatch is a major concern.
- Recovery of data blocks, hash collision, fingerprint table of hash for big data are major issues in centralized cloud storage [9].

Issues of cloud storage commonly in terms of security of data storage, hash collision for calculating the chunks, recovering the data blocks in the block level data deduplication, synchronization of client side and server side block checksum validation. The proposed system of block-level deduplication in the below section discuss to resolve the issues of dedupe cloud storage.

6 Proposed System of Block-Level Deduplication

Data deduplication protocol is method between the client and server where the server is storage for eliminating the data segments. The new proposed protocol aims to transfer the data to the server in an efficient manner which improves the bandwidth traffic while transferring the data, storage reduction in terms of block level. New hash

technology is to be included for inline data block comparison over the stored data on storage. The protocol enhances the more secure data storage with compressed data storage. A new system of storage is the centralized unique single instance data block transfer over the network.

Proposed protocol transfers the data blocks from the client to server over the network. Server storage system compares the data chunks of already stored data chunks on the server side by a new hash algorithm. Duplicate data is found by the new hash algorithm, so that data segment that was transferred for storage to the server is not considered for storage. Data chunk is referenced in the server storage with a new pointer.

7 Conclusion

In Cloud Storage, data deduplication is the elimination of redundant data based on the file-level, block-level, and byte-level techniques. An Efficient Protocol for transfer of data from sender to cloud server storage needs well designed reliable, a secured architecture for bandwidth utilization across the network. Various design of protocols in the current trends are based on the common traditional hashing techniques, centralized cloud storage transfer with compression and encryption with increasing complexity which effects performance in modern data backup. A proposed design of protocol ensures the data transfer with fast transfer data with fast hashing algorithm and removal of duplicate data in terms of block level across the cloud storage.

References

1. <http://searchstorage.techtarget.com/definition/cloud-storage>
2. https://docs.druva.com/Knowledge_Base/inSync/
3. <https://spinbackup.com/blog/cloud-storage-environment-impact/>
4. <http://searchdatabackup.techtarget.com>
5. Kim, D., Song, S., Choi, B.: Data Deduplication for Data Optimization for Storage and Network Systems. Springer, New York (2017)
6. Venish, A., Sankar, K.S.: Study of Chunking Algorithm in Data Deduplication, vol. 743, p. 343. Springer (2016). ISBN 978-81-322-2672-7
7. Douglis, F., Fu, M., Zhang, Y., Zhou, Y.: A Comprehensive Study of the Past, Present, and Future of Data Deduplication. IEEE Proceedings (2016)
8. <http://www.cambridge.org/core/journals/apsipa-transactions-on-signal-and-information-processing/article/survey-on-securing-data-storage-in-the-cloud/D3E3ED530F4E026E695E32BD71DDBD0D/core-reader>
9. Wang1, S., Zhang2, Y., Zhang3, Y.: A blockchain-based framework for data sharing with fine-grained access control in decentralized storage systems. In: IEEE Transactions and Content Mining (2018)
10. Wang, S., Du, J., Wu, J., Wang, R., Lv, J., Ma, S.: A storage solution for multimedia files to support data deduplication. In: CCIOT (2016)

11. Su, H., Zheng, D., Zhang, Y.: An efficient and secure deduplication scheme based on Rabin fingerprinting in cloud storage. In: IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC) (2017)
12. Haoran, W., Weiqin, T., Shengan, Z.: A data deduplication method in the cloud storage based on FP-tree. In: 4th International Conference on Computer Science and Network Technology (ICCSNT 2015) (2015)
13. Cho E.M., Sakura-ku, Koshiha, T.: Big data cloud deduplication based on verifiable hash convergent group signcryption. In: IEEE Third International Conference on Big Data Computing Service and Application (2017)
14. Wua, S., Lic, K.-C., Maob, B., Liaob, M.: DAC: improving storage availability with deduplication-assisted cloud-of-clouds. *Future Gener. Comput. Syst.* (2016)
15. Kai, H., Chuanhe, H., Hao, Z., Jiaoli, S., Xiaomao, W., Feng, D.: Public Auditing for Encrypted Data with Client-Side Deduplication in Cloud Storage, vol. 20 No. 4, pp. 291–298. Wuhan University and Springer, Berlin (2015)
16. Gai, K., Qiu, M.K., Sun, X.T.: Smart data deduplication for telehealth systems in heterogeneous cloud computing. *J. Commun. Inf. Netw.* **1**(4), 93–104 (2016)
17. Yan, Z., Zhang, L., Ding, W., Zheng, Q.: Heterogeneous data storage management with deduplication in cloud computing. *IEEE Trans. Big Data* (2017)
18. Xu, X., Hu, N., Tu, Q.: Two-side data deduplication mechanism for non- center cloud storage systems. *IEEE* (2016)
19. Ma, J., Wang, G., Liu, X.: DedupeSwift: object-oriented storage system based on data deduplication. In: IEEE TrustCom-BigDataSE-ISPA (2016)
20. Zhou, B., We, J.-T.: Metadata feedback and utilization for data deduplication across WAN. *J. Comput. Sci. Technol.* **31**(3), 604–623 (2016)
21. Venish, A., Siva Sankar, K., Study of Chunking Algorithm in Data Deduplication. Springer, India (2016)
22. Zhou, B., Wen, J.-T.: Improving metadata caching efficiency for data deduplication via in-RAM metadata utilization. *J. Comput. Sci. Technol.* **31**(3), 604–623 (2016). Appendix: Springer-Author Discount

Computing Narayana Prime Cordial Labeling of Web Graphs and Flower Graphs



B. J. Balamurugan, K. Thirusangu and B. J. Murali

Abstract The process of assigning the binary numbers 0 and 1 to the edges of a graph $G = (V, E)$ through evaluating functions defined on the vertex set V and the edge set E of G using the concepts of prime and Narayana numbers by satisfying cordiality on the edges. This process on the graph G is known as Narayana prime cordial labeling of G and this graph G is called Narayana prime cordial graph. In this article, we compute the Narayana prime cordial labeling of Web graphs and Flower graphs.

Keywords Narayana numbers · Prime numbers · NPC graphs

1 Introduction

A labeling pattern of a graph G is a process of allocating numbers to the nodes of G or lines of G or both through mathematical functions [1]. The basic notion of graph labeling is found in [2]. The vital application of labeled graphs can be found in science, engineering, and technology and we refer [3] for the same. For graph labeling literature, we refer [4]. We refer the textbook Harary [5], for notations, concepts, and terminology in graph theory.

B. J. Balamurugan (✉)
School of Advanced Sciences, VIT University, Chennai Campus, Chennai 600127,
Tamil Nadu, India
e-mail: balamurugan.bj@vit.ac.in

K. Thirusangu
Department of Mathematics, SIVET College, Gowrivakkam, Chennai 600073,
Tamil Nadu, India
e-mail: kthirusangu@gmail.com

B. J. Murali
Research and Development Centre, Bharathiar University, Coimbatore 641046,
Tamil Nadu, India
e-mail: muralibjgpm@gmail.com

The Narayana numbers [6], a recent development in number theory occurs in various combinatorial problems. The applications of Narayana numbers play an important role in various topics of mathematics, especially in cryptography. The Narayana numbers have been used in multiple input and output communication systems, RNA Secondary Structure configuration and in the partition of graphs in terms of trees. Since the Narayana numbers and labeled graphs have many interesting practical applications, we are concerned in computing the Narayana prime cordial labeling of graphs. We introduced this labeling pattern in [7] and proved that the graphs, viz., (i) paths (ii) cycles, and (iii) helm graphs are Narayana prime cordial graphs. In this article, we compute the Narayana prime cordial labeling of Web graphs and Flower graphs.

2 Preliminaries

We refer [1, 8] for the definitions of the Web graph and flower graph and for the concept of cordiality in graphs we refer [9–12]. The concept of Narayana numbers and their properties are given in [6].

Definition 1 [6] Let \mathbb{N}_0 be the set of nonnegative integers and let $k, n \in \mathbb{N}_0$. The Narayana numbers can be defined as

$$N(n, k) = \frac{1}{n} \binom{n}{k} \binom{n}{k+1}; 0 \leq k < n \text{ where } \binom{n}{k} = \frac{n!}{(n-k)!k!}$$

The Narayana numbers discovered by Narayana are highly associated with the Catalan numbers [13]. That is, $C_n = \frac{1}{n+1} \binom{2n}{n}$ and $\sum_{k=0}^{n-1} N(n, k) = C_n$ where C_n is a Catalan number.

For example, the Narayana numbers $N(n, k)$ where $0 \leq k < n \leq 8$ are given in the following triangular array. Here, the sum of each row is a Catalan number [13].

$n \setminus k$	0	1	2	3	4	5	6	7
1	1	1						
2	1	1						
3	1	3	1					
4	1	6	6	1				
5	1	10	20	10	1			
6	1	15	50	50	15	1		
7	1	21	105	175	105	21	1	
8	1	28	196	490	490	196	28	1

3 Narayana Prime Cordial Labeling of Web Graphs

This section introduces the Narayana prime cordial labeling [7] of a graph $G = (V, E)$ and shows that the Web graphs are Narayana prime cordial graphs.

Definition 2 Let $G(V, E)$ be a simple graph. A 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ is called a Narayana prime cordial labeling of G if there exists an edge map $\ell^* : E \rightarrow \{0, 1\}$ satisfying the two conditions, viz.,

(i) For every $uv \in E, u, v \in V$

$$\begin{aligned} \ell^*(uv) &= 1 \text{ if } p|N(\ell(u), \ell(v)), \text{ where } \ell(u) > \ell(v) \text{ and } \ell(u) = p^m \\ &\quad \text{for some } m \in \mathbb{N}_0; 1 \leq \ell(v) \leq \ell(u) - 2 \\ &= 1 \text{ if } p|N(\ell(v), \ell(u)), \text{ where } \ell(v) > \ell(u) \text{ and } \ell(v) = p^m \\ &\quad \text{for some } m \in \mathbb{N}_0; 1 \leq \ell(u) \leq \ell(v) - 2 \\ &= 0 \text{ if } p \nmid N(\ell(u), \ell(v)), \text{ where } \ell(u) > \ell(v) \text{ and } \ell(u) = p^m - 1 \\ &\quad \text{for some } m \in \mathbb{N}_0; 0 \leq \ell(v) \leq \ell(u) - 1 \\ &= 0 \text{ if } p \nmid N(\ell(v), \ell(u)), \text{ where } \ell(v) > \ell(u) \text{ and } \ell(v) = p^m - 1 \\ &\quad \text{for some } m \in \mathbb{N}_0; 0 \leq \ell(u) \leq \ell(v) - 1 \end{aligned}$$

where p is a prime number.

(ii) $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$ where $e_{\ell^*}(0)$ and $e_{\ell^*}(1)$ denote respectively the number of edges having the number 0 and the number of edges having the number 1.

Definition 3 Let $G = (V, E)$ be a simple graph. If G assumes a Narayana prime cordial labeling, then it is known as a Narayana prime cordial graph.

Remark 1 We call the Narayana prime cordial labeling of a graph as NPC labeling of a graph for simplicity in this paper.

Theorem 1 *The web graph $W(2, n)$ is a NPC graph.*

Proof Let $W(2, n)$ be the web graph with $3n + 1$ vertices and $5n$ edges. Let $V = \{v_0\} \cup V_1 \cup V_2 \cup V_3$ be the vertex set of $W(2, n)$ where v_0 is the central vertex, $V_1 = \{v_{1,i} | 1 \leq i \leq n\}$, $V_2 = \{v_{2,i} | 1 \leq i \leq n\}$ are the vertices on the cycles of the web graph $W(2, n)$ and $V_3 = \{v_{3,i} | 1 \leq i \leq n\}$ is the vertex set of pendant vertices.

Let $E = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5$ where $E_1 = \{v_0v_{1,i} | 1 \leq i \leq n\}$, $E_2 = \{v_{1,i}v_{1,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{1,n}, v_{1,1}\}$, $E_3 = \{v_{1,i}v_{2,i} | 1 \leq i \leq n\}$, $E_4 = \{v_{2,i}v_{2,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{2,n}, v_{2,1}\}$ and $E_5 = \{v_{2,i}v_{3,i} | 1 \leq i \leq n\}$.

Define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ such that $\ell(v_0) = 1, \ell(v_{1,i}) = p_1^{i+1}; 1 \leq i \leq n, \ell(v_{2,i}) = p_2^{i+1} - 1; 1 \leq i \leq n, \ell(v_{3,i}) = p_3^{i+1}; i \equiv 1 \pmod{2}$ and $1 \leq i \leq n$ and $\ell(v_{3,i}) = p_3^{i+1} - 1; i \equiv 0 \pmod{2}$ and $1 \leq i \leq n$ and an edge map $\ell^* : E \rightarrow \{0, 1\}$ as in the Definition 2. In this type of labeling pattern,

Case (i): When $n \equiv 0 \pmod{2}$, the edge set E has $\frac{5n}{2}$ lines with label 0 and $\frac{5n}{2}$ lines with label 1. That is, $e_{\ell^*}(0) = \frac{5n}{2}$ and $e_{\ell^*}(1) = \frac{5n}{2}$. Hence $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Case (ii): When $n \equiv 1 \pmod{2}$, the edge set E has $\frac{5n-1}{2}$ lines with label 0 and $\frac{5n+1}{2}$ lines with label 1. That is, $e_{\ell^*}(0) = \frac{5n-1}{2}, e_{\ell^*}(1) = \frac{5n+1}{2}$ and hence, the condition $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Therefore in both cases, we get the condition $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$. Hence, the web graph $W(2, n)$ is a NPC graph.

The following example shows a NPC labeling of the web graph $W(2, 4)$ which is given in Fig. 1.

Example 1

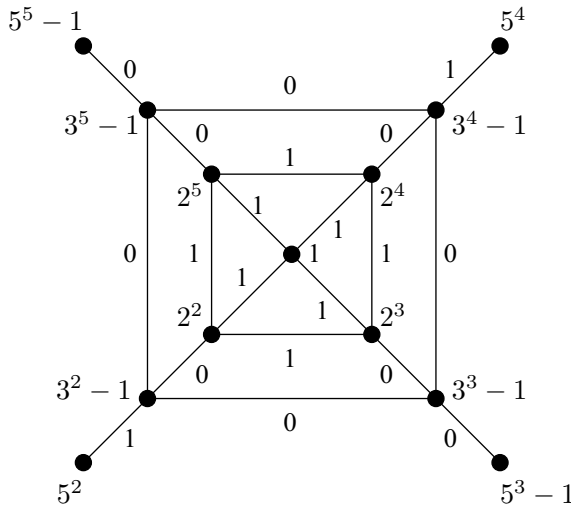
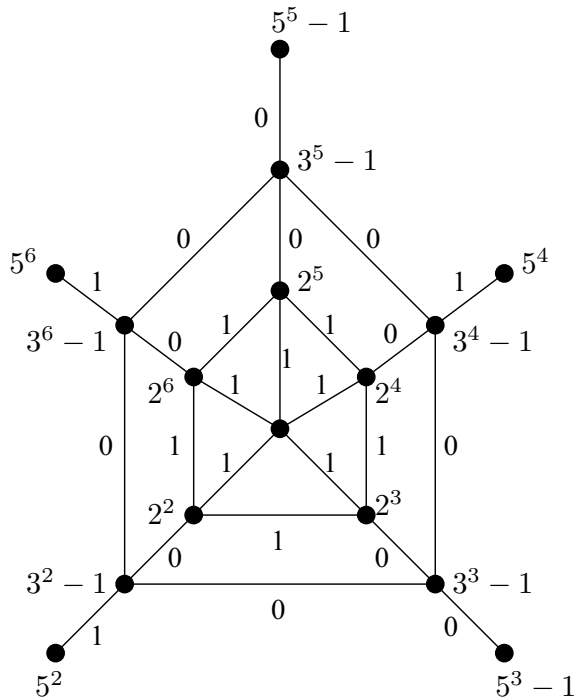


Fig. 1 NPC labeling of $W(2, 4)$

Fig. 2 NPC labeling of $W(2, 5)$



Example 2 A NPC labeling of $W(2, 5)$ is given in Fig. 2.

Theorem 2 *The generalized web graph $W(m, n)$ is a NPC graph.*

Proof Let $W(m, n)$ be the generalized web graph with $(m + 1)n + 1$ vertices and $(2m + 1)n$ edges. Let $V = \{v_0\} \cup V_1 \cup V_2 \cup \dots \cup V_m \cup V_{m+1}$ be the vertex set of $W(m, n)$, where v_0 is the central vertex set $V_1 = \{v_{1,i} | 1 \leq i \leq n\}$, $V_2 = \{v_{2,i} | 1 \leq i \leq n\} \dots V_m = \{v_{m,i} | 1 \leq i \leq n\}$ are the set of vertices on the cycles of $W(m, n)$ and $V_{m+1} = \{v_{m+1,i} | 1 \leq i \leq n\}$ is the vertex set of pendant vertices.

Let $E = E_1 \cup E_2 \cup \dots \cup E_n$ be the edge set of the generalized web graph $W(m, n)$ where $E_1 = \{v_0 v_{1,i} | 1 \leq i \leq n\}$, $E_2 = \{v_{1,i} v_{1,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{1,n} v_{1,1}\}$, $E_3 = \{v_{1,i} v_{2,i} | 1 \leq i \leq n\}$, $E_4 = \{v_{2,i} v_{2,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{2,n} v_{2,1}\}$, \dots , $E_{2m-1} = \{v_{m-1,i} v_{m,i+1} | 1 \leq i \leq n\}$, $E_{2m} = \{v_{m,i} v_{m,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{m,n} v_{m,1}\}$ and $E_{2m+1} = \{v_{m,i} v_{m+1,i} | 1 \leq i \leq n\}$.

Case 1: When $m \equiv 0 \pmod{2}$, define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ such that

- $\ell(v_0) = 1$
- $\ell(v_{1,i}) = p_1^{i+1}$; for some $i + 1 \in m, 1 \leq i \leq n$
- $\ell(v_{2,i}) = p_2^{i+1} - 1; 1 \leq i \leq n$

$$\begin{aligned} \ell(v_{3,i}) &= p_3^{i+1}; 1 \leq i \leq n \\ \ell(v_{4,i}) &= p_4^{i+1} - 1; 1 \leq i \leq n \\ \ell(v_{m-1,i}) &= p_{m-1}^{i+1}; 1 \leq i \leq n \\ \ell(v_{m,i}) &= p_m^{i+1} - 1; 1 \leq i \leq n \text{ and} \\ \ell(v_{m+1,i}) &= \begin{cases} p_{m+1}^{i+1}; & i \equiv 0 \pmod{2} \\ p_{m+1}^{i+1} - 1; & i \equiv 1 \pmod{2} \end{cases} \end{aligned}$$

where $p_1, p_2, \dots, p_{m+1} \dots$ are distinct primes such that $p_1 < p_2 < p_3 \dots < p_m < p_{m+1}$ and an edge map $\ell^* : E \rightarrow \{0, 1\}$ is given as in Definition 2.

Case 1(a): When $n \equiv 0 \pmod{2}$, E has $mn + \frac{n}{2}$ edges with 0 labeling and $mn + \frac{n}{2}$ edges with 1 labeling. That is $e_{\ell^*}(0) = mn + \frac{n}{2}$ and $e_{\ell^*}(1) = mn + \frac{n}{2}$. Therefore $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Case 1(b): When $n \equiv 1 \pmod{2}$, E has $mn + (\frac{n+1}{2})$ edges with 0 labeling and $mn + (\frac{n-1}{2})$ edges with 1 labeling. That is $e_{\ell^*}(0) = mn + (\frac{n+1}{2})$ and $e_{\ell^*}(1) = mn + (\frac{n-1}{2})$. Hence we have $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Case 2: When $m \equiv 1 \pmod{2}$, define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ as

$$\begin{aligned} \ell(v_0) &= 1 \\ \ell(v_{1,i}) &= p_1^{i+1}; 1 \leq i \leq n \\ \ell(v_{2,i}) &= p_2^{i+1} - 1; 1 \leq i \leq n \\ \ell(v_{3,i}) &= p_3^{i+1}; 1 \leq i \leq n \\ \ell(v_{4,i}) &= p_4^{i+1} - 1; 1 \leq i \leq n \\ &\dots \\ \ell(v_{m,i}) &= \begin{cases} p_m^{i+1} - 1; & i \equiv 1 \pmod{2} \\ p_m^{i+1} - 1; & i \equiv 0 \pmod{2} \end{cases} \\ \ell(v_{m+1,i}) &= \begin{cases} p_{m+1}^{i+1} - 1; & i \equiv 1 \pmod{2}; 0 \leq i \leq n - 2 \\ p_{m+1}^{i+1}; & i \equiv 0 \pmod{2}; 0 \leq i \leq n - 2 \end{cases} \\ \ell(v_{m+1,n-1}) &= p_{m+1}^{i+1}; i = n - 1 \\ \ell(v_{m+1,n}) &= p_{m+1}^{i+2}; i = n. \end{aligned}$$

Case 2(a): When $n \equiv 1 \pmod{2}$, E has $(m - 1)n + (\frac{n+1}{2}) + (\frac{n+1}{2}) + (\frac{n-3}{2})$ edges having the label 0 and $(m - 1)n + (\frac{n-1}{2}) + (\frac{n-2}{2}) + (\frac{n-3}{2}) + 3$ edges having the label 1. Therefore $e_{\ell^*}(0) = mn + \frac{n}{2} - \frac{1}{2}$ and $e_{\ell^*}(1) = mn + \frac{n}{2} + \frac{1}{2}$. Hence we get $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Case 2(b): When $n \equiv 0 \pmod{2}$, define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ as

$$\begin{aligned} \ell(v_0) &= 1 \\ \ell(v_{1,i}) &= p_1^{i+1}; 1 \leq i \leq n \\ \ell(v_{2,i}) &= p_2^{i+1} - 1; 1 \leq i \leq n \\ \ell(v_{3,i}) &= p_3^{i+1}; 1 \leq i \leq n \\ \ell(v_{4,i}) &= p_4^{i+1} - 1; 1 \leq i \leq n \\ &\dots \\ \ell(v_{m,i}) &= \begin{cases} p_m^{i+1} - 1; & i \equiv 1 \pmod{2} \\ p_m^{i+1}; & i \equiv 0 \pmod{2} \end{cases} \\ \ell(v_{m,i}) &= \begin{cases} p_{m+1}^{i+1} - 1; & i \equiv 1 \pmod{2} \\ p_{m+1}^{i+1}; & i \equiv 0 \pmod{2} \end{cases} \end{aligned}$$

In this type of labeling, we have E has $\frac{n}{2}(2m + 1)$ edges with label 0, E has $\frac{n}{2}(2m + 1)$ edges with label 1. Therefore $e_{\ell^*}(0) = \frac{n}{2}(2m + 1)$, $e_{\ell^*}(1) = \frac{n}{2}(2m + 1)$ and $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

In all cases, we get the condition $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$. Hence, $W(m, n)$ admits a NPC labeling.

4 Narayana Prime Cordial Labeling of Flower Graphs

Theorem 3 *The sunflower graph SF_n is a NPC graph.*

Proof Let SF_n be the sunflower graph with $3n + 1$ vertices and $5n$ edges. Let $V = \{v_0\} \cup V_1 \cup V_2 \cup V_3$ be the node set where v_0 is the central node, $V_1 = \{v_{1,i} | 1 \leq i \leq n\}$, $V_2 = \{v_{2,i} | 1 \leq i \leq n\}$, $V_3 = \{v_{3,i} | 1 \leq i \leq n\}$ is vertex set of pendant vertices. Let $E = E_1 \cup E_2 \cup E_3 \cup E_4 \cup E_5$ be the edge set of SF_n where $E_1 = \{v_0v_{1,i} | 1 \leq i \leq n\}$, $E_2 = \{v_{1,i}v_{1,i+1} | 1 \leq i \leq n - 1\} \cup \{v_{1,n}v_{1,1}\}$, $E_3 = \{v_{1,i}v_{2,i} | 1 \leq i \leq n\}$, $E_4 = \{v_0v_{2,i} | 1 \leq i \leq n\}$ and $E_5 = \{v_0v_{3,i} | 1 \leq i \leq n\}$.

Case 1: When $n \equiv 0 \pmod{2}$, define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ as

$$\ell(v_0) = 1, \ell(v_{1,i}) = 2^{i+1}; 1 \leq i \leq n$$

$$\ell(v_{2,i}) = 3^{i+1} - 1; 1 \leq i \leq n$$

$$\ell(v_{3,i}) = \begin{cases} 5^{i+1}; & 1 \leq i \leq \frac{n}{2} \\ 5^{i+1} - 1; & \frac{n}{2} < i \leq n \end{cases}$$

This vertex function with the induced edge function as in the Definition 2, the edge set E receives the values 0 and 1. The set E has $\frac{5n}{2}$ lines with label 0 and $\frac{5n}{2}$ lines with label 1. That is, $e_{\ell^*}(0) = \frac{5n}{2}$, $e_{\ell^*}(1) = \frac{5n}{2}$ and hence, we get the condition $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Case 2: $n \equiv 1 \pmod{2}$, define a 1-1 map $\ell : V \rightarrow \mathbb{N}_0$ such that

$$\ell(v_0) = 1, \ell(v_{1,i}) = 2^{i+1}; 1 \leq i \leq n$$

$$\ell(v_{2,i}) = 3^{i+1} - 1; 1 \leq i \leq n$$

$$\ell(v_{3,i}) = \begin{cases} 5^{i+1}; & 1 \leq i \leq \frac{n+1}{2} \\ 5^{i+1}; & \frac{n+1}{2} < i \leq n \end{cases}$$

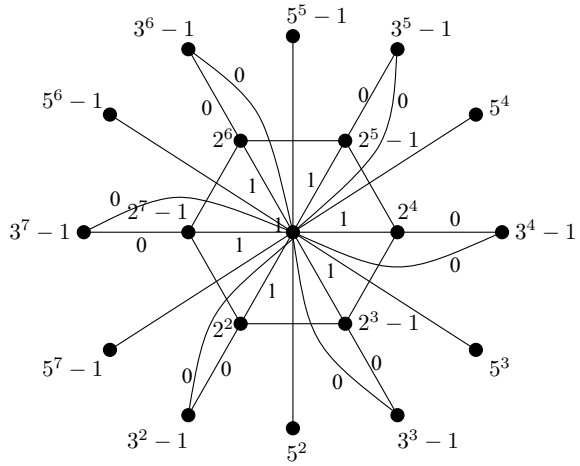
This function with the induced edge function as in Definition 2, the edge set E receives the values 0 and 1. The set E has $\frac{5n-1}{2}$ edges with label 0 and $\frac{5n+1}{2}$ edges with label 1. That is, $e_{\ell^*}(0) = \frac{5n-1}{2}$ and $e_{\ell^*}(1) = \frac{5n+1}{2}$. Therefore, $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$.

Hence in both cases, the condition $|e_{\ell^*}(0) - e_{\ell^*}(1)| \leq 1$ is satisfied. Hence, SF_n is a NPC graph.

The NPC labeling of SF_6 is shown in Fig. 3.

Example 3

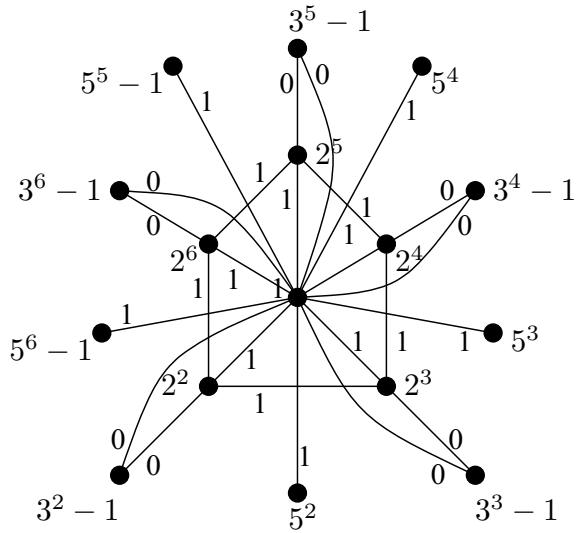
Fig. 3 NPC labeling of SF_6



The NPC labeling of SF_5 is shown in Fig. 4.

Example 4

Fig. 4 NPC labeling of SF_5



5 Conclusion

In this manuscript, we proved that the graphs, viz., (i) $W(2, n)$, (ii) the generalized Web graph $W(m, n)$, and (iii) the sunflower graph SF_n admit a *NPC* labeling with suitable examples. The process of finding a family of graphs which admits the *NPC* labeling is an interesting and potential area of research. Finding such family of *NPC* graphs is the future scope for the researchers in this field.

References

1. Acharya, B.D., Hegde, S.M.: Arithmetic graphs. *J. Graph Theory*. **14**(3), 275–299 (1990)
2. Rosa, A.: On certain valuations of the vertices of a graph. In: *Theory of graphs (Internat. Sympos. Rome: 349–359 (1967) Gordan and Breach. Newyork. Dunod, Paris (1966)*
3. Lakshmi Prasana, N., Saravanthi, K., Sudhakar, N.: Applications of graph labeling in major areas of computer science. *Int. J. Res. Comput. Commun. Technol.* **3**(8) (2014)
4. Gallian, J.A.: A Dynamic Survey of Graph Labeling. *Electron. J. Comb.* **17** (DS6) (2016)
5. Harary, F.: *Graph Theory*. Addison-Wesley, Reading Mass (1972)
6. Bona, M., Sagan, B.E.: On divisibility of Narayana numbers by primes. *J. Integer Sequences*. **8** (2005)
7. Murali, B.J., Thirusangu, K., Balamurugan, B.J.: Narayana prime cordial labeling of graphs. *Int. J. Pure Appl. Math.* **117**(13), 1–8 (2017)
8. Miller, Mirka, Phanalasy, Oudone, Ryan, Joe: Antimagicness of some families of generalized graphs. *Australas. J. Comb.* **53**, 179–190 (2012)
9. Cahit, I.: Cordial graphs: a weaker version of graceful and harmonics graphs. *Ars Combin.* **23**, 201–207 (1987)
10. Ho, Y.S., Lee, S.M., Shee, S.C.: Cordial labelings of unicyclic graphs and generalized Petersen graphs. *Congr. Number*. **68**, 109–122 (1989)
11. Vaidhya, Samir K., Shah, Nirav H.: Prime cordial labeling of some graphs. *Open J. Discrete Mathe.* **2**(1), 11–16 (2012)
12. Yousef, M.Z.: Graph operations and cordiality. *Ars Combin.* **97**, 161–174 (2010)
13. Koshy, T.: *Catalan Numbers with Applications*. Oxford University Press (2009)

Reliability-Aware Green Scheduling Algorithm in Cloud Computing



Chesta Kathpal and Ritu Garg

Abstract Nowadays, to a significant extent, cloud computing usage is increasing because of its enormous features such as resource sharing, on-demand resource provisioning, and virtualization. To provide the resources to the user according to their requirement, the client's applications must be scheduled in an optimized way. A rapidly increasing number of users causes a huge amount of energy consumption while executing the task. The temperature of the system increases as there is a drastic increment in power density. Energy and temperature both are related to power consumption. Moreover, cloud servers are prone to failure, so it needs extra computation time to handle the failure. In this work, we proposed a scheduling algorithm which optimizes three conflicting objectives, i.e., reliability maximization, minimum energy consumption, and temperature consolidation in the cloud. The failure model used in our work is Weibull failure distribution that considers the effect of aging on the performance of the system. The simulation results show that the algorithm reduces the energy consumption while scheduling the task to the reliable virtual machine with less temperature consolidation.

Keywords Reliability · Energy consumption · CCS

1 Introduction

Cloud is a representation of Internet-based computing environment which is an open, secure, and more powerful solution available today. Cloud computing provides reliable computing services to users based on the usage of the pay-as-you-go model. The offered services are classified as Infrastructure-as-a-service (IAAS), Platform-as-a-service (PAAS), Software-as-a-service (SaaS) [1]. Due to the enormous feature

C. Kathpal (✉) · R. Garg
National Institute of Technology, Kurukshetra, India
e-mail: chestakathpal93@gmail.com

R. Garg
e-mail: Ritu.59@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_38

of cloud computing, more number of users are shifting their applications to cloud servers, hence data size is constantly growing on the servers [2]. It is a necessity of the cloud computing systems (CCS) to schedule the tasks in an efficient way because of its highly scalable and complex architecture. Normally, the main idea behind scheduling is to optimize the makespan or minimize the schedule length with the assumption that the node never fails during execution. Hence, the properties of real parallel computation system are not reflected, because it needs additional computing time due to some unexpected failures. The resource failure and failure of scheduled tasks on these resources may be the prime reason for the economic loss for both the service providers and users. To make the reliable system, understanding of the failure and repair characteristics is necessary before scheduling the task to a virtual machine. Garraghan et al. [3] investigated the Google Cloud Trace Log to know the failure characteristics of a number of different resources such as processors and number of different tasks by operating the 12,500 servers which took 29 days to operate. And it concluded that the cloud server's failure follows the Weibull failure distribution. To handle the failure, a number of different resources such as servers and disks are needed to process the data within a required period of time, through which data centers utilize more amount of energy. According to Moore's law [4]; to maintain the resource availability and performance, power consumption in CCS increases. The temperature of the system increases while the power density of the system increases. Energy and temperature both are related to power consumption [5], but also different to each other with different physical entities. Further, experiments in the study [6, 7] showed that the lifetime of the system may get reduced to half with 10–15 °C increase in temperature. Thus, the cloud service provider must provide the services according to QoS [8] requirements of the users.

However, a number of resources are used to handle the failure which eventually maximizes the reliability but for executing the redundant resources, more energy will be consumed which eventually increases the temperature. Our work primarily focuses on the scheduling of workflow applications to reliable virtual machines in large-scale cloud data centers with three conflicting objectives, i.e., reliability maximization, minimum energy consumption, and makespan. The algorithm consists of three phases. First, the establishment of task priority phase sorts the application tasks in topological order. The second phase consists of selecting the frequency for efficient energy consumption and considering the temperature of the virtual machine under the threshold limit. Then, the assignment phase assigns the tasks to the reliable virtual machine to maximize the reliability for executing the application under the constraint of late finish time.

2 Related Work

Task scheduling is considered a problem of task mapping, in which the best resource will be selected for mapping the task. In general, the nature of the task scheduling problem is NP-Hard [9]. Thus, the scheduling problem can be optimally solved

within polynomial time with the help of techniques list based heuristics [10], meta-heuristics [11], or approximations algorithms. For instance, critical-path-on-a-processor (CPOP) algorithm and heterogeneous earliest finish time (HEFT) algorithm [12] are well-known heuristics that schedules the task to the processor to achieve minimal scheduling length with the lowest complexity and better performance. Many heuristic algorithms have been used in scheduling problem which also considers other objectives like minimal energy consumption, reliability, makespan, etc.

Recently, business users are shifting their data to the cloud through which data is growing increasingly to the server. So, it is the necessity of the system to keep performing its function correctly. Reliability-aware scheduling algorithms schedule the tasks to the reliable virtual machine by calculating the failure probability of the virtual machine. Tang et al. [13] present a reliability-aware scheduling algorithm with duplication (RASD) for heterogeneous distributed computing systems that aim at achieving high reliability and reducing the makespan of the workflow application. Tang et al. [14] also present a hierarchical reliability-driven scheduling algorithm (HRDS) that includes both a local scheduler and a global scheduler for maximizing the reliability of the system. In [15]; the author presents a reliability-aware scheduling algorithm for MPSoC platforms. Tasks are allocated to processors using the simulated annealing technique. The reliability of the system also depends upon the usage of power consumption. Hence, while focusing on the reliability less power must be consumed by the system. In [16]; the authors use the dynamic voltage frequency scaling (DVFS) technique to minimize energy consumption. In [17], authors consider that during stall time (waiting time), processor frequency can be used according to voltage scaling in order to save energy. Many meta-heuristic algorithms have been developed in a scheduling problem, which also considers the optimal energy consumption in a cloud environment uses particle swarm optimization [18], genetic algorithm [19]. Due to high power consumption, energy and temperature of the systems is also affected. In [20]; authors schedule the tasks on a processor using efficient energy consumption technique considering the thermal effect of the processor as a constraint. This factor results in more failure rate of the system and thus it affects the reliability and performance of the system which eventually affects the financial losses to both service providers and users.

The aforementioned algorithms consider either system reliability or energy consumption at a time. But, it is very important to jointly optimize the power consumption, temperature and lifetime reliability of the system. The approaches used in [13, 14, 21, 22] algorithm to maximize the reliability considers the Poisson failure distribution. In [23] authors developed the novel approach to schedule the task to the reliable node under consideration of Poisson failure distribution but with the assumption, the failure rate remains constant for a specific period of time. In [24], the authors proposed a bi-objective workflow scheduling algorithm that accounts for system reliability as well as energy consumption by using Pareto front optimization [25] technique to obtain a non-dominated solution. Here, Poisson failure distribution has been used for maximizing reliability. Tang et al. [26] proposed the linear programming based mathematical model that considers simultaneous management of

energy consumption with high system reliability and better system performance. In [27], the authors proposed the algorithm for joint optimization of lifetime reliability of system, power, and temperature to meet the predefined threshold value. However, failure of cloud servers follows a Weibull distribution with a varying failure rate as per aging [28]. Due to the growing demand of computing in a heterogeneous system, it is desirable to consider the other objectives energy, reliability and temperature during the scheduling of the tasks. Thus, the approach used in our algorithm considers green computing. Weibull failure distribution is used for maximizing the reliability because it incorporates the aging effects of cloud nodes based on the different values of scale and shape parameters.

3 Model

3.1 Workflow Model

A Workflow application is a set of dependent tasks represented as a directed acyclic graph (DAG). An application (DAG), $G = \{V, E\}$ submitted by the user, consists of a set of n parallel and precedence-constrained tasks $V = \tau_i \{1 \leq i \leq n\}$ to be executed and $E = \varepsilon_{ij} \{1 \leq i \leq n, 1 \leq j \leq n\}$ represents the dependencies between the tasks τ_i and τ_j where the task τ_j is dependent on task τ_i . We consider data dependent tasks where the output data from the parent task is used as an input for the child task. The set of all immediate predecessors of a task can be represented as a predecessor-set. The set of all immediate children of a task can be represented as a successor-set. We can add dummy entry and exit tasks to satisfy the assumption that each workflow application is having a single-entry task as well as a single-exit task. The Latest finish time, $LFT(\tau_i)$ of the task t_i is the time at which task finish its execution. $LFT(\tau_i)$ depends upon mean execution time and mean communication time and it can be calculated as defined in [16]. The Latest finish time, $LFT(\tau_i)$ of the task t_i is the latest time at which task finishes its execution while satisfying the deadline constraint D . Total workflow deadline D is the total execution time of a workflow application.

3.2 Reliability Model

During execution of an application, a fault may be hard to avoid that arises for several reasons such as software bugs, hardware failures, etc. However, one of the main reasons is when any device is exposed to extreme temperature. Hence, it is required to find the reliability of the node before scheduling the task. In our approach, Weibull failure distribution is used to describe the wear-out effects of a system. Lifetime reliability of single node at time t is defined as the probability of occurrence of correct operation of the node and is given as [15];

$$R(t, T) = e^{-t/\alpha(T)^\beta} \quad (1)$$

where T represents the temperature of the node. $\alpha(T)$ and β are the scale parameter with respect to temperature and slope parameter respectively in Weibull distribution. Value of temperature is not taken as a fixed value. The currently accepted model for mean time to failure (MTTF) of a Weibull distribution is given as

$$\text{MTTF}(T) = \alpha(T) \cdot \Gamma(1 + 1/\beta) \quad (2)$$

So, we have

$$\alpha(T) = \frac{\text{MTTF}(T)}{\Gamma(1 + 1/\beta)}$$

Currently accepted model for failure mechanism is taken because of electro-migration [29] and based on Black's equation MTTF is;

$$\text{MTTF}(T) \propto (J - J_{\text{critical}})^{-n} \cdot e^{(E_a/k \cdot T)} \quad (3)$$

where J is current density, J_{critical} is the critical current density required for electro-migration. E_a is activation energy and k is the Boltzmann's constant. Here $J \gg J_{\text{critical}}$, i.e., J tends to be much higher than J_{critical} hence $(J - J_{\text{critical}}) \approx J$. Here we set current density $J = 1.5 \times 10^6$ A/cm² as used in [15]. Hence,

$$\text{MTTF}(T) = A_0 \cdot J^{-n} \cdot e^{(E_a/k \cdot T)} \quad (4)$$

where A_0 is material-related constant. From Eq. (2), the Scale Parameter is

$$\alpha(T) = A_0 \cdot J^{-n} e^{(E_a/k \cdot T)} / \Gamma(1 + 1/\beta) \quad (5)$$

3.3 Energy Model

In modern data centers, power consumption depends upon the type of application executed on hardware platform. Most of the high-performance computing (HPC) resources are DVFS-enabled resources so that it can reduce the power consumption by adjusting voltage according to the frequency at which machine is operating. The total power consumption of a virtual machine is a combination of dynamic power (PO_d) and static power (PO_s) [30]. Dynamic power is consumed when machine inputs like capacitance corresponding to logic gates are active.

$$PO_d = C_{\text{eff}} * V^2 * f \quad (6)$$

where C_{eff} is the switching capacitance per clock cycle, V and f is the operating voltage and frequency, respectively. Equation (7) specifies that voltage V is the dominant factor; therefore, there are mostly reductions in power consumption with voltage reductions. Normally, the static power is directly proportional to dynamic power [31], which is usually less than 30%. Thus, dynamic power is considered for scheduling application in our proposed algorithm. Thus, total power PO_T is considered as dynamic power PO_d . Hence, energy consumption for executing the precedence-constrained task on the virtual machine VM_j is calculated as

$$ER(i, j) = PO_T * EET(i, j) \quad (7)$$

$EET(i, j)$ is estimated execution time of task τ_i on the virtual machine VM_j .

3.4 Temperature Model

During task execution, there is some heat dissipation between processors. Due to this reason, it is needed to know the thermal behavior of the task. But in our approach, a thermal model proposed by Skadron et al. [32] is used to predict that temperature at which task will get execute which is heat independent. An RC-equivalent-based thermal model mainly depends upon the power consumption and instantaneous temperature of task t_i executing at frequency f on virtual machine VM_j is calculated as

$$\text{Temp}_{i,j}(t) = \text{Temp}_{i,j}^{\text{steady}} - \left(\text{Temp}_{i,j}^{\text{steady}} - \text{Temp}_i^{\text{init}} \right) * e^{-t/R_j C_j} \quad (8)$$

where $\text{Temp}_{i,j}^{\text{steady}}$ is steady-state temperature of task τ_i on VM_j . Temperature will reach to steady state when $\tau \rightarrow \infty$, i.e., more number of tasks will execute continuously on the VM. $\text{Temp}_i^{\text{init}}$ is initial temperature of task τ_i . R_j, C_j is thermal resistance and thermal capacitance of virtual machine VM_j .

It is associated with certain input power and formulated as

$$\text{Temp}_{i,j}^{\text{steady}} = PO_{i,j} * R_j + \text{Temp}_{\text{amb}} \quad (9)$$

where Temp_{amb} is the ambient temperature in the sleep state. $PO_{i,j}$ denotes the power consumption of task τ_i on the VM_j .

4 Proposed Algorithm

This section describes the proposed reliability-aware green scheduling algorithm (RAGS) for scheduling the workflow application tasks to the various number of

virtual machines in a cloud environment with the motive to maximize the reliability and minimization of energy consumption with temperature consolidation.

Here, we considered that before scheduling the tasks to available virtual machines, the reliability of the virtual machine VM_i will be calculated to minimize the effect of failure in the future. After submission of tasks to the broker, sort the tasks based on the priority and then schedule to a currently available virtual machine with minimum energy consumption. Priority of the tasks will be decided on the basis of their scheduling time on virtual machines and precedence relation and communication cost or data transfer cost (DTC) between tasks using the *upward rank* as defined in [10]. After calculating the priority, the scheduler schedules the task to a virtual machine with minimum power consumption without violating the reliability and temperature of the system.

Algorithm: Reliability Aware Green Scheduling Algorithm

1. Compute $priority_i$ of execution for each task starting from exit node
 2. Sort the tasks in increasing order using $priority_i$ of those tasks into a task-queue $queue_{priority}$
 3. While the $queue_{priority}$ is not empty **do**
 - a. Remove the first task from the list.
 - b. for each virtual machine $vm_j \in M$ **do**
 - for each $f_k \in F$
 - Calculate the energy consumption $ER(i, j)$ using equation (7) and mark the frequency with minimum energy consumption.
 - End for
 - c. Add the $ER(i, j)$ to ER_{set} .
 - End for
 - d. while ($ER_{set} \neq \emptyset$)
 - $min_ER = \min\{ER(i, j); \text{where } ER(i, j) \in ER_{set}\}$
 - Calculate the $EET(i, j)$ where virtual machine vm_j having minimum energy i.e. min_ER .
 - If ($EET(i, j) \leq LFT$)
 - Add the VM_j to V_{set} .
 - $ER_{set} = ER_{set} - ER(i, j)$
 - else
 - $ER_{set} = ER_{set} - ER(i, j)$
 - end while
 - e. for each virtual machine in V_{set}
 - Calculate the temperature $Temp_{i,j}(t)$ using the execution time $EET(i, j)$ of task τ_i as per eq. (8).
 - Compute the MTTF Find the virtual machine with minimum Mean Time to Failure(MTTF).
 - End for
 - f. Assign the task to a virtual machine with minimum energy and minimum MTTF.
 - g. Delete the task from the head node of $queue_{priority}$.
 4. Compute reliability of a system by summation of each task reliability after assigning all the tasks to virtual machines using Eq. (1).
-

5 Experimental Results and Analysis

The experiments are conducted on a personal computer with Intel Core I7 (8core) and 8 GB RAM using window 10. The implementation of the proposed algorithm is performed using CloudSim [31] Tool. Based on this framework, we have used Randomly generated DAG as a workflow application to implement our proposed RAGS algorithm and verified the reliability and energy consumption during the execution of tasks. In our experiments, energy consumption and reliability is calculated for different sizes of input graphs and a varied number of virtual machines. To verify the efficacy of our proposed algorithm, we compared with two algorithms HRDS [14] and RDLS [33]. The main objective of both of the algorithm is to schedule the task to the reliable virtual machine with a shorter schedule length. In [33]; Tasks are assigned to scheduler dynamically where the number of virtual machines is not fixed.

Experiment 1: At different sizes of Input Graph:

The input tasks' graph size is varying from 20 to 100 in the difference of 20 steps as {20, 40, 60, 80, 100}. Lifetime reliability of the system is calculated using Weibull failure distribution in which scale parameter $\alpha(T)$ is dependent upon the temperature and slope parameter β is taken as 2. Here, we set the activation energy $E_a = 0.48$ eV and cross-sectional area $A_c = 6.4 \times 10^{-8}$ cm². The total number of 8 virtual machines are used for executing the tasks where computing capacity is generated randomly for each virtual machine (Fig. 1).

Experiment 2: At different numbers of virtual machines:

The number of virtual machines is varying from 8 to 64 in the difference of 8 steps for randomly generated task graph with 100 tasks for each set of virtual machines. It is observed that lifetime reliability of the system increases as the number of the virtual machines increases for executing the tasks. If we increase the number of virtual machines, there are more available choices for selecting the reliable virtual machine for execution (Fig. 2).

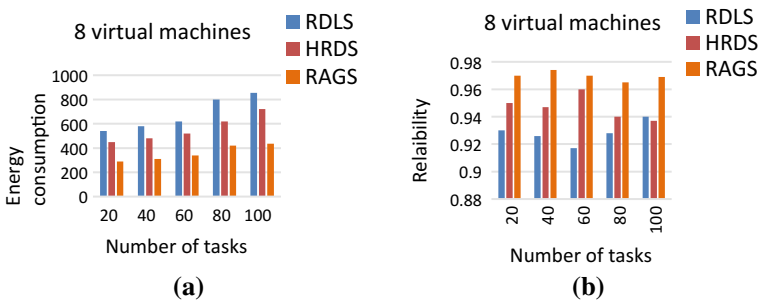


Fig. 1 **a** Lifetime reliability of a system for different sizes of the task graph. **b** Energy consumption (watt/sec) of the system for different sizes of task graph

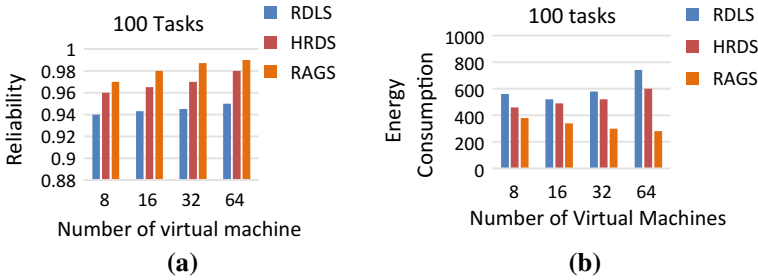


Fig. 2 **a** Lifetime reliability of a system for varying number of virtual machines. **b** Energy consumption (watt/sec) of the system for varying number of virtual machines

6 Conclusion

In this paper, we aim at developing the reliability-aware green scheduling algorithm that incorporates the temperature consolidation for scheduling the tasks. For failure modeling, Weibull failure distribution is taken into account that considers the aging effect. Here scale parameter depends upon the temperature of the system which affects the reliability of application execution. Various number of tasks and virtual machines have been taken for doing the simulation experiment of our algorithm. Our results clearly indicate that our proposed RAGS algorithm outperforms the RDLS and HRDS algorithm. In the future, we can extend our work for multiple workflow applications.

References

1. Sadiku, M.N., Musa, S.M., Momoh, O.D.: Cloud computing: opportunities and challenges. *IEEE Potentials* **33**(1), 34–36 (2014)
2. Wikipedia, Big data. (2014). http://en.wikipedia.org/wiki/Big_data
3. Garraghan, P., Townend, P., Xu, J.: An empirical failure-analysis of a large-scale cloud computing environment. In: 2014 IEEE 15th International Symposium on High-Assurance Systems Engineering (HASE), pp. 113–120. IEEE, New York (2014, January)
4. Wikipedia Moore's Law. (2012). http://en.wikipedia.org/wiki/Moore's_law
5. Skadron, K., Stan, M.R., Huang, W., Velusamy, S., Sankaranarayanan, K., Tarjan, D.: Temperature-aware microarchitecture. In: International Symposium on Computer Architecture (2003)
6. Brooks, D., Martonosi, M.: Dynamic thermal management for high-performance microprocessors. In: International Symposium on High-Performance Computer Architecture (2001)
7. Chantem, T., Dick, R.P., Hu, X.S.: Temperature-aware scheduling and assignment for hard real-time applications on MPSoCs. In: Design, Automation and Test in Europe (2008)
8. Patel, P., Ranabahu, A.H., Sheth, A.P.: Service level agreement in cloud computing (2009)
9. Khan, M.A.: Scheduling for heterogeneous systems using constrained critical paths. *Parallel Comput.* **38**(4), 175–193 (2012)
10. Garg, R., Singh, A.K.: Adaptive workflow scheduling in grid computing based on dynamic resource availability. *Eng. Sci. Technol. Int. J.* **18**(2), 256–269 (2015)

11. Jadon, S.S., Bansal, J.C., Tiwari, R., Sharma, H.: Artificial bee colony algorithm with global and local neighborhoods. *Int. J. Syst. Assur. Eng. Manag.* pp. 1–13 (2014). <https://doi.org/10.1007/s13198-014-0286-6>
12. Topcuoglu, H., Hariri, S., Wu, M.: Performance-effective and low-complexity task scheduling for heterogeneous computing. *IEEE Trans. Parallel Distrib. Syst.* **13**(3), 260–274 (2002)
13. Tang, X., Li, K., Li, R., Veeravalli, B.: Reliability-aware scheduling strategy for heterogeneous distributed computing systems. *J. Parallel Distrib. Comput.* **70**(9), 941–952 (2010)
14. Tang, X., Li, K., Qiu, M., Sha, E.H.M.: A hierarchical reliability-driven scheduling algorithm in grid systems. *J. Parallel Distrib. Comput.* **72**(4), 525–535 (2012)
15. Huang, L., Yuan, F., Xu, Q.: Lifetime reliability-aware task allocation and scheduling for MPSoC platforms. In: *Design, Automation & Test in Europe Conference & Exhibition, 2009. DATE'09*, pp. 51–56. IEEE, New York (2009, April)
16. Garg, R., Singh, A.: Energy-aware workflow scheduling in grid under QoS constraints. *Arab. J. Sci. Eng.* **41**(2) (2016)
17. Bingulac, S.P.: On the compatibility of adaptive controllers. In: *Proceedings of the 4th Annual Allerton Conference on Circuits and Systems Theory*, New York, p. 816 (1994)
18. Xu, A., Yang, Y., Mi, Z., Xiong, Z.: Task scheduling algorithm based on PSO in cloud environment. In: *2015 IEEE 12th International Conference on Ubiquitous Intelligence and Computing and 2015 IEEE 12th International Conference on Autonomic and Trusted Computing and 2015 IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, pp. 1055–1061. IEEE, New York (2015, August)
19. Salido, M.A., Escamilla, J., Giret, A., Barber, F.: A genetic algorithm for energy-efficiency in job-shop scheduling. *Int. J. Adv. Manuf. Technol.* **85**(5–8), 1303–1314 (2016)
20. Wang, S., Chen, J.J., Shi, Z., Thiele, L.: Energy-efficient speed scheduling for real-time tasks under thermal constraints. In: *15th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications, 2009. RTCSA'09*, pp. 201–209. IEEE, New York (2009, August)
21. Mei, J., Li, K., Zhou, X., Li, K.: Fault-tolerant dynamic rescheduling for heterogeneous computing systems. *J. Grid Comput.* pp. 1–19 (2015)
22. Guo, S., Huang, H.Z., Wang, Z., Xie, M.: Grid service reliability modeling and optimal task scheduling considering fault recovery. *IEEE Trans. Reliab.* **60**(1), 263–274 (2011)
23. Das, A., Kumar, A., Veeravalli, B.: Reliability and energy-aware mapping and scheduling of multimedia applications on multiprocessor systems. *IEEE Trans. Parallel Distrib. Syst.* **27**(3), 869–884 (2016)
24. Zhang, L., Li, K., Li, C., Li, K.: Bi-objective workflow scheduling of the energy consumption and reliability in heterogeneous computing systems. *Inf. Sci.* **24379**, 241–256 (2017)
25. Deb, K., Pratap, A., Agarwal, S., Meyarivan, T.A.M.T.: A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)
26. Tang, X., Tan, W.: Energy-efficient reliability-aware scheduling algorithm on heterogeneous systems. *Sci. Program.* **2016**, 14 (2016)
27. HYSTERY: a hybrid scheduling and mapping approach to optimize temperature, energy consumption and lifetime reliability of heterogeneous multiprocessor systems
28. Zhang, L., Li, K., Xu, Y., Mei, J., Zhang, F., Li, K.: Maximizing reliability with energy conservation for parallel task scheduling in a heterogeneous cluster. *Inf. Sci.* **319**, 113–131 (2015)
29. Srinivasan, J., Adve, S.V., Bose, P., Rivers, J.A.: The case for lifetime reliability-aware microprocessors. In: *ACM SIGARCH Computer Architecture News*, vol. 32, No. 2, p. 276. IEEE Computer Society (2004, June)
30. Kim, K.H., Buyya, R., Kim, J.: Power aware scheduling of bag-of-tasks applications with deadline constraints on DVS-enabled clusters. In: *CCGrid*, vol. 7, pp. 541–548 (2007, May)
31. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., Buyya, R.: CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. *Softw.: Pract. Exp.* **41**(1), 23–50 (2011)

32. Skadron, K., Stan, M., Sankaranarayanan, K., Huang, W., Velusamy, S., Tarjan, D.: Temperature-aware microarchitecture: modeling and implementation. *ACM Trans. Arch. Code Optim.* **1**(1), 94–125 (2004)
33. Dogan, A., Ozguner, F.: Matching and scheduling algorithms for minimizing execution time and failure probability of applications in heterogeneous computing. *IEEE Trans. Parallel Distrib. Syst.* **13**(3), 308–323 (2002)

An Empirical Study to Classify Website Using Thresholds from Data Characteristics



Ruchika Malhotra and Anjali Sharma

Abstract The advent of web had resulted in a plethora of information and data. However, its volume heterogeneity and unstructured organization makes information retrieval difficult. To the existing practice where website categorization is largely based on style rather than text, addition of an extra dimension in form of genre is expected to significantly improve the search outcome. Keeping this in view, we attempt to build a novel classification model to categorize websites into genres using thresholds of the web metrics. Statistical measures of central tendency are assumed to render a value that distinguish websites from a sample space containing News, Travel and Tourism, Entertainment and Social media. Through the statistical analysis of the data we find that the data distribution of all metrics which constitute the website properties are highly skewed. Hence, conventional analysis based on normal distribution statistics fails to apply. Adopting to a systematic empirical approach, we find that the classification performance measure identified through the Area Under the Curve is maximized around a threshold value which is twice the value of the “median-absolute-deviation” of the web metrics.

Keywords Web genre · HTML metrics · Threshold · Median-Absolute-Deviation · Naive Bayes

1 Introduction

The semi-structured, dynamic and heterogeneous nature of websites make information classification increasingly challenging [1, 2]. As a result, even the most versatile search engines provide lesser accurate results to very specific information sought on the web. In fact, the current search engines return the ranked list of documents

R. Malhotra (✉) · A. Sharma
Delhi Technological University, Bawana Road, New Delhi 110042, India
e-mail: ruchikamalhotra2004@yahoo.com

A. Sharma
CSIR-NPL, Dr K S Krishnan Marg, New Delhi 110012, India

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_39

depending upon textual similarity, together with an independent measure of each web page's importance [3]. In such cases, the search outcome is a myriad of web pages, which then further requires more user defined search filters, thereby making the search task more difficult. For more complex cases, it may even require that the user would need to visit each web page and apply a manual search for the desired information.

However, dealing with genres pose two important aspects, (i) What are most important and relevant genres to be considered? and (ii) and what factors explicitly help distinguish the genres. In general, the complexity of a web page is heterogeneous in nature, i.e., despite prescription of what is genre, it may be found that genres tend to overlap and mix [4, 5]. For instance, the Google search engine has tabs for 'Maps', 'News', 'Images', 'Videos' etc., to segregate content according to presentation style, rather than topic. However, the web page information are not classified according to genres, and as a result the user is unable to specify the category of search content style like 'Report' or 'Wikipedia' or 'e-Commerce', which may be more specific in accordance to the search interest itself. Thus, classification of web pages into genre is a challenging task as the information sought is specific to certain appropriate features that describe the web page in context of a genre.

However, beset with difficulties the need has been realized to have an extra classification scheme, particularly in terms of genre. As of date, the web genre classification distinguishes between pages by means of their style, presentation layout, form and meta-content features rather than topic. Genre adds an extra dimension to the classification of web pages, with improved search results. In order to classify the web pages we require specific features through which the genres can be classified. This forms the objective of the current work.

Assuming that a set of independent HTML metrics can be identified and that a combination of these metrics can aptly describe the genre of a given website, we propose that there exists a threshold value associated with each metric, on the basis of which one can classify the web pages. Provided that there exists such a scheme, and that the underlying methodology is simple, it may be then guaranteed that information retrieval from web would be fast and accurate and be quite simpler to handle large databases. As a case study, we try to identify websites related to "Travel and Tourism" and "Social Media" from a sample space containing genres such as "E-Commerce", "Social Media", "Entertainment", and "News". Using statistical analysis of the data distribution, a systematic methodology is being developed for identifying the thresholds of the web metrics by exploiting inherent internal characteristics of the HTML metrics. The thresholds of selected web metrics are then utilized for building models to predict genre of website using machine learning techniques.

The remainder of the paper is organized as follows. Section 2 provides a brief review of the related works in the estimation of thresholds in software engineering. Section 3 explains our threshold-based feature selection methodology. In Sect. 4, we show our results which include the statistical description of the metric data distribution, the classifier performance values against a range of values. From the trend that follows, we find that a reasonable threshold associated with each web category is the value corresponding to twice the median absolute deviation. In Sect. 5 we present

the limitations of the work. Finally, we conclude the work in Sect. 6 and suggest improvements for future studies.

2 Literature Survey

The web genre classification depends upon the input metrics, machine learning technique and the genre class. In earlier research work, web pages were considered to represent a single web genre [6, 7] for web genre classification. In contrast with this assumption, many scholars and researchers argued that a single genre classification scheme is inappropriate for web pages [8–11].

In one of the original attempts, Crowston and Williams [12] from a set of randomly selected web pages, with certain set of objectives identified, proposed four types of genres - Reproduced genres, Adapted genres, Novel genres and Unclassified web pages. The study revealed that genres cannot be simply cached and stored in a repository, but evolves. Similar credence was supported by Shepherd and Watters [6, 7]. The latter authors introduced a new terminology “Cybergenre”, which is currently popular as “web genre”. Accordingly, the genre is characterized by a base triplet namely, {<content>, <form>, <functionality>}. While both <content> and <form> represented the traditional genres, <functionality> defined the capabilities offered by the web. Significantly, along with the functionality genre, an important attribute was soon realized, i.e., based on the use of hypertext and/or HTML. Each hypertext corresponded to a genre. It may be noted that although a website may be a collection of web pages, the genre analysis is basically done for the entire website [13, 14]. A super-genre classification of websites [15] was done by using structure, content and their combination to improve the classification accuracy.

Overall, the research work discussed above is in agreement that structure and functionality attributes of a web page represent useful information which can be used to identify the genre of a website. Therefore, we have focused on the quantitative web metric set of <Structural> and <Functionality> attributes represented by text formatting, navigation and external object HTML tags.

In general, the threshold in software systems can be estimated from statistical deductions and mathematical models. The statistical methodology provides qualitative thresholds, however, to improve the validity of the results it is important to study the relation between the data characteristics, underlying assumptions and nature of the problem. Here we briefly review the threshold estimation studies based on statistical inference for software.

The study conducted by Erni and Lewerentz [16] estimated the threshold to be in the range of statistical mean (μ) and standard deviation (σ), represented as $T_{\pm} = \mu \pm \sigma$, assuming data to be distributed normally. However, the technique assumed the input metric data to be normally distributed. Usually such distributions are seldom common in software projects and hence the applicability of the technique is limited. The work by French [17] included Chebyshev’s inequality theorem along with μ and σ for threshold calculation but distribution nature of data was again not

considered and the methodology suffered to the data outliers. The recent work of de Siqueria et al. [18], have suggested three similarity thresholds, using arithmetic or the weighted mean, k -means clustering and silhouette coefficient maximization, for the genre aware focused crawling.

Shatnawi [19] used ROC characteristics to identify threshold values and analyzed its association with different error severity levels. The relevant threshold values were found for high and medium risk categories of ordinal classification but could not find practical threshold values for binary classification. In a following study [20], the author calculated the thresholds corresponding to the C&K metrics using Bender's approach [21] based on logistic regression and it was found that risk levels can be used to identify metric thresholds. Similarly, Malhotra et al. [22] also used Bender's approach to calculate the metrics threshold and determined the effects of threshold on change prediction with inter-project studies. Their results showed that the transferability of the threshold is limited rather to a narrow confidence interval in inter-project comparisons. In a more recent work, Shatnawi [23] proposed data transformation method to reduce skewness in the data and the threshold values were estimated using the statistical parameters such as μ and σ , similar to the works of Erni et al. [16]. However, what limits the underlying methodology is the shift of values by a constant value prior to the data transformation. Alves et al. [24] investigated data distribution properties of object oriented metrics to derive threshold values and the estimated metrics threshold values were insensitive towards data outliers. Similarly, Ferreira et al. [25] statistically analyzed the data to calculate the threshold range of certain metrics. The authors found that most of the metrics followed a heavy-tailed distribution and argued that a general threshold could not be applied to the object oriented software projects. On the other hand, Hussain et al. [26], compared the effect of thresholds derived using Bender's approach and those mentioned by Alves et al. [24] and concluded that thresholds cannot be generalized for all the systems due to variation in data characteristics.

The studies discussed above have emphasized the importance of the data characteristics and statistics to be considered before estimating the thresholds. Hence, in this work we estimate the threshold of web metrics using the statistical measures of central tendency after analyzing the data distribution. The threshold estimates are used for categorizing the websites according to their genre.

3 Methodology

The methodology we follow in this work is schematically shown in Fig. 1. The Web Metric Collection and Reporting System (MCRS) [27], crawls URL to collect HTML, NLP and text metrics for web genre classification. The HTML metric collector extracts all the links in the web page and collects various HTML web metrics namely, Text Formatting tags, document structure tags, external object tags, instruction and navigation tags. As highlighted in [28], the combination of lexical, functional and structural attributes shall be used for genre classification. Therefore,

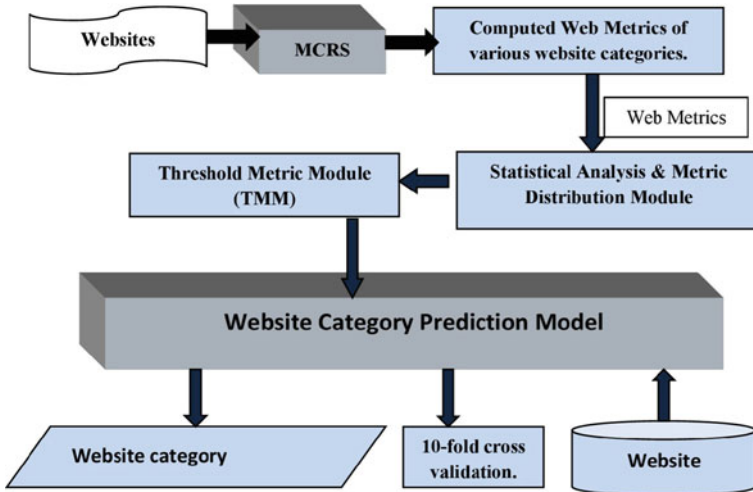


Fig. 1 Schematic representation of the methodology adopted in the study

we have used the “Structural” features of web page represented by text formatting `
`, `<div>`, ``, `<p>`, ``, `` and navigation `<a>` tags, while the external object tags ``, `<script>` are used to define “Functionality”. Therefore, these nine web metrics, listed in Table 1, constitute the independent variables in the study, which are used to categorize the website as “Travel and Tourism”, “E-Commerce”, “Social Media”, “Entertainment” or “News”.

The set of nine metrics and website category serve as input to the statistical analysis and metric distribution module. The statistical parameters of central tendency for sample space including range (R_{max}, R_{min}), mean (μ) and median (x_m) are calculated in this module. Also the histogram plots are investigated to identify the distribution characteristics of the input metric space. These statistical parameters along with the sample space serve as input to the Threshold Metric Module (TMM), which estimates the threshold values for the “Travel and Tourism” and “Social Media” website category.

The website category prediction model is built using Naive Bayes classifier, with input from the TMM and renders the classification performance measure of the web category in terms of the AUC values. The selection of the Naive Bayes algorithm is not only because of the common use in data mining applications, but also due to its reliable performance for small dataset [29]. The default parameter settings were used for the learners as specified in Weka. A priori, AUC is chosen as the performance measure, due to the inherent class imbalance observed in the dataset. By definition, AUC is the probability a classifier ranks a randomly chosen positive instance higher than its negative instance counterpart. In the Receiver Operating Characteristics (ROC), the magnitude of AUC varies in between 0 and 1. Note that ROC analysis helps in decision making, by relating the performance and non-performance of a classification model.

Table 1 The HTML metrics used in the study

Table	HTML tag	Description
Break	 	It breaks the line and bring the statement in next line. It is an empty tag which means it has no end tag
Division	<div>	It defines a section in html tag and act as a container for other elements to style them using CSS and perform various tasks
List		It is used in ordered list, menu or unordered list to list various items
Paragraph	<p>	To write any paragraph we use <p> tag. Spacing is provided automatically by browser before and after paragraph
Span		It provides no change by itself and is used to group elements in a document
Unordered list		 tag is used with tag to create unordered list. It defines the list with bullets
Image	<image>	It defines image in html. It adds image to the html page by using two attributes 'src' (source) and 'alt' (alter). Src provides the path of the image. Alt alters the size of the image
Script	<script>	It defines client side scripting. It contains scripting statements and also it points external file through "src" (source) attribute
Anchor	<a>	This tag is used to include hyperlinks to the html page. It links one page to another. The main attribute of <a> tag is <href> which include link destination

4 Results and Discussion

4.1 Data Characteristics

In Table 2 we describe the statistics of the metric data, the latter which includes all five web categories. The range of the metrics with its upper limit, designated as R_{\max} are shown along with the measures of central tendency, i.e., the mean (μ) and median (x_m).

It is evident from Table 2 that the range of the metrics are quite different. We also find significant difference in the mean and median values thereby inferring a non-normal distribution of the metrics. All metrics distribution are found positively skewed, since $\mu > x_m$. Empirically, by considering the difference ($\mu - x_m$), as a measure of skewness, the data reveal that the distribution associated with the <div> and <a> metrics are relatively more skewed, while <script> is least skewed.

We first attempt to construct the threshold parameters for the "Travel and Tourism" web category. For the same, we first analyze the statistical parameters associated with the web category with respect to the sample space. In Table 2, we show the statistical description of the metrics associated with the "Travel and Tourism" web category. The overall characteristics of the sub-space remains similar to that of the sample space, i.e., the metrics distribution are skewed with mean being greater than the

Table 2 The statistical description of the selected HTML metrics for the sample space, travel and tourism and social media categories

Metrics	Sample space			Travel and tourism				Social media			
	R_{max}	x_m	μ	R_{max}	x_m	μ	MAD	R_{max}	x_m	μ	MAD
 	113	4	12.24	94	5	13.23	19.5	113	2	10.2	2
<div>	1229	90	155.03	472	56	71.97	65	1229	65	147.8	43
	907	48	97.97	315	36	51.34	72	652	11	60.3	11
<p>	203	6	20.65	153	3.5	11.71	9	203	3	20.4	3
	506	29	68.72	403	11.5	37.47	39	376	10	43.8	9
	190	8	17.45	56	6.5	9.36	14.5	122	3	11.2	3
	179	22	34.01	90	15	20.34	24.5	179	8	27.9	8
<script>	108	16	19.46	42	10	13.81	14.75	66	12	17.7	7
<a>	958	101	175.55	353	76	88.81	69	760	44	111.3	35

The R_{max} , x_m , μ , MAD values represent the Upper, Median, Mean and Median-Absolute-Deviation values, respectively. The Lower value (R_{min}) for all the metrics is 0

median. Besides, we also note that the category resides well inside the sample space with no range maximum of any of its nine metrics with that of the sample itself.

The basic statistical description of the data pertained to the ‘‘Social Media’’ web category is shown in Table 2. It is found that the distribution of data are very different from <script> metrics and spans the entire range in the ‘‘Social Media’’ category, which is not the case for ‘‘Travel and Tourism’’. The <div> metrics range from [0, 472] in ‘‘Travel and Tourism’’, but shows a wider range of [0, 1229] in the ‘‘Social Media’’ category. We also note that four out of nine metrics, namely
, <div>, <p> and representing the ‘‘Social Media’’ category are spread all across the entire range, which is in contrast to the data distribution associated with ‘‘Travel and Tourism’’ category.

For a better understanding of the category wise metric distribution with respect to the sample, and also among the five categories, we analyzed the data in terms of frequency plots as shown in Figs. 2 and 3. The wide difference among the web categories in the metric space is very evident. Not only that we find the frequencies associated with the metric values to be very different, but also that certain metrics distribution were found to be continuous for some categories, while for others it looked non-uniform and discontinuous. For instance, for <div> in ‘‘Travel and Tourism’’ the frequency of data in the range [0, 50] was found to be 17, while in ‘‘Social Media’’ it was determined to be 30. On the other hand, both metrics and shows a continuous and decreasing trend with increasing range in the ‘‘Social Media’’ category, while in ‘‘Travel and Tourism’’, the distribution is discontinuous. In fact, based on our inter-quartile analysis, we find that the data in the ‘‘Travel and Tourism’’ category for metric in the range [300, 350] and that for in [50, 60] are representation of being outliers. Thus, statistical analyses show that the web categories in the sample space are widely different in terms of the metrics that define each category.

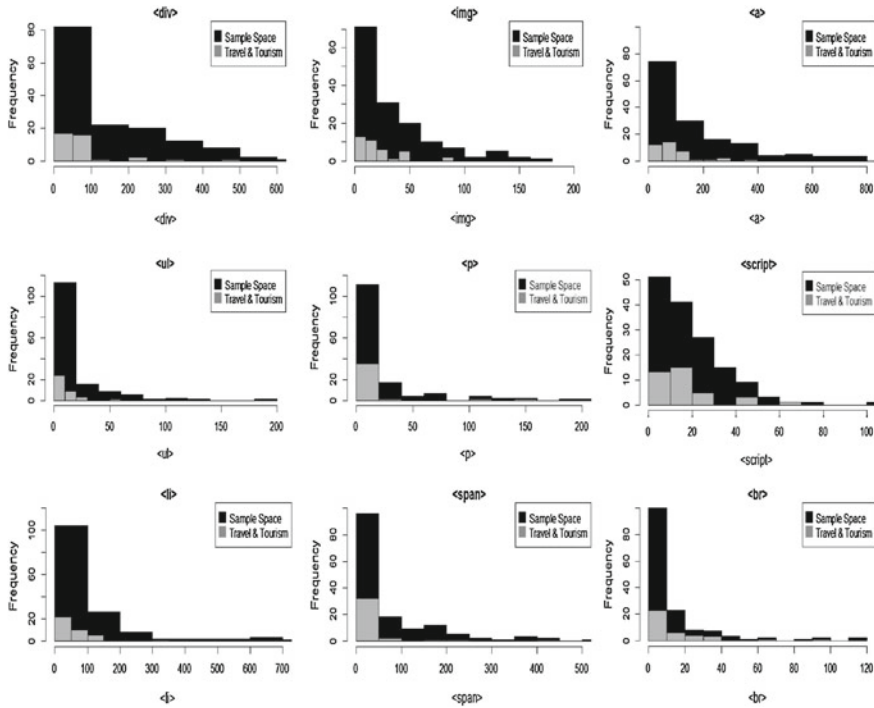


Fig. 2 Histogram representation of the sample space (shaded black), with data on “Travel and Tourism” (shaded grey) projected, of the nine metrics used in the study

These statistical observations of the metric data incite to look for a threshold value, or a set of values, which differentiate each web category in a given sample space.

4.2 Threshold Calculation

The problem at hand, therefore is to determine the threshold upon which one can classify the web categories from a given sample space. For this, one need to have an initial guess to the threshold value, upon which the performance measure of certain chosen category can be calculated. Further assuming that there exists a unique set of threshold parameters to determine the threshold, we vary the guess parameter in increments so as to obtain an optimal value of performance. As obvious, the range is quite different for all metrics within a given web category, and also among various categories. Thus, the minimum and maximum values of a metric distribution are not very good statistical parameter to use, since they can fluctuate greatly from sample to sample. Besides, the distribution as mentioned above have significant deviations from that of a normal distribution.

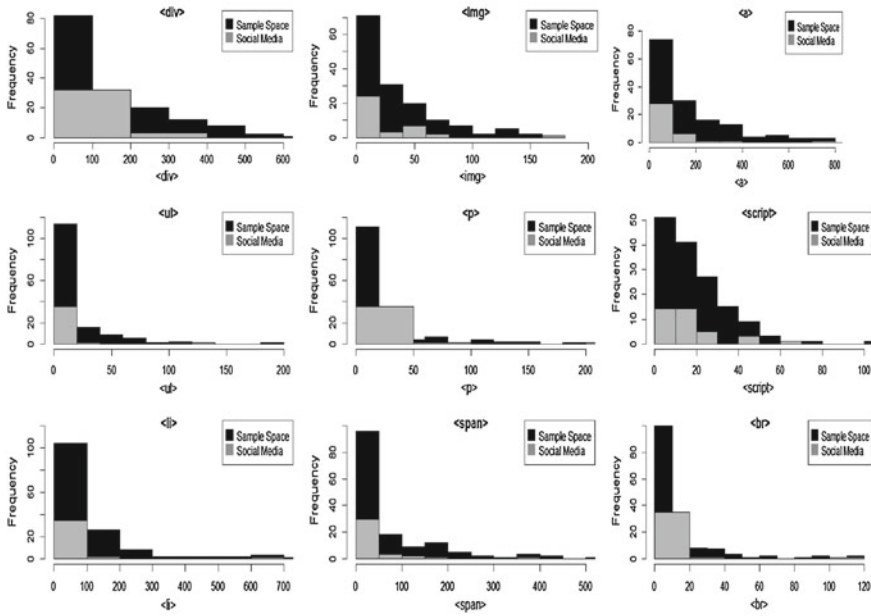


Fig. 3 Histogram representation of the sample space (shaded black), with data on “Social Media” (shaded grey) projected, of the nine metrics used in the study

As a result, we argue that neither the limiting range parameters nor the mean value of the distribution is a good choice to be considered as an initial guess for the threshold. For a simple, nonparametric statistic to represent variability of a skewed data, we therefore consider median (x_m) as our reference measure of central tendency, which also forms as our initial guess to the threshold value. In analogy to the role of standard deviation (σ) in normal distribution. A widely used parameter for variance in skewed dataset is the statistical quality referred as “Median Absolute Deviation” (MAD). Much similar to the relevance of $\mu \pm 2\sigma$ in normal statistic dataset, here we use $x_m \pm (2 \times MAD)$ as a range for the calculation of the threshold value. Mathematically, MAD is defined in Eq. (1) as,

$$MAD = \text{median} \times (|X_i - \text{median}_j(X_j)|) \tag{1}$$

By definition, MAD represents a measure of statistical dispersion. For non-normal dataset, MAD is a robust estimator of scale than the conventional variance or standard deviation. MAD also is a much better statistical quantity for distributions that have neither mean nor variance, such as that for Cauchy distribution, and thus includes as a universal statistical quantity for any metric space irrespective of its nature. Furthermore, an advantage of using MAD as a statistical estimator is due to its insensitiveness towards outliers. We define threshold as a boundary which differentiates radically different regions. In this context, we anticipate a change in the variation of AUC as a

Table 3 The performance of the web category prediction model with AUC measure, with and without threshold

	Travel and tourism	Social media
Without threshold	0.71	0.74
x_m	0.66	0.94
$x_m + \delta$	0.55	0.94
$x_m + 2\delta$	0.62	0.89
$x_m + 3\delta$	0.74	0.89
$x_m + 4\delta$	0.97	0.87
$x_m + 5\delta$	0.92	0.87
$x_m + 6\delta$	0.92	0.87

Here x_m represents the median value, and $5\delta = (2 \times \text{MAD})$

function of $x_m + n\delta$, where δ is an increment and “ n ” a positive integer. To determine the maximum range up to which $n\delta$ values will be varied, $2 \times \text{MAD}$ is considered.

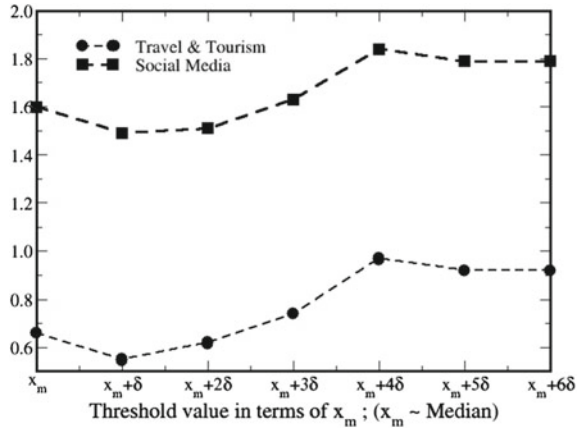
4.3 Website Category Prediction Model Using Threshold

The median and $2 \times \text{MAD}$ values of all metrics are first calculated (Refer Table 2). To calculate the performance of web category prediction model using thresholds, we transform each metric into a binary form. The metric values below threshold are transformed to “zero” and those above as “one”. Thereafter, the binary transformed dataset is fed as input to the Naive Bayes algorithm with stratified remove folds as class balance technique. The corresponding AUC value is computed which are shown in Table 3.

The AUC with median as threshold for the transformed data is determined to be 0.66, which is relatively lower to the AUC computed for the original (untransformed) dataset, i.e., 0.71. Note that for the value of $x_m + n\delta$, which shows a characteristic change in the AUC value would be referred as the threshold value associated with the web category.

In Fig. 4 we show the stacked line graph of the performance measure variation with respect to $x_m + n\delta$, obtained for both “Travel and Tourism” and “Social Media”. The results are shown in this form mainly because stacked line representation enable us to capture the trend in the variation with variable threshold range assumed in the calculation. Besides, since such a graph is cumulative at each point the data does not overlap. It is very evident that the performance measures of both web categories are very similar. With increase in the δ , the graph initially decreases by 16 and 7% for “Travel and Tourism” and “Social Media”, respectively. Thereafter, we find a gradual increase in the performance measure up to $x_m + 4\delta$, beyond which the values saturate. Behold, the definition of threshold as the boundary which separate two region in the variation of performance, we find that the boundary in this study points to $x_m + 5\delta$. Interestingly, this also correspond to the $(2 \times \text{MAD})$ value. Thus, with the

Fig. 4 The stack plot showing the variation of relative performance measure of “Travel and Tourism” and “Social Media”. Note that for the exact AUC measure corresponding to “Social Media”, the value corresponding to a given $x_m + n\delta$, has to be subtracted from the “Travel and Tourism” value



observation that two very different skewed metric data distribution associated with “Travel and Tourism” and “Social Media”, not only exhibiting similar trend but also inferring $(2 \times MAD)$ as threshold help us formulate the following hypothesis; i.e., there exists a close relationship between the threshold value and statistics in the web category distinction, with $(2 \times MAD)$ value corresponding to the threshold itself.

5 Threats to Validity

One of the basic question is how well the present experiment has been done. In this perspective, one of the confounding issue is in the selection of the input metrics. Whether or not the chosen metrics forms a complete metric space, and/or whether there exists a linear dependence between the metrics is an internal threat to the validity of the results. As a check, it would suffice to use feature selection techniques and calculate the optimal threshold value.

In this study, we have used the Naive Bayes algorithm. For a wider understanding, the use of a single machine learning algorithm could be a possible threat to the conclusion validity of this study. However, As mentioned above Naive Bayes has been found to yield reliable results for smaller dataset and also yet being a simple model the algorithm has found numerous applications providing high performance for a large variety of datasets. However, as a future work we will be evaluating the performance with several other machine learners such as Bagging and Boosting algorithms.

The observations following the study is limited in generalization as to similar studies which may span other networking sites. This pose a possible external validity threat. For instance, the design of websites can significantly depend on the culture and tradition of various communities and public across the globe. To minimize these

local effects, it is important to collect data from various other networking sites across the globe and investigate to what extent the results can be generalized.

6 Conclusion

Identification of proper web genres are expected to ease classification at both organizational and at user level. Given that the evaluation of web sites would thereby become plausible at lower cost, development of web genres are becoming increasingly important for the developers to adopt measures so as to ease search queries. In this regard, we propose a model based on threshold to distinguish various web categories based on the statistical measures of central tendency. Since the metrics that define the metric space are found skewed, we guess that the threshold would be more related to the median value than the more widely used mean. Setting the definition of threshold as the boundary that differentiates the classification performance rendered by a machine learning algorithm, we vary the threshold value in increments, from median towards the skewed part of the spectra. The trend as captured by the AUC values clearly shows that beyond certain optimum value, the magnitude of the performance measure saturates. We argue that the set of metric values that put the magnitude of the performance in saturation can be termed as the threshold. In statistical realms, our study shows that the threshold is $(x_m + 2 \times MAD)$, where x_m and MAD represents the median and “median absolute deviation”, respectively. In analogy with the standard deviation which is commonly used for dataset with normal distribution, we conclude that the proposed threshold estimate evaluated lies within 95% confidence interval. The use of Median-Absolute-Deviation has never been proposed in any earlier works related to threshold determination in website categorization, and hence require more experiments over a wider range of website classifications.

References

1. Chetry, R.: Web genre classification using feature selection and semi-supervised learning (2011)
2. Gatto, M.: Web as Corpus: Theory and Practice. Bloomsbury Academic, London (2014)
3. Stein, B., Zu Eissen, S.M., Lipka, N.: Web genre analysis: use cases, retrieval models, and implementation issues. In: Genres on the Web, pp. 167–189. Springer, Dordrecht (2010)
4. Ponzanelli, L., Mocci, A., Lanza, M.: Summarizing complex development artifacts by mining heterogeneous data. In: Proceedings of the 12th Working Conference on Mining Software Repositories, pp. 401–405. IEEE Press, New York (2015)
5. Wu, L., Du, L., Liu, B., Xu, G., Ge, Y., Fu, Y., Li, J., Zhou, Y., Xiong, H.: Heterogeneous metric learning with content-based regularization for software artifact retrieval. In: IEEE International Conference on Data Mining (ICDM), pp. 610–619. IEEE, New York (2014)
6. Shepherd, M., Watters, C.: The functionality attribute of cybergenres. In: Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences, HICSS-32, p. 9. IEEE, New York (1999)

7. Shepherd, M., Watters, C.: Identifying web genre: hitting a moving target. In: Proceedings of the WWW Conference. Workshop on Measuring Web Search Effectiveness: The User Perspective, vol. 18, New York (2004)
8. Rosso, M.A.: Using genre to improve web search. Doctoral dissertation, University of North Carolina, Chapel Hill (2005)
9. Williams, K.C.M.: Reproduced and emergent genres of communication on the World Wide Web. *Inf. Soc.* **16**, 201–215 (2000)
10. Santini, M.: Characterizing genres of web pages: genre hybridism and individualization. In: HICSS 40th Annual Hawaii International Conference on System Sciences, pp. 71–71. IEEE, New York (2007)
11. Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X.: Genre based navigation on the web. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences, p. 10. IEEE, New York (2001)
12. Crowston, K., Kwasnik, B.H.: A framework for creating a faceted classification for genres: addressing issues of multidimensionality. In: Proceedings of the 37th Annual Hawaii International Conference on System Sciences, p. 9. IEEE, New York (2004)
13. Copestake, A.: Errors in wikis. In: Proceedings of the Workshop on NEW TEXT Wikis and Blogs and Other Dynamic Text Sources (2006)
14. Mehler, A.: Text linkage in the wiki medium: a comparative study. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy, April 3–7, 2006 (EACL 2006): Workshop on New Text—Wikis and blogs and other dynamic text sources, pp. 1–8 (2006)
15. Lindemann, C., Littig, L.: Classification of web sites at super-genre level. In: Genres on the Web, pp. 211–235. Springer Netherlands (2010)
16. Erni, K., Lewerentz, C.: Applying design-metrics to object-oriented frameworks. In: Proceedings of the 3rd International on Software Metrics Symposium, pp. 64–74. IEEE, New York (1996)
17. French, V.: Establishing software metric thresholds. In: Proceedings of the 9th International Workshop on Software Measurement (1999)
18. de Siqueira, G.O., de Assis, G.T., Almeida Ferreira, A., Mangaravite, V., Cardeal P'adua, F.L.: Strategies for automatic determination of similarity threshold for genre-aware focused crawling processes. In: IADIS International Journal on WWW/Internet, vol. 15 (2017)
19. Shatnawi, R., Li, W., Swain, J., Newman, T.: Finding software metrics threshold values using ROC curves. *J. Softw. Maint. Evol.: Res. Pract.* **22**, 1–16 (2010)
20. Shatnawi, R.: A quantitative investigation of the acceptable risk levels of OO metrics in open-source systems. *IEEE Trans. Softw. Eng.* **36**, 216–225 (2010)
21. Bender, R.: Quantitative risk assessment in epidemiological studies investigating threshold effects. *Biom. J.: J. Math. Methods Biosci.* **41**, 305–319 (1999)
22. Malhotra, R., Bansal, A.J.: Fault prediction considering threshold effects of object-oriented metrics. *Expert. Syst.* **32**, 203–219 (2015)
23. Shatnawi, R.: Deriving metrics thresholds using log transformation. *J. Softw.: Evol. Process.* **27**, 95–113 (2015)
24. Alves, T.L., Ypma, C., Visser, J.: Deriving metric thresholds from benchmark data. In: IEEE International Conference on Software Maintenance (ICSM), pp. 1–10. (2010)
25. Ferreira, K.A., Bigonha, M.A., Bigonha, R.S., Mendes, L.F., Almeida, H.C.: Identifying thresholds for object-oriented software metrics. *J. Syst. Softw.* **85**, 244–257 (2012)
26. Hussain, S., Keung, J., Khan, A.A., Bennin, K.E.: Detection of fault-prone classes using logistic regression based object-oriented metrics thresholds. In: IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C), pp. 93–100 (2016)
27. Malhotra, R., Sharma, A.: A web metric collection and reporting system. In: Proceedings of the Third International Symposium on Women in Computing and Informatics, pp. 661–667. ACM, New York (2015)
28. Malhotra, R., Sharma, A.: Quantitative evaluation of web metrics for automatic genre classification of web pages. *Int. J. Syst. Assur. Eng. Manag.* **8**, 1567–1579 (2017)

29. Frman, G., Cohen, I.: Learning from little: comparison of classifiers given little training. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 161–172. Springer, Berlin (2004)

Part V
Intelligent Image Processing

Assessment of Spectral-KMOD Composite Kernel-Based Supervised Noise Clustering Approach in Handling Nonlinear Separation of Classes



Ishuita SenGupta, Anil Kumar and Rakesh Kumar Dwivedi

Abstract The paper propounds a concept of incorporating composite kernel methods with fuzzy-based image classifiers. The study incorporates noise classifier as a fuzzy classifier. The work demonstrates how nonlinearity among the different classes of remote sensing data with uncertainty is handled with noise classifier without entropy (fuzzy classifier) using composite kernel technique for land use/land cover maps generation. It also showcases the comparative study between the performance of Noise Classifier with Euclidean Distance and Noise Classifier with Composite Kernel functions. This study has incorporated the composition of two prominent kernels: Spectral and KMOD Kernel. The performance of both the classifier is evaluated in supervised mode and, image-to-image assessment of accuracy has been carried out using FERM (Fuzzy Error Matrix).

Keywords Fuzzy classifier · Mixed pixel · Composite kernel · Noise classifier · FERM

I. SenGupta (✉) · R. K. Dwivedi
College of Computing Sciences and Information Technology, Teerthanker Mahaveer
University, Moradabad, India
e-mail: ishuitasengupta8@gmail.com

R. K. Dwivedi
e-mail: principal.computers@tmu.ac.in

A. Kumar
Photogrammetry and Remote Sensing Department, Indian Institute of Remote Sensing,
Dehradun, India
e-mail: anil@iirs.gov.in

1 Introduction

The conventional classification techniques typically classify maps into hard, discrete categories or classes (like River, Forest). Heterogeneous features in the digital image are higher resulting in the existence of mixed pixel [1]. Soft Classifiers are widely used to handle mixed pixel. Fuzzy Classifiers are based on the idea of fuzzy set logic by introducing a degree of vagueness or fuzziness with membership function [2]. Several studies have been done on various fuzzy-based classifiers. Fuzzy classifiers are effective only in classifying the data by linear boundaries, and in order to extend the functionalities by nonlinear boundaries composite kernels are used. The major objective is to develop an objective function for kernel-based Noise classifier to handle nonlinear class separation by selecting composite Spectral-KMOD kernel. Supervised Noise Clustering approach has been taken as the base fuzzy classifier. The noise clustering algorithm was introduced by R. N. Dave, as an alternative to overcome the sensitivity of FCM (Fuzzy C mean) and PCM (Possibilistic C mean) algorithm to noisy data, resulting to be more robust classifier [3, 4]. It stressed the concept of having a separate cluster (noise cluster) into which, all the noisy data points/outlier in the data may be fling out.

A kernel function maps data from original input feature space to a higher dimensional feature space where the problem of nonlinearity can be resolved and can perform analysis without further information from the original data set (illustrated in Figs. 1a, b, and 2a, b) [4–6].

The composite kernel concept is introduced to merge the efficiency of two different kernel functions. The composite kernel function is formed by merging kernel function from two different kernel families like a global kernel and a local kernel or a local kernel or a spectral kernel [7]. The composite kernel function may demonstrate both improved classification accuracy as compared to primitive single kernel approach and it provides the flexibility to adjust between the influences of the kernels by including weight factor [8]. There are different methods for combining kernels such as stacked approach, direct summation kernel; weighted summation kernel and cross-information kernel [9]. The present study has followed weighted kernel summation approach as defined in Eq. (1.3) for input feature vector x_i and x_j . The weight

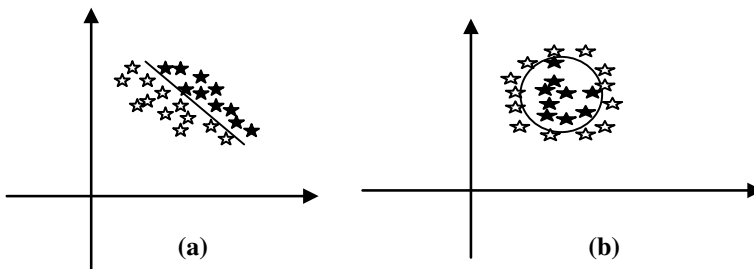


Fig. 1 Two classes in feature space. **a** Linearly separable classes. **b** Nonlinearly separable classes

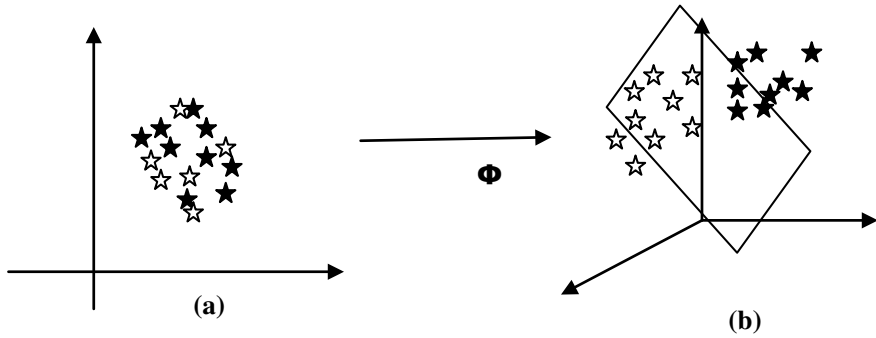


Fig. 2 Feature space transformation using kernel function. **a** Nonlinearly separable classes in input feature space. **b** Linearly separated classes in transformed kernel feature space

factor λ varies between (0, 1) and is optimized to get the best mixing between two kernels. Here, K_a and K_b are two different kernel functions that are used to form the composite kernel K . Kernel methods used in this research work are Spectral Kernel fused with KMOD Kernel.

The kernel function (K) implicitly computes the dot product between two vectors x and x_i in higher dimensional feature space without explicitly transforming x and x_i to that higher dimensional feature space, this technique is known as “Kernel trick” [10]. In Eq. (1.1), the feature map (φ) is the mapping function that nonlinearly maps the data to a higher dimensional feature space.

$$\Phi : R^p \rightarrow R^q \quad , \quad \text{where } p < q \tag{1.1}$$

$$K(\vec{x}, \vec{x}_i) = \Phi(x) \cdot \Phi(x_i) \tag{1.2}$$

$$K(\vec{x}_i, \vec{v}_j) = \lambda K_a(x_i, x_j) + (1 - \lambda) K_b(x_i, x_j) \tag{1.3}$$

1.1 KMOD—(Kernel with Moderate Decreasing)

KMOD is a type of local kernel and it is the distance-based kernel function introduced by Ayat et al. [5] as shown in Eq. (1.4). It shows a better result in classifying closely related datasets (highly correlated) and has shown better accuracy than Radial Basis Function (RBF) and polynomial kernel.

$$K(\vec{x}_i, \vec{v}_j) = e^{\left(\frac{\gamma}{\sigma^2 + \|x_i - v_j\|^2}\right)} - 1, \quad \text{where } \sigma, \gamma > 0 \tag{1.4}$$

The parameters γ and σ control the decreasing speed of the kernel function and the width of the kernel, respectively. In this study, the value of γ was taken to be one.

1.2 Spectral Kernel

The spectral kernel takes into consideration the spectral signature concept [11] as shown in Eq. (1.5). These kernels are based on the use of spectral angle (α) to measure the distance between the feature vector x and the mean vector of the class v_i . It is expressed as follows:

$$\alpha(x, v_i) = \arccos\left(\frac{(x \cdot v_i)}{\|x\| \|v_i\|}\right) \quad (1.5)$$

2 Study Area and Dataset Used

The dataset used in the research work has been acquired from Landsat-8 and Formosat-2 satellites [12, 13].

The site for the study work is situated in Haridwar district in the state of Uttarakhand, India. The area extends from $29^\circ 52' 49''$ N to $29^\circ 54' 2''$ N and $78^\circ 9' 43''$ E to $78^\circ 11' 25''$ E. The site is identified with five land cover classes (Fig. 3), i.e., Water, Wheat, Forest, Riverine Sand, Fallow Land.

3 Adopted Methodology

The methodology that has been adopted in order to achieve the desired objective has been shown diagrammatically through a block diagram in Fig. 4.

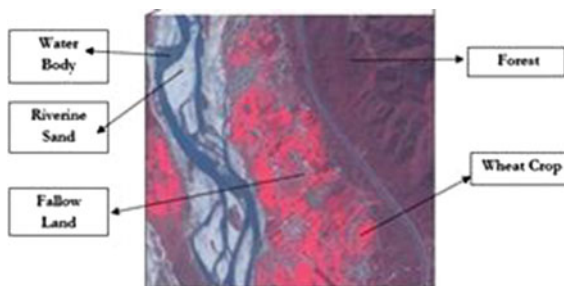


Fig. 3 Location of the area under study

Implementation has been done through a tool developed in Java programming. All the steps mentioned in Fig. 4 are briefly explained below.

(a) **Input Image and the training data**

Initially, the images of Formosat-2 and Landsat 8 were geometrically rectified and geo-registered. The training data or signature files for each class (water, wheat, forest, riverine sand, and fallow land) has been taken from simulated Formosat-2 image.

(b) **Developing the objective function for COMPOSITE-KNC**

The Composite Kernel-based NC (Noise Clustering) classifier was formed by replacing the Euclidean distance norm present in the Noise classifier with kernel metric. Spectral and KMOD Kernel have been considered for the composition.

(c) **Parameter Optimization**

The parameter estimation is one of the most important steps in the classification process. Choosing the optimal parameter guarantees the best results from the classifier. Here, the developed COMPOSITE-KNC classifier was executed on the available data set for different values of fuzzy parameter m , ranging between [1.1, 5.0] and the resolution parameter δ has been taken ranging from 1 to 10^9 . The optimal value of $\delta = 10,000$ and that of m was optimized in [2.7, 5.0] along with this composite kernel weight factor found to be at $\lambda = 0.9$. These values were selected on the basis of the accuracy assessment of classification.

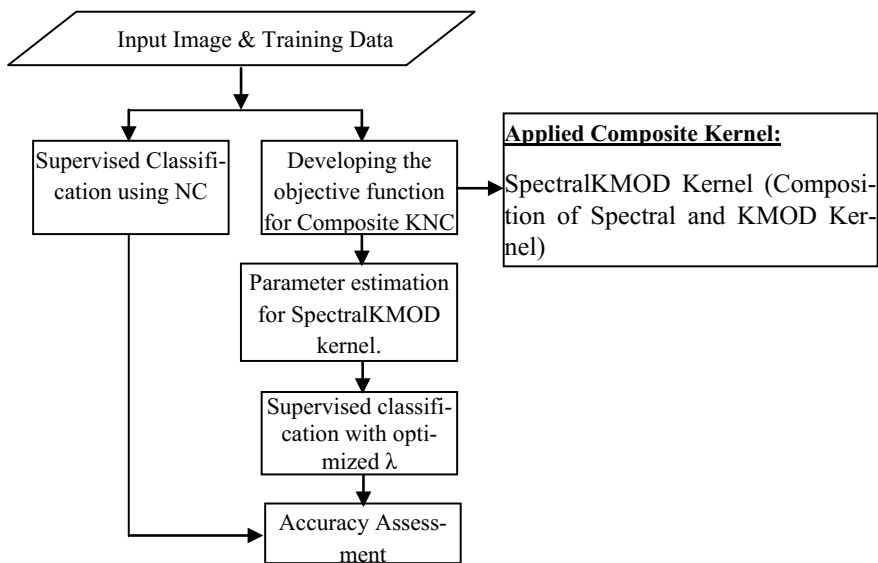


Fig. 4 Overview of methodology

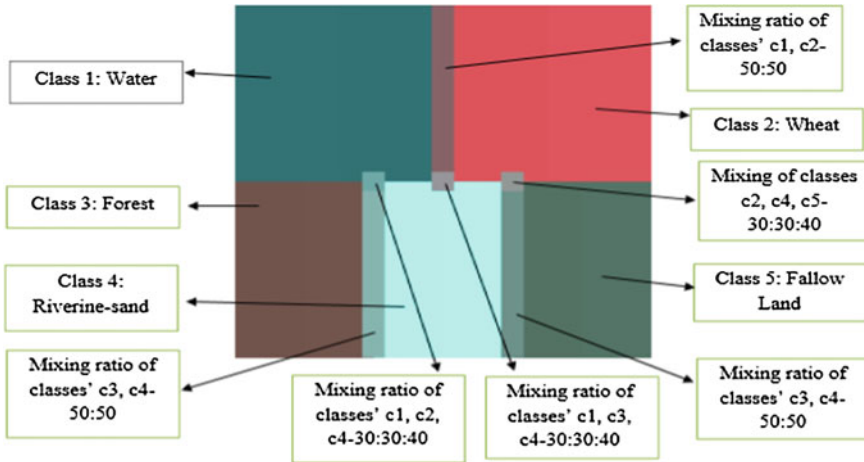


Fig. 5 Simulated image of Formosat-2 (class distribution)

The weight factor (λ) for the composite kernel was optimized using FERM within the range of [0.1, 0.9]. Class-wise user and producer accuracy has been shown in Table 1, against every value of λ . The class-based accuracy was analyzed for each. It was observed that maximum producer's accuracy in Water (95.16%) and Riverine Sand (96.81%) at $\lambda = 0.8$, Forest (96.78%), and Fallow Land (93.01%) at $\lambda = 0.9$, Wheat (96.19%) class were obtained for the composite kernel at $\lambda = 0.7$. The value of overall accuracy increases as the value of λ tends to 1 (95% at $\lambda = 0.9$), that is when the composite kernel is composed of the only spectral kernel.

(d) Supervised classification with Spectral-KMOD Kernel

The optimized parameter has been used for classification using Spectral-KMOD Noise Clustering Classifier without entropy.

(e) Accuracy Assessment

Evaluating the accuracy of classification has been done jointly with the simulated image technique illustrated in Fig. 5, and FERM (Fuzzy Error Matrix) method, to strengthen the findings of simulated image technique. It is a modification of the traditional error matrix for accuracy assessment of the soft classifier [2, 8].

4 Simulated Image Technique

The basis is of assigning the fuzzy membership values to feature vectors based on the distance measured from the mean vector of the classes (mean vector). The simulated image is generated based on the sample data for each class with the desired number

Table 1 The tabular representation of the user and producer accuracy using FERM of individual of Spectral-KMOD composite kernel corresponding to λ

Class-wise accuracy	FERM accuracy percentage (Spectral-KMOD composite kernel)									
	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.3$	$\lambda = 0.4$	$\lambda = 0.5$	$\lambda = 0.6$	$\lambda = 0.7$	$\lambda = 0.8$	$\lambda = 0.9$	
<i>Water</i>										
UA (%)	94.76	95.59	95.72	96.33	96.53	97.22	97.64	96.99	97.10	
PA (%)	93.24	94.03	94.14	94.69	94.18	94.50	94.90	95.16	95.01	
<i>Wheat</i>										
UA (%)	84.23	86.91	87.10	87.16	88.03	88.55	89.01	89.50	90.32	
PA (%)	93.35	93.09	93.17	94.35	95.29	94.89	96.19	95.81	96.01	
<i>Forest</i>										
UA (%)	94.59	94.61	94.50	94.72	94.70	94.59	95.19	94.86	94.56	
PA (%)	89.18	91.99	92.83	93.16	94.76	95.35	96.10	96.19	96.78	
<i>Riverine sand</i>										
UA (%)	87.14	89.05	89.97	91.11	92.85	93.45	94.99	96.20	96.65	
PA (%)	93.44	95.43	96.15	95.52	95.01	96.11	96.21	96.81	95.60	
<i>Fallow land</i>										
UA (%)	96.35	96.75	96.92	96.96	97.17	97.47	97.43	97.76	97.95	
PA (%)	88.36	88.46	88.06	88.69	89.64	89.89	90.66	91.18	93.01	
Avg. UA (%)	91.41	92.58	92.84	93.26	93.85	94.26	94.85	95.06	95.31	
Avg. PA (%)	91.51	92.60	92.87	93.28	93.78	94.15	94.81	95.03	95.28	
Overall accuracy (%)	91.21	92.35	92.61	93.06	93.65	94.04	94.70	94.94	95.27	

UA user accuracy, PA producer accuracy

of bands. With the simulated image, it is easy to compare the outcome of the classifier with the expected known input at a particular location. Also, it makes easy to identify the behavior of classifier with the mixed pixels. The mixed pixels can be simulated with varying proportions of different classes.

In this study, the simulated image of multi-spectral data of Formosat-2 (4 bands) has been taken to study the performances of all the Kernels. In this simulated image, we have intentionally mixed classes in a specific ratio and also have created an intra-class variation. Based on these controlled conditions the ability to handle the mixed pixel problem and detecting the intra-class pixel value variation was tested on the simulated image, shown in Fig. 5.

The mixed pixels were simulated with two variations, one with the composition of 50:50 [shown as Mixed Pixel (50:50)] between two different classes and other with the composition of 30:30:40 [(shown as Mixed Pixel (30:30:40)] in Fig. 5. The optimal range for the fuzzy parameter was obtained in the close interval of [2.7, 5.0].

5 Composite Kernel-Based Noise Classifier (COMPOSITE-KNC)

The KNC classifier is formed by using kernel methods with Noise Classifier. It is expected to handle nonlinearity in the data by the implementation of kernel methods.

The objective function of Noise classifier without entropy is mentioned below:

$$J_{\text{NC}}(U, V) = \sum_{i=1}^N \sum_{j=1}^C (u_{ij})^m D(\vec{x}_i, \vec{v}_j) + \sum_{l=1}^N (u_{i,c+1})^m \delta \quad (5.1)$$

C is the number of classes, N is the total number of pixels in the image, m is the fuzzification factor, u_{ij} represent the membership value of i th pixel in the j th class, $u_{i,c+1}$ represents the membership values of the noise class, v_j is the mean value (cluster center) of the j th class, x_i is the vector value of the i th pixel, D is the Euclidean distance between \vec{x}_i and \vec{v}_j and δ is a positive constant called the Noise distance.

In KNC the kernel metric is used to compute the distance between the cluster prototype (the mean value of the cluster) and the feature vector (pixel). This distance can be calculated in a kernel higher dimension feature space without actual transformation of the feature vector to that higher dimensional feature space.

The distance between two vectors in higher dimensional feature space can be expressed as

$$D(\vec{x}_k, \vec{v}_i) = \|\varphi(x_k) - \varphi(v_i)\| \quad (5.2)$$

In the higher dimensional feature space, the KNC objective function and the membership function μ_{ij} can be expressed as in Eqs. (5.3) and (5.4) respectively.

The objective function for Kernel-Based Noise Classifier (KNC) is derived as

$$J_{\text{COMPOSITE-KNC}}(U, V) = \sum_{i=1}^c \sum_{k=1}^n u_{ki}^m \left\| \varphi(x_k) - \varphi(v_i) \right\|^2 + \sum_{k=1}^n u_{k,c+1}^m \delta \quad (5.3)$$

Here, Membership Values

$$u_{ij} = \left[\sum_{j=1}^c \left(\frac{\|\varphi(x_k) - \varphi(v_i)\|}{\|\varphi(x_k) - \varphi(v_j)\|} \right)^{\frac{2}{m-1}} + \left(\frac{\|\varphi(x_k) - \varphi(v_i)\|^2}{\delta} \right)^{\frac{1}{m-1}} \right]^{-1}, \quad 1 \leq i \leq c \quad (5.4)$$

$$u_{k,c+1} = \left[\sum_{j=1}^c \left(\frac{\delta}{\|\varphi(x_k) - \varphi(v_i)\|^2} \right)^{\frac{1}{m-1}} + 1 \right], \quad 1 < i < c \quad (5.5)$$

Instead of Euclidean Distance, D , as shown in Eq. (5.1) the mapping function has been replaced by kernel function as

$$K_{\text{Composite}}(\vec{x}_i, \vec{v}_j) = \lambda K_{\text{Spectral}}(x_i, x_j) + (1 - \lambda) K_{\text{KMOD}}(x_i, x_j) \quad (5.6)$$

$$D(\vec{x}_k, \vec{v}_i) = \|\Phi(x_k) - \phi(v_i)\| = K_{\text{Composite}}(x_k, v_i) \quad (5.7)$$

Thus, composite-KNC objective function has been generated by replacing the Euclidean distance metric by kernel distance metric in the NC objective function.

6 Accuracy Assessment

In this study, FERM has been applied upon Landsat 8 data for classified output and Formosat-2 data has been taken as a reference map, and in both the data kernel-based noise classifier has been applied. Simulated Image technique is incorporated for evaluating the accuracy of classification, to estimate the results from NC and Composite-KNC classifier.

Table 2 illustrates the overall accuracy computed composite Spectral-KMOD kernel with varying m and λ , with the highest accuracy of 95.26% at $m = 5.0$ and $\lambda = 0.9$. Similarly, Table 1 demonstrates the class-wise accuracy including user and producer accuracy as well as overall accuracy of Spectral-KMOD-based Noise Classifier. A comparative study between the accuracy results of both Spectral-KMOD-based Noise Classifier and of Noise Clustering Classifier with Euclidean distance has been done, Table 3 shows the percentages calculated corresponding to varying m . The results clearly marked out the incorporation of the composite kernel have shown better results.

Table 2 The tabular representation of overall accuracy using FERM of Spectral-KMOD Kernel

<i>m</i>	λ								
	0.1 (%)	0.2 (%)	0.3 (%)	0.4 (%)	0.5 (%)	0.6 (%)	0.7 (%)	0.8 (%)	0.9 (%)
1.5	74.22	75.15	76.02	75.71	75.69	77.18	78.31	76.62	78.16
2	74.90	77.64	78.56	79.41	80.31	80.69	82.35	81.81	82.92
2.5	80.52	81.50	83.32	83.32	84.61	84.67	86.07	86.66	87.82
3	83.55	84.86	85.71	87.00	87.30	88.39	89.23	89.67	90.33
3.5	86.21	87.73	88.93	89.39	89.81	90.35	91.28	91.96	92.46
4	88.83	89.26	90.01	90.82	91.48	92.04	92.63	93.21	93.64
4.5	90.24	90.77	91.38	92.11	92.67	93.26	93.80	94.22	94.70
5	91.21	92.35	92.61	93.06	93.65	94.04	94.70	94.94	95.27

Table 3 Tabular representation of the best two kernels along with noise classification with Euclidean distance at varying *m*

<i>m</i>	NC (Euclidean distance) (%)	Spectral-KMOD ($\lambda = 0.9$) (%)
1.5	0.08181	78.15966
2	1.01052	82.92201
2.5	4.08068	87.82319
3	11.75677	90.32541
3.5	21.07052	92.45899
4	31.09292	93.64158
4.5	40.08072	94.69621
5	47.74982	95.26698

To identify the advantage of KNC classifier over NC classifier, the classified outputs of KNC classifier were compared with outputs from NC classifier.

7 Conclusion

The objective of this work was to develop a composite kernel-based Noise clustering classifier without entropy classifier (NC) for handling the non-linear data. The incorporation of kernel function into NC involves replacing the Euclidean distance measure in NC to kernel distance measure that takes care of the complicated non-linear boundaries between the classes. Fusion of Spectral Kernel with KMOD Kernel (Local Kernel) has been undertaken to assess the results and to perform a comparative analysis with Noise Classifier with Euclidean distance and have proven that fusing kernel methods with traditional classifier can lead to increase in performance of classification. To access the accuracy of classification Fuzzy Error Matrix has been

adopted, the output of this classifier shows an overall accuracy of 95.27%. Therefore, it is evident that SpectralKernel classifier has shown stable accuracy results.

References

1. Bezdek, J.C., Ehrlich, R., Full, W.: FCM: the fuzzy *c*-means clustering algorithm. *Comput. Geosci.* **10**(2–3), 191–203 (1984)
2. Upadhyay, P., Ghosh, S.K., Kumar, A: A brief review of fuzzy soft classification and assessment of accuracy methods for identification of single land cover. *Stud. Surv. Mapp. Sci. (SSMS)*, **2**(Mlc), 1–13 (2014)
3. Davé, R., Sen, S.: Noise clustering algorithm revisited. In: Annual Meeting of the North American Fuzzy Information Processing Society, 1997. NAFIPS'97, pp. 199–204. IEEE, New York (1997)
4. Choti wattana, W.: Noise clustering algorithm based on kernel method. In: Advance Computing Conference. IACC 2009, pp. 56–60. IEEE, New York (2009)
5. Ayat, N.E., Cheriet, M., Remaki, L., Suen, C.Y: KMOD-a new support vector machine kernel with moderate decreasing. In: Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 1215–1219 (2001)
6. Gu, Y., Jocelyn, C., Jia, X., Benediktsson, J.A.: Multiple kernel learning for hyperspectral image classification: a review. *IEEE Trans. Geosci. Remote. Sens.* (2017)
7. Vidnerova, P., Neruda, R.: Evolving sum and composite kernel functions for regularization networks. In: Adaptive and Natural Computing Algorithms, pp. 180–189 (2011)
8. Byju, A.P.: Non-Linear Separation of classes using a Kernel based Fuzzy *c*-Means (KFCM) Approach. M.Sc. Thesis, ITC, University of Twente, The Netherlands (2015)
9. Camps-Valls, G., Gomez-Chova, L., Mñnoz-Marí, J., Vila-Francés, J., Calpe-Maravilla, J.: Composite kernels for hyperspectral image classification. *IEEE Geosci. Remote Sens. Lett.* (2005)
10. Mittal, D., Tripathy, B.K: Efficiency analysis of kernel functions in uncertainty based *C*-means algorithms. In: 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 807–813 (2015)
11. Li, W., Ong, K.L., Ng, W.K., Sun, A.: Spectral kernels for classification. In: International Conference on Data Warehousing and Knowledge Discovery, DAWAK 2005, pp. 520–529 (2015)
12. Landsat-8 Specifications. <https://landsat.usgs.gov/landsat-8>
13. Formosat-2 Specification. <https://earth.esa.int/web/eoportal/satellite-missions/f/formosat-2>

Multilevel Steganography for Data Protection



Sourabh Tyagi, Priyanka Anurag and Smitha N. Pai

Abstract Nowadays, the world is changing very fast. In this world, there are many technologies that are being developed to protect the data from the outer world. Here the outer world is referred to a hacker or the third man who is in between one end user and the other end user. There is many software's which protects and transfers the data from the one machine to another machine or from one user to the other user, as in mobile phones, laptop, IoT devices, cloud, etc. There are also many algorithms and techniques, which protect data from unauthorized accesses. One of the techniques is called steganography. Two levels of data protection are carried out in the current proposed model. One is image steganography and another one is video steganography. Here, the text is encrypted by using digital encryption system (DES). The encrypted data is embedded in the image by using the LSB technique. The stegano image is converted to the base 16 in text format and embedded on to the video by using the frame converting technique. This process is called the multilevel steganography where the data is processed by two steganography techniques, i.e., image steganography and video steganography. This increases the security of data and the data is easily transmitted from one machine to the other machine providing multilevel security.

Keywords Encryption system · Base 16 · Security · Data · Image steganography · Video steganography · LSB

S. Tyagi (✉) · P. Anurag · S. N. Pai
Department of Information and Communication Technology, Manipal Institute
of Technology, MAHE, Manipal University, Manipal, Karnataka, India
e-mail: tyagi.sourabh77@gmail.com

P. Anurag
e-mail: priyankaraou432@gmail.com

S. N. Pai
e-mail: smitha.pai@manipal.edu

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_41

1 Introduction

The technology has revolutionized the way we look at things these days. Data is abundantly sent and received in large amount. Protection of data is of immense importance. A large number of software are being developed to mitigate the problems of phishing, manipulation, denial of service, etc. Steganography is one such technique which involves hiding the data in an image, video, audio file, etc. The Greek word steganography where “stegos” means to cover and “graphy” means writing together is called steganography. Steganography is the technique of hiding message from unauthorized access during transmission between transmitter and receiver [1]. Hiding the data has taken an important step towards transferring the multimedia contents in safe side, where data is transferred securely and secretly. Today there are many tools which are available in the market to encrypt and decrypt the steganography method. So using all the available tools in the market is not safe to use and sending the data by using them. “So there should be such a system which uses the same method but having slight change so that it will be different from the existing system.” New system with additional security is developed, which provide more security to data. By using base papers as the existing system, a new system can be generated. There are many software which can read data from the image or the video steganography by simply reading the LSB or MSB the data can be easily determined. The steganography plays an important role in sending the data by embedding data into video, audio, or the image file, which cannot be determined by the human eyes easily. Hiding secret message in video clip or file is called as video steganography. There are many techniques involved in the video steganography, i.e., Least Significant Bit (LSB) and Most Significant Bit (MSB). In LSB, the last bits are changed according to message, whereas in MSB last bit is changed. As compared to LSB and MSB, the LSB is more efficient for data hiding. In MSB, the large amount is changed in the pixel can be observed. In LSB only a single bit is changed. So in this proposed system, the LSB is used as to hide data. There lots of complexity is involved in the video steganography as compared to the image and audio steganography. This enhances the security provided to message [2]. The main advantage of the video steganography is that the human cannot identify the slight change in the videos color during playing it. An image is a collection of pixels of 8-bit gray scale and 24-bit color pixel that is represented in a uniform grid with different combinations of colors, i.e., pixels are displayed horizontally. The color schema counter is the smallest part, which is called the depth bit and is of 8 bits which is 1 pixel. The color of each pixel is present in 8 bits; each pixel is displaying 256 different shades of and color [3]. The 24-bit color format with Red, Green, Blue (RGB) each representing 8 bits is used. Over 6 million combinations of colors are used. The base 16 format or text format uses the different character and symbol using the true RGB color format. Binary value from a single pixel value is converted to ASCII value. This value is the index value used to convert to the base 16 format. Images are converted to characters. In this proposed system, the image and video steganography is used. In the early reviewed paper, the data is encrypted and embedded into the video or the image. In this proposed system, there

is a link between the image and video steganography how the data is embedded into the video from the image by converting into the text format. This process enhances the security of data by twofold. And data is more secure as compared to existing system in the market.

2 Literature Survey

Various frameworks have been proposed in the area of steganography, which are combination of various techniques like audio, video, and audio also exist. DES by Rijndael [4] algorithms provides more security than Advance Encryption Standard (AES) because of the complexity in structure. It is an encryption and decryption of plain text file, which creates one of the encrypted files; the encrypted file will be decrypted and converted into the plain text. One means of ensuring data security is use of cryptography. The cipher which uses a secret key for both encryption, as well as decryption, makes the data or the message more secure. In this proposed method the data is encrypted by using Discrete Cosine Transform (DCT), i.e., data is copied into the image by using LSB algorithm [5]. Image stenography algorithm, which uses a number of techniques to hide secret data or message in the image cover, is called stereo-image. The utilized methods are encryption with Discrete Cosine Transform (DCT) and data substitution by Least Significant Bit (LSB). Securing the image during transmission from one machine to other machine was the one of the problems which is address in this proposed paper. In this method, the image is encrypted by using DCT algorithm and use the hash file with secret key for securing the image. In the next step, the output is embedded in the image. Further, the image is converted to a text file by using base 64 encoding and it is sent through several SMS messages. As per the rule of GSM03.40 [6] standard, 16 characters is the length of the exchanged message. The message is a combination of letters, digits and further converted into a binary message called the picture. The Base 64 encoding is to convert image files to a text file. Base 64 encoding encodes every three bytes of binary data to four English character and SMS also can send non-text binary data. With the use of same binary messages, pictures can be also sent which is named as SMS Picture Message [7]. The other method to hide data is in the video file. This process can be carried out by embedding text file or any data in video, such a way that the video does not lose its originality. This uses the LSB method. This proposed method strives for high security of data without being able to detect hidden information [8]. There are many different approaches that have been made by different authors. In every approach, the authors are embedding text in video files or into image and some are converting the data in the random noise. In the previous approaches, there is no connectivity in image and video file. In this proposed system there is a link between the image and video.

3 Methodology

In this proposed system, steganography uses more than one level of steganography processing for securing the data. Currently in this method the image and video steganography are used. Figure 1 shows the entire process involved in this proposed system. Rijndael or Data Encryption Standard (DES) algorithm [9] is used for providing more security to data as compared to Advance Encryption Standard (AES), because they are more computationally intensive and less time taking for execution. In the first stage, the data is encrypted Fig. 2 shows the plain text encrypted using the DES algorithm [9] and converted into cipher text. Note that it will not decrypt the data if there is wrong key on the other side, which is used during the encryption of data. The encryption process of data and the key for the first stage shown in Fig. 3 [10] is the same key used on both the sides [9] The DES algorithm follows the following steps:

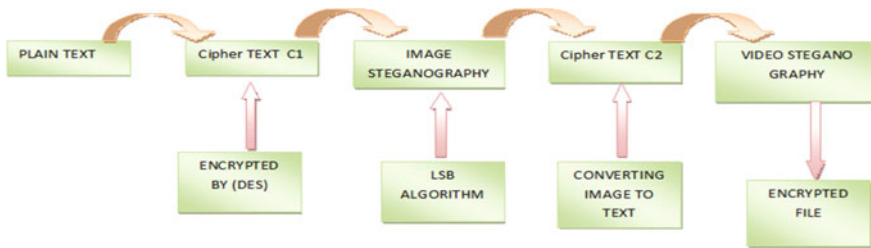


Fig. 1 Steps of encryption

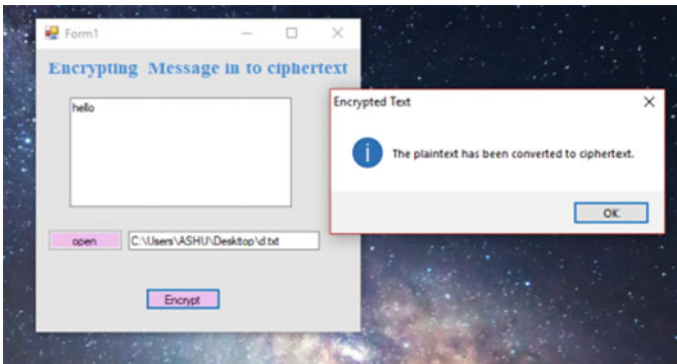
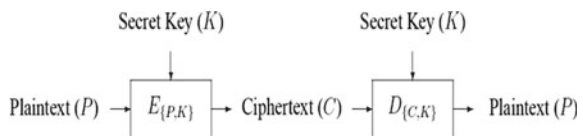


Fig. 2 Encryption of text

Fig. 3 Diagram of encrypting and [10]



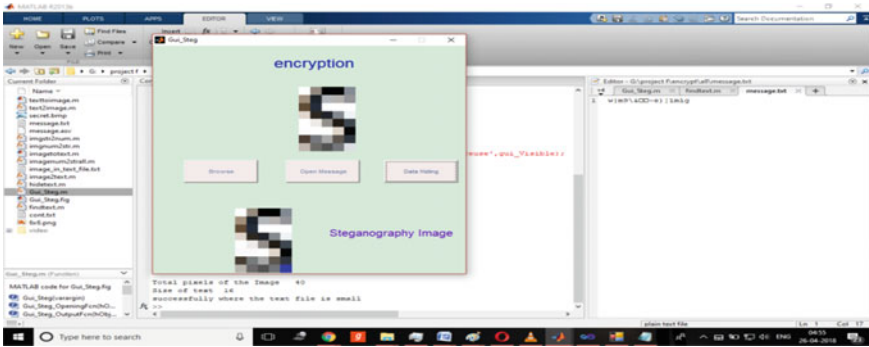


Fig. 6 Hiding cipher text in image

Fig. 7 Original image

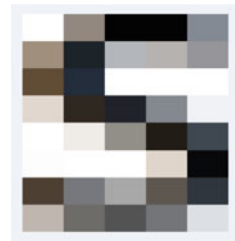
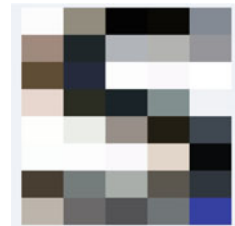


Fig. 8 Stegano-image



The process of hiding the cipher text onto the image is shown in Fig. 6. The resulting file obtained in stage two after applying the LSB algorithm is of higher size than that obtained from the stage one (using the DES algorithm). The number of pixels in the image should be larger than the size of the encrypted message obtained from stage one. So that we required 1 byte of pixel to hide 1 bit of data. There will be insignificant differences between the original and the stegano-image, Fig. 7, shows the original image as carrier of data and Fig. 8 shows the stegano-image where data is hidden.

By this process, the cipher text is embedded into the image. After embedding text into the image the third stage comes where the conversion of image to text come into picture, stegano-image is converted into text format. The conversation of image to text where each pixel of image is read. From each pixel, the least significant bit is retrieve and further converted into 8-bit character.



Fig. 9 Diagram of image-to-text conversation

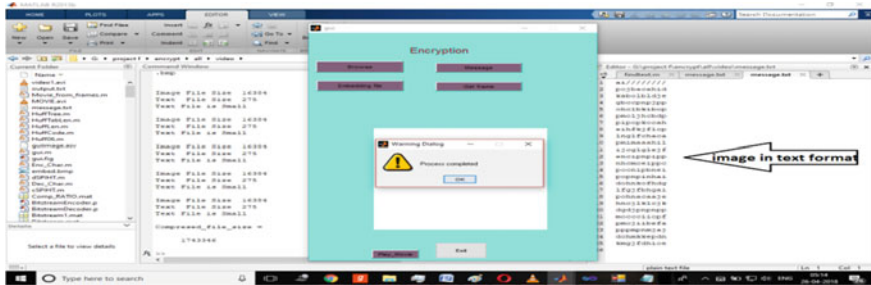


Fig. 10 Text in video file

There are many encoding methods by which the image is converted to the text format, but here only one of the methods is used to convert the image to text format. In this encoding method, the uppercase letters lower case letters, digits, and other special characters are used. This method is basically used by the RAiO. The RA8835A [11] graphics display which is also used to store text character code and bitmap in external frame of buffer memory. The process and steps of image is converted to text format, shown in Fig. 9.

In this method, the image is first broken down into each pixel, that will converted into pixel value. After converting to pixel value, it is converted to string. Third, String value is converted to base 16 string. The string is further divided into their largest integer value less than or equal to number of pixel value. Replacement will take place when the number of each pixel value is replaced with two characters. The basic idea is to convert from string value to character. After converting the image into the text format the fourth stage comes where text is embedded into the video. The working of hiding text in the video is simply replacing the LSBs of each frame value of the image with each bits of message. Video file is simply converting into the frames, each frames are considered as a image. The text will be embedded into each image of video by using the LSB method, which is discussed in image steganography process. Next steps in encrypting or hiding the text into the fragmented images of video are shown in Fig. 10. Where the encrypted image is embedded into the video file. The roles of different blocks are used in this proposed system are explained as follows:

- **Video File:** AVI (Audio Visual Interleaved) format video, which is used for covering the message.
- **Text File:** It is the converted image in text format.
- **Frame Conversion:** Convert video frames to <filename> .bmp format.
- **Text Compression:** Convert the text file into the bytes format.

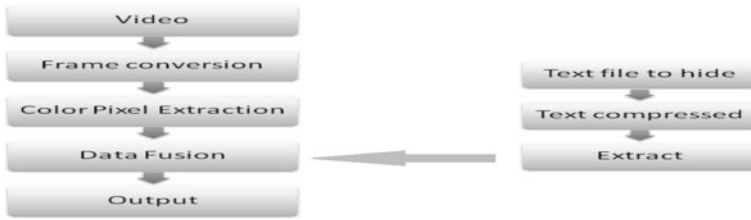


Fig. 11 Diagram of video steganography methodology

- **Colore Pixel Extraction:** In this section, the color value is calculated in form of pixel RGB format (0...255, 0...255, 0...255).
- **Output Video:** Output is generated to transmit over the network.
- **Comparison Ratio:** Compiling of a text file and video file.

In Fig. 11, the entire working of video steganography algorithm is explained in the form of block diagram. The two parameters are of importance, imperceptibility and capacity. Imperceptibility encompasses embedding data, which is not visible and cannot be analyzed. Capacity is total amount of data that is essential. Imperceptibility compares the original video frame with the stegano frames of video. The difference can determine by one of the parameters called Data Quality (DQ). DQ consists of two parameters; one is Peak Signal-to-Noise Ratio (PSNR), and other is Mean Squared Error (MSE).

- Data quality
- Measurement of the Accuracy of frames.

$$\text{Accuracy} = \frac{\text{Recovered_data}}{\text{Orginal_data}} \tag{1}$$

Mean Squared Error (MES)

$$\text{MES} = \frac{1}{H * W} \sum P((i, j) - S(i, j)) \tag{2}$$

H and *W* are Height and Width where $P_{(ij)}$ is original frame and $S_{(ij)}$ is stegano frame, respectively. *i* and *j* are the pixel index values in column wise and row wise, respectively.

Peak Signal-To-Noise Ratio (PSNR), It measures the Peak Signal-To-Noise ratio in two different images, where the PSNR is basically is used to determine the quality of the image. Higher the PSNR the higher will give the quality of image. The value calculated is calculated by

$$10 \log_{10} \frac{L^2}{\text{MES}} \tag{3}$$



Fig. 12 Text-to-image conversion

L is peak signal level for 8-bit gray scale image. The value used in this equation is 255.

During decryption of video, the LSB method is used to decrypt the text form the video file where the output will be the same as the input of stage three. The output will be in text format. The process of text is decrypted and again converted to image as shown in Fig. 12 In this, process two stages are triggered, one converts the text file to a specific value and the other converts the value to the pixel.

Finally, the image is converted into text format, where the conversation from text to image is simply the reverse process of converting the image to text. In the conversation of image to text, the cell of two characters is printed, where each cell will give the pixel value which is converted to the character. Similarly in the reverse process by getting character and converted to the pixel value. The pixel value is compared to RGB (Red, Green, Blue) value, where RGB is converted into image. On the other hand, the data is retrieved from the image by same algorithm called LSB. The cipher text which is retrieved from image as shown in Fig. 13, which is stored in the text file. The text file is converted into readable format decrypting cipher text by using DES algorithm, as shown in Fig. 14. Before decrypting the decryption

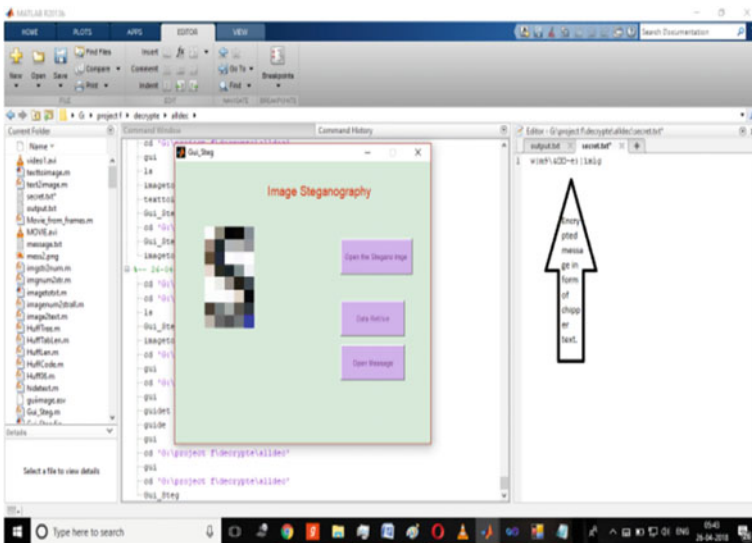
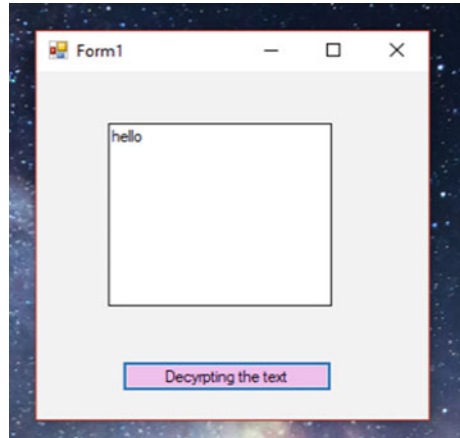


Fig. 13 Image to cipher text

Fig. 14 Readable message



key should be same as it is used for encryption. If the key is not same what is use for encryption, the data or the message will not be able to decrypted. Finally, the original message is retrieved from the cipher text.

4 Result Analysis

In this proposed system, it has been made some of the parameters as a constant. Video is fixed with *.avi format, size of the video is 62.8 MB (658, 93,888 bytes). After encryption of a message, the message will only embed into the 5 of the video frames. The maximum payload is 2.05 Kb can be embedded into these 5 frames. There should be minimum 5 frames, because for any video after decryption the video should be in running condition, i.e., to play video it is necessary to have at least five frames as per the results obtained in Tables 1 and 2.

In the bar in Fig. 15 it is noticed that as the payload is decreasing PSNR is increasing. It says about the quality of the video if there is 100% PSNR then the

Table 1 Results obtained from various payloads

S. No.	1	2	3	4
PSNR	52.4877	52.4875	57.7519	57.6761
Avg MSE	0.37	0.37	0.11	0.11
Payload	2.05 Kb	1.13 Kb	1 bytes	12 bytes
C.R	0.3758	0.3758	0.37593	0.37593
Compressed file size	1,943,478	1,768,222	1,743,324	1,743,326
No. of words	2049	1138	1	12

Table 2 Accuracy before sending and after sending

Analyses of image					
S. No.	Size of image	Memory of payload	Memory of image before sending	Memory of image after sending	No. of words
1	224	40	182	182	40
2	224	30	182	182	30
3	224	10	182	182	10
4	224	1	182	182	1

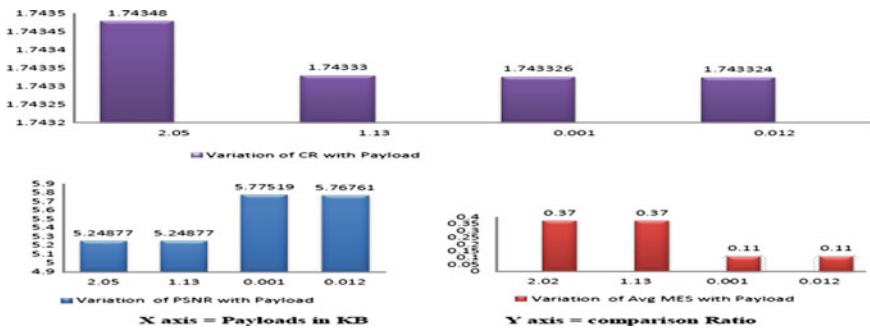


Fig. 15 Change in parameters as payload varies

quality of the image is very low. Coming to image sending and receiving, the accuracy is 100% at receiving end. There is no change in memory, dimension of an image as well as in quality.

5 Conclusion

Steganography is one of the secure methods of sending data technique where it sends data from one machine to the other machine. We have seen that in other method it is directly is embedded into the image or to the video. Which can be easily turned down by online available tools? In the current proposed method, data is encrypted and embedded in the image. Then the image is converted to the text format, the file with the text format is embedded onto the video. Converting the image to the text format provides one more level of the security. This makes data more secure and the safe to the end user.

In this proposed system, a new approach can be seen that image also be embedded into video file by converting into text format. To make it more secure for use the data or message is encrypted. Using this technique, the data is more secure over the communication network. Where in this proposed system the data is safe and secure

on each of the stages over the network. The data or the message cannot be retrieved by using any of the tools available online or in the market.

References

1. Rabah, K.: Steganography-the art of hiding data. *Inf. Technol. J.* **3**(3), 245–269 (2004)
2. Joseph, P., Vishnukumar, S.: A study on steganographic techniques. In: 2015 Global Conference on Communication Technologies (GCCT), pp. 206–210. IEEE, New York (2015)
3. Cheddad, A., Condell, J., Curran, K., Mc Kevitt, P.: Digital image steganography: survey and analysis of current methods. *Signal process.* **90**(3), 727–752 (2010)
4. Yi, G.: Notice of violation of IEEE publication principles research on a novel MAS for automated negotiation method. In: 3rd International Conference on Intelligent System and Knowledge Engineering, 2008. ISKE 2008, vol. 1, pp. 635–639. IEEE, New York (2008)
5. Jassim, F.A.: A novel steganography algorithm for hiding text in image using five modulus method. arXiv preprint [arXiv:1307.0642](https://arxiv.org/abs/1307.0642) (2013)
6. Shirali-Shahreza, M.H., Shirali-Shahreza, M.: An anti-SMS-spam using CAPTCHA. In: ISECS International Colloquium on Computing, Communication, Control, and Management, 2008. CCCM'08, vol. 2, pp. 318–321. IEEE, New York (2008)
7. Singh, A., Jauhari, U.: Data security by preprocessing the text with secret hiding. *Adv. Comput.* **3**(3), 63 (2012)
8. Al-Afandy, K.A., Faragallah, O.S., ElMhalawy, A., El-Rabaie, E.-S M., El-Banby, G.M.: High security data hiding using image cropping and lsb least significant bit steganography. In: 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), pp. 400–404. IEEE, New York (2016)
9. Haney, J.D.: The use of cryptography to create data file security: with the rijndael cipher block. *J. Comput. Sci. Colleges* **21**(3), 30–39 (2006)
10. [Online]. Available: <http://www.facweb.iitkgp.ernet.in/sourav/DES.pdf>
11. Zhuoran, Z., Daolian, X., Yunyun, S., Jikun, D., Lei, X.: The LCD interface design based on pic16f877 and ra8835. *Microcomput. Appl.* **17**, 010 (2011)

Feature Extraction of Normalized Colorectal Cancer Histopathology Images



Alok Kumar Jain and Shyam Lal

Abstract This paper presents different types of feature extraction of normalized colorectal cancer histopathology images. These highlights are exceptionally helpful for separating epithelium and stroma in colorectal cancer (CRC) histopathology images. It is also useful for selecting features and its analysis. In this paper, 27 features are extracted in which 5 are the visual texture features and 22 are the other features such as GLCM, run length and intensity-based features to separate epithelium from the stroma of Colorectal Cancer histopathology images. The utilized component has straightforwardly identified with the human recognition which makes it conceivable to distinguish the nearness of tissue based on parameters. The quantity of utilized highlights is little to differentiate the epithelium from the stroma of CRC histopathology images. In the simulation, we use well-defined and verified histopathology images of stroma and epithelium to correctly differentiate epithelium from stroma. The textural features measure provides the excellent result for 16 typical texture patterns. The issue emerges between the human vision, and modernized strategies that are experienced in this examination show the central point in dissecting of the surface. In which, some of them has removed by using better techniques. In conclusion, perception-based features work well in comparison to previously features used. Some modification like colour normalization of epithelium and stroma image and some of the new features are added because the classification of perception-based features is less.

Keywords Histopathology images · Perception-based feature · Textural features · Intensity-based features

A. K. Jain · S. Lal (✉)

Department of E&C Engineering, National Institute of Technology Karnataka,
Surathkal, Mangaluru 575025, India
e-mail: shyam.mtec@gmail.com

A. K. Jain

e-mail: jain.alok46@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_42

473

1 Introduction

Colorectal cancer (CRC) is a leading cause of death worldwide. Colorectal cancer is one of the most prevalent cancer types, tumour architecture changes during tumour progression and is related to patient prognosis [1]. The incidence of colorectal cancer has been increased more in several developing and developed Asian countries than the United States and other Western European countries [1]. Colorectal cancer (CRC) is an inside malignancy, and it is the development of cancer that is in the colon which is the part of the large intestine, i.e. located at the digestive lower tract's end. In the textural features, we used a perception-based feature to differentiate the epithelium and the stroma in the colorectal cancer histopathology images. CRC depends on the size and location of cancer, and it may be a malignant or benign or non-cancerous. Benign cancer does not spread in another part of the body, but a malignant tumour can diffuse into the other part of the body and damage them.

Epithelium is a type of animal tissue, along with the nervous, muscle and connective tissues and it is lining in the inner surface of the cavity or the outer surface of the organs and the blood vessels through the body. Epithelium layer does not contain any blood vessels and they get food from the diffusion of the substance from the connective tissue through the basement membranes. Stromal cells are the supportive tissue that consists of the connective tissue or the blood vessels or any part of the body organs. Stroma consists of the benign cell or non-malignant cell but they can provide an extracellular matrix on which tumour cells can grow. The size of the benign tumour is more in comparison to the malignant tumour. There are five perception-based features which are used to differentiate epithelium from stroma such as contrast, coarseness, roughness, directionality and line-likeness [2]. We are identifying the degree of correlation between all five perception-based features. There are two advantages of perception-based features (1) Based on the set of values, it is possible to access the tissue appearance by utilization of highlights straightforwardly identified with human perception (2) Numbers of features are very small, so the accuracy will be less.

In digital histopathology image recognition, edge detection and region growing analysis has used as a basis approaches for the analysis and feature extraction of the histopathology images. Tumour–stroma ratio (TSR) is used an independent prognostic factor for the determination of oncologic disease. To identifying the boundaries of the textural region, further segmented techniques are required. Texture edge detector techniques have used between the regions with different averaged grey level. Texture feature cannot be directly used for the region segmented problems. Indirectly, we can use this method for the measurement of the grey level in the feature process. For feature extraction, we choose a set of properties for identifying two or more textured region. Texture properties are used for region or sub-images, not for a point. In this situation, if the two regions differ only in contrast, scale, orientation or shape. Then, we detect the textured boundary between the two regions. In this paper, we have to use the features that are used to differentiate the epithelium from the stroma in CRC histopathology images. Further, we have to use those features which are

directly identified with the human recognition that makes the evaluation of tissues very simple, but the numbers of features are less to separate epithelium from the stroma.

The outline of this paper is as follows. Section 2 describes related work. Section 3 describes types of feature extracted of normalized colorectal cancer histopathology images. Section 4 gives simulation results and discussions to demonstrate the performance of different features. Finally, the conclusion is drawn in Sect. 5.

2 Background Details and Related Work

In the literature, many researchers put forward to discrimination between tumour epithelium and stroma by using various features. Bianconi et al., presented perception-based features for discrimination between tumour epithelium and stroma [2]. Kylberg and Sintorn presented local binary patterns (LBP) based features for medical image analysis because due to the high discrimination capability to differentiate epithelium from stroma, low computational cost and implementation is very easy [3, 4]. Tumura et al., propose five visual features that are related to the human perception and these features are used to separate epithelium from stroma in the array of colorectal cancer (CRC) such as contrast, roughness, directionality, coarseness and line-likeness [5]. Linder et al., uses local binary patterns (LBP) and contrast measures to differentiate epithelium from stroma [6]. This approach is given approximately 97% accuracy between the computerized approach and human observer. The disadvantage of this method is that LBP-based approach is quite difficult to interpret. Dalal et al. presented histograms of oriented gradients (HOGs) feature for human detection [7]. Haralick texture features have been used to differentiate epithelium from stroma [8], but some of the Haralick texture features are irrelevant and not directly related to the human perception, i.e. higher order statistics. So, the accuracy of the Haralick texture feature is less but the implementation of Haralick texture features is easy. Fouad et al., extract the colour features of the superpixels which are segmented by clustering algorithms [9]. The accuracy of this method is approximately 90% [9]. The disadvantage of this method is that it used the concept of neural network which is more difficult to interpret.

3 Types of Features

In this paper, 27 features are extracted in which 5 are the visual texture features and 22 are the other features such as GLCM, run length and intensity-based features to separate epithelium from stroma of colorectal cancer histopathology images.

A. Perception-based features

Tamura et al. proposed the image features which correspond to the visual perception [5]. Based on the experiment, Tamura et al. decided the five textual features such as contrast, roughness, directionality, coarseness and line-likeness which are related to the human perception. Practical application of these features is very complicated. The output range of these features can be different from one another, and these features can also be used for the classification of cancer, i.e. cancerous or non-cancerous. The range of all the above features is in the interval [0, 1]. For the classification of the histopathology images, all the five features share the same weight. Image coordinate system can be decided from the origin of the images. The upper left pixel of the image coordinate system determines q- and p-axis which is pointing rightwards and downwards, respectively.

1. Coarseness:

Coarseness depends upon the original size of the texture element [2, 10]. Higher the size of the texture element, coarser the surface and smaller the size of the texture element, coarseness will give very less value, i.e. not desirable or the value of the rudeness is near to zero.

Let us assume that $M_{a,v}$ and $M_{a,h}$ are the matrixes in the direction of vertical and horizontal axis for all a [11]. Now, we evaluate the value of a in every point that minimizes $M(p, q)$ value in both directions that is given below.

$$\bar{a}(p, q) = \arg \max\{M(p, q)_{a,h}, M(p, q)_{a,v}\} \quad (1)$$

Lastly, assess the coarseness value, i.e. the normal window measure in every point that maximizes the estimation of $M(p, q)$ in one of the direction. Mathematically, coarseness can be defined as

$$FV_{\text{crs}} = \frac{1}{2^a} \frac{1}{MN} \sum_{p=1}^M \sum_{q=1}^N 2^{\bar{a}(p,q)} \quad (2)$$

where M and N represents the input images dimension. The factor $\frac{1}{2^a}$ is used to normalize the coarseness in the range of interval [0, 1].

2. Directionality:

The concept of directionality is connected with the probability [2]. Directionality tells about the probability that the variation in the pixel values happen or occur in a predefined orientation. If the parallel lines form an image, it will give the strong directionality, but the randomly scattered points form an image, it will provide the weak directionality. For the evaluation of directionality, the gradient is defined as

$$\theta(p, q) = \arctan \left[\frac{G_q(p, q)}{G_p(p, q)} \right] \tag{3}$$

The value of θ and B_o is normalized in the angular interval of $\frac{2\pi}{B_o}$. Mathematically, directionality is defined as

$$FV_{dir} = \sqrt{\frac{B_o}{B_o - 1} \sum_{i=1}^{B_o} \left[h(\theta_i) - \frac{1}{B_o} \right]^2} \tag{4}$$

where $h(\theta_i)$ denotes the likelihood of every predefined orientation θ_i . The factor $\frac{B_o}{B_o-1}$ represents the distance and it is used to normalize in the range of interval $[0, 1]$. By observation, we set the value of $B_o = 12$ that is corresponds to the angular span of 15° [5].

3. Contrast:

Contrast can be defined as the ratio of standard deviation with respect to the kurtosis. Based on the proposed perception, features of Tamura et al., tells that the contrast is related to the histogram of the grey levels or the distribution of the grey level in the histogram, the period of repeating patterns and edge sharpness [2, 5, 7]. Mathematically, the contrast can be defined as

$$FV_{con} = 2 \frac{\sigma}{\alpha_4^n} \tag{5}$$

where σ and α_4 denotes the standard deviation and the kurtosis and n represent the positive number which is set to be $\frac{1}{4}$ by the observation [5] and the factor 2 is used to normalized the output value in the interval $[0, 1]$.

4. Line-likeness:

In line-likeness, let us consider that the image is recognized by the line [2]. The concept of line-likeness is related to the probability. The line-likeness tells about the likelihood that the angle or gradient keeps a similar course at each neighbouring pixel. If the direction is nearly constant, then the group of pixels near the direction acts as a line [5]. It can be accessed from the joint likelihood. If the two-gradient vector probability is higher and collinear, then the line-likeness will be higher and vice versa. Mathematically, line-likeness can be defined as

$$FV_{lin,l} = \frac{\sum_{i=1}^{B_o} \sum_{j=1}^{B_o} \left| \cos \left[(i - j) \frac{2\pi}{B_o} \right] \right| CM_{\theta,l}}{\sum_{i=1}^{B_o} \sum_{j=1}^{B_o} CM_{\theta,l}} \tag{6}$$

where $CM_{l,\theta}$ signifies the co-event grid of inclination course with the given relocation vector. Cosine in Eq. (6) indicates the weights. If the cosines take the value one, it

means the two gradients vectors are collinear but it takes the value zero, it means both the gradient vectors are orthogonal to each other. The average result of the circularly orientated line-likeness is given as follows.

$$FV_{\text{lin}} = \frac{1}{4} \sum_{l=1}^4 F_{\text{lin},l} \quad (7)$$

where $l_1(0, 2)$, $l_2(-l, l)$, $l_3(-l, 0)$, $l_4(-l, -l)$. In the experiment, set the value of $l = 4$ (by observation) [3]. In the Eq. (6), the modulus of the cosine indicates that the output is normalized to the interval [0, 1].

5. Roughness:

The concept of roughness is linked with the completing of a textual surface [2]. Roughness indicates that the ridges, protuberances, and valleys, i.e. opposed to smoothness which is free from the irregularities. The roughness of the image indicates that the rate of inconstancy and the degree of the intensity of the neighbouring pixels. According to the Tamura et al., roughness is related to the ‘root mean square height of the scale-limited surface’ $-s_q$. Mathematically, roughness can be defined as

$$FV_{\text{rgh}} = 2 \sqrt{\frac{1}{MN} \sum_{p=1}^M \sum_{q=1}^N [I(p, q) - C]^2} \quad (8)$$

where I denotes the source image and the source image is in greyscale only. C is the correction factor. The factor 2 is used to normalize the output in the interval [0, 1].

B. Other features

In this work, some of the features are added such as grey level co-occurrence matrix (GLCM) [11], intensity-based and run length features [12–15]. These features are added because the classification accuracy of the perception-based features is less that is why we have added some new features. In GLCM features [11], seven statistical features are extracted in the possible direction ($0^\circ, 45^\circ, 90^\circ, 135^\circ$), and these features are: Angular second moment (ASM), inverse difference moment (IDM), difference entropy (DEFF), entropy (E), maximum probability (MP), cluster prominence (CP) and cluster shade (SADE). In run length features, ten-second-order statistical features are extracted in every possible direction ($0^\circ, 45^\circ, 90^\circ, 135^\circ$). The run length features are: Short-run emphasis (SRE) and long-run emphasis (LRE), high grey level run (HGRE) and low grey level run emphasis (LGRE), grey level non-uniformity (GLN) and run length non-uniformity (RLN), short-run low grey level (SRLGE) and short-run high grey level emphasis (SRHGE), long run low grey level (LRLGE) and long run high grey level emphasis (LRHGE).

The five intensity-based features are extracted, and these features are interquartile range (IQR), median absolute deviation (MAD), standard deviation (SD), skewness (S) and kurtosis (K) [8]. Ten run length, five intensity-based and eight GLCM features are extracted for the three channels red, green and blue of RGB colour image. All the extracted features can be used for the classification of cancerous and non-cancerous or discriminate epithelium from the stroma in the CRC histopathology images. In all features, some of the features are redundant and irrelevant for the classification of cancerous and non-cancerous cells [16]. We use the principal component analysis (PCA) to select a subset of features that can be used for the classification of cells. A total of 22 masked feature values have been calculated for the 20 images of epithelium and stroma. Combining all these features with the perception-based feature, and the classification accuracy is high in comparison to perception-based features.

Another modification is colour normalization of the segmented image (epithelium and stroma). Colour normalization of the source can be done by transferring the mean colour of the target image into the source image. The source image is that which is going to be processed and target image is that which is suggested by the pathologist. After colour normalization, feature value has been extracted. The extracted feature value gave the better result in comparison to previous features values shown in Figs. 2 and 3. Normalized feature value has high classification accuracy in comparison to the non-normalized feature value that is why we are using the colour normalization.

4 Experimental Setup and Results

This section presents experimental results and discussion about various feature extracted from colorectal cancer histopathology images. In this paper, there are total of 27 features which are considered for analysis and there feature values are extracted from CRC histopathology images. The perception-based feature values of an epithelium and stroma tumour are given in Figs. 1, 2 and 3. Run length feature values, grey level co-occurrence matrix (GLCM) feature values and intensity-based feature values are given Tables 1, 2 and 3, respectively.

Table 1 Run length features value

Number of features	SRE	LRE	HGRE	LGRE	SRLGE	SRHGE	GLN	RNL	LRLGE	LRHGE
Epithelium without normalization	0.48	11.70	130.7	0.02	0.07	0.83	380.23	961.39	113.34	127.71
Stroma without normalization	0.52	12.62	111.5	0.03	0.08	0.93	347.7	5.83	128.64	136.49
Epithelium with normalization	0.49	525.5	187.9	0.14	0.077	0.89	413.6	43.28	147.46	164.42
Stroma with normalization	0.67	245.9	144.9	0.06	0.083	0.99	554.5	117.9	144.23	167.84

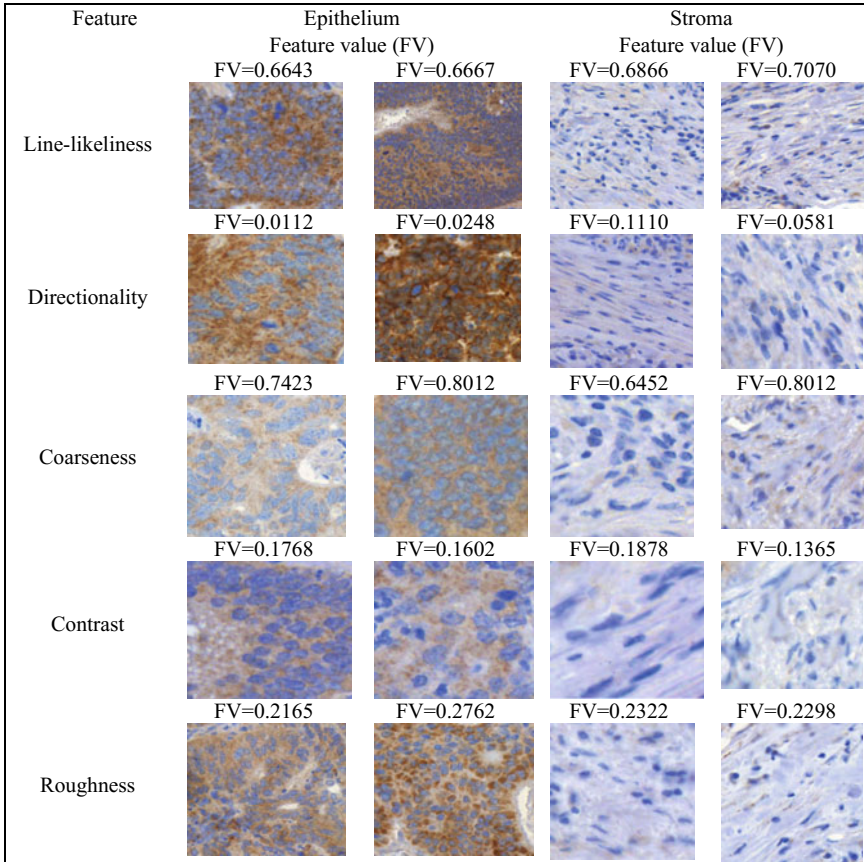


Fig. 1 Sample image of epithelium tumour and stroma tumour along with the feature values

Table 2 Grey level co-occurrence matrix (GLCM) features value

Number of features	ASM	IDM	DEFF	E	MP	CP	SHADE
Epithelium without normalization	11.25	0.70	2.40	1.20	0.29	3.24	0.88
Stroma without normalization	43.43	0.81	2.51	1.25	0.35	6.53	0.97
Epithelium with normalization	43.57	0.99	2.61	1.80	0.94	6.10	0.99
Stroma with normalization	57.40	0.83	3.13	1.82	0.94	7.53	0.98

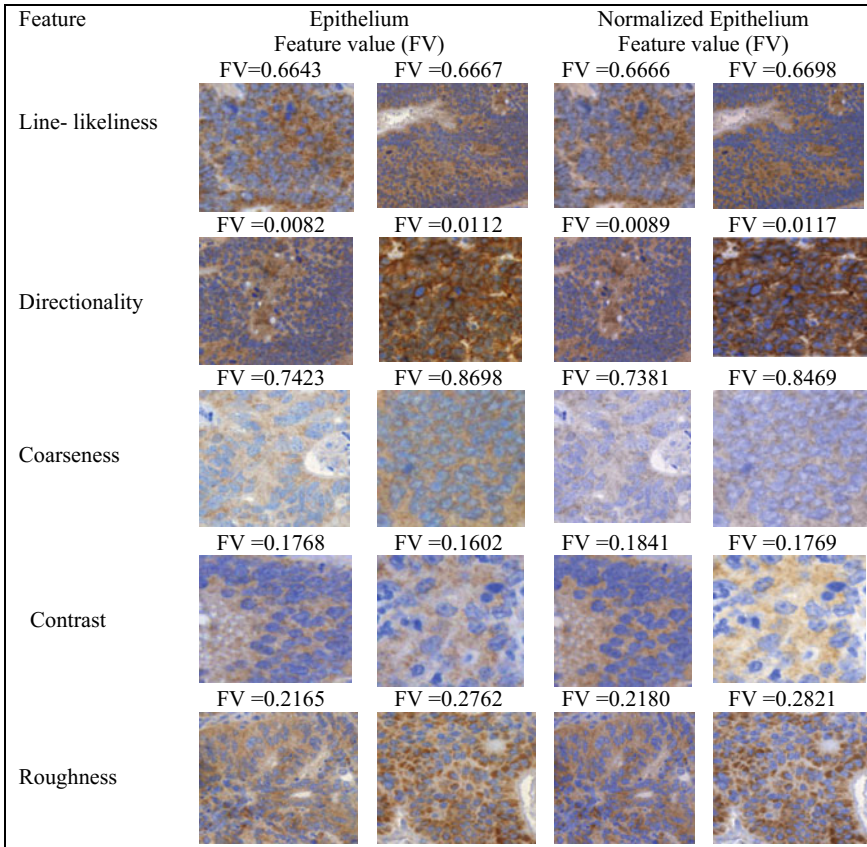


Fig. 2 Sample image of tumour epithelium and normalized tumour epithelium along with the corresponding feature values

Table 3 Intensity-based feature value

Number of features	IQR	MAD	SD	S	K
Epithelium without normalization	25.02	15.29	19.45	0.57	3.44
Stroma without normalization	30.21	19.85	25.81	0.87	3.95
Epithelium with normalization	38.23	28.41	23.67	0.71	3.61
Stroma with normalization	26.52	23.61	27.15	0.92	3.97

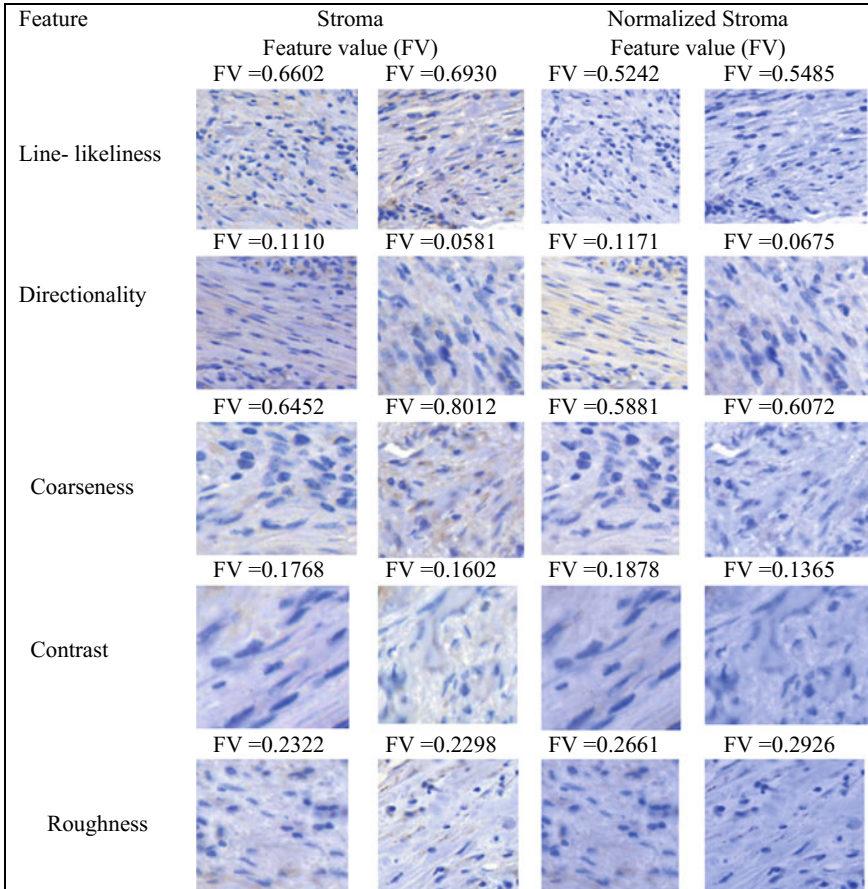


Fig. 3 Sample image of tumour stroma and normalized tumour stroma along with the feature values

A. Database

Colorectal cancer (CRC) database consists of 1332 image of epithelium and stroma tissue sample extracted from the patients and this database is available on the website <http://fimm.webmicroscope.net/supplements/epistroma> [17]. There are 820 images of tumour epithelium and 522 images of stroma tumour.

B. Discussions

Figure 1 shows that the perception-based feature values of an epithelium and stroma tumour. Figure 2 shows that the perception-based feature values of epithelium and normalized epithelium tumour. Here, the comparison between the features values of the epithelium tumour and the stroma tumour with the normalized feature value of an epithelium tumour and stroma tumour have been presented. In Fig. 1, the feature

value (FV) of the stroma is greater than the epithelium. It means perception-based features will probably happen in the stroma than the epithelium. By observation, in Fig. 1, we can say that the stroma has more tendencies to formed cluster in comparison to the epithelium. In line-likeness, the feature value of stroma is higher than the epithelium, and it means that the line-likeness will probably happen in stroma in comparison to the epithelium.

Similarly, directionality will probably happen in stroma in comparison to the epithelium. So, we can say that the perception-based features will probably happen in the stroma in comparison to the epithelium. Hence, the stroma has more tendencies to form a cluster than the epithelium but between the epithelium tumours, the feature value of the epithelium between the same classes is more isotropic than the stroma, and it does not show any orientation. In coarseness, the feature value is approximately the same for both a tumour (epithelium and stroma). So, coarseness has fewer tendencies to differentiate the epithelium from the stroma.

After epithelium and stroma extraction from the tissue, colour normalization of epithelium and stroma have been done by transferring the mean colour of target image into the source image by using complete colour normalization algorithm [18]. In Figs. 2 and 3, extract the feature value of normalized epithelium and stroma to compare with the original epithelium and stroma feature value. It is clear from the figures that the feature value of the normalized image of epithelium and stroma is more than the feature value of the original image of stroma and epithelium. So, the normalized image has high classification accuracy in comparison to the original image of stroma and epithelium.

From the above result, normalized feature value provides a better result in comparison to the non-normalized image. Instead of non-normalized feature value, we have to use the normalized feature value for the classification of cancerous and non-cancerous cells. The result of feature value is the average of 20 images in that we showed only 10 images of tumour epithelium and tumour stroma. In the figure, we showed 60 images in which 10 images of epithelium and stroma and other images are the normalized images of the 10 images of epithelium and stroma.

5 Conclusions

After colour normalization, features were extracted based on visual perception which is used to differentiate of epithelium from stroma. Here, we have to compute the perception-based features to differentiate the epithelium from the stroma of colorectal cancer histopathology image. Experimental results demonstrated that the feature value of stroma tumour is more than then the epithelium tumour. So, cluster formation is more in stroma than the epithelium but between the epithelium tumours, the feature value of epithelium is isotropic. Hence, it does not show any orientation. The perception-based features provided a better result in comparison to all the existing features. This method has two advantages; first, if the dimension of the features vector is very low, then the perception-based features work well. Second, the interpretation

is very easy by using perception based features in common sense. For this purpose, we have to use the publically available database. The database consists of epithelium and stroma image which is histological verified case of colorectal cancer. The result obtained from the perception-based features is used in another type of cancer is a conceivable heading for future research. The use of neural network and machine learning is the possible direction of the future research.

Acknowledgement This work is supported by a part of Grant of Young Faculty Research Fellowship under Visvesvaraya PhD Scheme for Electronics & IT from Digital India Corporation (formerly Media Lab Asia), A Research & Development Company of Ministry of Communications & Information Technology, Govt. of India.

References

1. Ferlay, J., et al.: GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]. International Agency for Research on Cancer, Lyon, France (2013)
2. Bianconi, F., et al.: Discrimination between tumour epithelium and stroma via perception-based features. *Neurocomputing* **154**, 119–126 (2015)
3. Nanni, L., Lumini, A., Brahnam, S.: Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intell. Med.* **49**, 117–125 (2010)
4. Kylberg, G., Sintorn, I.M.: Evaluation of noise robustness for local binary pattern descriptors in texture classification. *EURASIP J. Image Video Process.* **17** (2013)
5. Tamura, H., Mori, T., et al.: Textural features corresponding to visual perception. *IEEE Trans. Syst. Man Cybern.* **8**, 460–473 (1978)
6. Linder, N., Konsti, J., Turkki, R., et al.: Identification of tumor epithelium and stroma in tissue micro arrays using texture analysis. *Diagn. Pathol.* **7**(22), 1–11 (2012)
7. Dalal, N., et al.: Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), vol. 1, pp. 886–893 (2005)
8. Haralick, R.M., et al.: Statistical and structural approaches to texture. *Proc. IEEE.* **67**, 786–804 (1979)
9. Fouad, S., Randell, D., Galton, A., Mehanna, H., Landini, G.: Unsupervised superpixel-based segmentation of histopathological images with consensus clustering. In: Annual Conference on Medical Image Understanding and Analysis MIUA 2017: Medical Image Understanding and Analysis, pp. 767–779 (2017)
10. Demir, C., et al.: Automated cancer diagnosis based on histopathological images: a systematic survey. Technical Report, TR-05–09, pp. 1–16 (2009)
11. de Siquira, F.R., et al.: Multi-scale gray level occurrence matrices for texture description. *Neurocomputing* **120**, 336–345 (2013)
12. Hamilton, P.W., et al.: Automated location of dysplastic fields in colorectal histology using image texture analysis. *J. Pathol.* **182**, 68–75 (1997)
13. Weyn, B., et al.: Automated breast tumor diagnosis and grading based on wavelet chromatin texture description. *Cytometry* **33**, 32–40 (1998)
14. Irshad, H., et al.: Automated mitosis detection using texture, SIFT features and HMAX biologically inspired approach. *J. Pathol. Inform.* **1** (2013)
15. Gurcan, M.N., et al.: Histopathological image analysis: a review. *IEEE Rev. Biomed. Eng.* **2**, 147–168 (2009)
16. Ishikawa, M., et al.: Automatic quantification of morphological features for hepatic trabeculae analysis in stained liver specimens. *J. Med. Imaging* **3**(2), pp. 027502-(1–13) (2016)

17. Web microscope—web-based virtual microscopy, available online at www.webmicroscope.net/. Last accessed on 16 May 2014
18. Li, X., Plataniotis, K.N.: Complete color normalization approach to histopathology images using color cues computed from saturation weighted statistics. *IEEE Trans. Bio-med. Eng.* **62**(7), 1862–1873 (2015)

An Efficient License Plate Text Extraction Technique



Anuj Kumar, Anuj Sharma and R. K. Singla

Abstract In vehicle recognition, a machine-based system is required to recognize vehicles. To accomplish this, extraction of character regions are required to perform with high degree of accuracy in real-life images datasets. In this letter, we present an innovative framework for character regions extraction from the license plate of vehicles. Our framework includes needful preprocessing steps to remove noise in images. The necessary checks are applied to structural properties of regions to result in character regions as an outcome. The proposed algorithms of character region extraction work with intra-regions and inter regions-dependent structural properties of characters. Our developed framework achieves results at par with respect to results reported in the literature. The character regions extraction and recognition results are 97 and 95%, respectively. These results are for images where all the regions are correctly extracted and recognized.

Keywords Preprocessing · Characters extraction · Plate detection · Character recognition

1 Introduction

The machine-based object identification is ubiquitous in real world and analysis of machine recognition has been studied for more than two decades now. This field of research is now a promising area in image processing and computer vision. One application of real-world use is Automatic Vehicle License Plate Recognition (AVLPR)

A. Kumar (✉) · A. Sharma · R. K. Singla
Department of Computer Science and Applications, Panjab University,
Chandigarh, India
e-mail: anuj_gupta108@rediffmail.com

A. Sharma
e-mail: anujs@pu.ac.in

R. K. Singla
e-mail: rksingla@pu.ac.in

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_43

which requires a vision system to identify number printed on the license plate of vehicle. This automated system results in manual efforts reduction and faster processing. Although AVLPR has been investigated in literature, it always demands new efficient algorithms to meet current challenges subject to the respective real-life environment [1]. The variety and volume of data are other unavoidable issues that demand algorithms to meet state-of-the-art challenges. The major challenge in this field with respect to the availability of latest findings is tracing of the area of interest in image through machine. Besides tracing, the vision system should be capable of extracting subregions with high amount of accuracy.

2 Related Work

To design such vision system, we have discussed selective literature close to our work in this paper. Such type of vision system includes preprocessing, extraction of region of interest in image, and recognition as common phases [1]. The role of preprocessing in vision system has been performed, mainly in binary format understanding and de-noising the image. Also, median filter, Gaussian filter, and bottom hat filter are such techniques discussed in recent past [2–4]. The unwanted line estimation algorithm and histogram equalization are other preprocessing techniques employed in these vision systems [5, 6]. The extraction of license plate region offers multiple challenges as discussed in the literature [5]. The color-based feature is one such common method [7]. Recent work indicates the role of fuzzy techniques for localization of license plate area [8]. Techniques based on morphological operations and Harris corner method is successful examples in AVLPR [9, 10]. A thresholding method based on features of license plate further proved a suitable application to localize license plate [11]. Few other hybrid techniques are based on support vector machines [12]. The edge clustering and region-based method are other recent work in this direction [13, 14]. The localization of license plate further needs extraction of subregions where desired number plate characters are printed. The recent literature indicates connected component analysis and maximally stable extreme region are such suitable techniques. Jia et al. proposed a region-based method for license plate detection. This technique first found candidates regions using mean shift and then these candidate regions are analyzed and classified to decide whether a candidate region contains a license plate [15]. Discrete Wavelet Transform (DWT) based techniques are employed in [16] for plate detection. The recognition phase is followed by extraction where extracted subregions as characters are identified [10, 13]. The work in AVLPR with respect to recognition phase has been observed in recent past [8, 11, 17, 18]. Feature salience classifier is used in [19] for character recognition. A multi-style license plate recognition framework is proposed in [20] for Automatic Vehicle License Plate Detection. The AVLPR challenges vary with respect to spacing between characters on number plates as discussed in previous work [21]. We have studied state of the art in this area and propose an efficient robust and accurate

technique to meet the challenges of real-life scenarios. Our contributions include the following:

- (1) An innovative framework that integrates all three phases as preprocessing, extraction and recognition to accomplish real-life data handling.
- (2) A novel algorithm that extracts subregions with accuracy in multiple situations. The algorithm based on mathematical concepts that further prove to be suitable for other real-life computer vision based system.
- (3) An extensive analysis for in-house data set collected in the real-life environment has been performed successfully.

3 System Design

We present a framework of AVLPR as depicted in Fig. 1. The framework includes three phases. Phase I is preprocessing that perform size normalization, binarization, and noise removal of images. We have used state-of-the-art techniques as Gaussian filter in noise removal. The size normalization is based on aspect ratio and binarization is two steps process as gray scale and binary form conversion. Phase II is the extraction of license plate area and further extracting subregions as character regions. Here, we have employed our proposed algorithm based on structural features that include extraction of regions based on specific features. We have used these regions filtration based on its features inter- and intra-regions dependency. The details of this phase including algorithm as explained in following paragraphs. Phase III is recognition of identified regions based on similarity score. We have used statistical property as correlation in calculating the similarity score. There are other classifiers as SVMs, HMMs, and k-nn that have been avoided in the present study as our prime focus is the identification of correct character regions.

In Algorithm 1, step 1 input image as I and its preprocessing is performed in step 2. In preprocessing, size normalization changes image width and height subject to target aspect ratio. Binarization is done in two steps, the first step converts image to gray scale image using established formula as [22], second step convert gray scale to binary using mean value comparison as discussed in [23]. We have used Gaussian technique to remove noise as discussed in [24]. The regions are extracted in step 3 where binarized image finds connected components in neighborhood. An implementation of extracting regions has been explained in [25]. Step 4 make groups based on structural properties using common properties as corner coordinates of the image pattern, width, height, area, and perimeter as explained in [25]. Step 5 filter collected regions from step 4 that do not meet preset lower and upper bound criteria. This step helps to remove unwanted very small or large regions. Step 6 performs inter- and intra-dependent structural regions properties checks to find regions of interests. This step filters regions that are not likely to be character regions. This step is important as our desired extractions of regions are characters area in the license plate. For example, one region property dependence on neighboring regions

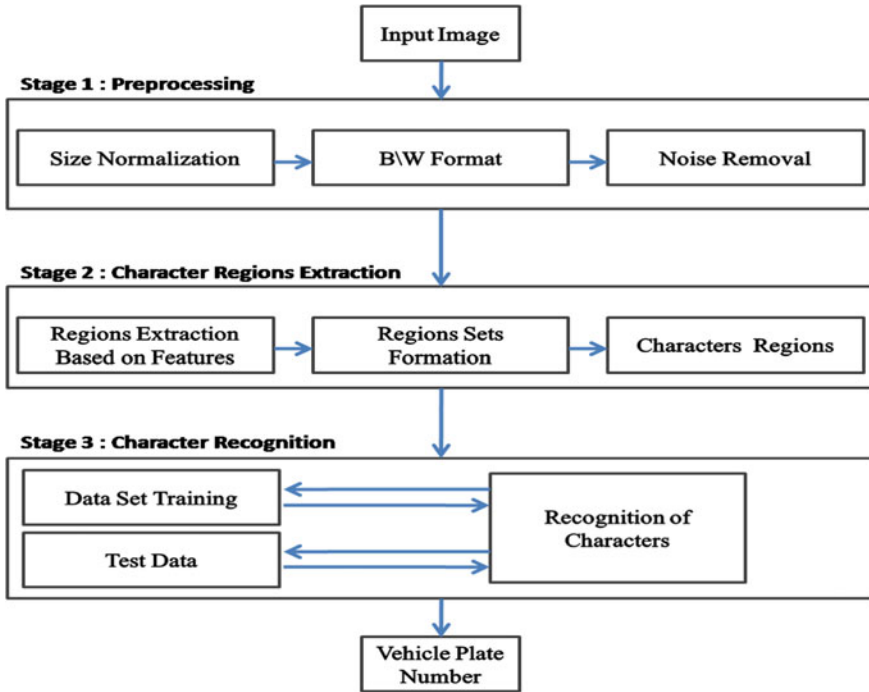


Fig. 1 Phases in AVLPR framework

properties with respect to distance between regions, comparable aspect ratios of regions, positions, and orientation. The collected regions from step 6 are finally character regions. If training of data happens, characters regions are labeled with respective classes, whereas character regions are recognized if the test process of character regions happens. Recognition has been performed using matching score based on correlation. The image that results in maximum score is the recognized class of character region. The correlation between two images A and B are computed as Eq. (1)

$$r = \frac{\sum_m \sum_n (A_{mn} - \bar{A})(B_{mn} - \bar{B})}{\sqrt{(\sum_m \sum_n (A_{mn} - \bar{A})^2)(\sum_m \sum_n (B_{mn} - \bar{B})^2)}} \tag{1}$$

where, \bar{A} and \bar{B} are the mean values for m by n images.

The complexity of our algorithm for extraction of character regions is $O(n)$, where n is number of regions in step 3. Its best case is $O(m)$, where m is number of regions in step 6 and $m \leq n$. The recognition complexity is $O(N^2)$, where N is total number of images in train dataset. The recognition complexity could be reduced with other

better classification techniques but the prime focus of our study is region extraction as discussed in the previous section.

```

Algorithm 1
begin
1. Input image  $I$ .
2. Preprocess image  $I$  as
    $I \leftarrow \text{size normalization}(I)$ ;  $I \leftarrow \text{binarization}(I)$ ;
    $I \leftarrow \text{denoising}(I)$ ;
3. Extract regions of  $I$  and  $R$  is set of  $m$  isolated
   regions are:
        $R = \{r_1, r_2, \dots, r_m\}, m \geq$ 
4. for each region as  $r_i, 1 \leq i \leq m$  :
       make groups based on structural properties  $S$ .
       end for
5. for each group based on step 4,
       filter the regions based on respective preset lower
       and upper bounds. Update region set  $R$ .
       end for
6. for each group based on step 5,
       make new groups based on structural properties
       that satisfy inter and intra dependent regions
       properties. Update region set .
       end for
7. for each region as  $r_i, 1 \leq i \leq m$  :
       if training then
           label region  $r_i$ ;
       endif
       if test then
           recognize region  $r_i$ ;
       endif
       end for
end of algorithm

```

4 Experiments and Discussion

To evaluate the performance of the proposed framework, we conduct a set of experiments in which we estimated images based on different real-life scenarios as different angles of camera, weather conditions and traffic. The developed system application is available online to use at (<https://sites.google.com/view/anujsharma>). Our target is to identify correctly character regions and their recognition. We have used 500 images of different types of vehicles in a real-life environment. The set of 300 images out of 500 images are used as training dataset and rest 200 images are test dataset. These training and test dataset images are chosen randomly to perform unbiased experi-

Table 1 Character region extraction results

Data	Images	x_1	x_2	x_3
Trained	300	291	295	296
Test	200	192	194	196

x_1 all character regions extracted correctly

x_2 character regions extracted correctly except one region

x_3 character regions extracted correctly except two region

Table 2 Character region recognition results

Data	Images	y_1	y_2	y_3
Trained	300	287	291	293
Test	200	190	191	193

y_1 all character regions recognized correctly

y_2 character regions recognized correctly except one region

y_3 character regions recognized correctly except two region

mentation. Table 1 presents results for character region extraction. We have noted results for extracted characters in train dataset and test dataset both. In addition, we noted results for number plates with all correct regions; all correct regions except one and all correct regions except two. For example, a number plate includes 12 character regions, all correct regions are 12, all correct regions except one is 11, and all correct regions except two is 10. Our character extraction regions accuracy for train datasets are 97% (all correct), 98.33% (all correct except one), and 98.66% (all correct except two). Similarly, test dataset accuracy percentages are 96% (all correct), 97% (all correct except one), and 98% (all correct except two).

The recognition results are presented in Table 2. As we have used the correlation technique, results are noted for both train and test datasets. As discussed, higher accuracy classifiers as SVMs, HMMs, and k-NN are avoided in view to see the focus of this paper. We are mainly interested in correct regions identification as the first step of this study. We have presented results in Table 2 for all correct character regions recognized in number plate, all correct recognized except one and all correct recognized except two. Our recognition accuracy percentages for train datasets are 95.66% (all correct), 97% (all correct except one), and 97.66% (all correct except two). Similarly, test dataset accuracy percentages are 95% (all correct), 95.50% (all correct except one), and 96.50% (all correct except two). The experimentation for selective images has been presented in Fig. 2 that include developed application results based on Algorithm 1. Figure 2 presents one vehicle in each row and four columns are original image, preprocessed image, extracted character regions, and recognized number plate.

In this paper, we proposed an innovative AVLPR that extract character regions and their recognition with high accuracy for the real-life dataset. We adopted a novel region extraction technique based on image structural properties. Such image subregions properties are intra- as well interdependent. Our approaches covered both

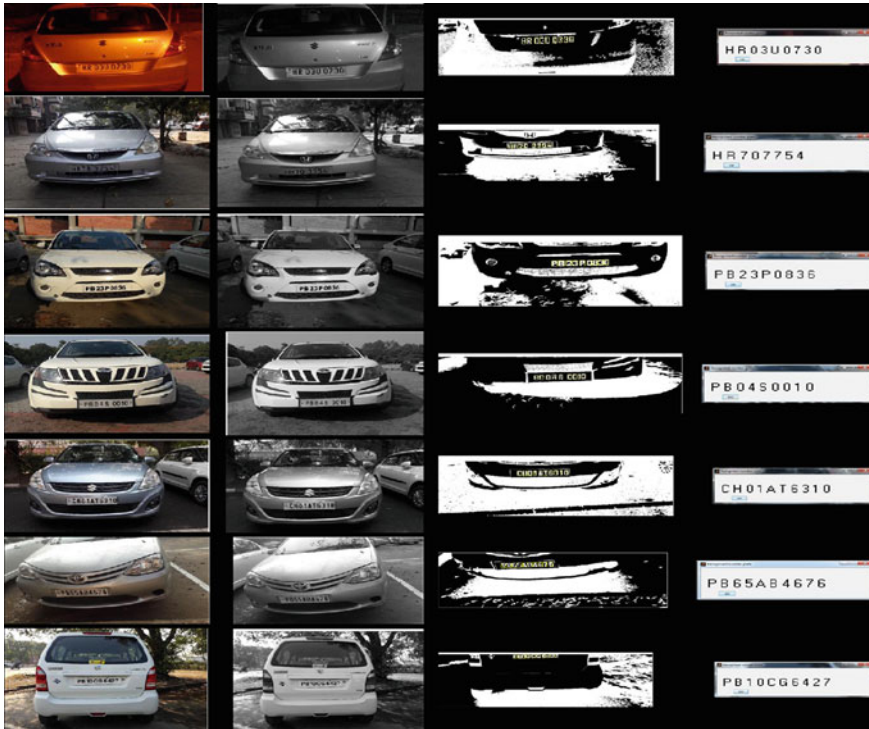


Fig. 2 One row for one vehicle results where the first column is original image, second column is preprocessed image, third column is character regions extracted, and fourth column is recognized vehicle number plate

intra-regions and inter regions properties. The correlation technique applied to match characters regions to complete the working of proposed character regions technique efficiency. We applied the developed system successfully and it could be helpful to use in other real-life applications.

5 Conclusion

This paper proposed a robust vehicle plate characters extraction and recognition algorithm. We used the features of characters like area, height, width, perimeter, and aspect ratio to extract the vehicle number characters from the image of the vehicle. Experimental results showed that the proposed technique gives 97 and 95% accurate results for characters extraction and character recognition, respectively, that are at par with the literature work. Although the proposed system is trained for English license plates, this algorithm can be extended to use with license plates of other scripts.

References

1. Du, S., Ibrahim, M., Shehata, M., Badawy, W.: Automatic license plate recognition (ALPR): a state-of-the-art review. *IEEE Trans. Circuits Syst. Video Technol.* **23**(2), 311–325 (2013)
2. Karwal, H., Girdhar, A.: Vehicle number plate detection system for Indian vehicles. In: 2015 IEEE International Conference on Computational Intelligence & Communication Technology (CICT), pp. 8–12. IEEE, New York (2015, February)
3. Mousa, A.: Canny edge-detection based vehicle plate recognition. *Int. J. Signal Process. Image Process. Pattern Recognit.* **5**(3), 1–8 (2012)
4. Sarfraz, M.S., Shahzad, A., Elahi, M.A., Fraz, M., Zafar, I., Edirisinghe, E.A.: Real-time automatic license plate recognition for CCTV forensic applications. *J. Real-Time Image Proc.* **8**(3), 285–295 (2013)
5. Gao, D.S., Zhou, J.: Car license plates detection from complex scene. In: 5th International Conference on Signal Processing Proceedings, 2000. WCCC-ICSP 2000, vol. 2, pp. 1409–1414. IEEE, New York (2000)
6. Al-Ghaili, A.M., Mashohor, S., Ismail, A., Ramli, A.R.: A new vertical edge detection algorithm and its application. In: International Conference on Computer Engineering & Systems, 2008. ICCES 2008, pp. 204–209. IEEE, New York (2008, November)
7. Ashtari, A.H., Nordin, M.J., Fathy, M.: An Iranian license plate recognition system based on color features. *IEEE Trans. Intell. Transp. Syst.* **15**(4), 1690–1705 (2014)
8. Chang, S.L., Chen, L.S., Chung, Y.C., Chen, S.W.: Automatic license plate recognition. *IEEE Trans. Intell. Transp. Syst.* **5**(1), 42–53 (2004)
9. Gou, C., Wang, K., Yao, Y., Li, Z.: Vehicle license plate recognition based on extremal regions and restricted Boltzmann (2016)
10. Panchal, T., Patel, H., Panchal, A.: License plate detection using Harris corner and character segmentation by integrated approach from an image. *Procedia Comput. Sci.* **79**, 419–425 (2016)
11. Öztürk, F., Özen, F.: A new license plate recognition system based on probabilistic neural networks. *Procedia Technol.* **1**, 124–128 (2012)
12. Kim, K., Jung, K., Kim, J.: Color texture-based object detection: an application to license plate localization. *Pattern Recognit. Support. Vector Mach.* pp. 321–335 (2002)
13. Hsu, G.S., Chen, J.C., Chung, Y.Z.: Application-oriented license plate recognition. *IEEE Trans. Veh. Technol.* **62**(2), 552–561 (2013)
14. Abolghasemi, V., Ahmadyard, A.: An edge-based color-aided method for license plate detection. *Image Vis. Comput.* **27**(8), 1134–1142 (2009)
15. Jia, W., Zhang, H., He, X.: Region-based license plate detection. *J. Netw. Comput. Appl.* **30**(4), 1324–1333 (2007)
16. Wang, Y.R., Lin, W.H., Horng, S.J.: A sliding window technique for efficient license plate localization based on discrete wavelet transform. *Expert Syst. Appl.* **38**(4), 3142–3146 (2011)
17. Massoud, M.A., Sabeel, M., Gergais, M., Bakhit, R.: Automated new license plate recognition in Egypt. *Alex. Eng. J.* **52**(3), 319–326 (2013)
18. Menotti, D., Chiachia, G., Falcão, A.X., Neto, V.O.: Vehicle license plate recognition with random convolutional networks. In: 2014 27th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 298–303. IEEE, New York (2014, August)
19. Chen, Z.X., Liu, C.Y., Chang, F.L., Wang, G.Y.: Automatic license-plate location and recognition based on feature saliency. *IEEE Trans. Veh. Technol.* **58**(7), 3781–3785 (2009)
20. Jiao, J., Ye, Q., Huang, Q.: A configurable method for multi-style license plate recognition. *Pattern Recogn.* **42**(3), 358–369 (2009)
21. Ramalingam, S., Rhead, M., Gurney, R.: Impact of character spacing on the performance of Automatic Number Plate Recognition (ANPR) systems through simulation. In 2014 International Carnahan Conference on Security Technology (ICCST), pp. 1–6. IEEE, New York (2014, October)
22. <https://in.mathworks.com/help/matlab/ref/rgb2gray.html>

23. <https://in.mathworks.com/help/images/ref/im2bw.html>
24. <https://in.mathworks.com/help/images/noise-removal.html>
25. <https://in.mathworks.com/help/images/ref/regionprops.html>

Toward Recognition and Classification of Hindi Handwritten Document Image



Shalini Puri and Satya Prakash Singh

Abstract With the increased demand of digitization of Indic scripts in today's world, many Devanagari printed and handwritten text recognition and extraction techniques have been developed and are used in industries, corporate, and institutional domain areas. Because of the script and character structure difficulties, and handwritten content-based criticalities, Hindi handwriting processing is considered as a big bottleneck in recognition systems. This paper introduces NMC handwriting types and complexity evaluators for Hindi language. The inherent challenges of handwriting are discussed further. Although several Hindi-based handwritten character recognizers have been developed, they are limited to the segmentation and identification of character images only. So, this paper proposes a new idea of offline Hindi handwritten document classification, which first recognizes and classifies the character images, and then classifies the document image into the predefined category. In support of this concept, this paper provides a case study using a set of Hindi handwritten documents and shows their segmentation and classification results. The proposed system puts a step ahead in the direction of automatic document image classification.

Keywords Hindi handwriting · Document analysis · Character recognition · Text identification · Template matching · Document classification

1 Introduction

Over the last three decades, offline Text Document Analysis and Classification (TDAC) is used as the important and most demanding research area. The TDAC system first analyzes the documents, then preprocesses and postprocesses them through multiple stages, and finally, classifies them into predefined mutually exclusive categories [1, 2]. A variation of these classification systems is the *Hindi Handwritten Document Classification Systems (HHDCS)*, which accept Hindi handwritten docu-

S. Puri (✉) · S. P. Singh

Department of Computer Science, BIT, Mesra, Ranchi, Jharkhand, India
e-mail: eng.shalinipuri30@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_44

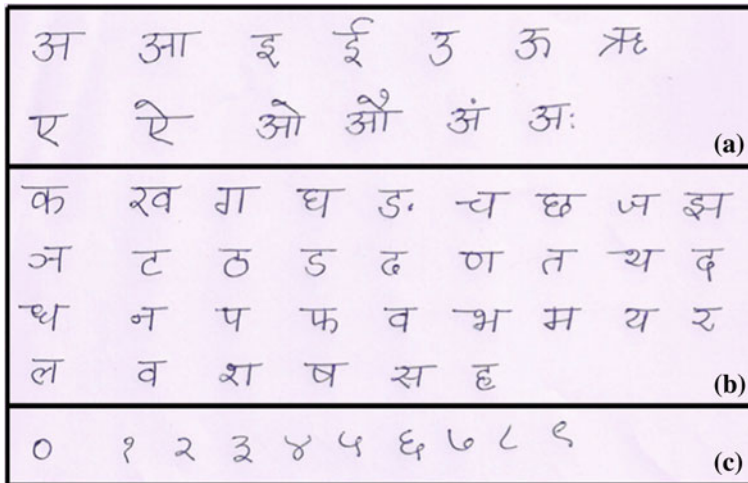


Fig. 1 Hindi handwritten **a** Vowels, **b** Consonants, **c** Numerals

ment images and classify them into distinct predefined classes. The HHDCS model analyzes the image lines, words, and characters, extracts features at each level, and processes image step by step for accurate document classification. Its performance and complexity may vary for different Devanagari languages, which become more challenging for unstructured, difficult, noisy and degraded documents. The applications of such systems include classification of handwritten notes, legal papers, and Government and authorized documents etc. These documents are generally generated and collected in large quantity and may also exist in low and degraded quality.

To have more understanding of HHDCS aspects, the basics of Hindi language and handwriting are discussed. Based upon Devanagari script and mother language Sanskrit, the most popular and widely used Hindi is the primary and official language of India. All the properties of Devanagari are applicable to Hindi, and it consists of 11 vowels, 33 consonants, and 10 numerals as depicted in Fig. 1. The structure of a Hindi word contains three zones, called as, core (the busiest zone), upper and lower zones. The character behavior is learned through its zones, and then its features are extracted for further processing. The Hindi word pronunciation is predicted from its written form only. Like some Indic languages, Sanskrit, Bangla, Punjabi, Marathi, Konkani, and Oriya, Hindi uses shirorekha (head bar) in writing. Many other Indic languages, say Kannada, Tamil, Telugu, Gujarati, and Malayalam do not use head bar.

Several researchers have contributed works toward Hindi handwritten character analysis, identification, and recognition. A critical parameter observed in handwritten character recognition is the writing style of a writer. Every writer has its own style of writing, and many a time, the writer is the only one who understands the handwritten contents. An example is the writer's signature, where the writer uses its own stylish way to sign the document. Other parameters, such as offline or online

documents, writing device, image resolution, Depth Per Inch (DPI), etc., also introduce additional challenges in handwriting recognition. Therefore, it is important that the ability to understand the handwriting must be increased, so that all its words and characters can be recognized with high accuracy. The next section demonstrates a review tour on existing Hindi handwritten processing systems. This study discusses about their Optical Character Recognition (OCR) systems, processing steps, feature extraction methods, and performance results. Such study concludes with the limitations of these recognition systems and the need to develop advanced HHDCS. HHDCS uses the underlying script and language concepts and classifies the offline scanned handwritten images.

This paper is organized as follows. The next section discusses several Hindi handwritten text recognition and extraction approaches in detail. Its following section provides an overview of related concepts, and demonstrates NMC handwriting types, complexity evaluators, and inherent challenges. A case study on HHDCS is explained in the next section. The final section includes the conclusion.

2 A Review Tour on Existing Hindi Handwriting Processing Systems

Although the printed character recognition is found simpler than handwritten ones, yet many good contributions of handwritten OCR are available with Latin script. In the past two decades, several techniques and research works have also been successfully developed and contributed toward printed mono or multi-lingual Indic OCR. However, it is always challenging to handle Indic, especially Devanagari, handwritten OCR, and there are only a few research works on it. This section provides review of Hindi handwritten processing systems from the year 2009 to 2018.

A top-down bilingual Hindi-English and Bangla-English script identification method [3] at the page, block, paragraph, and word levels was proposed using Support Vector Machine (SVM) and Adaptive Boosting (AdaBoost). The English and Hindi handwritten character recognition methods [4] simulated the best features of characters by combining script-independent character structural features and curve fitting. The handwritten Devanagari word segmentation approach proposed in [5] first used morphological operations in segmentation, and then neighborhood tracing algorithm to get segmented objects from three zones. It achieved 57, 55, and 52% segmentation accuracies for the top, bottom, and core zone characters, respectively. Next, Hindi handwritten word recognition method was proposed through segmentation and lexicon-test word matching [6], which used two-pass dynamic programming and directional element features, and achieved 91.23–79.94% accuracy on 10–30 vocabulary words. Further, [7] proposed a preprocessing-less Hindi handwritten line segmentation method through header and baseline detections and contour following technique. This approach worked upon fluctuating and variable skew text lines. The Hindi handwritten character segmentation of conjuncts and overlapping char-

acters through cluster detection [8] and of unconstrained words using writing style structural patterns [9] were proposed, where later method achieved 96.93% average accuracy. The shape feature-based and fuzzy logic-based method [10] recognized Devanagari handwritten characters through the steps of thin character segmentation using basic structure features, character coding using average compressed direction code, shape classification, character identification by applying script grammar, locating the characters using fuzzy classification, and finally, character pre-classification by tree classifier; therefore, it achieved 86.4% average recognition accuracy.

The comparative study on Devanagari handwritten OCR [11] discussed 12 classifiers, such as projection distance, subspace method, linear discriminant function, SVM, modified quadratic discriminant function, mirror image learning, Euclidean distance, nearest neighbor, K-Nearest Neighbor (KNN), modified projection distance, compound projection distance, and compound modified quadratic discriminant function, and computed the features through curvature and gradient information of binary- and gray-scale images. Another study [12] reviewed different script identification structure-based and visual appearance-based methods along with the script identification of online data and video texts. The handwritten Hindi OCR using curvelet transform [13] first segmented the image, obtained curvelet features by calculating statistics of thick and thin images, and trained the KNN, and achieved more than 90% recognition accuracy. Another handwritten OCR [14] used Artificial Neural Network (ANN) classifier and achieved 75.6% accuracy on noisy characters. The handwritten OCR [15] preprocessed and binarized the image, removed shirorekha, extracted the features, used K-means clustering, and then classified the characters using linear kernel-based SVM. Another such OCR [16] implemented Radial Basis Function (RBF) neural network and directional group feature extraction and achieved low recognition accuracy with less training and classification times than back propagation neural network. The unconstrained handwritten Devanagari OCR [17] categorized the characters into smaller groups through fuzzy inference system and character structure parameters, and recognized them through feed forward neural network with 96.95% accuracy. The offline Devanagari handwritten OCR [18] used strength, angle, and histogram of gradient directional features along with a combined quadratic and SVM classifier, and achieved 95.81% accuracy with threefold cross-validation.

It is observed here that these existing systems are limited in scope, and identified and recognized characters only. They were primarily oriented toward script identification, OCR, and line, word, and character (basic and conjuncts) segmentation. They used ANN, SVM, KNN, fuzzy, and many other classifiers for character recognition and classification. These observations find that there is much more to do with these results and recognized characters, which can go beyond the OCR. Such limitations motivated the authors to propose the advanced Hindi handwritten document analysis and classification systems.

3 Handwriting Types, Complexity Evaluation, and Challenges

For the technological development and advancements in Devanagari, nowadays researchers are paying a good deal of attention in analyzing and recognizing Hindi-based document images. Although Hindi consists of very strong and well-structured grammar rules for reading and writing texts, yet its processing is found difficult for unconstrained handwritten images [19]. This section introduces NMC handwriting classification types for Hindi, that discriminates the handwriting on the basis of space available between two adjacent lines, words and characters. Further, it discusses complexity evaluators to assess the processing difficulties of handwriting, and then the inherent challenges of handwritten text extraction and recognition.

3.1 Handwriting Types

Hindi text is written as the combination of shirorekha and characters. While writing Hindi words and characters, suitable space must be provided between two lines, between two shirorekhas of two adjacent words, and between two adjacent characters of a single word, so that it can avoid and reduce unnecessary vertical and horizontal overlaps. Therefore, three Threshold Values TV-I, TV-II, and TV-III are defined to show the minimum space in inter-line, inter-shirorekha (or inter-word), and intra-word. Here, TV-I represents minimum gap between two adjacent lines, TV-II represents minimum gap between two shirorekhas of two adjacent words, and TV-III represents minimum gap between two adjacent characters of the same word. Table 1 depicts *Normal–Moderate–Complex (NMC) handwriting classification model* along with their conditions, consequences, and risk factors. The observations from Table 1 find that normal handwriting is better than the other two, and uses appropriate spaces to achieve good recognition results.

3.2 Handwriting Complexity Evaluation

The complexity evaluation of Hindi handwriting recognition system consists of two sets of parameters, known as, *human-based parameters and writing-based parameters*. These parameters are further subdivided into more refined categories as shown in Fig. 2. Both of these parameters are correlated with each other and their coexistence increases the system complexity as well as text recognition challenges. The human-based parameters consist of writer features, such as his/her behavior and nature, and physical features, such as writer identification, personality, age, and gender. It is found that the writer behavior and nature are directly reflected by his/her handwriting, say, if the writer's mood, behavior or nature is good, then his/her hand-

Table 1 NMC handwriting classification with conditions, consequences, and risks

Handwriting type	Conditions	Consequences and risk factors
Normal	$InterLineSpace_{Normal} \geq TV-I,$ $InterWordSpace_{Normal} \geq TV-II$ & $IntraWordSpace_{Normal} \geq TV-III$	Easy interpretation. Approx. correct recognition 80–100%. Low/negligible risks
Moderate/semi normal	$InterLineSpace_{Normal} >$ $InterLineSpace_{Moderate} >$ $InterLineSpace_{Complex},$ $InterWordSpace_{Normal} >$ $InterWordSpace_{Moderate} >$ $InterWordSpace_{Complex}$ & $IntraWordSpace_{Normal} >$ $IntraWordSpace_{Moderate} >$ $IntraWordSpace_{Complex}$	Approx. correct recognition 40–80%. Includes mixture of correct, confused and wrong interpreted words and characters
Complex	Random, unconstrained and improper spacing	Approx. Recognition 0–39%. Very hard to find start/end of a word/character. Very less/negligible interpretation & has many overlaps

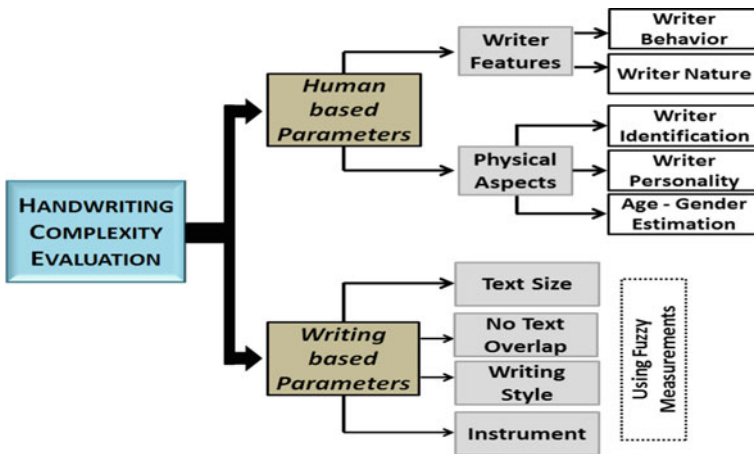


Fig. 2 Handwriting complexity evaluators

writing is found favorable, otherwise he/she can produce worst handwriting. The writer’s physical aspects are measured on the basis of his/her handwriting features, so the writer’s identity, personality, age, and gender can be decided by evaluating his/her good or bad handwriting.

The writing-based parameter describes the way in which the contents are written and embedded in a document. These parameters are called *Text Size and No Overlaps (TSNO)*, *Style (S)*, and *Instrument (I)* as shown in Fig. 2 and Table 2. Each parameter uses a fuzzy span of seven values, given as, {Extreme {High Value}, Very {High

Table 2 Writing-based complexity evaluators and their criteria

Writing-based complexity evaluators	Evaluation criteria
Text size and no overlaps	Small, medium, and large
	Proper inter-line, inter-word, and intra-word spacing
	Average length of words and letters within shirorekha margin
	No incomplete or overwritten word, letter, or shirorekha
	Includes straight, cursive, and circular characters
Style	Left/right text orientation
	Needs appropriate writing speed
	Skew angle (slant) determination of written text
	Writing non-confusing and complete letters
Instrument (Pen/Pencil/Paper/Scanner)	Instrument quality
	Pen pressure
	Word and character thickness

Value}, {High Value}, Normalized Average, {Low Value}, Very {Low Value}, and Extreme {Low Value}}. In TSNO, the fuzzy range of handwriting measurement is {Extremely Large, Very Large, Large, Average, Small, Very Small, Extremely Small}. Some writers use to write long and big characters, which may increase vertical character overlapping, and may reduce the correct character identification. On the opposite side, many writers write very small text, which may also increase recognition issues. Stylish handwriting primarily consists of curves and results from writer’s mood and nature. Its fuzzy measures vary among {Extremely High Stylish, Very High Stylish, High Stylish, Average, Less Stylish, Very Less Stylish, Extremely Less Stylish}. The fuzzy measures of instrument category vary among {Extremely Thick, Very Thick, Thick, Moderate, Thin, Very Thin, Extremely Thin}. The extremely thick handwriting is very much dark, whereas extremely thin is very much light.

3.3 Challenges in Handwriting Recognition

Hindi handwriting OCR includes a series of challenging steps, which makes character interpretation and recognition tasks difficult. Toward disappearance of handwriting concept, the recent article “Handwriting versus typing: is the pen still mightier than the keyboard?” [20] by Anne Chemin on December 16, 2014 provided a good comparison between handwriting and typing skills. The typing is considered as simple, fast, clear, easy to read that has graphic freedom and can easily use MS word features. On the other side, writing with a hand develops one’s cognition, learning, basic and

special skills, motor skills, creativity, body memory, grasp on alphabets, and document edit records. The British survey of 2000 people in June 2014 and the Los Angeles Times editorial comments on September 4, 2013 also focused on the need of handwriting. Another article “*How can I convert my handwritten notes into Word documents?*” [21] by Jack Schofield on December 18, 2015 discussed the problems and need of converting handwritten documents into word documents using OCR, so that it could recognize printed and handwritten characters. According to him, the handwriting must be extremely neat, clear, and consistent, and OCR must be trained to recognize such inputs. Some online available OCRs accept non-connected handwritten and printed characters, and some other, say, Tesseract, needs preprocessing before handwriting recognition. Along with these handwriting issues, many other Hindi OCR challenges occur due to the acquisition device, its large set of characters and conjuncts, character structure, modifier shape, presence of modifiers, writing speed, style variations, distortions, etc. They are discussed below.

1. *Character Issues*: Irregular character writing; broken and incomplete characters; filled loops; shape variations of same character; conjunct characters separation and recognition; conjunct recognition; stroke identification; adjacent characters overlapping horizontally; characters overlapping vertically; overlapping of two or more consonants in core area; and half and full characters overlapping.
2. *Header Line Issues*: Inter-line distance; irregularities in shirorekha height, shirorekha width and shirorekha leftmost position; fluctuating, curved, touching and overlapped shirorekha; space between words and shirorekhas; vertical overlapping of word's bottom modifier with other word's top modifier of line below; shirorekha and word slant; presence of multiple and nonuniform shirorekha; absence of shirorekha; and character, and shirorekha overlapping.
3. *Modifiers Issues*: Improper modifier's size, broken and incomplete modifiers, presence of multiple modifiers, conjunct formation, overlapping of two or more modifiers, consonant and modifier overlapping in core region, and overlapping of 'Nukta' (◌̣) with shirorekha or other top modifiers.
4. *Segmentation Issues*: Top and bottom modifiers separation; line, word, character and conjuncts segmentation; and shirorekha extraction and removal.
5. *Skew-based Issues*: Due to scanning and small angle creation.
6. *Instrument-based Issues*: Printing, quantization, binarization, pen width, ink color, improper paper placement, and bad quality paper.

4 Case Study on HHDCS

Hindi Handwritten Document Classification System (HHDCS) is an advanced image classification system, which first extracts and recognizes the image characters and then classifies the document images into mutually exclusive predefined categories. It includes character and word analysis, matching, recognizing, and classification stages. This was observed in the review section that the existing handwritten Hindi

OCRs were limited to the script identification and character classification only. The proposed HHDCS provides a step ahead toward handwritten document classification through preprocessing, character recognition and classification, word formation and recognition via template matching, and finally, document classification.

This section discusses a case study of HHDCS with four scanned documents of categories ‘शिक्षा’ and ‘खेल’ in two different handwritings. These documents are written by two writers. The segmentation results of ‘शिक्षा’ and ‘खेल’ documents are shown in Figs. 3a, b, and 4a, b correspondingly. Here red, green, and pink bounding boxes contain lines, words, and characters and modifiers, respectively.

After performing character recognition and classification, all the character images are associated back to make the words. The system stores the relevant and important words, and ignores stop symbols (., :), stop words (‘पर’, ‘ना’, ‘कर’) and irrelevant words. The primary keywords obtained are ‘शिक्षा’, ‘शिक्षक’, ‘विद्यार्थी’, and ‘प्रश्न’ in Fig. 3a, and are ‘शिक्षा’, ‘विद्यार्थी’, and ‘अनुसंधान’ in Fig. 3b. Other keywords are ‘जगत’, ‘केंद्र’, and ‘परिधि’ in Fig. 3a, and are ‘विभाग’, ‘विश्वस्तरीय’, and ‘अंतर्राष्ट्रीय’ in Fig. 3b. The primary keywords obtained are ‘खेल’, ‘ओलिंपिक’, and ‘जैक्स’ in Fig. 4a and are ‘खेल’ and ‘टीम’ in Fig. 4b, whereas, other keywords are ‘भारत’, ‘अंतरराष्ट्रीय’, ‘समिति’, ‘अध्यक्ष’, ‘मंत्रालय’, and ‘ऑस्ट्रेलियाई’ in Fig. 4a, and are ‘पेशेवर’, ‘करियर’, ‘युवाओं’, ‘रुचि’, and ‘सहयोग’ in Fig. 4b. HHDCS uses template matching to compare the relevant and primary words with the predefined classes, and then the classifier classifies the document into a suitable category. In this way, HHDCS classifies the Hindi handwritten document images and provides an enhancement over existing systems.

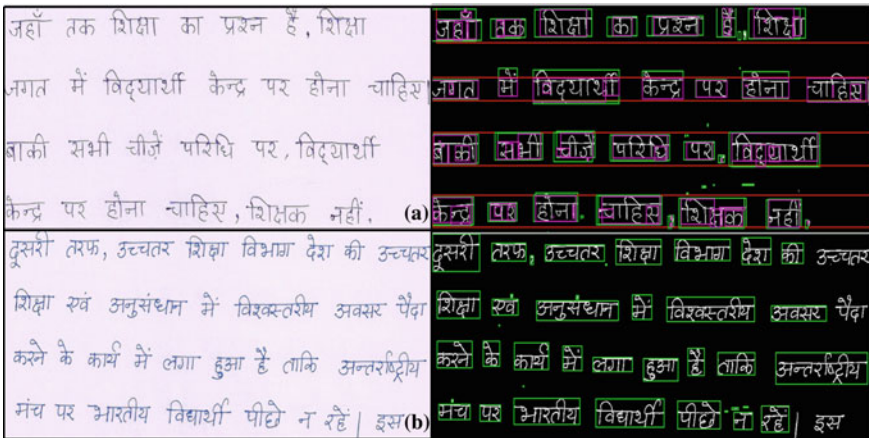


Fig. 3 Handwritten documents a and b on class ‘शिक्षा’ and their segmented results

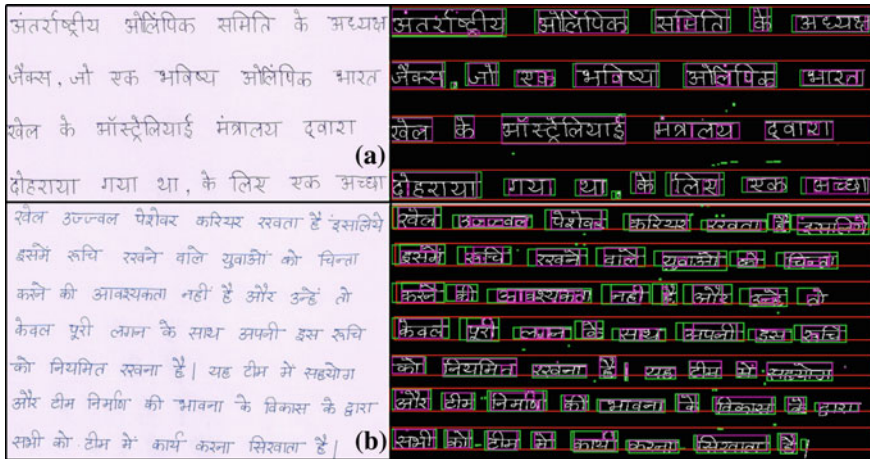


Fig. 4 Handwritten documents **a** and **b** on class 'खेल' with their segmented results

5 Conclusion

This paper provided the discussion on recent handwritten Hindi character recognition and classification systems and proposed the idea of Hindi handwritten document classification systems. The review demonstrated the existing handwritten OCR methods along with in-depth discussion of their processing steps, contents, features, classifiers, and accuracies. In addition to this, NMC handwriting classification, complexity evaluators, and inherent challenges were discussed. Further, a case study was discussed on HHDCS in the direction of handwritten document image classification, and showed the segmentation and document classification results. Authors are implementing HHDCS and have achieved success to some extent.

References

1. Puri, S., Singh, S.P.: A technical study and analysis of text classification techniques in N-lingual documents. In: International Conference on Computer Communication and Informatics, pp. 1–6. IEEE Press, New York (2016)
2. Toselli, A.H., Juan, A., Vidal, E.: Spontaneous handwriting recognition and classification. In: Proceedings of the 17th International Conference on Pattern Recognition, vol. 1, pp. 433–436. IEEE Press, New York (2004)
3. Hassan, E., Garg, R., Chaudhury, S., Gopal, M.: Script based text identification: a multi-level architecture. In: Proceedings of the Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data, pp. 1–8. ACM, New York (2011)
4. Khanduja, D., Nain, N.: Script independent feature set for handwritten text recognition. In: 37th International Convention on Information and Communication Technology, Electronics and Microelectronics, pp. 1147–1152. IEEE Press, New York (2014)

5. Ladwani, V.M., Malik, L.: Novel approach to segmentation of handwritten Devnagari word. In: 3rd International Conference on Emerging Trends in Engineering and Technology, pp. 219–224. IEEE Press, New York (2010)
6. Ramachandrula, S., Jain, S., Ravishankar, H.: Offline handwritten word recognition in Hindi. In: Proceeding of the workshop on Document Analysis and Recognition, pp. 49–54. ACM, New York (2012)
7. Garg, N.K., Kaur, L., Jindal, M.K.: A new method for line segmentation of handwritten Hindi text. In: 7th International Conference on Information Technology: New Generations, pp. 392–397. IEEE Press, New York (2010)
8. Thakral, B., Kumar, M.: Devanagari handwritten text segmentation for overlapping and conjunct characters—a proficient technique. In: Proceedings of 3rd International Conference on Reliability, Infocom Technologies and Optimization, pp. 1–4. IEEE Press, New York (2014)
9. Bag, S., Krishna, A.: Character segmentation of Hindi unconstrained handwritten words. In: Barneva, R., Bhattacharya, B., Brimkov, V. (eds.) International Workshop on Combinatorial Image Analysis. Lecture Notes in Computer Science, vol. 9448, pp. 247–260. Springer, Cham (2015)
10. Mukherji, P., Rege, P.P.: Shape feature and fuzzy logic based offline Devnagari handwritten optical character recognition. *J. Pattern Recognit. Res.* **5**(1), 52–68 (2010)
11. Pal, U., Wakabayashi, T., Kimura, F.: Comparative study of Devnagari handwritten character recognition using different feature and classifiers. In: 10th International Conference on Document Analysis and Recognition, pp. 1111–1115. IEEE Press, New York (2009)
12. Ghosh, D., Dube, T., Shivaprasad, A.: Script recognition—a review. *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(12), 2142–2161 (2010)
13. Verma, G.K., Prasad, S., Kumar, P.: Handwritten Hindi character recognition using curvelet transform. In: Singh, C., Singh, Lehal G., Sengupta, J., Sharma, D.V., Goyal, V. (eds.) Information Systems for Indian Languages. Communications in Computer and Information Science, vol. 139, pp. 224–227. Springer, Berlin (2011)
14. Sahu, N., Raman, N.K.: An efficient handwritten Devnagari character recognition system using neural network. In: International Multi-Conference on Automation, Computing, Communication, Control and Compressed Sensing, pp. 173–177. IEEE Press, New York (2013)
15. Gaur, A., Yadav, S.: Handwritten Hindi character recognition using K-means clustering and SVM. In: 4th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, pp. 65–70. IEEE Press, New York (2015)
16. Singh, D., Saini, J.P., Chauhan, D.S.: Hindi character recognition using RBF neural network and directional group feature extraction technique. In: International Conference on Cognitive Computing and Information Processing, pp. 1–4. IEEE Press, New York (2015)
17. Shelke, S., Apte, S.: A fuzzy based classification scheme for unconstrained handwritten Devanagari character recognition. In: International Conference on Communication, Information & Computing Technology, pp. 1–6. IEEE Press, New York (2015)
18. Bhalerao, M., Bonde, S., Nandedkar, A., Pilawan, S.: Combined classifier approach for offline handwritten Devanagari character recognition using multiple features. In: Hemanth, D., Smys, S. (eds.) Computational Vision and Bio Inspired Computing. Lecture Notes in Computational Vision and Biomechanics, vol. 28, pp. 45–54. Springer, Cham (2018)
19. Smith, T.L.: Six basic factors in handwriting classification. *J. Crim. Law Criminol.* **44**(6), 810–816 (1954)
20. Chemin, A.: Handwriting vs typing: is the pen still mightier than the keyboard? <https://www.theguardian.com/science/2014/dec/16/cognitive-benefits-handwriting-decline-typing>. Last accessed 21 July 2018
21. Schofield, J.: How can I convert my handwritten notes into word documents? <https://www.theguardian.com/technology/askjack/2014/dec/18/how-can-i-convert-my-handwritten-notes-into-word-documents>. Last accessed 21 July 2018

Efficient Image Deblurring Using Alpha Plane Blending on Images Recovered with Linearly Varied Point Spread Function (PSF)



Poonam Sharma, Ashwani Kumar Dubey and Ayush Goyal

Abstract The medical imaging has seen a significant technological shift from analog-film-based imaging to high-end digital computer imaging. One of the problems this domain has always faced is the blurred images due to the natural motion in the body and mechanical motion of the capturing devices. This work utilizes the image blending technique to improve the image quality while deblurring through motion adaptive point spread functions. The idea is to first get a sample set of images restored through Lucy RC deconvolution techniques using varied PSFs and then applying the image blending over the image set in alpha plane to get deblurred sharp images.

Keywords Image blending · Deconvolution · Point spread function · Deblurring · Alpha plane · Lucy RC deconvolution

1 Introduction

The medical imaging has a great role toward effective diagnosis and analysis of a physiological aspect of the human body. The medical experts heavily depend and count on the imaging results obtained from the image sensors [1]. Unfortunately, there is unavoidable motion in the human body such as a moving fetus in the womb or an elderly person who cannot stabilize his position. Even sometimes, the image capturing sensor or device may not have a sharp exposure time resulting in the blurred images. The blurred medical images are difficult to analyze the physiology and reach

P. Sharma (✉) · A. K. Dubey
Amity University Uttar Pradesh, Noida 201301, Uttar Pradesh, India
e-mail: sh.poonam15@gmail.com

A. K. Dubey
e-mail: akdubey@amity.edu

A. Goyal
Texas A&M University, Kingsville, TX 78363, USA
e-mail: ayush.goyal@tamuk.edu

a good diagnostic result. There have been many efforts in the image processing techniques like deconvolution algorithms such as Blind deconvolution [2], Wiener deconvolution [3], and Lucy RC deconvolution [4] to name a few. Many researchers have suggested the Lucy RC deconvolution techniques with point spread function (PSF) to be the closest approximation as far as motion deblurring [5] is considered. But still it is an approximation and there is scope of improvement [6].

A close look into the deconvolved images obtained from linearly varied PSF function [7] hints toward the change in edges of the restored images in a random fashion.

This work achieves the objective of enhancing the quality of restoration of motion-blurred images. The proposed method attempts a novel idea where it uses blending of images in the alpha plane to merge the variations in the edges of deconvolved images and smoothen them out to provide a crisper and sharper image obtained after alpha plane blending of PSF [8] varied deconvolved images using Lucy RC deconvolution.

1.1 The Principle Architecture

The proposal and the idea behind the work start from consideration of motion blurring in linear fashion of length and angle. The deconvolution is chosen to be Lucy RC algorithm and close approximation of PSF is done. Thus, with varied PSF [9] a set of images is obtained to be operated with alpha plane blending. Figure 1 shows the architecture behind approximating the PSF used with Lucy RC deconvolution.

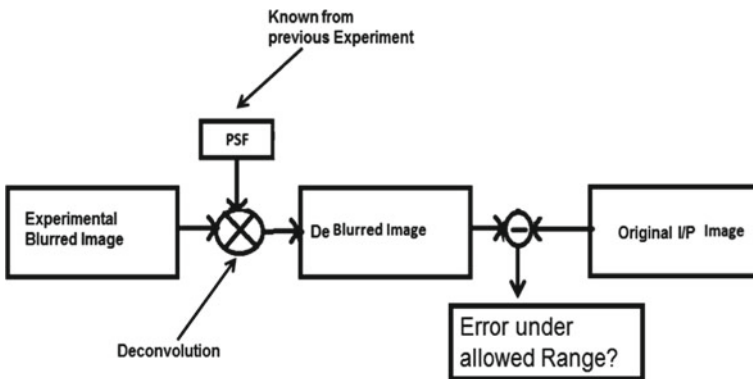


Fig. 1 Varied PSF deconvolution

2 Method

In medical imaging, the computed tomography (CT) scans are done using many X-rays to build a cross-sectional view of body, whereas in magnetic resonance imaging (MRI) strong magnetic fields are used to create cross-sectional views for anatomy. In both the cases, the images created may have motion blurring effects. In order to obtain pre-blending, deconvolved images from two computed tomography (CT) and three magnetic resonance imaging (MRI) images, and Lucy RC deconvolution algorithm is used. The sample images are first convolved with a specified point spread function (PSF) factor to obtain the same set of images in the blurred form to create a sample set to be deconvolved with Lucy RC. The images are random but are chosen on the basis of gray scale [10] variation to match the realism in practical medical scenario [11]. The sample input of 2 CT and 2 MRI images are shown in Figs. 2 and 3, respectively.

All Images Courtesy by—<https://medpix.nlm.nih.gov/>.

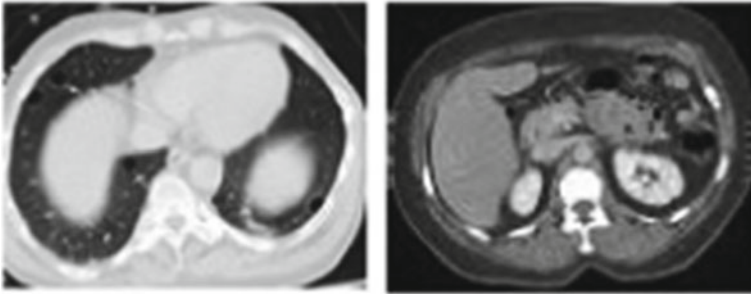


Fig. 2 Sample computed tomography images (CT1 and CT2)

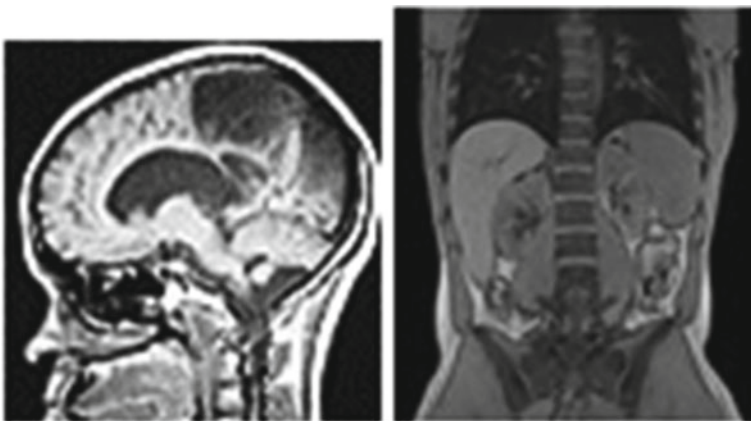


Fig. 3 Sample magnetic resonance images (MRI1 and MRI2)

The sample images are processed through the steps as mentioned hereafter.

2.1 Blurring of Sample Images Using Convolution

Here, the sample image is blurred with a known and varied PSF using circular convolution and noise is also introduced to prototype a real-time approximation of blurred images as shown in Figs. 4 and 5. The steps to blur are given below:

```
Read an image [15] –  
IP = im2double(imread(sample image));  
Get PSF -  
PSF = motion with length = 21 and theta = 11.  
Introduce blurring –  
Blurred Image = imfilter(smple, PSF, 'conv', 'circular');  
Noise Introduction –
```



Fig. 4 Blurred MRI1 and CT1 images, respectively



Fig. 5 Blurred MRI2 and CT2 images, respectively

```
noise_mean = 0;
noise_var = 0.0001;
blurred + noisy = imnoise(blurred1, 'gaussian', noise_mean, noise_var);
```

2.2 Deconvolution Using Lucy RC Algorithm with Linearly Varying PSFs

In this step, the images which are blurred in previous step to obtain the samples for deblurring are deconvolved with varied PSF.

```
PSF1 length = 19,
PSF2 length = 20,
PSF3 length = 21,
PSF4 length = 22,
PSF5 length = 23, and
PSF6 length = 24.
```

$$D1 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF1}); \quad (1)$$

$$D2 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF2}); \quad (2)$$

$$D3 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF3}); \quad (3)$$

$$D4 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF4}); \quad (4)$$

$$D5 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF5}); \quad (5)$$

$$D6 = \text{deconvlucy}(\text{blurred_noisy1}, \text{PSF6}). \quad (6)$$

The deblurred images obtained from PSF1 are shown in Fig. 6.

Figure 7 shows the Lucy RC deconvolved image with PSF length = 20, i.e. 2 images for CT and MRI input images under consideration.

Similarly, Figs. 8, 9, 10, and 11 show Lucy RC deconvolved images with PSF2, 3, 4, 5, and where length is varied linearly from 21 to 24.

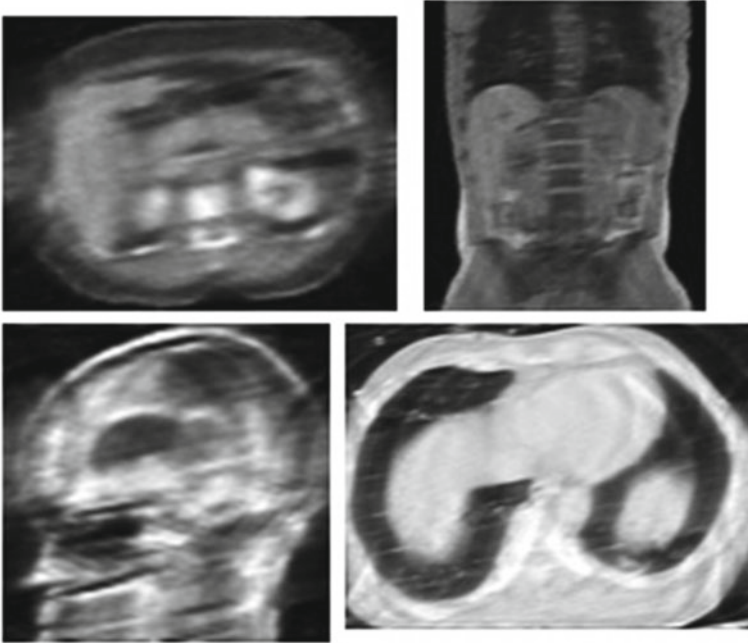


Fig. 6 Lucy RC deconvolved images with PSF1

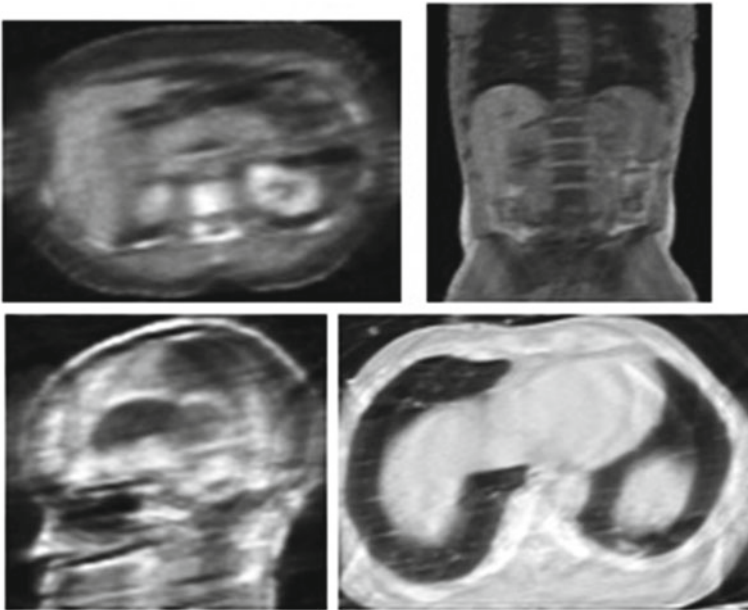


Fig. 7 Lucy RC deconvolved images with PSF2

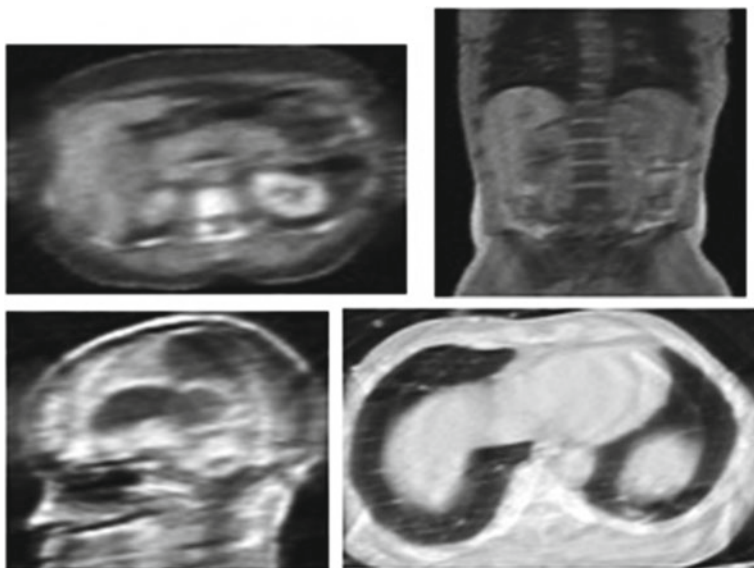


Fig. 8 Lucy RC deconvolved images with PSF3

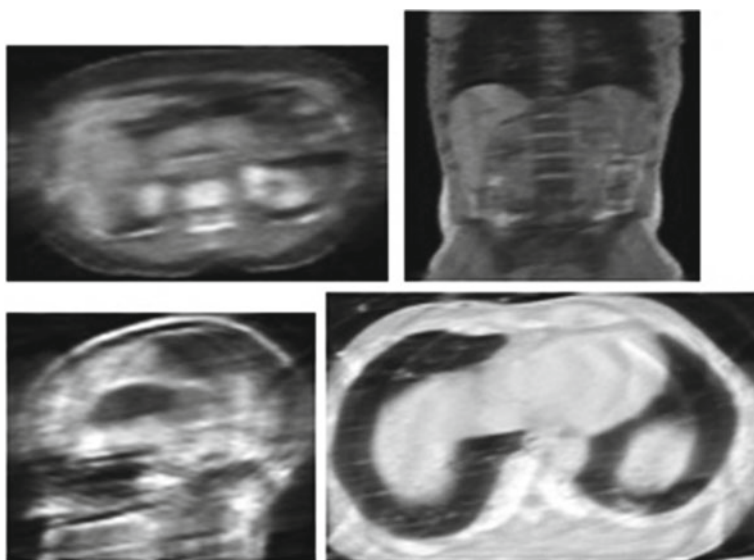


Fig. 9 Lucy RC deconvolved images with PSF4

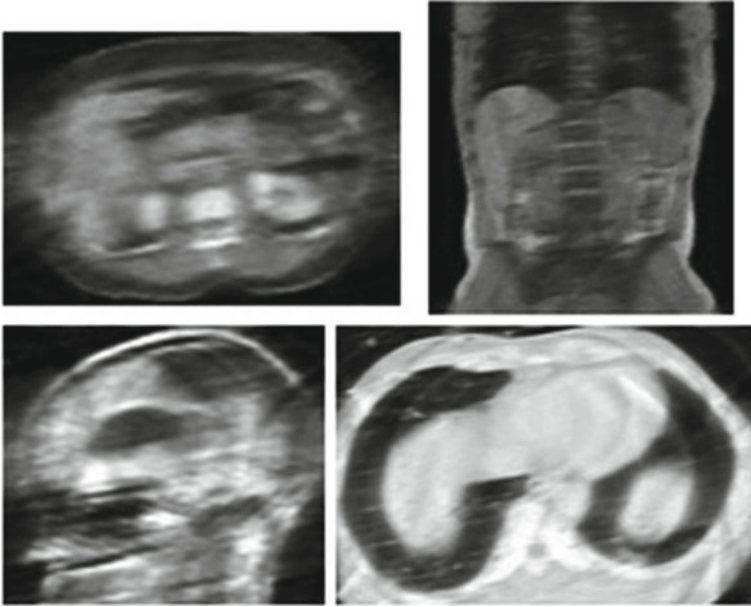


Fig. 10 Lucy RC deconvolved images with PSF5

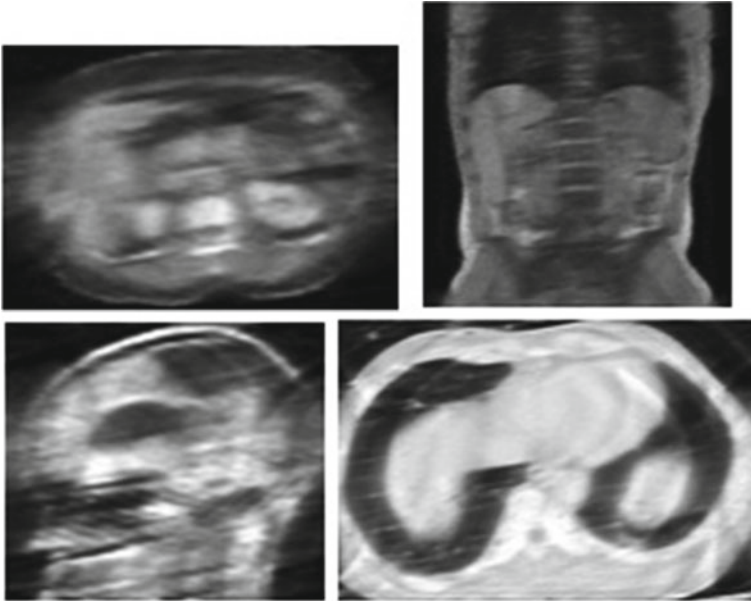


Fig. 11 Lucy RC deconvolved images with PSF6

2.3 Mean Square Error Calculation with Varied PSF Deconvolved Image

To approximate the performance of the deconvolution, the mean square error [12] is calculated with Dn values as obtained from Eqs. (1)–(6).

$$\text{Error, } E_n = IP - D_n \tag{7}$$

$$\text{MSE}_{D_n} = \text{MSE}(E_n) \tag{8}$$

where n is the respective PSF taken from linear value 19–24, en is error for respective PSF, IP is the original image, and Dn is the deconvolved image with respective PSF.

Equation (7) gives the instantaneous spatial errors. Equation (8) calculates the mean square error. For each PSF, PSF1–PSF6, with which the Lucy RC deconvolution is done to obtain a nonblurred image, the MSE on each image, i.e., three MRIs and two CTs are calculated with respect to original true image.

The mean square error for each case of deconvolution is given in Tables 1, 2, 3, 4, 5 and 6 in the fourth column.

The results with MSE calculations are depicted in Tables 1, 2, 3, 4, 5 and 6 gives the markers for deviation from ideal deconvolutions.

Table 1 MSE calculated for PSF1

S. No.	Test images	Blended after Lucy	Lucy PSF1
1	mri1_original	0.04165395	0.046467383
2	mri2_original	0.00418494	0.004898615
3	mri3_original	0.004729498	0.005688372
4	ct1_original	0.009255153	0.010538349
5	ct2_original	0.006904503	0.00801705

Table 2 MSE calculated for PSF2

S. No.	Test images	Blended after Lucy	Lucy_PSF2
1	mri1_original	0.04165395	0.04342134
2	mri2_original	0.00418494	0.004221456
3	mri3_original	0.004729498	0.00492314
4	ct1_original	0.009255153	0.011342673
5	ct2_original	0.006904503	0.0097254274

Table 3 MSE calculated for PSF3

S. No.	Test images	Blended after Lucy	Lucy_PSF3
1	mri1_original	0.04165395	0.04213472
2	mri2_original	0.00418494	0.005168202
3	mri3_original	0.004729498	0.00596357
4	ct1_original	0.009255153	0.0153420
5	ct2_original	0.006904503	0.00821754

Table 4 MSE calculated for PSF4

S. No.	Test images	Blended after Lucy	Lucy_PSF4
1	mri1_original	0.04165395	0.0542681
2	mri2_original	0.00418494	0.006723976
3	mri3_original	0.004729498	0.006543876
4	ct1_original	0.009255153	0.02074562
5	ct2_original	0.006904503	0.008208537

Table 5 MSE calculated for PSF5

S. No.	Test images	Blended after Lucy	Lucy_PSF5
1	mri1_original	0.04165395	0.042397463
2	mri2_original	0.00418494	0.00493462
3	mri3_original	0.004729498	0.005743629
4	ct1_original	0.009255153	0.0128534
5	ct2_original	0.006904503	0.008119063

Table 6 MSE calculated for PSF6

S. No.	Test images	Blended after Lucy	Lucy_PSF6
1	mri1_original	0.04165395	0.04654329
2	mri2_original	0.00418494	0.004973523
3	mri3_original	0.004729498	0.00574536
4	ct1_original	0.009255153	0.010372578
5	ct2_original	0.006904503	0.008018945

2.4 Image Blending in Alpha Plane

The image blending [13] or compositing in alpha plane is a means of taking weighted RGB values of each pixel in an image and adding in fraction to the corresponding pixel of second image RGB. For example, if we have two images as p1 and p2, for each pixel, RGB values are calculated for a blended or composited image [14] say c_out as shown below:

$$c_{out_R} = p1_R * (A) + p2_R * (1 - A), \quad (9)$$

$$c_{out_G} = p1_G * (A) + p2_G * (1 - A), \quad (10)$$

$$c_{out_B} = p1_B * (A) + p2_B * (1 - A). \quad (11)$$

In Eqs. (9)–(11), alpha factor [15] A (range from 0 to 1) is the blend depth for the images which decides the foreground and background for the composited image.

Similarly, the six deconvolved images are blended together with near equal alpha level to obtain a composited final deconvolved image.

$$O1 = ((0.16 * L11) + (0.16 * L21) + (0.16 * L31) + (0.16 * L41) + (0.16 * L51) + (0.2 * L61)); \quad (12)$$

$$O2 = ((0.16 * L12) + (0.16 * L22) + (0.16 * L32) + (0.16 * L42) + (0.16 * L52) + (0.2 * L62)); \quad (13)$$

$$O3 = ((0.16 * O13) + (0.16 * O23) + (0.16 * O33) + (0.16 * O43) + (0.16 * O53) + (0.2 * O63)); \quad (14)$$

$$O4 = ((0.16 * O14) + (0.16 * O24) + (0.16 * O34) + (0.16 * O44) + (0.16 * O54) + (0.2 * O64)); \quad (15)$$

Equations (12)–(15) gives the calculations for image blending pixel by pixel.

$L11$ is the deconvolved Lucy RC image for MRI 1 and with PSF1, $L21$ is the deconvolved Lucy RC image for MRI 2 and with PSF1 and so on, $L12$ is the deconvolved Lucy RC image for MRI 1 and with PSF2 and $L22$ is the deconvolved Lucy RC image for MRI 2 and with PSF2 and so on.

$O1$ is the final blended image with all six PSF functions over near equal weightage and so on is $O2$ – $O4$.

To measure the effectiveness of deconvolution plus blending scheme, the MSE for all blended resultant images [16] for MRI and CT samples are calculated and a comparison is done with Lucy RC deconvolved images with respect to blended images. Table 7 shows the comparison and gives a minimum 10% improvement in the recovered images if they are blended after deconvolution (the worst case is taken where minimum 10% improvement is observed).

Table 7 MSE comparison blended versus original—worst case

S. No.	Test images	Blended after Lucy	Lucy PSF1	% Error reduction
1	mri1_original	0.04165	0.04646	10.3587
2	mri2_original	0.00418	0.00489	14.5689
3	mri3_original	0.00472	0.00568	16.8567
4	ct1_original	0.00925	0.01053	12.1764
5	ct2_original	0.00690	0.00801	13.8773

3 Result

The error calculation done with original recovery through various PSFs varied over linear scale gives a good approximation but still with a significant MSE and the MSE calculated with blending the recovered images gives significant improvement as given in Table 7. These results are measured in a simulation environment using MATLAB tool over random available CT and MRI image samples. The standard mean square error formula is used for error comparison.

Figure 12 shows the resultant deblurred images obtained after blending operation which is applied to Lucy RC deconvolved images.

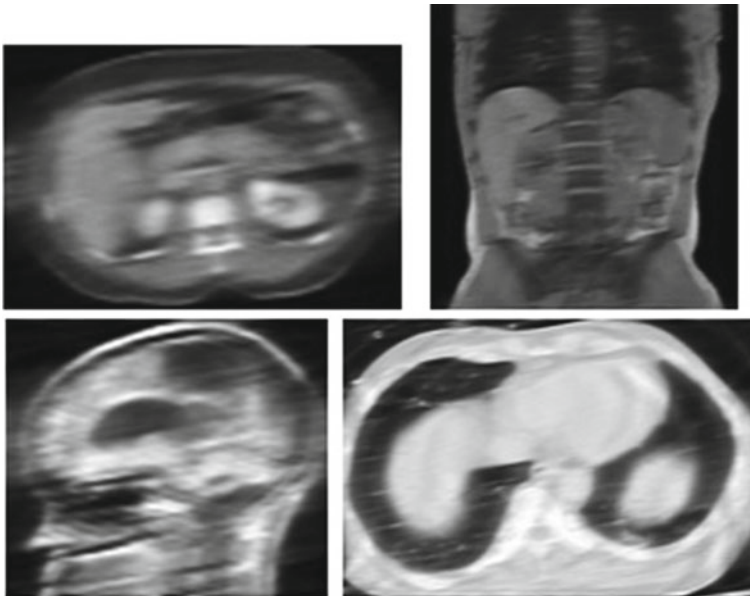


Fig. 12 Deblurred images post blending

4 Conclusions

This has been evidently proven that though Lucy RC provides a good deblurred images where motion blurring is present still there is scope of significant improvement which here is obtained through alpha plane blending done on deblurred images obtained from linearly varied PSF values used with Lucy RC deconvolution. The blending factor in compositing the images has been used as approximately equal weightages. The enhancement in deblurring of MRI and CT images is clearly notified after involving the image blending on deconvolved images. Thus, the work has a strong conviction to contribute advancements in medical imaging through this novel idea of varied PSF deconvolution followed by alpha plane blending.

References

1. Fowler, B., El Gamal, A., Yang, D.X.D.: A CMOS area image sensor with pixel-level A/D conversion. In: 41st ISSCC, IEEE International Conference (1994)
2. Mane, A.S., Pawar, M.M.: Removing blurring from degraded image using blind deconvolution with canny edge detection technique. *Int. J. Innov. Res. Adv. Eng. (IJIRAE)* **1**(11) (2014). ISSN: 2349–2163
3. Lil, W., Meunier, J., Soucy, J.-P.: A 3D adaptive Wiener filter for restoration of SPECT images using MRI as reference images. In: Proceedings of the 27th IEEE Conference in Medicine and Biology, Shanghai, China, 1–4 Sept 2005
4. Jiunn-Lin, W., Chang, C.-F., Chen, C.-S.: An adaptive Richardson-Lucy algorithm for single image Deblurring using local extrema filtering. *J. Appl. Sci. Eng.* **16**(3), 269276 (2013)
5. Li, B., Zhan, Z.: Research on motion blurred image restoration. In: 5th International Congress on Image and Signal Processing (CISP) (2012)
6. Sharma, P., Sharma, S., Goyal, A.: An MSE (mean square error) based analysis of deconvolution techniques used for deblurring/restoration of MRI and CT Images. In: ACM—ICPS Proceedings ISBN No 978-1-4503-3962-9 (2016)
7. Dong, B., Xie, S.: A method of point spread function estimation for wideband electromagnetic distribution. In: Asia-Pacific International Symposium Electromagnetic Compatibility (APEMC) (2016)
8. Schretter, C., Bundervoet, S., Blinder, D.: Ultrasound imaging from sparse RF samples using system point spread functions. *IEEE Trans Ultrason. Ferroelectr. Freq. Control* **65**. <https://doi.org/10.1109/tuffc.2017.2772916>
9. Angelis, G.I., Gillam, J.E., Kyme, A.Z.: Modelling the motion dependent point spread function in motion corrected small animal PET imaging. In: Nuclear Science Symposium, Medical Imaging Conference and Room-Temperature Semiconductor Detector Workshop (NSS/MIC/RTSD) 2016
10. Yim, P.J., Choyke, P.L., Summers, R.M.: Gray-scale skeletonization of small vessels in magnetic resonance angiography. *IEEE Trans. Med. Imaging* **19**(6) (2000)
11. Pizurica, A., Philips, W.: A versatile wavelet domain noise filtration technique for medical imaging. *IEEE Trans. Med. Imaging* **22**(3) (2003)
12. Guo, D., Wu, Y., Shitz, S.S.: Estimation in Gaussian noise: properties of the minimum mean-square error. *IEEE Trans. Inform. Theor.* **57**(4) (2011)
13. Wang, W., Ng, M.K.: A variational method for multiple-image blending. *IEEE Trans. Image Process.* **21**(4) (2012)
14. Pandey, A., Pati, U.C.: Development of saliency-based seamless image compositing using hybrid blending (SSICHB). *IET Image Process.* **11**(6) (2017)

15. Igarashi, K., Yanagisawa, M., Togawa, N.: Image synthesis circuit design using selector-logic-based alpha blending and its FPGA implementation. In: 11th IEEE International Conference ASIC (ASICON) (2015)
16. Bao, J., Fan, J., Hu, X.: An effective consistency correction and blending method for camera-array-based microscopy imaging. In: WSSIP International Conference on Systems, Signals and Image Processing (2017)

Biometric Authentication-Based Data Encryption Using ECG Analysis and Diffie–Hellman Algorithm



Archana Bhardwaj, Shikha Chaudhary and Vijay Kumar Sharma

Abstract Data security is becoming the prime concern these days. In the daily and day-to-day activities, data is required to be shared, whether it is a simple chat message or even the crucial business message or mail. In this paper, the unique concept of using the ECG reports as the authentication medium is proposed. First, the ECG report of the person is analyzed using sin charts and compared with the database. Before processing to the next phase, the user has to validate the transaction ID and SHA key generated using the. The next phase of the message transfer is further secured by providing the transaction id and encryption key. After that, the fingerprints of the users are validated again using the SHA-based concept of validation, and then to transfer data, Diffie–Hellman algorithm is used. In the dissertation, the double level security is maintained and implemented by making use of ECG reports and using the fingerprints. From the previous approaches and base papers, we have speed up the comparison process and reduce the time involved by making use of SHA-based analysis, and the double authentication helps us to deal with hackers in the better ways as compared to the previous approaches as now double efforts will be requirement by the hackers to break through the whole process. The uniqueness of the ECG report will act as uniquely validating the identity of two persons.

Keywords ECG analysis · Heartbeat charts · Message transfer

A. Bhardwaj (✉) · S. Chaudhary · V. K. Sharma
Department of CSE, Rajasthan Institute of Engineering and Technology, Jaipur, Rajasthan, India
e-mail: archanabhardwaj37@gmail.com

S. Chaudhary
e-mail: Shikha.chaudhary18@gmail.com

V. K. Sharma
e-mail: Vijaymayankmudgal2008@gmail.com

© Springer Nature Singapore Pte Ltd. 2019
Y.-C. Hu et al. (eds.), *Ambient Communications and Computer Systems*,
Advances in Intelligent Systems and Computing 904,
https://doi.org/10.1007/978-981-13-5934-7_46

1 Introduction

Electrocardiography is the route toward recording the electrical activity of the heart over some vague time period using anodes on the skin. These anodes perceive the minor electrical changes on the skin that rise up out of the heart muscle's electrophysiological, case of depolarizing and repolarizing in the midst of each heartbeat. It is more often than not cardiology test is performed [1].

In a customary 12-lead ECG, ten terminals are attached to the patient's limbs and on the surface of the chest. The general degree of the heart's electrical potential is then assessed from 12 particular focuses ("leads") and is recorded over some stretch of time (commonly ten seconds). Thusly, the general degree and course of the heart's electrical depolarization are gotten consistently through the cardiovascular cycle. The diagram of voltage versus time made by this noninvasive restorative system is suggested as an electrocardiogram [1].

In the midst of each heartbeat, a strong heart has a systematic development of depolarization that starts with pacemaker cells in the sinoatrial center, spreads out through the chamber, experiences the atrioventricular center down into the load of His and into the Purkinje fibers, spreading down and to the other side all through the ventricles. This sorted out case of depolarization offers climb to the trademark ECG following. To the clinician, an ECG passes on a great deal of information about the structure of the heart and the limit of its electrical conduction system.

The fingerprint is made of a few edges and valley on the surface of the finger, which is unique to each human. Edges are the upper skin layer segments of the finger and valleys are the lower divides. The edges outline two particular centers: edge endings—where the edges end, and edge bifurcations—where the edges split into two. The edges and wrinkles are responsible for the uniqueness of a fingerprint [2].

To get the surface of the fingerprint for checking in the midst of the unmistakable evidence of clients, new progressions are arranged with devices, for instance, optical and ultrasound. There are two essential figurings which are used to see fingerprints—points of interest coordinating and configuration coordinating.

The three fundamental examples of fingerprint edges are the curve, circle, and whorl:

1. Arch: The edges enter from one side of the finger, climb in within encircling a roundabout section, and after that leave the contrary side of the finger.
2. Loop: The edges enter from one side of a finger, shape a twist, and a while later exit on that same side.
3. Whorl: Edges shape circularly around a basic issue on the finger.

The ECG database is used for storing details like user name, password, ECG report, email ID, and photo. The datasets for ECG are available for the implementation purpose and in the research, we have used the ECG reports from such databases; some of them are as follows:

[Class 1; core] ANSI/AAMI EC13 Test Waveforms. These 10 short records are controlled by the present American National Standard for testing different contraptions that measure heart rate.

[Class 1; core] European ST-T Database. The makers of this database and the European Society of Cardiology have contributed each one of the 90 two-hour records of this database completely.

[Class 2; core] BIDMC Congestive Heart Failure Database. It contains whole bargain ECGs (around 20 h each) from 15 subjects with bona fide CHF (NYHA class 3–4).

[Class 2] ECG impacts of dofetilide, moxifloxacin, dofetilide + mexiletine, dofetilide + lidocaine and moxifloxacin + diltiazem in healthy subjects.

2 Related Work

The data of Kumari [3] are secured and advanced. Security is about the protection of focal points. Information security implies securing the data from unauthorized access. Cryptography is evergreen for data security. Cryptography guarantees security to customers by offering encryption of information and approval of various customers. Weight is the path toward diminishing the number of bits or bytes anticipated that would address a given plan of information. It helps in saving more information.

Cryptography is the standard technique for sending basic data. There are various cryptographic strategies available, and among them, AES is the most powerful method. The circumstance of the present day of data security framework joins mystery, validity, trustworthiness, and non-disavowal. The security of correspondence is a basic issue on the World Wide Web. It is about characterization, dependability, approval in the midst of access, or changing of mystery internal archives.

To secure the data, the pressure is utilized in light of the fact that it utilizes less plate space (saves cash), and more data can be exchanged by means of the web. It expands the speed of data exchanged from circle to memory [3].

Dalvi and Wakde [1] visual cryptography is another creation that is making this stream the moreover solidifying. In this work, shares are framed in encryption and decryption side, and by utilizing serialization check which shares are ambiguous, and in decoding side, every last mixed up share is consolidated and packaged in the main pictures.

For each offer, they give the key, so it gives more prominent security. It is the most secured idea inferable from encrypted shares at various levels using the keys without which one can never decrypt the picture [1].

Agrawal Pal [2] cryptography is an essential bit of data security instrument over the web. Cryptography makes data distorted to an unapproved individual. Cryptography surrenders arrangement and keeps trustworthiness to true blue clients. Cryptography is the examination of numerical strategies related to data security points, for instance, mystery, data uprightness, substance validation, and data confirmation. The commitment to an encryption method is typically called the plain substance, and the yield the figure content. A cryptographic algorithm works in a blend with a keyword, number, or expression to encrypt the plaintext. The same plaintext encrypts different figure contents with multiple keys. In this paper, another approach which gives more

veritable and strong encryption and decryption process for short message has been proposed [2].

The whole work of Agrawal Patankar [4] derived and proposed the course of action that will give an elective security display than SSL and digital to keep up security on the intranet. SSL requires HTTPs protocol, whereas the proposed course of action does not require any kind of protocol relationship in applications. In addition, the key use of the proposed plan is a mix of security approach with nearby system-based applications. It may be steady in such applications where insurance, authentication, and trustworthiness all are basic requests. Solid security approaches are required to give a legitimate level of security procedures to achieve arrangement, authentication, and respectability. To take care of mystery, digital envelope, which is the mix of the encrypted message and signature with the encrypted symmetric key, is moreover used. This investigation paper proposed a crossover model to achieve mystery, authentication, and genuineness in the same way [4].

A covert operative that breaks into the message will reestablish a stupid message. Encryption and decryption are uncommon contrasted with different strategies for covering the ramifications of a message from intruders in a system circumstance. The proposed system has been planned and made with straightforward joining and change to take the full favored point of view of future advances. There are a couple of restrictions in the present system to which plans will be given as a future change, for instance, a modest number of keywords use just keyword monoalphabetic substitution algorithm and system data transmission. In the future, a public-key encryption plan will be installed in the secure informing system.

In [5], security analysis of an encryption algorithm which targets to encrypt the data with the help of ECG signals and chaotic functions was performed by making use of the logistic map in case of the text encryption and Henon map in case of the image encryption. In the proposed algorithm, text and image data are simultaneously encrypted.

In [6], the authors proposed a fast, secure, and robust scheme for the digital image encryption by making use of the chaotic system of Lorenz, 4D hyper-chaotic system, and the Secure Hash Algorithm SHA-1. The encryption process comprises the three layers, namely, sub-vectors, confusion, and two-diffusion process. This results in better security and good encryption speed.

Tian et al. [7] propose the concept of the public-key encryption scheme, which is provable chosen-ciphertext (CCA) security that relies on the gap computational linear Diffie–Hellman assumption in the standard model. This scheme is quite efficient and results in the implementation of tight security.

3 Proposed Work

The base papers are mentioned in [5–7], and in order to improve the performance and efficiency of the cryptosystem, in the proposed work, the research is extended toward the double authentication using the ECG and fingerprint and the usage to

SHA hash codes for the faster comparison and adding the another layer of security in the whole system of secure text and image transfer. The proposed work functions in the two phases: first is the user validation process and the other one is message sending process.

Before starting with the process, we will discuss in brief the algorithms which are involved in the proposed work. The proposed work revolves around the three main algorithms:

- a. **Blowfish:** Blowfish is a symmetric key block cipher and it has a 64-bits block size and the key which is involved in it is a variable key and its length varies from 32 to 448 bits. The Blowfish is the 16 rounds-based Feistel cipher, and it is a good algorithm which is used for the encryption purpose, and we have used this algorithm for the encryption of the images of the users which are involved in the encryption process.
- b. **SHA:** SHA stands for Secure Hash Algorithm and is referred as the cryptographic hash function which receives input and produces the 40 digits long hexadecimal string. In the proposed algorithm, extracted characters from this hexadecimal code will be used as the key for encrypting the images using the Blowfish algorithm. And this SHA code is generated the corresponding ECG report files.
- c. **Diffie–Hellman:** Diffie–Hellman is the concept of the digital encryption that makes use of the numbers raised to some of the specific powers in order to produce the decryption keys which are on the basis of integral parts that are never directly transmitted, and it will make the hackers’ task difficult to break the code mathematically.

3.1 User Validation Process

In the user validation, we will select the photo of the user 1 and user 2 involved in the sending process.

The ECG records then validated in the user’s databases, and the username and details are fetched. After using the encryption algorithms, the image is encrypted and sent.

Along with that, a unique transaction key and the encryption key are saved in the database. The encryption key will act us the key to encrypt images (Fig. 1).

Steps in the User Validation Algorithm



Fig. 1 Block diagram of user’s validation

Step 1: Read user 1 and user 2 ECG reports. And the SIN graph is generated for the diagrammatic visualization of the ECG reports. The ECG reports are taken in the text files and on the basis of the data contained in the text files this visualization of the graph is done.

Step 2: Search for ECG Reports in the database in order to verify the authorized users. If the ECG report entry is found, then the images of the related users are displayed.

Step 3: Now the role of the Blowfish algorithm will come; it will perform the encryption of the images of the users and the keys which are used for the encryption of the images are generated using the SHA algorithm, and the key generation is on the basis of the SHA code which is generated using the ECG report file.

Step 4: After all the process, the details are then stored in the database.

3.2 Message Exchange

First, we will enter the transaction key and key for encrypting image. The entries are validated from databases, and then images are decrypted and shown on screen; then only we can proceed to message sending step (Fig. 2).

Now, fingerprints are input, and unique random number on the basis of the fingerprints is generated, and after that, the message is then sent or exchanged using the Diffie–Hellman algorithm. This whole process can be explained in the following steps:

Step 1: Fingerprints of the user are provided as an input and unique keys are generated corresponding to the fingerprints, which will be considered as the private keys for the message transferring transaction.

Step 2: And the common public-key is used.

Step 3: After that, the Diffie–Hellman algorithm is used for the simple text message transfer and if we talk about the images further the SHA algorithm will play role and the SHA code corresponding to the image file is transferred and then stored in the database and then the receiver has to enter that SHA code to get access to the image shared.



Fig. 2 Block diagram for message exchange

4 Implementation

The implementation is done in Java language by making the use of IDE Eclipse, which enables us to program the Java Code very easily, and the designing part of the proposed work is done in Java Foundation classes often known as Swings. From the code part, the database related to ECG analysis is created using the Microsoft Access 2007, and related data is stored in this database.

5 Results and Analysis

The SHA code which is related to the ECG is generated in which the ECG file is read and the SHA code which is a 40 characters Hexadecimal value is used and every fifth character of the hexadecimal code is used to generate an eight characters key related to the keys of both the users. The information will be stored in the database with the following information. The database, which we have used, contains the two tables, figdata, and encdata.

1. Figdata table: This table is used for storing the fingerprint of the persons. This table consists of the username, five fingerprint information, and ECG file for the person.
2. Encdata table: This table is used for storing the encryption related information; it contains the username, encryption keys, and details of the users involved in the data transmission.

The ECG reports of the users are browsed as shown in Fig. 3, and the Sin graph of the ECG reports is generated to show the ECG report in the graphical pattern.

In the case of sending the message, the form is opened in which transaction key and the encryption keys are required to be verified and entered by the user and after that, the users' interaction images will be shown.

In the next level, we have to verify the fingerprint and send the message from user 1 to user 2 using the Diffie–Hellman algorithm (Fig. 4).

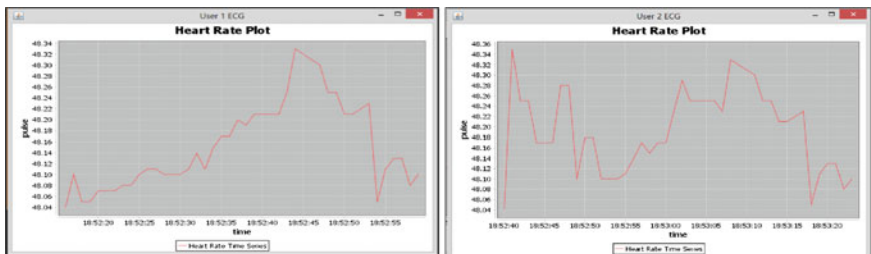


Fig. 3 ECG graph for user 1 and user 2 ECG reports

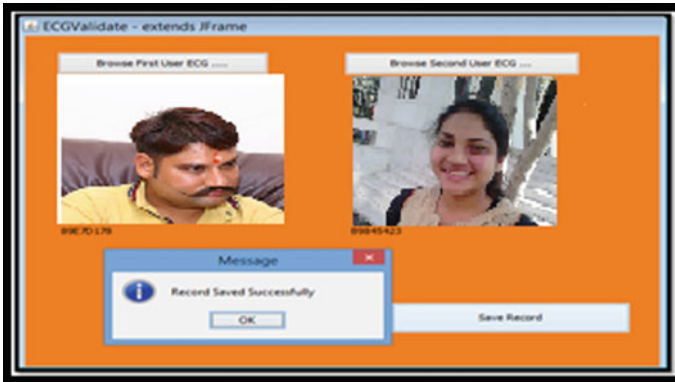


Fig. 4 Message sending (photos used are of authors and her relative)

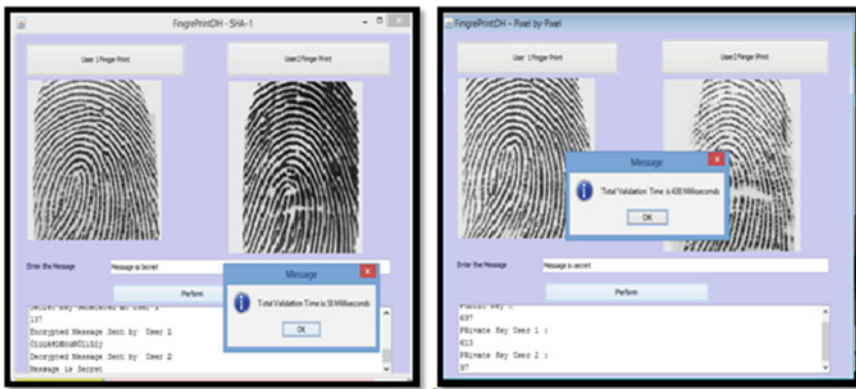


Fig. 5 Fingerprint comparison proposed and pixel by pixel

Table 1 Time comparison table for case II

	Pixel by pixel approach	Proposed work
User 1 and user 2	672 ms	63 ms

The user 1 is first required to enter the fingerprint details, and then the fingerprint is validated in the database by making use of the SHA algorithm concept, and after that, the user two is required to submit the fingerprint which is again confirmed in the database using the SHA-1 algorithm, and then the Diffie–Hellman algorithm is followed for the process (Fig. 5; Table 1).

Figure 6 shows the comparison graph between two approaches, using the data which we have obtained in the analysis, where ms stands for milliseconds.

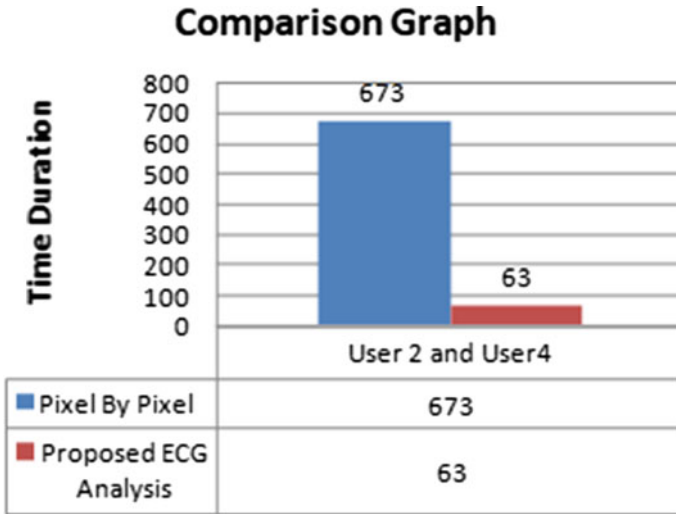


Fig. 6 Fingerprint comparison proposed and pixel by pixel

6 Conclusion

The security of a business’ IT systems is central and ought to never be ignored. The security is the main concern in the transaction; the proposed dissertation makes use of the ECG-based analysis in order to increase the security by encrypting the photos of a user, which are interacting in the session for sending the data using the SHA-based encryption key, on which is generated corresponding to the ECG reports, and the encrypted images are first decrypted at the time of the sending of the message, and after the encryption and transaction key are validated, the message is further transferred. Using the SHA in the image comparison will speed up the process of image comparison. Thus, both security and speed have been enhanced. Thus, we can say that our proposed implementation provides a better way to share the data securely. In the further studies, we will like to extend our research to use the real-time password like live pictures, video, and retina verification concepts for sharing the file to enhance the security in the suggested framework further.

7 Statement of Consent

I am providing the statement of consent for publishing the image in Fig. 4. The images used in Fig. 4 of this paper are of the authors of this paper, i.e., Archana Bhardwaj and her brother Shanker Bhardwaj.

References

1. Dalvi, G.D., Wakde, D.G.: Facial images authentication in visual cryptography using sterilization algorithm. In: 2nd International Conference for Convergence in Technology (I2CT) (2017)
2. Agrawal, E., Pal, P.R.: A new and more authentic cryptographic based approach for securing short message. *Int. J. Adv. Res. Comput. Sci.* **8** (2017)
3. Kumari, S.: A research paper on cryptography encryption and compression techniques. *Int. J. Eng. Comput. Sci. IJECS* **6** (2017)
4. Agrawal, A., Patankar, G.: Design of hybrid cryptography algorithm for secure communication. *Int. Res. J. Eng. Technol. IRJET* **3** (2016)
5. Gençoğlu, M.T.: Mathematical cryptanalysis of personalized information encryption using ECG signals with chaotic functions. In: 2017 International Conference on Computer Science and Engineering (UBMK), Antalya (2017)
6. Slimane, N.B., Bouallegue, K., Machhout, M.: A novel image encryption scheme using chaos, hyper-chaos systems and the secure Hash algorithm SHA-1. In: 2017 International Conference on Control, Automation and Diagnosis (ICCAD), Hammamet (2017)
7. Tian, F., Xue, H., Haiyang, X.: A secure public key encryption from computational linear Diffie-Hellman problem. In: 2012 Eighth International Conference on Computational Intelligence and Security, Guangzhou (2012)
8. Shah, D.: Digital security using cryptographic message-digest algorithm. *Int. J. Adv. Res. Comput. Sci. Manage. Stud. IJARCSMS* **3** (2015)
9. Javheri, S., Kulkarni, R.: Secure data communication and cryptography based on DNA based message encoding. *Int. J. Comput. Appl. IJCA* **98** (2014)
10. Chaudhari, M.P., Patel, S.R.: A survey on cryptography algorithms. *Int. J. Adv. Res. Comput. Sci. Manage. Stud. IJARCSMS* **2** (2014)
11. Durairajan, M.S., Saravanan, R.: Biometrics based key generation using Diffie Hellman key exchange for enhanced security mechanism. *Int. J. Chem. Tech. Res.* **6** (2014)
12. Arora, S., Singh, L., Kumar, A.: Comparison of images using MIAC algorithm. KIET, IJICI (2014)
13. Mushtaque, M.A., Dhiman, H., Hussain, S.: A hybrid approach and implementation of a new encryption algorithm for data security in cloud computing. *Int. J. Electron. Electr. Eng.* (2014)
14. Aggarwal, S., Goyal, N., Aggarwal, K.: A review of comparative study of MD5 and SHA security algorithm. *Int. J. Comput. Appl. IJCA* **104** (2014)
15. Singh, P., Singh, K.: Image encryption and decryption using Blowfish algorithm in Matlab. *Int. J. Sci. Eng. Res.* **4** (2013)
16. Al-Hazaim, O.M.A.: A new approach for complex encrypting and decrypting data. *Int. J. Comput. Netw. Commun. IJCNC* **5** (2013)

Author Index

A

Agarwal, Shiv Kumar, 125
Agrawal, Sharad, 135
Alam, Shadab, 387
Anjali, T., 161
Anurag, Priyanka, 461
Arora, Sangeeta, 265

B

Balamurugan, B. J., 411
Bansal, Abhay, 257
Barde, Chetan, 47
Bharathi, M. A., 401
Bhardwaj, Archana, 523
Bharti, Sunil Kumar, 237

C

Chaudhary, Shikha, 523
Choubey, Arvind, 47

D

Dadhich, Ajay, 37
Das, Madhabananda, 175
Debnath, Subir K., 285
Deegwal, J. K., 37
Dembla, Deepak, 103
Deshpande, D. S., 11
Deshpande, S. P., 11
Dev, Chethan, 161
Dubey, Anil Kumar, 113

Dubey, Ashwani Kumar, 509
Dubey, Sanjay Kumar, 237
Dutta, Ajoy K., 285
Dwivedi, Rakesh Kumar, 449

G

Garg, Ritu, 421
Goutham, S., 377
Goyal, Ayush, 509
Gupta, Purnima, 353
Gupta, Shubhi, 55
Gupta, Suneet Kumar, 3

H

Hanmandlu, Madasu, 187

J

Jain, Alok Kumar, 473
Jain, Aman, 275
Jishag, A. C., 329

K

Kar, Amitava, 285
Karande, Shubham, 319
Kasdekar, Dinesh Kumar, 135
Kathpal, Chesta, 421
Kaushal, Sakshi, 295
Kohli, Rashi, 55
Kumar, Anil, 449
Kumar, Anuj, 487

Kumar, Avadhesh, 257
 Kumar, Deepak, 223
 Kumar, Kripa, 161

L

Lal, Shyam, 473

M

Mahajani, Abhishek, 339
 Maheshwari, Ekansh, 307
 Mahto, Santosh Kumar, 47
 Malhotra, Ruchika, 433
 Mallik, Pankhurhi, 307
 Maria, Isaac, 339
 Menon, Maya, 329
 Mewada, Pradeep, 67
 Mishra, Bhabani Sankar Prasad, 175
 Mishra, Shambhavi, 25
 Mohan, Suraj, 329
 Munjal, Geetika, 187
 Murali, B. J., 411

N

Nagwani, N. K., 195
 Nair, Lekshmi S., 329
 Navamani, T. M., 247
 Negi, Richa, 209

P

Pai, Smitha N., 461
 Palathil, Arjun, 161
 Pal, Harendra, 135
 Pandey, Adesh Kumar, 265
 Pandey, Manjusha, 307
 Pandya, Vinay, 339
 Panicker, Vinitha, 161
 Parnami, Pooja, 275
 Pattanaik, Anshuman, 175
 Prasad, Ritu, 67, 367
 Puri, Shalini, 497

R

Raina, Priya, 295
 Rakesh, Vishnu, 329
 Ranjan, Abhishek, 247
 Ranjan, Prakash, 47
 Rautray, Siddharth, 307
 Roy, Chandrima, 307

S

Sagnika, Santwana, 175
 Sahay, Sanjay K., 149
 Saini, Gurpreet Singh, 237
 Samad, Abdus, 387
 Samdani, Preeti, 37
 Saurabh, Praneet, 67, 367
 SenGupta, Ishuita, 449
 Sen, Priya, 367
 Shabu, Vaisakh, 329
 Sharma, Anjali, 433
 Sharma, Anuj, 487
 Sharma, Ashu, 149
 Sharma, Deepak, 339
 Sharma, M. M., 37
 Sharma, Navneet, 275
 Sharma, Poonam, 509
 Sharma, Sarika, 223
 Sharma, Vijay Kumar, 523
 Shuaib, Mohammed, 387
 Shyamala, L., 319
 Siddiqui, Shams Tabrez, 387
 Siji Rani, S., 377
 Singhal, Abhishek, 257
 Singh, Jaiteg, 79
 Singh, Nar, 93
 Singh, Nitin, 209
 Singh, Pavan Kumar, 209
 Singh, Pawan, 25
 Singh, Satya Prakash, 497
 Singla, R. K., 487
 Sinha, Rashmi, 47
 Srivastava, Akhilesh Kumar, 3
 Srivastava, Sangeet, 187
 Srivastava, Saumya, 93
 Suresh, L., 401

T

Thakare, V. M., 11
 Thirusangu, K., 411
 Tiwari, Anil Kumar, 25
 Tiwari, Ashima, 103
 Tyagi, Rajesh Kumar, 353
 Tyagi, Sourabh, 461

U

Upadhyay, Kamal K., 93

V

Verma, Deepak Kumar, [353](#)

Verma, Kesari, [195](#)

Verma, Shrish, [195](#)

Vidhyotma, [79](#)

Vinayak Varma, N., [329](#)

Y

Yadav, Surendra, [125](#)

Z

Zaidi, Uroosa, [67](#)