

Lecture Notes on Data Engineering  
and Communications Technologies 9

Nabendu Chaki  
Agostino Cortesi  
Nagaraju Devarakonda *Editors*



# Proceedings of International Conference on Computational Intelligence and Data Engineering

ICCIDE 2017

 Springer

# **Lecture Notes on Data Engineering and Communications Technologies**

Volume 9

## **Series editor**

Fatos Xhafa, Technical University of Catalonia, Barcelona, Spain  
e-mail: [fatos@cs.upc.edu](mailto:fatos@cs.upc.edu)

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

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

More information about this series at <http://www.springer.com/series/15362>

Nabendu Chaki · Agostino Cortesi  
Nagaraju Devarakonda  
Editors

# Proceedings of International Conference on Computational Intelligence and Data Engineering

ICCIDE 2017

 Springer

*Editors*

Nabendu Chaki  
Department of Computer Science and  
Engineering  
University of Calcutta  
Kolkata, West Bengal  
India

Nagaraju Devarakonda  
Department of Information Technology  
Lakireddy Bali Reddy College of  
Engineering  
Mylavaram, Andhra Pradesh  
India

Agostino Cortesi  
DAIS  
Ca' Foscari University of Venice  
Venice  
Italy

ISSN 2367-4512                      ISSN 2367-4520 (electronic)  
Lecture Notes on Data Engineering and Communications Technologies  
ISBN 978-981-10-6318-3              ISBN 978-981-10-6319-0 (eBook)  
<https://doi.org/10.1007/978-981-10-6319-0>

Library of Congress Control Number: 2017949505

© Springer Nature Singapore Pte Ltd. 2018

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, express 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.

Printed on acid-free paper

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

# Preface

The First International Conference on Computational Intelligence & Data Engineering (ICCIDE 2017) took place on 14 and 15 July 2017 in Lakireddy Bali Reddy College of Engineering (LRBCE), Autonomous at Mylavaram, Krishna District, Andhra Pradesh, India. Ca' Foscari University of Venice, Italy and the University of Calcutta, India were the academic partners for ICCIDE 2017.

ICCIDE is conceived as a forum for presenting and exchanging ideas, which aims at resulting in high-quality research work in cutting edge technologies and most happening areas of computational intelligence and data engineering.

The conference solicited latest research ideas on computational intelligence and data engineering, thus inviting researchers working in the domains of machine learning, Bayesian network, computational paradigms and computational complexity, rough sets, semantic web, knowledge representation, data models, ubiquitous data management, mobile databases, data provenance, workflows, scientific data management and security.

The sincere effort of the program committee members coupled with indexing initiatives from Springer have drawn a large number of high-quality submissions from scholars all over India and abroad. A thorough peer-review process has been carried out by the PC members and by external reviewers. While reviewing the papers, the reviewers mainly looked at the novelty of the contributions, besides the technical content, the organization and the clarity of the presentation. The entire process of paper submission, review and acceptance process was done electronically.

The Technical Program Committee eventually could identify 30 papers for publication out of 311 submissions. The resulting acceptance ratio is 9.65%, which is healthy and quite good in the very first year of the conference. The program also includes four invited talks, by Dr. Venu Govindaraju (University at Buffalo, USA), Dr. D. Janakiram (IIT Madras, India), Prof. Stephen Marsland (Massey University, New Zealand) and Prof. Agostino Cortesi (Ca' Foscari University, Italy).

We eventually extend our gratitude to all the members of the Program Committee and the external reviewers for their excellent and time-bound review work. We thank all the sponsors, who have come forward towards organizing this

symposium. We submit our indebtedness to Sri Lakireddy Bali Reddy, Chairman; Sri Lakireddy Jaya Prakash Reddy, Co-Chairman; and Sri Lakireddy Prasada Reddy, Vice-Chairman. We are thankful to the entire management of LBRCE for their patronage and continual support to make the event successful. We appreciate the initiative and support from Mr. Aninda Bose and his colleagues in Springer Nature for their strong support towards publishing this volume in the Lecture Notes on Data Engineering and Communications Technologies (LNDECT) series of Springer Nature. Finally, we thank all the authors without whom the conference would not have reached the expected standards.

Kolkata, India  
Venice, Italy  
Mylavaram, India  
May 2017

Nabendu Chaki  
Agostino Cortesi  
Nagaraju Devarakonda

# Program Committee

Kavita Agarwal, Integral University, Lucknow  
Shish Ahmad, Integral University, Lucknow  
Yahia Sabri Mustafa Al-Halabi, PSUT  
Mohammad Al-Shamri, King Khalid University  
Andrea Albarelli, Department of Computer Science, University “Ca’ Foscari” in Venice  
Rashid Ali, Aligarh Muslim University  
Annappa B., National Institute of Technology Karnataka, Surathkal  
Debdutta Barman Roy, Calcutta Institute of Engineering and Management  
Debotosh Bhattacharjee, Jadavpur University  
Bhogeswar Borah, Tezpur University  
Narayan C. Debnath, Winona State University, Winona  
Nabendu Chaki, University of Calcutta  
Rituparna Chaki, West Bengal University of Technology  
Samiran Chattopadhyay, Jadavpur University  
Sankhayan Choudhury, University of Calcutta  
Agostino Cortesi, Ca’ Foscari University of Venice  
Ranjan Dasgupta, Tata Consultancy Services Limited  
Nagaraju Devarakonda, Lakireddy Bali Reddy College of Engineering, Mylavaram, Vijayawada  
Sushil Kumar Dohare, Jawaharlal Nehru University, Delhi  
Asif Ekbal, Indian Institute of Technology Patna  
Faten F. Kharbat, Al Ain University of Science and Technology, Abu Dhabi  
Rudolf Fleischer, Fudan University  
Kumaravelan G., Pondicherry University, Karaikal  
Mallikarjuna Rao G., RCI, DRDO, Hyderabad  
Venu Govindaraju, State University of New York, Buffalo  
Sagar Gurram, Alghurair University, Dubai  
D.S. Guru, University of Mysore  
Raju Halder, Indian Institute of Technology Patna  
Meena Jha, CQ University



Sanjay Jha, Central Queensland University, Sydney, Australia  
 Sudan Jha, Kalinga Institute of Industrial Technology University, Bhubhaneswar  
 Sukhendu Kanrar, University of Calcutta  
 Raees Ahmad Khan, Babasaheb Bhimrao Ambedkar University, Lucknow  
 Chawngsangpuii Khiangte, Mizoram University  
 Andrea Marin, Università Ca' Foscari Venezia  
 Pinaki Mitra, Indian Institute of Technology Guwahati  
 M. Hamdan Mohammad, Yar Mouk University, Irbid, Jordan  
 Khanum Mohammadi, Integral University, Lucknow  
 Ghose Mrinal K., Sikkim Manipal Institute of Technology  
 H.S. Nagendra Swamy, University of Mysore  
 Ravishankar Nanduri, Lakireddy Bali Reddy College of Engineering  
 Prem Nath, Mizoram University  
 Ngoc Tu Nguyen, Missouri University of Science and Technology  
 Sandhya P., Pooja Bhagavat Memorial Mahajana Education Centre, Mysore  
 Shanthi Bala P., Pondicherry University, Karaikal  
 Carla Piazza, Università di Udine  
 Piotr Porwik, University of Silesia  
 Partha Pratim Ray, Sikkim University  
 Srn Reddy, Indira Gandhi Delhi Technical University for Women, Delhi  
 Khalid Saeed, Bialystok University, Bialystok, Poland  
 Anirban Sarkar, National Institute of Technology Durgapur, West Bengal, India  
 713209  
 Bidyutbiman Sarkar, Techno India  
 Vipin Saxena, Babasaheb Bhimrao Ambedkar University Lucknow  
 Soumya Sen, University of Calcutta  
 Sabnam Sengupta, Department of Computer Science, Jadavpur University, Kolkata  
 700032  
 Utpal Sharma, Tezpur University, Assam  
 Birmohan Singh, Sant Longowal Institute of Engineering and Technology(DU),  
 Longowal, Panjab.  
 Jyoti Prakash Singh, National Institute of Technology Patna  
 Karan Singh, Jawaharlal Nehru University, Delhi  
 Pradeep Kumar Singh, Jaypee University of Information Tecchnology  
 Noor Mahammad S.K., Indian Institute of Information Technology, Design and  
 Manufacturing, Kancheepuram, Tamil Nadu  
 T. Sobha Rani, University of Hyderabad  
 D.V.L.N. Somayajulu, National Institute of Technology, Warangal  
 Ramesh T., National Institute of Technology, Warangal  
 Suresh Thommandru, Lakireddy Bali Reddy College of Engineering  
 Umopathy Venugopal, Central Queensland University, Sydney, Australia  
 Veeraganesh Yalla, Faraday Future, Los Angeles  
 Ping Yu, University of Wollongong Australia

## **Additional Reviewers**

Agrawal, Kavita  
Al-Shamri, Mohammad  
Ali, Rashid  
B.V., Subba Rao  
Bala, Shanthi  
Bhattasali, Tapalina  
Borah, Bhogeswar  
Caiazza, Gianluca  
Calzavara, Stefano  
Chakraborty, Manali  
Chakraborty, Supriya  
Chakravarthy, Srinivasa  
Chepuri, Samson  
Datta, Soma  
Deb, Dipankar  
Deb, Novarun  
DeiRossi, Gian-Luca  
Dey, Ayan  
DiGiacomo, Francesco  
Dirisinapu, Lakshmisreenivasareddy  
Donepudi, Kavitha  
Dr. Bvb, Vijay  
Fleischer, Rudolf  
G., Varaprasad  
Ghosh, Ammlan  
Jha, Sudan  
Khanum, Mohammadi Akheela  
Khatua, Aparup  
Mahmood, AliMirza  
Mandal, Amit  
Marin, Andrea  
Martin, Nathaniel  
Mishra, Rakesh  
Mitra, Pinaki  
Patel, Birajkumar  
Porwik, Piotr  
Ragupathy, Rengaswamy  
Rao, Dr. K.V.R.  
S., Murugan  
S. Murugan  
SeshaSai, Dr. M. Srinivasa  
Shah, Dr. Saurabh

Silvestri, Claudio  
Singh, JyotiPrakash  
Spanò, Alvise  
Srinivas, Edara  
Srivastava, Navita  
Suryadevara, Nagender  
T., Sobha Rani  
Varma, Sandeep  
Venugopal, Umapathy  
Vuppala, Anil Kumar  
Zollo, Fabiana

# Contents

<b>Energy-Efficient Data Route-in-Network Aggregation with Secure EEDRINA</b> .....	1
B. Sujatha, Chala Tura Jilo and Chinta Someswara Rao	
<b>Prediction Models for Space Mean Speed on Urban Roads</b> .....	11
Mariangela Zedda and Francesco Pinna	
<b>SRAM Design Using Memristor and Self-controllable Voltage (SVL) Technique</b> .....	29
Naman S. Kumar, N.G. Sudhanva, V. Shreyas Hande, Mallikarjun V. Sajjan, C.S. Hemanth Kumar and B.S. Kariyappa	
<b>A Study on Certificate-Based Trust in MANETs</b> .....	41
K. Gowri Raghavendra Narayan, T. Srinivasa Rao, P. Pothu Raju and P. Sudhakar	
<b>User Interest Modeling from Social Media Network Graph, Enriched with Semantic Web</b> .....	55
Bansari Shah, Ananda Prakash Verma and Shailendra Tiwari	
<b>Swarm and Artificial Immune System-Based Intelligence Techniques for Geo-Spatial Feature Extraction</b> .....	65
Lavika Goel, Mallikarjun Swamy and Raghav Mantri	
<b>Network Intrusion Detection System to Preserve User Privacy</b> .....	85
Sireesha Rodda and Uma Shankar Rao Erothi	
<b>Hiding Encrypted Multiple Secret Images in a Cover Image</b> .....	95
Prashanti Guttikonda, Hemanthi Cherukuri and Nirupama Bhat Mundukur	
<b>Achieving Higher Ranking to Webpages Through Search Engine Optimization</b> .....	105
B. Swapna and T. Anuradha	

<b>Parallel Computing Algorithms for Big Data Frequent Pattern Mining</b> . . . . .	113
Subhani Shaik, Shaik Subhani, Nagaraju Devarakonda and Ch. Nagamani	
<b>Writer Identification System for Handwritten Gurmukhi Characters: Study of Different Feature-Classifer Combinations</b> . . . . .	125
Sakshi, Naresh Kumar Garg and Munish Kumar	
<b>Modified CSM for FIR Filter Realization</b> . . . . .	133
N. Udaya Kumar, K. Durga Teja, K. Bala Sindhuri and P. Rakesh	
<b>Toward Design and Enhancement of Emotion Recognition System Through Speech Signals of Autism Spectrum Disorder Children for Tamil Language Using Multi-Support Vector Machine</b> . . . . .	145
C. Sunitha Ram and R. Ponnusamy	
<b>Aadhaar Card Voting System</b> . . . . .	159
Lingamallu Naga Srinivasu and Kolakaluri Srinivasa Rao	
<b>An Innovative Security Model to Handle Blackhole Attack in MANET</b> . . . . .	173
MD. Sirajuddin, Ch. Rupa and A. Prasad	
<b>Knowledge Discovery via SVM Aggregation for Spatio-temporal Air Pollution Analysis</b> . . . . .	181
Shahid Ali	
<b>A Case Study in R to Recognize Human Activity Using Smartphones</b> . . . . .	191
Kella Bhanu Jyothi and K. Hima Bindu	
<b>Big Data Collection and Correlation Analysis of Wireless Sensor Networks Yielding to Target Detection and Classification</b> . . . . .	201
M. Giri and S. Jyothi	
<b>Time- and Cost-Aware Scheduling Method for Workflows in Cloud Computing Systems</b> . . . . .	215
G. Narendrababu Reddy and S. Phani Kumar	
<b>A Novel Statistical Feature Selection Measure for Decision Tree Models on Microarray Cancer Detection</b> . . . . .	229
Janardhan Reddy Ummadi, B. Venkata Ramana Reddy and B. Eswara Reddy	
<b>A Novel Region Segmentation-Based Multi-focus Image Fusion Model</b> . . . . .	247
Garladinne Ravikanth, K.V.N. Sunitha and B. Eswara Reddy	

**OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation . . . . .** 265  
 G. Leena Giri, Gerard Deepak, S.H. Manjula and K.R. Venugopal

**A Deep Autoencoder-Based Knowledge Transfer Approach . . . . .** 277  
 Sreenivas Sremath Tirumala

**Performance Preview on Image Super Resolution Using Wavelets Transform Based on Samples . . . . .** 285  
 Vicharapu Balaji, Ch. Anuradha, P.S.R. Chandra Murty and Grandhe Padmaja

**Handwritten Symbol Recognition Using Hierarchical Shape Representation Model Based on Shape Signature . . . . .** 293  
 M. Raja Babu, T. Gokaramaiah and A. Vishnuvardhan Reddy

**Determination of Optimal Clusters for a Non-hierarchical Clustering Paradigm K-Means Algorithm . . . . .** 301  
 T.V. Sai Krishna, A. Yesu Babu and R. Kiran Kumar

**Nonlinear Tensor Diffusion Filter Based Marker-Controlled Watershed Segmentation for CT/MR Images . . . . .** 317  
 S.N. Kumar, A. Lenin Fred, H. Ajay Kumar and P. Sebastian Varghese

**Improved Ensemble Methods to Solve Multi-class Imbalance Problem Using Adaptive Weights . . . . .** 333  
 K. Vasantha Kokilam and D. Ponmary Pushpa Latha

**Opinion Content Extraction from Web Pages Using Embedded Semantic Term Tree Kernels . . . . .** 345  
 Veerappa B. Pagi and Ramesh S. Wadawadagi

**Combining Fuzzy C-Means and KNN Algorithms in Performance Improvement of Intrusion Detection System . . . . .** 359  
 B. Sujata and P. Ravi Kiran Varma

**Author Index . . . . .** 371

## About the Editors

**Nabendu Chaki** is a Professor in the Department Computer Science and Engineering, University of Calcutta, Kolkata, India. Dr. Chaki did his first graduation in Physics from the legendary Presidency College in Kolkata and then in Computer Science and Engineering from the University of Calcutta. He completed Ph.D. in 2000 from Jadavpur University, India. He is sharing six international patents including four U.S. patents with his students. Prof. Chaki has been quite active in developing international standards for software Engineering and Cloud Computing as a member of Global Directory (GD) member for ISO-IEC. Besides editing more than 25 book volumes, Nabendu has authored six text and research books and has more than 150 Scopus Indexed research papers in Journals and International conferences. His areas of research interests include distributed systems, image processing and software engineering. Dr. Chaki has served as a research faculty in the Ph.D. program in Software Engineering in U.S. Naval Postgraduate School, Monterey, CA. He is a visiting faculty member for many Universities in India and abroad. Besides being in the editorial board for several international journals, he has also served in the committees of over 50 international conferences. Prof. Chaki is the founder Chair of ACM Professional Chapter in Kolkata.

**Agostino Cortesi**, Ph.D., is a Full Professor of Computer Science at Ca' Foscari University, Venice, Italy. He served as Dean of the Computer Science Studies, as Department Chair, and as Vice-Rector for quality assessment and institutional affairs. His main research interests concern programming languages theory, software engineering and static analysis techniques, with particular emphasis on security applications. He published more than 110 papers in high-level international journals and proceedings of international conferences. His h-index is 16 according to Scopus, and 24 according to Google Scholar. Agostino served several times as member (or chair) of program committees of international conferences (e.g. SAS, VMCAI, CSF, CISIM, ACM SAC) and he is in the editorial boards

of the journals ‘Computer Languages, Systems and Structures’ and ‘Journal of Universal Computer Science’. Currently, he holds the Chairs of ‘Software Engineering’, ‘Program Analysis and Verification’, ‘Computer Networks and Information Systems’ and ‘Data Programming’.

**Nagaraju Devarakonda** received B.Tech. from Sri Venkateswara University, M.Tech from Jawaharlal Nehru University, New Delhi and Ph.D. from Jawaharlal Nehru Technological University, Hyderabad. He had published 30 research papers in international conferences and journals. He is presently working as Professor and HOD of IT Department at Lakireddy Bali Reddy College of Engineering. His research areas are data mining, soft computing, machine learning and pattern recognition. He has supervised 25 M.Tech. students and is currently guiding eight Ph.D.s.



# Energy-Efficient Data Route-in-Network Aggregation with Secure EEDRINA

B. Sujatha, Chala Tura Jilo and Chinta Someswara Rao

**Abstract** Wireless sensor network (WSN) is one in every of the rising technologies inside the place of networking, wherein a set of spatially dispersed self-sufficient sensor nodes work together to perform some given project-important task. In WSN, performing data aggregation and routing protocol is identified as the principle additives. In designing WSN, limitation of strength, data aggregation, and security are the major challenges. To reduce strength consumption at the same time as transferring data from source to destination, data must be aggregated. However, data aggregation or routing protocols requires adequate security features to maintain data safe both in aggregation or transmission. The set of rules we used for routing protocols and safety additionally has an impact in energy. Extended Data Routing for In-Network Aggregation (EDRINA) and Energy-Efficient Data Routing for In-Network Aggregation (EEDRINA) are some of data aggregation protocols that designed with the aim of reducing the number of transmissions and saving energy. However, they lack required level of security. This paper presents performance evaluation of EEDRINA and proposes Secure EEDRINA (SEEDRINA). SEEDRINA has efficient energy aggregation and identity-based digital signature feature. Efficient energy aggregation balances the energy of WSN. Whereas, Identity-Based Digital Signature is used to provide security in the data transmission in EEDRINA. SEEDRINA outperforms the existing algorithm.

**Keywords** Wireless sensor network · Data aggregation · Energy consumption IBOOS · Security

---

B. Sujatha

Department of CSE, University College of Engineering, Osmania University, Hyderabad, TS, India

C.T. Jilo

University College of Engineering, Osmania University, Hyderabad, TS, India

C.S. Rao (✉)

Department of CSE, S R K R Engineering College, Bhimavaram, Andhra Pradesh, India

e-mail: chinta.someswararao@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_1](https://doi.org/10.1007/978-981-10-6319-0_1)

## 1 Introduction

To maximize the advantage of WSNs, electricity consumption and security have to be achieved further to different parameters along with complexity overhead and memory usage. Researchers have accomplished an independent work on each electricity consumption and protection. They have advanced the electricity consumption with the aid of making use of data aggregation routing protocols. However, regarding protection, plenty of work must be carried out.

Additionally, growing protocols or algorithm memory trouble has to be taken into consideration. Memory utilization and processing speed are without delay proportional to energy consumption. Accordingly, it has to be deliberate in advance earlier than implementation.

Therefore, we took the initiative to assess how lots power consumption is progressed with or without security characteristic.

## 2 Related Work

Security problems of WSNs are labeled into five categories. These are cryptography, key management, secure routing, secure data aggregation, and intrusion detection. To obtain those protection issues, protection offerings must be supplied. A number of protection offerings consist of confidentiality, authenticity, integrity, availability, non-repudiation, freshness, ahead secrecy, and backward secrecy [1]. Comparable work was carried out in [2], aiming evaluation of Wi-Fi sensor network security demanding situations, goals, attacks, and defense mechanism used. Even though each of them advises necessity of safety, they did no longer endorse how it needs to be carried out.

Short of wi-fi sensor networks and routing protocols and security troubles were given in [3]. The paper highlights feasible attacks and counter measures. Besides these protective mechanisms together with authentication, encryption, redundancy, probing, tracking, flexible routing; three-manner handshaking, and authentication were noted. Finally, they concluded that routing protocols did not fulfill protection intention.

Thangaraji and Ponmalar [4] proposed secured hybrid information aggregation tree (SHDT). For strength-efficient information aggregation, they carried out genetic algorithmic rule (GA), which was carried out simply at sink. Aside from this, they used synthetic Bee Colony (ABC) to collect secured aggregate data. The authors define a set of rules from categories of bees with respect to sensor node. The GB-ABC algorithm has higher result in extending the life of WSN. However, utilization intelligence based totally on algorithm like ABC consumes extra power.

Ranjan and Karmore [5] survey focused on integrated SDRA and ESPDA data aggregation protocols. From their survey, they integrated dictated a number of securities violate and disadvantages. Then they proposed self-test (BIST) technique.

Accord built integrated them, this method can discover the susceptible node, balance the strength built-intake of sensor nodes built-independent built integrated on the residual power tiers, and rotate the function of facts aggregator among sensor. Moreover, the method can provide protection. However, evaluation and implementation are nonetheless built-incomplete.

Related survey on protection problems in wireless sensor network was performed by means of Pratihari [6]. The survey stresses on comfortable routing that is important to the popularity and use of insecure routing protocols. In connection to this, link layer encryption and authentication mechanisms might be possible for protection in opposition to mote magnificence outsiders, but cryptography alone is not always sufficient. Finally, it mentions the existence of pc-magnificence adversaries and insiders and the constrained applicability of end to give up protection mechanisms necessitates a proper layout protocol. But, the hassle nevertheless remains to satisfy protection goals.

Sadafal and Borhade [7] have in brief discussed DRINA set of rules and propose the way to implement safety on it. The device version starts with calculating fake information sent by using malicious sensor node. In addition, they anticipate information encryption earlier than transmission. They applied the concept of RSA and SHA set of rules. The proposed algorithm gives confidentiality and authentication by means of improving current device. Nevertheless, the protocols underneath which they carried out protection have positive drawbacks which include undynamic route, high overhead, and hyperlink failure. Furthermore, the function indicates most effective false packet dictated.

Patle and Satao [8] recommend identification based totally on line Offline digital Signature scheme primarily based Multi-Wight Clustering algorithm (MWBCA). The set of rules is carried out on LEACH routing protocol. Nevertheless, LEACH algorithm suffers from unavailability of direction restore mechanism and occasional scalability excessive power intake.

Data Routing In-network Aggregation (DRINA) [9] is conceived by Villas et al. This set of rules builds a routing tree with shortest direction to be able to be connecting all source nodes to sink together with statistics aggregation maximization. DRINA has three stages. The primary phase deals with a way to configure the hop tree and parameter required for it. The second phase deals with cluster formation. Then finally, the third phase deals with the leader selection algorithm that starts and ends with leader or coordinator election.

Power improvement of DRINA is given on prolonged DRINA and more suitable DRINA in [10] and [11], respectively. The goal of these algorithms is to enhance specifically the performance of the community even as maximizing records aggregation inside the network. However, none of them stated safety problems.

Shinde and Sonavane [12] describe Energy-Efficient Data Routing for In-network Aggregation (EEDRINA) algorithm which matches primarily based on path redirection. Instinct in the back of this idea is to use residual electricity of different nodes to shield node engaged with forwarding traffic no longer to drain out battery life earlier than anticipated time.

### 3 Implementation

In this work, we propose Secure Energy-Efficient Data Routing In-network aggregation (SEEDRINA). SEEDRINA algorithm is expected to improve the existing protocol by using offering protection. Among safety to be had, symmetric key management suffers from orphan node problem [13] which happens while a node does no longer proportion a pair wise key with others in its preloaded key ring. The orphan node has hassle of increasing overhead, strength intake, and lowering the opportunity of a node becoming a member of a cluster head [14]. Therefore, it is not convenient for WSNs. Public key or asymmetric key control is predicted to be possible. Consequently, it overcomes challenge of symmetric key management security to enforce in WSN. Security feature comes from comfortable and efficient information transmission (SET) protocols for EEDRINA is referred to as SET-IBS and SET-IBOOS [14] through using the identity-based virtual signature (IBS) scheme and the identification primarily based on line/Offline digital Signature (IBOOS) scheme, respectively. This algorithm is to start with carried out on LEACH. Now we applied in EEDRINA.

### 4 Simulation Setup

The simulation randomly has on place of  $1400 \times 700 \text{ m}^2$ . Nodes are simulated in any such manner that each one of the nodes is completely consumed in network. All nodes are static and one among them is sink node. In the proposed network architecture, constant bit rate (CBR) is used for traffic generation in network. Parameters required to perform experiment are given in Table 1.

**Table 1** Simulation parameters and nodes configuration of network topologies

NO.	Parameter	Values
1	Simulation tool	Ns 2.34
2	Traffic pattern	CBR
3	Network size	$1400 \times 700$
4	Channel used	Channel/Wireless Channel
5	Initial energy	12 J
6	Mobility speed	10,000 ms
7	Simulation time	20, 40, 60, 80, 100 s
8	Queue length	50
9	Propagation model	Propagation/Two Ray Ground
10	Packet rate transmission	0.05
11	Routing protocols	EEDRINA, SEEDRINA
12	MAC protocol	MAC/IEE 802.11
13	Transmission protocol	TCP

## 5 Results and Discussions

In this experimental work, the results of proposed SEEDRINA set of rules have compared with the prevailing EEDRINA set of rules. Overall performance assessment parameter used had been quantity of nodes alive average throughput, average strength intake, packet shipping ratio, cease to quit postpone, and complexity overhead.

### 5.1 Number of Nodes Alive

The ability of sensing and gathering data in a WSN rely upon the set of alive nodes or the wide variety of nodes which can be working nicely. Right here, capability of the WSNs is evaluated by counting the variety of alive nodes in the network after given time. As a end result, SEEDRINA algorithm progressed the wide variety of alive nodes in comparison to EEDRINA. For this reason, SEEDRINA set of rules is preferable for large length of network (Fig. 1).

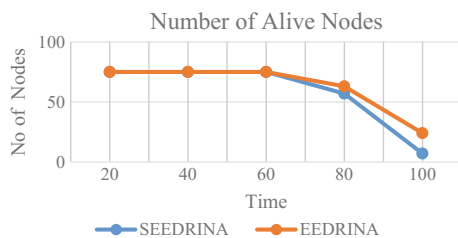
### 5.2 Average Throughput

Average throughput is the ratio of distinction between wide variety of data packets obtained at destination node and quantity of statistics packets misplaced in among to the total time required from first data packet to final records packet [15]. The unit of throughput is bits/sec [16]. Throughput of SEEDRINA outperforms that of EEDRINA (Fig. 2).

### 5.3 Average Energy Consumption

The percentage electricity ate up by means of all of the nodes is calculated as the average in their man or woman power consumption of the nodes [17]. The much

**Fig. 1** No of alive nodes for EEDRINA and SEEDRINA

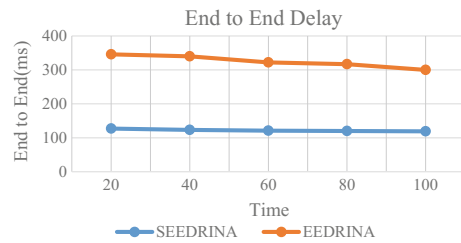


less average power intake the greater network lifestyles is extended. Therefore, it is indirect proportional to community lifespan. The simulation result suggests average energy consumption of proposed algorithm is much less as compared to EEDRINA as proven in Fig. 3.

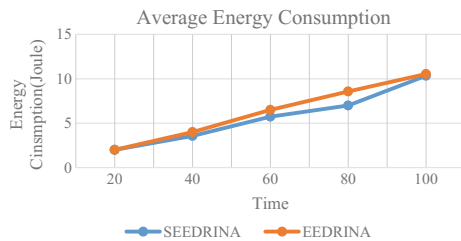
#### 5.4 Packet Delivery Ratio

Packet delivery ratio (PDR) network overall performance metric [18, 19] is described as the ratio among the numbers of statistics packets efficiently brought to the destination and the number of packets transmitted via the source. PDR is calculated as acquired packet/send packet \*one hundred. The higher the fee of the PDR, the lesser the packet loss rate and more efficient will be the routing protocol from delivery point of view. For that reason, proposed algorithm surpasses over the other algorithms as shown in Fig. 4.

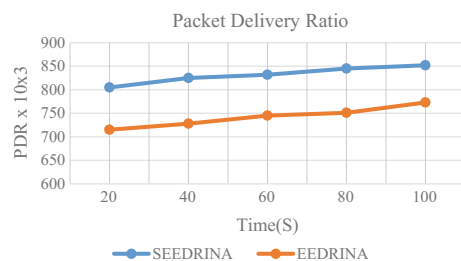
**Fig. 2** Average throughput for EEDRINA and SEEDRINA



**Fig. 3** Energy consumption for EEDRINA and SEEDRINA



**Fig. 4** PDR for EEDRINA and SEEDRINA



### 5.5 End-to-End Delay

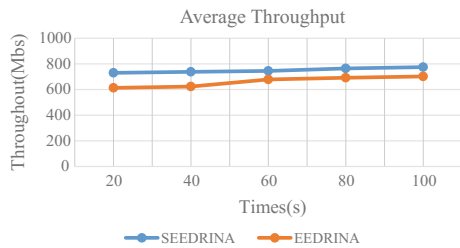
End-to-end delay [19, 20] is the period it takes for a packet to travel from the source to the destination (Fig. 5).

It calculates the postponing of the packet that is efficaciously transmitted from the source to the destination. This consists of all viable delays because of buffering at some point of direction discovery latency, queuing in the interface queue, retransmission delays on the MAC, propagation and switch times [21]. In WSN, quit-to-quit put off needs to be minimum in among source and destination. As an end result, SEEDRINAs’ put off is nearly negligible.

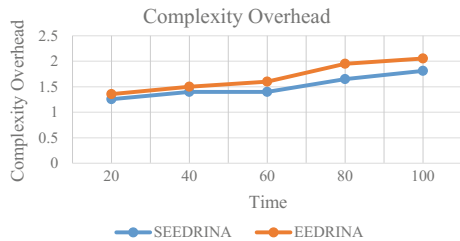
### 5.6 Complexity Overhead

Complexity overhead can also motive extra computation time, memory, and bandwidth, consume strength or other resources which might be required for the node particularly if attacker injects fake information or continuously sends fake packet; it affects directly this parameter. therefore, in sensor node complexity overhead needs to be saved minimum as many as possible. As a result, the end result of simulation for SEEDRINA set of rules is minimum in comparison to EEDRINA. Figure 6 portrays comparison of complexity overhead.

**Fig. 5** E<sub>2</sub>E delay for EEDRINA and SEEDRINA



**Fig. 6** Complexity overhead for EEDRINA and SEEDRINA



## 6 Conclusions

In this paper, appearing information aggregation and secure routing protocol are recognized as the main components for WSNs. However, there are few set of rules devised considering this two functions. Among routing protocols, EDRNA and EEDRINA are analyzed with appreciate to information aggregation and routing protocols. Even though they may be right in terms of records aggregation, they still lack security. To cope with this problem, we proposed SEEDRINA. SEEDRINA offers safety to EEDRINA. SEEDRINA algorithm changed into expansively in comparison with EDRINA and EEDRINA with respect to range of alive nodes, common throughput common strength intake, packet shipping ratio, quit to cease postpone, and complexity overhead. SEEDRINA has exact capacity to store electricity which gives fairly top wide variety of alive nodes. It has excellent improvement of average throughput. It can limit complexity, consequently, we have discovered that the set of rules we chose for routing protocol and security has huge effect on battery existence of nodes. Future work is recommended to plan and put into effect in Mica2 sensors and analyze its protection against numerous assaults.

## References

1. M. Teymourzadeh, R. Vahed, "Security in wireless sensor networks: Issues and challenges," *International Journal of Computer Networks and Communications Security*, vol. 1.
2. K. CHELLI, "Security issues in wireless sensor networks: Attacks and countermeasures," in *Proceedings of the World Congress on Engineering*, vol. 1, 2015.
3. Anjali, Shikha, and M. Sharma, "Wireless sensor networks: Routing protocols and security issues," in *Computing, Communication and Networking Technologies (ICCCNT)*, 2014 International Conference on, pp. 1–5, IEEE, July 2014.
4. M. Thangaraj and P.P. Ponmalar, "Swarm intelligence based secured data aggregation in wireless sensor networks," in *Computational Intelligence and Computing Research (ICCIC)*, 2014 IEEE International Conference on, pp. 1–5, Dec 2014.
5. R.K. Ranjan and S.P. Karmore, "Survey on secured data aggregation in wireless sensor network," in *Innovations in Information, Embedded and Communication Systems (ICIIECS)*, 2015 International Conference on, March 2015.
6. H. Pratihari, "A survey on security issues in wireless sensor network," *International Journal of Computer Science and Mobile Computing*, vol. 2, pp. 55–58.
7. P.V. Sadafal and R. Borhade, "Secure routing using clustering algorithm," *International Journal of Computer Applications*, vol. 84, no. 9, 2013.
8. R.R. Patle and R. Satao, "Aggregated identity-based signature to transmit data securely and efficiently in clustered wsn," in *Computing Communication Control and Automation (ICCUBEA)*, International Conference on, pp. 138–142, Feb 2015.
9. L.A. Villas, A. Boukerche, H.S. Ramos, H.A.B.F. de Oliveira, R.B. de Araujo, and A.A.F. Loureiro, "Drina: A lightweight and reliable routing approach for in-network aggregation in wireless sensor networks," *IEEE Transactions on Computers*, vol. 62, pp. 676–689, April 2013.
10. N. Shrivastava and R. Kawitkar, "Edrina for more battery life in wireless sensor networks," in *Convergence of Technology (I2CT)*, 2014 International Conference for, pp. 1–6, April 2014.



11. S. Subhashini and T. Parani, "An efficient routing approach for aggregated data transmission along with performance improvement in wireless sensor networks," *International Journal of Research in Engineering and Technology*, vol. 03.
12. Y.Y. Shinde and S.S. Sonavane, "An energy efficient critical event monitoring routing method for wireless sensor networks," *International Journal of Computer Applications*, vol. 114, no. 10, 2015.
13. L. Anil Kumar K, "A study on secure data transfer in cluster-based wireless sensor networks," *International Journal of Advanced Research in Computer Engineering and Technology (IJARCET)*, vol. 3, pp. 2012–2016, June 2014.
14. H. Lu, J. Li, and M. Guizani, "Secure and efficient data transmission for cluster-based wireless sensor networks," *IEEE transactions on parallel and distributed systems*, vol. 25, no. 3, pp. 750–761, 2014.
15. S. Chavan, A. Kulkarni, and M.A. Kasar, "Energy efficient and optimal randomized clustering protocol for self-organization in wsn," *Energy*, vol. 4, no. 7, 2016.
16. C. He, "Throughput and delay in wireless ad hoc networks," Final report of EE359 Class Project, Stanford University [Online]. Available: <https://www.dsta.gov.sg/index.php/DSTA>, 2006.
17. Chunawale and S. Sirsakar, "Minimization of average energy consumption to prolong lifetime of wireless sensor network," in *Wireless Computing and Networking (GCWCN), 2014 IEEE Global Conference on*, pp. 244–248, IEEE, 2014.
18. Y. Liu, Y. He, M. Li, J. Wang, K. Liu, and X. Li, "Does wireless sensor network scale? a measurement study on greenorbs," *IEEE Transactions on Parallel and Distributed Systems*, vol. 24, no. 10, 1983–1993, 2013.
19. M.F. Khan, E.A. Felemban, S. Qaisar, and S. Ali, "Performance analysis on packet delivery ratio and end-to-end delay of different network topologies in wireless sensor networks (wsns)," in *Mobile Ad-hoc and Sensor Networks (MSN), 2013 IEEE Ninth International Conference on*, pp. 324–329, IEEE, 2013.
20. J. Liu, X. Jiang, H. Nishiyama, N. Kato, and X. Shen, "End-to-end delay in mobile ad hoc networks with generalized transmission range and limited packet redundancy," in *2012 IEEE Wireless Communications and Networking Conference (WCNC)*, 1731–1736, April 2012.
21. P. Manickam, T.G. Baskar, "Performance comparisons of routing protocols in mobile ad hoc networks," 2011.

# Prediction Models for Space Mean Speed on Urban Roads

Mariangela Zedda and Francesco Pinna

**Abstract** The research investigates the relationship between driving behavior and characteristics of the road environment in urban area. This study allows the identification of factors which influence driving speed. The purpose is to develop mathematical models which link driving behavior with infrastructure geometric characteristics. The parameter used to describe driving behavior is space mean speed. This is very important because it considers speed of vehicles traveling a given segment of roadway during a specified period of time and it is calculated using the average travel time and the length for the roadway segment. This speed is used to understand driving behavior in normal traffic flow and in daylight conditions. The data are collected on urban road tangents. These roads have the common characteristic to be single or dual carriageways with two lanes for each direction with a speed limit of 50 km/h. With a multiple linear regression, two models are developed and validated to predict speed. Statistically significant variables include traffic characteristics (flow, number of vehicles entering and leaving traffic stream) and geometric design attributes (lanes width, type of median, tangent length and type of left-lateral obstacle). This study can be useful to both traffic manager and road designers because developed models could implement design guidelines, especially regarding road tangents.

**Keywords** Urban road · Space mean speed · Multiple linear regression  
Model validation · Road design · Road tangent

---

M. Zedda · F. Pinna (✉)  
Department of Civil Environmental Engineering and Architecture,  
University of Cagliari, Cagliari, Italy  
e-mail: fpinna@unica.it

M. Zedda  
e-mail: zedda.mangy@gmail.com

## 1 Introduction

Speed is a parameter which involves two fields closely related: safety and road design. In the first case, it is the element that influences the occurrence and the severity of an accident; in the second case, it is the factor used to design road infrastructure elements. Therefore, speed affects both the probability the accident happens and the injury severity. So, first, speed defines risk level of being involved in an accident: people need time to process information, to decide how to react to situation in order to have a reaction able to avoid the crash. Second, the speed affects severity of accident and injury. In fact, at higher speed, more energy is released colliding with another vehicle, road user, or obstacle. Some road types, such as urban street, have traffic conditions and interaction of multiple urban users more complex than others [1]. This, for example, depends on road infrastructure configuration, absence or presence of pedestrians, cyclists, the type of car, etc. In addition, cause of accidents is due to both road users' lack of respect for traffic code rules, and presence of road infrastructure deficiencies. In fact, roads should induce drivers to adopt speeds which are designed and in most cases this does not happen. This is probably due to the fact that in Italy standard prescribes the use of the design speed in order to size the horizontal alignment elements. The design speed is an ideal speed defined with the aim to design geometric elements, and safety, regularity, and efficiency of vehicles movement depend on these. In Italy, traditional design approach does not consider the real driving behavior in relation to geometric and environmental characteristics of the road infrastructure. So to meet the road safety requirements, it would be appropriate to make the roadway congruent, at project level, with real speeds. For these reasons, the driving behavior is studied in order to develop predictive models that can be used as a tool to improve the Italian standard. To pursue this aim, these models should be validated with data about other urban roads.

## 2 Literature Review

The focus of the present study is on factors influencing drivers' speed, in normal traffic condition, while traveling on urban street tangents, with special attention on geometric design and urban environment. The models in literature, although provide important results, cannot be considered universally valid. The causes may be differences between one country and another (but also differences between city of the same country) in terms of habits, vehicle fleet, traffic rules, climatic conditions, and orography of the area. For example, the developed models for rural roads cannot be applied to urban areas because drivers assume a different style to travel urban or suburban roads; environment characteristics are often dissimilar; flow, traffic characteristics, and geometric design elements are different. As regards the developed models for urban area, many researchers use speed limits as independent

variable; this cannot be used in the case study because urban roads are distinguished by the same speed limit of 50 km/h, which, moreover, is the only one in Italy for urban areas. Finally, several researchers used instantaneous speed as dependent variable. These speeds are measured at one specific point of roadway and not considered running speed variations adopted by drivers. The behavioral dynamics of drivers in urban areas are highly complex. Perhaps this fact has deterred the construction of mathematical models capable of representing the driving behavior in relation to urban environment characteristics. In Italy, the developed models may concern only the suburban area, while at the international level, there is poor scientific production especially aimed at studies concerning the urban tangents. A number of models are available in literature, but none of these considers urban roads, tangent and space mean speed. In consideration of the above, a general overview of international scientific production, regard to urban roads, is shown below.

Tarris et al. [2] collect vehicle operating speed data along 27 urban collectors in Pennsylvania. Models are developed for individual and aggregate data. In both cases, the variable used to predict operating speed is degree of curve.

Poe and Mason [3] study operating speed on 27 urban collectors in Pennsylvania. The tracking is used to include only free flow passenger cars, which consist of vehicles with time headways of 5 s or more. They introduce a mixed model approach to analyze influence of geometric elements on operating speed. The significant variables are degree of curvature, longitudinal grade, lane width, and roadside characteristics.

Fitzpatrick et al. [4] examine operating speed on 78 urban/suburban sites in Arkansas, Missouri, Tennessee, Oregon, Massachusetts, and Texas. Free flow speed is obtained using time headway of 5 s or more and a tail-way of 3 s or more. The posted limit and access density are found to be statistically significant to predict operating speed.

Wang et al. [5] collect speed data, in tangent sections, using 200 vehicles equipped with GPS. It is established that roadside density, driveway density, intersection density, sidewalk presence, and parking presence are negatively associated with operating speed on urban streets while a number of lanes, curb presence, and commercial and residential land uses are positively associated with operating speed.

Karin F.M. Arosson [6] uses two methods for speed investigation on urban street links. The first approach (micro model) models average travel speed results produced by a microscopic traffic simulation model calibrated by using observed driver behavior. In the second approach (macro model), space mean speed data is modeled for three street link types based on aggregated speed data collected in the field. In particular, urban model shows a relationship between speed and flow, average number of crossing pedestrians and cyclists.

Ali et al. [7] investigate the relationship between free flow speed and geometric variable on 35 urban street segments located in Fairfax County Virginia. The study shows a statistically significant relationship between free flow speed and posted

speed limits, median type, and segment length. Free flow speed is obtained using time headway of 7–8 s or more and a tail-way of 4 or 5 s.

To recap the above discussion, Table 1 summarizes mathematical models.

**Table 1** Summary of speed models

Autor (s)	Models	R2
Tarris et al. (1996)	$V_{\text{mean}} = 53.5 - 0.265\text{DC}$ (aggregate mean speed)	0.82
	$V_{\text{mean}} = 53.8 - 0.272\text{DC}$ (individual mean speed model)	0.63
Poe e Mason (2000)	$V = 57,47 - 0,23 \text{ DEGCVR} - 3,17 \text{ LANWIDN} - 1,23 \text{ HZRT5 N}$	–
Fitzpatrick et al. (2003)	$V85 = 7,68 + 0,98 \text{ PSL}$	0.90
	$V85 = 7,68 + 0,83 \text{ PSL} - 0,05 \text{ AD}$	0.90
Wang et al. (2006)	$V85 = 31.564 + (6.491 \text{ lane.num}) - (0.101 \text{ roadside}) - (0.051 \text{ driveway}) - (0.082 \text{ intersection}) + (3.01 \text{ curb}) - (4.265 \text{ sidewalk}) - (3.189 \times \text{parking}) + (3.312 \times \text{land.use1}) + (3.273 \times \text{land.use2})$	0.67
Arosson (2006)	$V_{\text{obs}} = 39,8 - 0,20 \text{ Flow} - 0,24 \text{ Ped} - 5,24 \text{ Lanes} + 4,73 \text{ BicSep} - 5,54 \text{ Park}$	0.66
Ali et al. (2007)	$\text{FFSmean} = 39,3 + 8,6\text{PS45} + 3,7\text{PS40}$	0.76
	$\text{FFS85} = 42,3 + 10,4\text{PS45} + 3,8\text{PS40}$	0.77
	$\text{FFSmean} = 37,4 + 6,8\text{PS45} + 2,6\text{PS40} + 13,5\text{SL}$	0.87
	$\text{FFS85} = 37,4 + 8\text{PS45} + 2,1\text{PS40} + 3,6\text{MT} + 13\text{SL}$	0.86

where

$V_{\text{mean}}$	Mean operating speed
DC	Degree of curve
V	Mean speed
DEGCVR	Degree of curve
LANWIDN	Lane width
HZRT5 N	Roadside hazard rating
V85	85th percentile operating speed
PSL	Posted speed limit
AD	Access density
V85	85th percentile speed
lane.num	number of lanes
roadside	density of trees and utility poles divided by their average offset from roadway
driveway	density of driveways
intersection	density of T-intersections
curb	0 if there is no curb; otherwise 1
sidewalk	0 if there is no sidewalk; otherwise 1
parking	0 if there is no on-street parking; otherwise 1

(continued)

**Table 1** (continued)

Autor (s)	Models	R2
land.use1	1 if land use is commercial; otherwise 0 (baseline is park and office land use)	
land.use2	1 if land use is residential; otherwise 0 (baseline is park and office land use)	
Vobs	Observed space mean speed	
Flow	Observed average traffic flow in the studied direction of travel expressed in vehicles per 5 min	
Ped	Average number of crossing pedestrians and cyclists (summarized in groups of 5, 15 and 25 people per 15 min and 400 m)	
Lanes	Number of lanes in the studied direction (1 or 2)	
BicSep	Separated bicycle lane (yes = 1; no = 0)	
Park	Roadside parking permitted (yes = 1; no = 0)	
FFSmean	Mean free flow speed	
FFS85	85th percentile free flow speed	
PSi	Posted speed limit ( $i = 45$ mph, 40 mph, 35mph)	
MT	Median type (divided = 1, no median = 0)	
SL	Segment length ratio	

### 3 Data Collection and Database

The data are collected on seven urban road tangents. These roads have the common characteristic to be single or dual carriageways with two lanes for each direction and to have an urban speed limit of 50 km/h. The preliminary analysis of collected speeds shows that drivers do not respect speed limit: 68.6% of drivers exceeded 50 km/h. Also, operating speed is calculated. The AASHTO defined it as “the speed at which drivers are observed operating their vehicles during free-flow conditions” [8]. The computed value on seven urban roads is 71 km/h.

“Viale Colombo” is not considered in developing models because this road is chosen in order to check the accuracy of speed equations developed in this study. This road has the greatest number of independent variables which are identified on all roads. The principal characteristics of roads are described in Table 2.

The next research step is to identify road section and then to perform monitoring campaign. We start from some observations reported in literature. Fitzpatrick et al. [9] define the section as a portion of a suburban arterial between horizontal curves and/or control devices. The sections selected are at least 200 m from an adjacent horizontal curve and 300 m from adjacent signal or stop sign. In another study, Fitzpatrick et al. [10] establish that to avoid the effect of traffic control devices on vehicle speed, distances between study sections and a signalized intersection should be at least 200 m. Polus et al. [11] identify at least 500 m from the study section and any intersection to eliminate the effect of traffic control devices on vehicle

**Table 2** Pattern of road features

Urban road geographic coordinate	Variables					
	Section length (m)	Tangent length (m)	Bus lane (m)	Lane width (m)	Fast lane width (m)	Obstacle right-lateral width (m)
Diaz 39°12'18.75"N 9°7'56.57"E	106	450	3.3	3.0	3	3.25
Lungo Saline 39°12'20.76"N 9°9'44.11"E	119.2	680	No	3.3	3.3	0.4
Poetto 39°11'59.07"N 9°8'42.58"E	120.2	450	3.3	3.0	3	3.65
Marconi 39°14'17.92"N 9°8'40.46"E	135.1	670	No	2.75	2.75	0.3
Elmas 39°14'13.17"N 9°5'41.78"E	150.0	1000	No	3.75	3.75	0.5
Monastir 39°14'10.75"N 9°5'49.03"E	143.0	800	No	3.75	3.75	0.3
Colombo 39°12'34.79"N 9°7'4.40"E	103.0	430	No	3.25	3.25	1.00
	Access points <sup>a</sup>	Travel direction	Type of left-lateral obstacle <sup>b</sup>		Median type <sup>c</sup>	
Diaz 39°12'18.75"N 9°7'56.57"E	Yes	2	2		1	
Lungo Saline 39°12'20.76"N 9°9'44.11"E	No	2	2		3	
Poetto 39°11'59.07"N 9°8'42.58"E	No	2	2		2	
Marconi 39°14'17.92"N 9°8'40.46"E	Yes	2	1		0	
Elmas 39°14'13.17"N 9°5'41.78"E	Yes	1	3		-	

(continued)

**Table 2** (continued)

Urban road geographic coordinate	Variables					
	Section length (m)	Tangent length (m)	Bus lane (m)	Lane width (m)	Fast lane width (m)	Obstacle right-lateral width (m)
Monastir 39°14'10.75"N 9° 5'49.03"E	Yes	1	0		–	
Colombo 39°12'34.79"N 9° 7'4.40"E	Yes	2	1		1	

<sup>a</sup>*Access points* Elements that allow vehicles to enter and leave traffic stream

<sup>b</sup>*Type of obstacle left-lateral* Type of obstacle next to the fast lane that causes influences on driving behavior. The variable is:

3 if there is not obstacle; 2 if there is median; 1 if lanes in opposite direction are not separate; 0 = presence of parking

<sup>c</sup>*Median type* Three types of central reservation characterize the urban roads in exam. The variable is: 0 if there is not median; 1 if the median, decorated with trees or hedges, allows the U-turn; 2 if the median, decorated with trees or hedges, is continuous; 3 if the median is indicated by road markings

speeds. Selected study segment should be sufficiently distant from traffic control devices so that driver speed on the roadway is not influenced by acceleration and deceleration zones. We use the criteria indicated in Italian Standard [12] to determine stopping visibility distance and zone lengths for vehicles in deceleration or acceleration. Calculated values are less than 200 m and, in view of these considerations, we prefer to use a distance greater than or equal to 200 m between section of the studied tangent and curve, stop sign, traffic lights, and yield signs close of roundabouts. Therefore using this method, study tangent has a length that varies between 100 and 150 m.

The equipment used to detect all information associated to vehicular traffic is the Radar EasyData. It records for each vehicle: date (day/month/year), time (hours/minutes/seconds), spot speed (km/h), vehicle length (meters), and direction of travel. A digital camcorder is used to check if there are any measurement errors. Also, it is used to analyze user behaviors and to identify any external or internal traffic factors able to affect driver behavior. In particular, we use two radars placed at the road section beginning and end. Figure 1 shows the scheme used for the surveys campaign.

The data collection is performed under dry weather conditions. Six hours of data are collected at each road section: four hours of morning traffic between 8.00 and 12.00 a.m. and two hours of afternoon traffic between 4.00 and 6.00 p.m. In particular surveys, campaign is performed in the weekdays at random, excluding





**Fig. 1** Schematic representation of the observed road segment with marked lines for vehicle passage time readings

Saturdays, Sundays, public holidays, days of strike by public transport, days with road works and events of various kinds in the streets under study or near the study sections. In the first section (line 1) and in the second section (line 2), the instantaneous speed and the transit times are collected for more than 97,000 vehicles.

The collected data are processed to obtain information about driver behavior and traffic flow characteristics. In particular, database is organized as follows. Vehicles are numerated and for each travel time, vehicle type (car, bus, motorcycles, and heavy vehicles) and numerical code of driver behavior (travel lane, change lane, driving behavior influenced by crossing pedestrian, etc.) are assigned. The anomalies regarding instrument errors and/or behavior drivers that are not the subject of study (influence due to the presence of radar, Traffic Code infractions, etc.) are recorded using numerical code. Once the database is prepared, data are controlled and processed. A first check is to verify that anomalies do not exceed 10% to prevent these can affect the quality of information altering the final result. The space mean speed is calculated using the following equation:

$$\begin{aligned} \bar{v}_{sms} &= (\text{distance traveled})/(\text{average travel time}) = d / \left( \left( \sum t_i \right) / n \right) \\ &= (n \times d) / \left( \sum t_i \right), \end{aligned} \tag{1}$$

where

- $v_{sms}$  space mean speed;
- $d$  distance traveled or length of roadway segment;
- $n$  number of observations;
- $t_i$  travel time of the  $i$ -th vehicle.

The space mean speeds are aggregated in 5 min intervals. So, the 5 min intervals are removed if the anomalies exceed 10%.

## 4 Analysis Method

### 4.1 Data Analysis

The space mean speeds are analyzed with specific statistical tools to obtain more information about the data and to understand if they need further investigation. Then the dependent variable analysis is carried out to test normality assumption. The histogram reveals that vehicle speeds distribution is bell-shaped and approximately symmetric. Kolmogorov–Smirnov test shows that speed data do fit the Gaussian distribution as shown in Fig. 2.

### 4.2 Regression Models

The process to develop and validate mathematical model is carried out to predict real driver behavior. The method is based on the observation of speeds on urban road tangent. Specifically, a multiple linear regression is used to developed models and stepwise selection method is used to select independent variables. In particular, the procedure begins with no variables in the model and adds a variable in succession according to the criterion of partial F-statistic. At each phase, a variable is added, whose partial F-statistic yields the smallest p-value. Variables are entered as long as the partial F-statistic p-value remains below a 0.05. When the addition of any of the remaining variables yields a partial  $p$ -value  $> 0.05$ , procedure stops.

This procedure shows that some variables, commonly used in mathematical models, are not statistically significant ( $p$ -value  $> 5\%$ ). For example, traffic composition is investigated: vehicles are divided in different categories (motorcycles, cars, trucks, and buses). Probably these variables are not significant because traffic

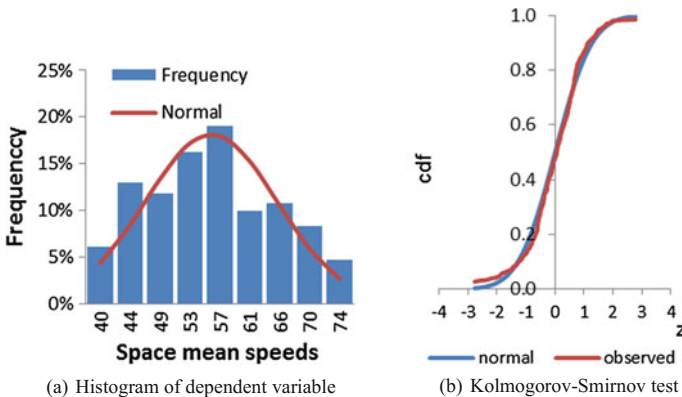


Fig. 2 Space mean speed analysis: **a** speed distribution **b** normality test

is homogenous (approximately 95% of vehicles are cars). Also, we consider access density, number of crossing pedestrians and cyclists, presence of sidewalk, curb, and parking. Moreover, posted speed limit does not effect on speed in fact, analyzing collected data, the speed limits are substantially exceeded by drivers: 68.6% of drivers has exceeded 50 km/h.

Six variables are used to develop models:

- flow ( $F$ ): number of vehicles over road section aggregated in 5 min intervals;
- number of vehicles entering and leaving traffic stream ( $F_{in,out}$ ): these are vehicles that travel partially road section, aggregated in 5 min intervals. This is due to the presence, in urban area, of driveways, parking areas, median which allows the U-turn
- presence of crosswalk ( $C$ ): 0 no crosswalk, 1 crosswalk;
- tangent length ( $TL$ ): is not the length of road section but the length of tangent between two intersections in which is located the study section;
- lane widths ( $LW$ ): is the sum of two lane widths in each direction (travel lane and fast lane);
- the left-lateral obstacle is interpreted in two ways:
  - i. type of left-lateral obstacle ( $O$ ): is the type of obstacle next to the fast lane and causes influences on driving behavior. This variable is used to develop model with collected data on every urban road. The variable is:
    - 3 if there is no obstacle;
    - 2 if there is median;
    - 1 if the lanes in opposite direction are not separate;
    - 0 = presence of parking;
  - ii. type of median ( $M$ ): the urban roads in exam are characterized by three types of central reservation. This variable is used to develop model with collected data on roads with median. The variable is:
    - 0 if there is not median;
    - 1 if the median, decorated with trees or hedges, allows the U-turn;
    - 2 if the median, decorated with trees or hedges, is continuous;
    - 3 if the median is indicated by road markings.

Two models are developed as shown in Table 3. The collected data on every urban road are used to develop the first model. The adjusted coefficient of determination ( $R^2$ ) is high, indicating that 81% of the variability is explained. The collected data on roads characterized by two lanes for traffic in each direction are used to develop the second model. This model is important to investigate the type of median. The adjusted  $R^2$  value is also high, indicating that 84.5% of the variability is explained.

All variables are significant at the 5% significance level (95% confidence level) for these two models. In other words,  $p$ -value is  $< 0.05$  for all independent variables. Variance inflation factor (VIF) is calculated to detect multicollinearity.

**Table 3** Multiple linear regression model results

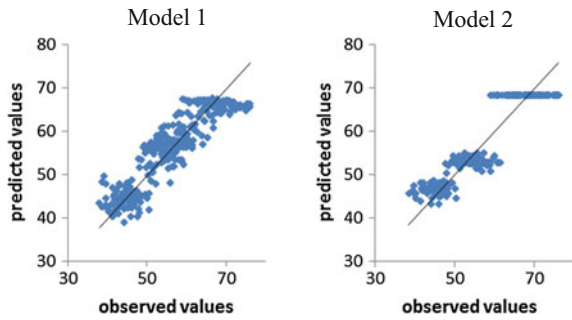
Model 1	Urban road: Diaz, Lungo Saline, Poetto, Marconi, Elmas, Monastir.									
$V_{ms} = 14.21 - 0.05 F - 0.64 F_{in,out} - 11.84 C + 0.03 TL + 3.84 O + 4.07LW$										
<i>Regression analysis</i>										
Multiple R	0.90									
R2	0.813									
R2 Adjusted	0.810									
Standard Error	4.03									
Observations	362									
ANOVA										Alpha 0.05
		gdl	SQ	MQ	F	p-value	Lower 95%	Upper 95%	VIF	
Regression		6	25200.21	4200.03	258.6	2.5E-06	8.36	20.06		
Residual		355	5765.21	16.24		1.6E-03	-0.08	-0.02	1.00	
Total		361	30965.42			3.4E-17	-0.78	-0.50	5.43	
						4.0E-42	-13.33	-10.35	3.20	
Intercept	14.21	Standard error	2.97	Stat t	4.78	5.1E-26	0.02	0.03	5.91	
F	-0.05		0.02	-3.18		4.5E-35	3.29	4.38	1.14	
$F_{in,out}$	-0.64		0.07	-8.88		3.6E-11	2.90	5.25	3.88	
C	-11.84		0.76	-15.59						
TL	0.03		0.00	11.44						
O	3.84		0.28	13.82						
LW	4.07		0.60	6.83						

(continued)

Table 3 (continued)

Model 2	Urban road: Diaz, Lungo Saline, Poetto, Marconi.						
Model 2	Urban road: Diaz, Lungo Saline, Poetto, Marconi.						
$V_{ms} = 21.80 - 0.60 F_{in,out} + 4.11 M + 0.05 TL$							
<i>Regression analysis</i>							
Multiple R	0.920						
R2	0.846						
R2 Adjusted	0.845						
Standard Error	3.806						
Observations	262						
ANOVA							Alpha 0.05
	gdl	SQ	MQ	F	p-value		
Regression	3	20583.37	6861.12	473.58	1.4106E-104		
Residual	358	3737.87	14.49				
Total	261	24321.23					
	Coeff.	Standard error	Stat t	p-value	Lower 95%	Upper 95%	VIF
Intercept	21.80	1.28	16.9	5.8E-44	19.28	24.33	
$F_{in,out}$	-0.60	0.16	-3.7	0.0001	-0.90	-0.29	1.00
M	4.11	0.34	12.2	1.6E-27	3.45	4.77	2.41
TL	0.05	0.00	23.1	6.0E-65	0.05	0.05	1.11

**Fig. 3** Scatter plot between observed values and predicted values



All VIF values are less than 10: these indicate that no collinear variables are present in both models.

The observed values are plotted versus predicted values as shown in Fig. 3: most of the points fall close to the 45° line demonstrating a linear relationship between the observed and predicted space mean speeds.

Diagnostic analysis and residual tests are performed to assess adequacy of models. Specifically, linearity, homoscedasticity, independence, and normality are verified [13]. Figure 4 shows that in both models, the following assumptions are satisfied:

- linearity: relationship between the predictors and the outcome variable is linear;
- homoscedasticity (homogeneity of variance): the error variance is constant;
- independence: the errors associated with one observation are not correlated with the errors of any other observation;
- normality: the errors are normally distributed. Kolmogorov–Smirnov test and Quantile–Quantile plot (Q–Q plot) are used to verify the condition of normality.

Leverage values and Cook’s distance are calculated to identify the presence of any anomalies observations.

The leverage statistic,  $h_i$ , measures the influence of  $y_i$  on its predicted value  $\hat{y}_i$ . The observed value  $y_i$  is influential if:

$$h_i > (2(k + 1))/n, \tag{2}$$

where

- $h_i$  is the leverage for the  $i$ -th observation;
- $k$  is the number of  $\beta$ ’s in the model (excluding  $\beta_0$ ).

In both models,  $h_i$  are less than 0.04, Fig. 5 shows that there are not observations which influence regression models.

Cook’s distance,  $D$ , is another measure of the influence of observations. Cook’s distance is defined as:

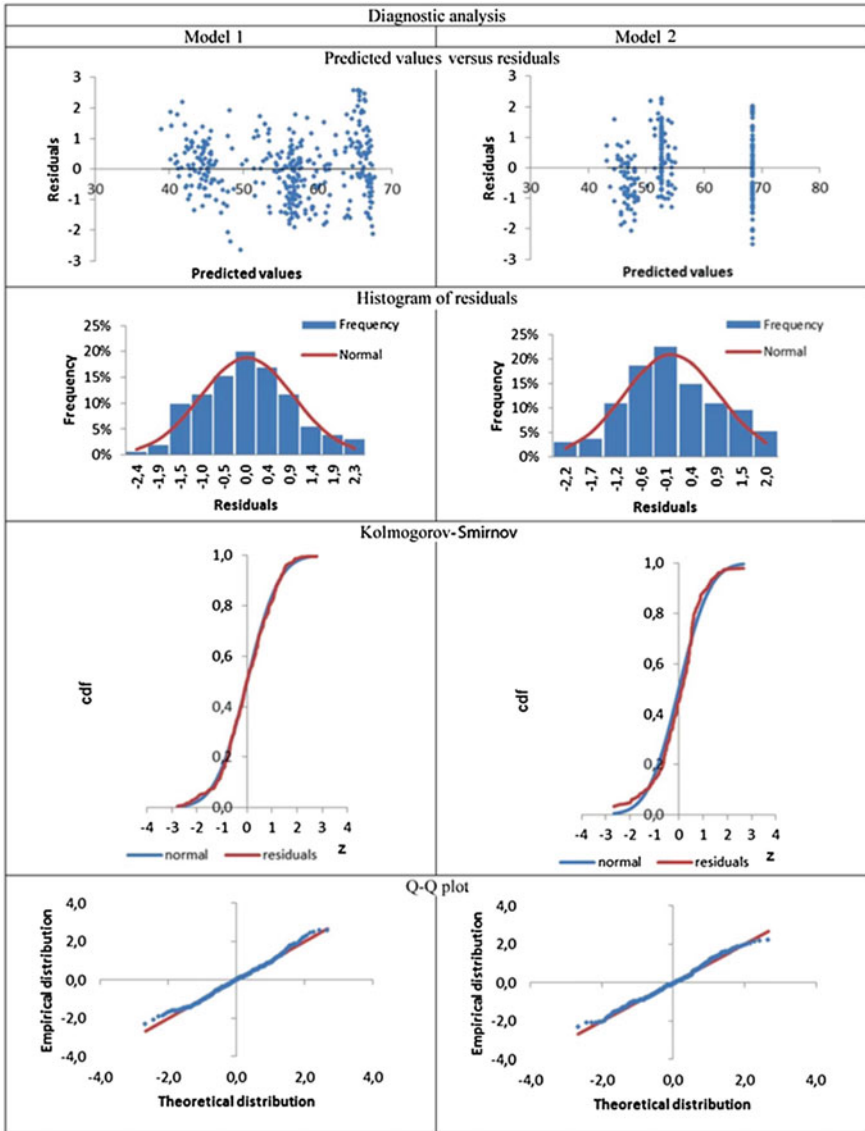


Fig. 4 Diagnostic analysis

$$D_i = \sum_{(j=1)}^n (y_j - \hat{y}_{j(i)})^2 / (k + 1) \text{MSE} \tag{3}$$

$y_j$  is the  $j$ -th fitted response value;

$\hat{y}_{j(i)}$  is the  $j$ -th fitted response value, where the fit does not include observation  $i$ .;

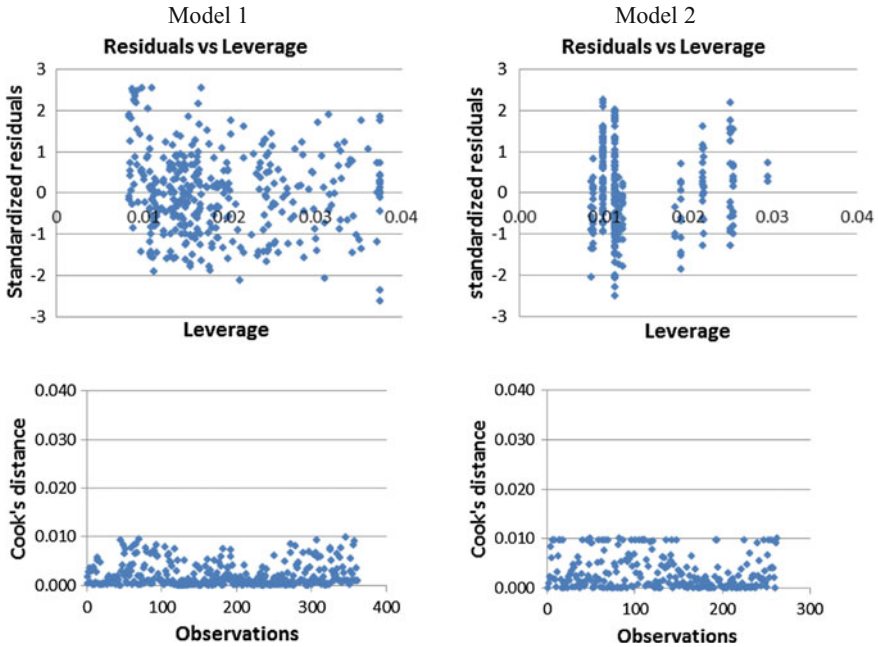


Fig. 5 Leverage values (*up*) and Cook’s distance (*down*)

$k$  is the number of coefficients in the regression model;  
 MSE is mean square error.

Values of Cook’s distance that are greater than 1 may be problematic. In the model 1,  $D$  is 0.0098 while in model 2,  $D$  is 0.0097 as shown in Fig. 5.

In view of all this, Table 4 illustrates the summary of models for space mean speeds. The first model is developed with collected data on every urban road. The second one is developed with data collected on roads characterized by two lanes for traffic in each direction.

## 5 Validation of Speed Models

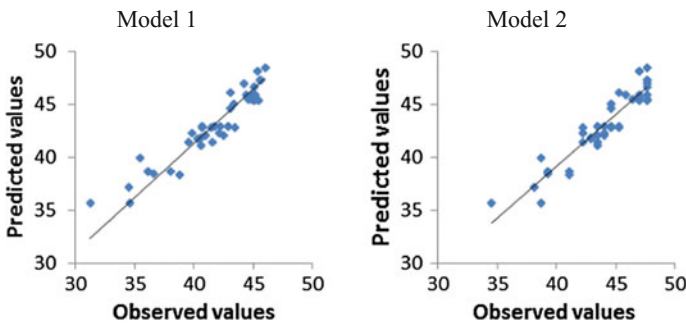
Validation is a useful process to test the predictive capacity of model using a new dataset. For this reason, all roads except one, “Viale Colombo”, are considered in developing models. This road is chosen for validation because it has the greatest number of independent variables identified on all roads.

In order to check the accuracy of speed equations developed in this study, the models are applied to new data. The observed values are plotted versus predicted values as shown in Fig. 6: most of points fall close to the 45° line demonstrating a good agreement between two sets of speed values.



**Table 4** Multiple linear regression model results

Regression equations	$R^2_{adj}$
$V_{ms} = 14.21 - 0.05 F - 0.64 F_{in,out} - 11.84 C + 0.03 TL + 3.84 O + 4.07 LW$	0.81
$V_{ms} = 21.80 - 0.60 F_{in,out} + 4.11 M + 0.05 TL$	0.84
Where	
$V_{ms}$ = space mean speed (aggregated in 5 min intervals) [km/h];	
$F$ = flow (aggregated in 5 min intervals) [v/min];	
$F_{in, out}$ = number of vehicles entering and leaving traffic stream (aggregated in 5 min intervals) [v/min];	
$C$ = presence of crosswalk: 0 no crosswalk, 1 crosswalk;	
$TL$ = length of tangent between two intersections in which is located the study section [m];	
$O$ = type of left-lateral obstacle. The variable is: 3 if there is not obstacle; 2 if there is median; 1 if the lanes in opposite direction are not separate; 0 = presence of parking.	
$LW$ = lane widths is the sum of two lane widths in each direction (travel lane and fast lane) [m];	
$M$ = type of median. The variable is: 0 if there is not median; 1 if the median, decorated with trees or hedges, allows the U-turn; 2 if the median, decorated with trees or hedges, is continuous; 3 if the median is indicated by road markings.	



**Fig. 6** Scatter plot predicted speeds and observed speeds

The Kolmogorov–Smirnov two-sample test is used to compare predicted speeds to observed speeds. The test confirmed that there is no significant difference between predicted speeds and observed speeds as shown in Table 5.

**Table 5** Kolmogorov–Smirnov two-sample test

Model 1		Model 2	
Alpha	0.05	Alpha	0.05
D-stat	0.27	D-stat	0.26
p-value	0.09	p-value	0.08
D-crit	0.29	D-crit	0.29
Observed	41	Observed	41
Predicted	41	Predicted	41
D-stat < D-crit	yes	D-stat < D-crit	yes
p-value > 0.05	yes	p-value > 0.05	yes

## 6 Conclusions

Several studies investigate relationships between driving behavior and roadway characteristics but only a few of these study the tangents of urban roads and consider the space mean speeds.

The proposed study allows a better understanding of driving behavior in urban areas. In particular, this paper addresses the modeling of space mean speeds on tangents. Specifically, multiple linear regression is used and two models are developed to predict speeds. These models include design characteristics that have a significant impact on space mean speeds such as tangent length, lane widths, type of left-lateral obstacle, median type, and presence of crosswalk. Flow and number of vehicles entering and leaving traffic stream are the variables of traffic characteristics.

This is due to the urban area which is characterized by the presence of a greater number of intersections, parking areas, and driveways. Moreover, diagnostic analysis and residual tests are performed to assess adequacy of models and the assumptions are satisfied in both models. Finally, the predictive capacity of models is tested using a new dataset. The test confirms that there is no significant difference between predicted speeds and observed speeds.

Developed models confirm that some variables have more influence on driving speed than other. The research shows that some variables, commonly used in mathematical models, are not statistically significant. For example, we investigate traffic composition: vehicles are divided in different categories (motorcycles, cars, trucks, and buses). Probably these variables are not significant because traffic is homogeneous (approximately 95% of vehicles are cars). Also, we consider access density, number of crossing pedestrians and cyclists, presence of sidewalk, curb, and parking. Moreover, posted speed limit urban roads are distinguished by the same speed limit of 50 km/h, which is the only one in Italy for urban areas that does not effect on speed. In fact, analyzing collected data, the speed limits are substantially exceeded by 68.6% of drivers.

The research contributes to literature on examining vehicle speeds on Italian urban roads. This study can be useful to both traffic manager and road designers because models developed could implement design guidelines, especially regarding

road tangents design. However, it would be appropriate to validate models in other urban roads to determine if drivers have similar behaviors. Furthermore, the results can help roadway designers to better understand expected speeds and, as a result, design the urban roads considering drivers' real driving behavior.

We are currently working on driver behavior study in free flow condition. Particularly, we are establishing time intervals that define isolated vehicle in Italian urban roads. This is important to provide methods to estimate free flow speed as a function of infrastructure geometric characteristics.

## References

1. European Road Safety Observatory (2006), 2007. Speeding, Available from Internet: [www.erso.eu](http://www.erso.eu).
2. Tarris, J.P., Poe C.M, Mason, J.M., Goulias, K.G.: Predicting operating speeds on low-speed urban streets: regression and panel analysis approaches. *Transportation Research Record: Journal of Transportation Research Board*. 1523, 46–54 (1996).
3. Poe C.M., Mason J.M. Analyzing influence of geometric design on operating speeds along low-speed urban streets: mixed-model approach. In *Transportation Research Record: Journal of Transportation Research Board*, No 1737, 2000, pp. 18–25.
4. Fitzpatrick, K., and Carlson, P. Selection of Design Speed Values. In *Transportation Research Record 81th, Annual Meeting, TRB, Washington, D.C., 2001*, pp. 3–11.
5. Wang J., Dixon K.K., Li H., Hunte M.P. Operating-speed model for low-speed urban tangent streets based on in-vehicle global positioning system data. In *Transportation Research Record: Journal of Transportation Research Board*, No 1961, 2006, pp. 24–33.
6. Aronsson Karin F.M.: Speed characteristics of urban streets based on driver behaviour studies and simulation. Doctoral Thesis in Infrastructure Royal Institute of Technology Stockholm, Sweden, 2006.
7. Ali, A.T., Flannery, A., Venigalla, M.M.: Prediction models for free flow speed on urban streets. In *Transportation Research Record 86th Annual Meeting*, 2007.
8. American Association of State Highway and Transportation Officials. *A Policy on Geometric Design of Highways and Streets*. Washington, D.C. (2001).
9. Fitzpatrick, K., Carlson, P., M. Brewer, and Woolridge M. Design Factors That Affect Driver Speed on Suburban Streets. In *Transportation Research Record 80th Annual Meeting*, Washington, D.C., 2001, pp. 18–25.
10. Fitzpatrick, K., Shamburger, C.B., Krammes, R.A., and Fambro, D.B. Operating Speed on Suburban Arterial Curves. In *Transportation Research Record No 1579, TRB, National Research Council, Washington, D.C., 1997*, 89–96.
11. Polus, A., Livneh, M., and Craus, J. Effect of Traffic and Geometric Measures on Highway Average Running Speed. In *Transportation Research Record No 960*, 1984.
12. Ministero delle Infrastrutture e dei Trasporti. *Norme Funzionali e Geometriche per la Costruzione delle Strade*. Decreto Ministeriale n.6792, November 5, Rome, Italy, 2001.
13. Washington, S.P., Karlaftis, M.G., Mannering, F.L. *Statistical and Econometric Methods for Transportation Data Analysis*, 2nd ed. Boca Raton, London & New York: CRC Press, ISBN 9781420082869, 2011, pp. 63–143.

# SRAM Design Using Memristor and Self-controllable Voltage (SVL) Technique

Naman S. Kumar, N.G. Sudhanva, V. Shreyas Hande,  
Mallikarjun V. Sajjan, C.S. Hemanth Kumar and B.S. Kariyappa

**Abstract** Power consumption is one of the major hurdles in scaling the technology of memories. Static Random Access Memories (SRAMs), in particular, contribute to major dissipation of static and dynamic power in processors that are utilized for building caches, buffers, reservation stations, portable devices, etc. [1]. The leakage current is increased to a greater extent because of technology scaling [2]. Introduction of nonvolatility into any system is advantageous. The time required to start a closed (due to abrupt power loss or forced shutdown) application decreases when nonvolatility exists. This paper proposes a design of SRAM cell utilizing memristors and Self-Controllable Voltage (SVL) techniques. Simulation on standard 45 nm CMOS technology is done in Cadence Virtuoso. The results obtained show that leakage power reduces by 90.59% than that of 7T cell [3].

**Keywords** Low power · SRAM · SVL · Memristor · Nonvolatile

---

N.S. Kumar (✉) · N.G. Sudhanva · V.S. Hande · M.V. Sajjan · B.S. Kariyappa  
Department of EC, RVCE, Bengaluru 560059, Karnataka, India  
e-mail: naman2204@gmail.com

N.G. Sudhanva  
e-mail: sudhanva.ng@gmail.com

V.S. Hande  
e-mail: shreyas.hande@gmail.com

M.V. Sajjan  
e-mail: mallikarjunvsaj@gmail.com

B.S. Kariyappa  
e-mail: kariyappabs@rvce.edu.in

C.S.H. Kumar  
Department of EC, GCE, Ramanagara 562159, Karnataka, India  
e-mail: hemanthcs@rediffmail.com

## 1 Introduction

Memories are a vital part of any processor. Any process requires usage of memory at one stage or the other for storage of data. SRAMs are vital part of processors as they are made use in building caches, buffers, reservation stations, portable devices, etc., and occupy about 94% of die-area [4] as per International Technology Road Map for Semiconductors (ITRS).

A new model [3] shows that after the standby mode, the data is retained even though there is no ground rail in the cell. The additional n-MOS used at the bottom of conventional 6T SRAM cell removes the ground connection when the device is not being used thereby reducing the static power consumed. The cell [5] designed by Shyam Akashe, Shishir Rastogi, and Sanjay Sharma requires a positive feedback to retain the cell data without refresh cycle. Though there is an advantage of decrease in the circuitry (for maintenance of charge within the system), the complexity of the cell has been increased. The model proposed by Shyam Akashe et al. [6] utilizes Self-controllable voltage technique.

According to Chua [7], Memristor (also known as memristance) is the fourth missing element. It is a two terminal passive element in which magnetic flux between the terminals is a function of amount of electric charge that has passed. The component was realized for the first time by Williams [8] in HP labs. A novel memristor based SRAM design is proposed [9] by Thangamani which shows improvement in the power consumption. Non-volatility is introduced into the SRAM cell with the introduction of memristor. As observed from the design, there is an overhead with respect to static noise margin. This paper proposes to use Memristor, to introduce nonvolatility and SVL technique, in order to reduce the leakage current. Basavaraj and Kariyappa [10] have designed a new model which uses low power while using a single bit-line. Here usage of single line reduces the requirement of additional power supply.

## 2 Methodology

### 2.1 Self-controllable Voltage Technique

There are many techniques which help to reduce leakage current but they have their own advantages and disadvantages. The two well-known technologies to reduce leakage current are Multiple Threshold CMOS (MTCMOS) and Variable Threshold CMOS (VTCMOS).

In MTCMOS, the leakage power is reduced by disconnecting the power supply using switches (SWs) having very high threshold voltage ( $V_{th}$ ) whereas in VTCMOS, the leakage power is reduced by increasing substrate bias. Considering the above techniques, a Self-Controllable Voltage Level (SVL) Circuit has been developed.

**Upper SVL (U-SVL) Technique** This consists of single n-MOS and p-MOS which are connected in series. A clock is given as input to gates of both n-MOS and

p-MOS. When clock is low, p-MOS is turned on and n-MOS is turned off. As p-MOS are strong one in nature, they allow maximum voltage to the load circuit without any drop. This mode is called as active mode. During standby mode, i.e., when clock becomes high, n-MOS is turned on and p-MOS is turned off. As n-MOS is weak one, there is some drop across it and voltage supplied to the load circuit is lesser than the maximum voltage ( $V_{dd}$ ). Thus, a drain-to-source voltage, i.e., drain voltage of the n-MOS is given by

$$V_{dsn} = V_{dd} - \text{drop across n-MOS}(mv), \quad (1)$$

where  $m$ —number of n-MOS

And  $v$ —voltage drop of single n-MOS

Decrease in  $V_{dsn}$  will increase the barrier height that in turn decreases the Drain-Induced Barrier Lowering (DIBL) effect which leads to rise in  $V_{th}$ . This results in decrease of sub threshold current of the n-MOS, hence leakage current through the inverter decreases.

**Lower SVL (L-SVL)** It consists of single n-MOS and p-MOS which are connected in series. A clock is given as input to gates of both n-MOS and p-MOS. When clock is high, n-MOS is turned on and p-MOS is turned off. As n-MOS are strong zero in nature, they allow low voltage (ground) to the load circuit. This mode is called as active mode. During standby mode, i.e., when clock becomes low, p-MOS is turned on and n-MOS is turned off. As p-MOS is weak zero, there is some drop across it and voltage supplied to the load circuit is not zero voltage. Thus, a substrate bias voltage, i.e., back-gate-bias voltage is given by

$$V_{sub} = -mv \quad (2)$$

Here, decrease in DIBL effect result in rise of back gate bias which leads to increase in  $V_{th}$ . Thus, the leakage current is decreased.

**Combined U-SVL and L-SVL circuit** Here U-SVL and L-SVL are connected to load circuit, so the effect of both U-SVL and L-SVL comes into the picture. And DIBL effect is further decreased due to drop in  $V_{dsn}$  which is given by

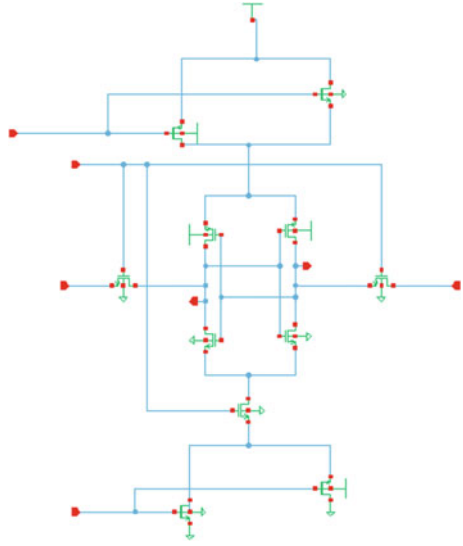
$$V_{dsn} = V_{dd} - 2mv \quad (3)$$

Figure 1 depicts transistor schematic of the usage of Upper SVL and Lower SVL to 7T SRAM cell. Layout schematic of the same is shown in Fig. 2.

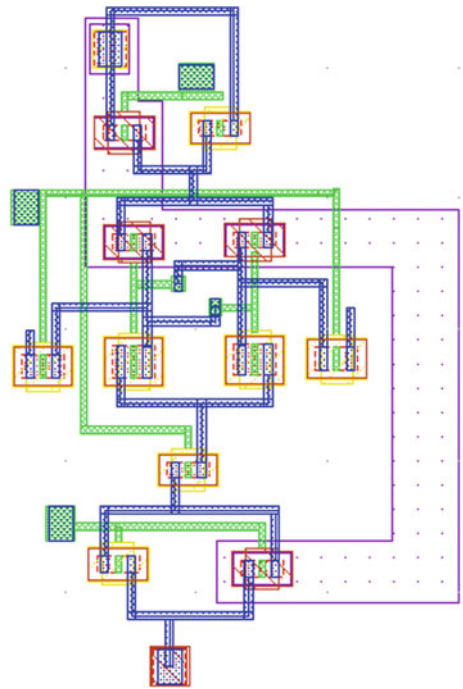
## 2.2 Memristor

Memristor is a missing component found recently in the field of electronics. There were only three basic elements namely resistor ( $R$ ), inductor ( $L$ ), and capacitor ( $C$ ). Also there are four basic quantities, voltage ( $V$ ), current ( $C$ ), flux ( $\emptyset$ ), and amount

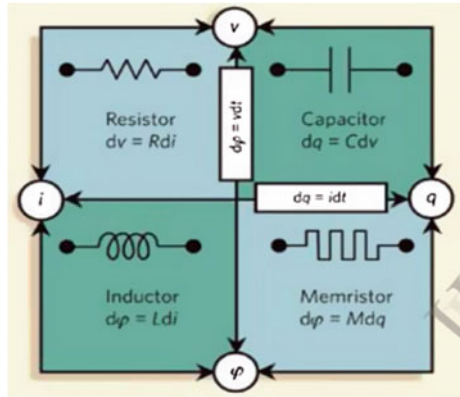
**Fig. 1** Transistor schematic of 7T SRAM with combined U\_SVL and L-SVL circuit



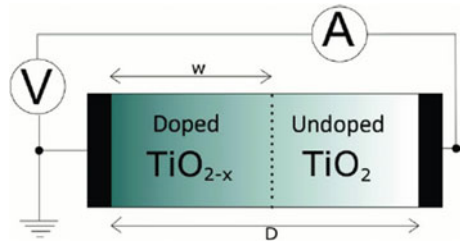
**Fig. 2** Layout schematic of 7T SRAM with combined U\_SVL and L-SVL circuit



**Fig. 3** Relationship between basic components and memristor [9]



**Fig. 4** Circuit using memristor [9]



of charge ( $q$ ). The relationship between  $\phi$  and  $V$  was previously unknown. The memristor gives this correlation

$$d\phi = M.dq \tag{4}$$

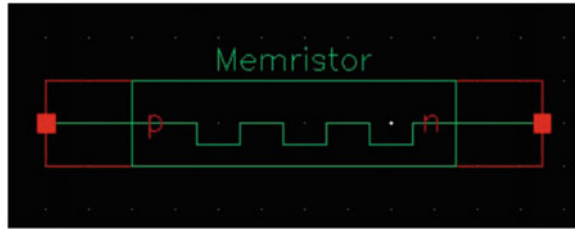
Hence, memristor is termed as fourth basic element along with resistor, inductor, and capacitor. Figure 3 shows the relation between the quantities.

**Working principle** The first memristor which was realized by HP in 2008 [8] consisted of a 50 nm of titanium dioxide ( $TiO_2$ ) sandwiched between two 5 nm electrodes. Initially, there are two layers of  $TiO_2$  films as shown in Fig. 4. One of them has a slight depletion of oxygen atoms (doped) and the other layer is undoped. The oxygen vacancies act as charge carriers. Hence, the doped layer will have a smaller resistance compared to the undoped region.

When electric field is applied, oxygen vacancies drift apart, and thus moving the boundary between low-resistance and high-resistance region. Hence, the resistance of the whole film depends on quantity of charges and the direction of the charges that have passed through it. Hence depending on the direction as well as quantity of charges that has passed through the memristor, its resistance changes.



**Fig. 5** Memristor cell obtained from Threshold Adaptive Memristor (TEAM) Model



**Fig. 6** Proposed cell

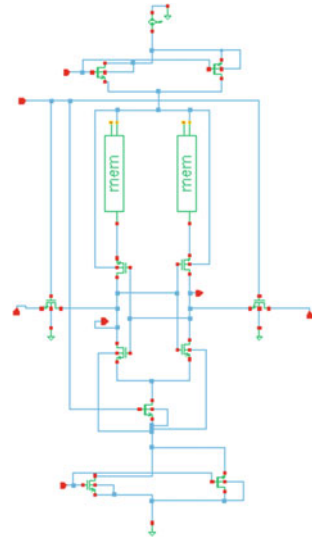


Figure 5 depicts the Threshold Adaptive Memristor (TEAM) Model that is used in the implementation.

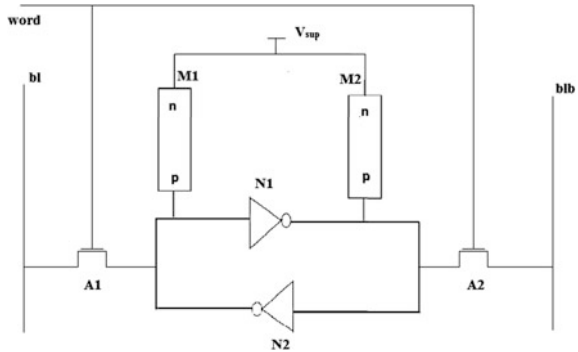
### 3 Proposed Method

Figure 6 shows the design of a SRAM with memristors and SVL technique. Two memristors are placed at the inputs of 2 NOT gates. Two access transistors are connected just as in a conventional 7T SRAM cell.

The principle involved here is that, during write phase, memristors act like resistors. So, it is similar to that of the write operation of conventional 7T SRAM cell. During read operation, depending on the previously written data, the resistance of the memristors will be changed. This will be the key to reading the data. That is depending on the resistance of the two cells, 0 or 1 is read.

The equivalent circuit of SRAM cell with memristors is shown in Fig. 7. The working operations of the circuit are explained below.

**Fig. 7** Equivalent of SRAM with Memristor



### 3.1 Write Operation

During write 1 operation, bl is made 1 and blb is made 0. After this, word line is asserted to enable a particular cell. Since current is flowing from p to n (positive direction) in M1 and it is flowing from n to p (negative direction) in M2, the resistance of the M1 is increased while the resistance of M2 remains the same. So during the next read operation (that is read 1 operation), resistance of M1 is higher compared to the resistance of M2. The write operation is complete by observing  $Q$  and  $Q_b$ . In this case,  $Q$  will be raised to high voltage ( $V_{dd}$ ) while  $Q_b$  will be pulled to gnd.

During write 0 operation, bl is made 0 and blb is made 1. After this, word line is asserted to enable a particular cell. Since current is flowing from n to p (negative direction) in M1 and it is flowing from p to n (positive direction) in M2, the resistance of the M2 is increased while the resistance of M1 remains the same. So during the next read operation (that is read 0 operation), resistance of M2 is higher compared to the resistance of M1. The write operation is complete by observing  $Q$  and  $Q_b$ . In this case,  $Q_b$  will be raised to high voltage ( $V_{dd}$ ) while  $Q$  will be pulled to gnd.

### 3.2 Read Operation

The first step in the read operation is to precharge the bl and blb to half of the  $V_{dd}$  (i.e.,  $V_{dd}/2$ ). Depending on the values stored in the cell, bl and blb will be raised to  $V_{dd}$  or pulled down to the gnd voltage. If bl is raised to high voltage, it means data read is 1, otherwise data read is 0.

During read 1 operation, the data stored in the cell will be 1. It means write 1 operation has been performed previously. Hence, the resistance of the memristor M2 is higher compared to the resistance of M1. Both bl and blb will be at a voltage

level of  $V_{dd}/2$ . Since the resistance of M1 is high, voltage drop of the supply voltage  $V_{sup}$  is higher across the memristor M1. The drop is very less across M2.

This high voltage drop acts as input to the NOT gate N1. So voltage at the input of N1 is  $V_{dd}$  while voltage at bl line is  $V_{dd}/2$  (because of pre-charge). So current flows from input of N1 to bl, driving bl to high voltage  $V_{dd}$ . The exact reverse process takes place at input of N2, driving voltage of blb to gnd voltage. In final stage, bl is high and blb is low. So read 1 operation will be completed successfully.

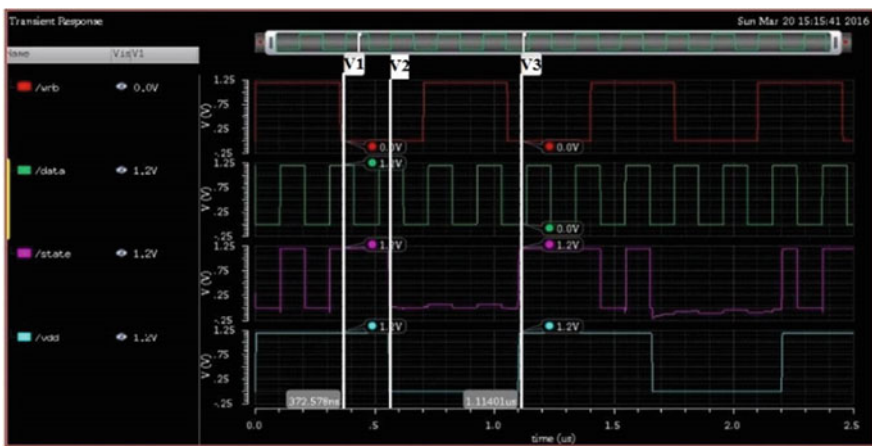
During read 0 operation, the data stored in the cell will be 0. It means write 0 operation has been performed previously. Hence, the resistance of the memristor M2 is higher compared to the resistance of M1. Both bl and blb will be at a voltage level of  $V_{dd}/2$ . Since the resistance of 2 is high, voltage drop because of the supply voltage  $V_{sup}$  is higher across the memristor M2. The drop is very less across M1.

This high voltage drop acts as input to the NOT gate N2. The voltage at the input of N2 is  $V_{dd}$  while voltage at blb line is  $V_{dd}/2$ . So current flows from input of N2 to blb, driving blb to high voltage,  $V_{dd}$ . The exact reverse phenomenon takes place at input of N2. In the final stage, blb is high and bl is low. Hence, read 0 operation will be completed successfully.

## 4 Results

Nonvolatility, as proposed by the introduction of memristor, can be observed from the transient analysis shown in Fig. 8.

The simulation consists of four waves namely wrb (write/read), data, state, and  $V_{dd}$ . “wrb” is an input signal which is used to determine mode of operation (i.e., write or read). If wrb = 1 it is write operation and if wrb = 0, it is read operation.

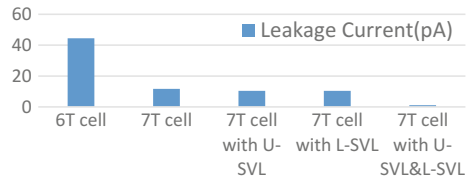


**Fig. 8** Transient analysis of novel SRAM cell

**Table 1** Leakage current values of various SRAM cells

SRAM cell	Leakage current (in pA)	Leakage power (in pW)
7T cell	11.69	14.028
7T cell with U-SVL	10.36	12.432
7T cell with L-SVL	10.43	12.516
Proposed SRAM cell	1.10	1.32

**Fig. 9** Comparison of leakage powers of SRAM cells



“Data” is an input signal which is the data given to the SRAM cell.  $V_{dd}$  is an input signal, which is the supply voltage to the SRAM cell. “State” is the output signal which is the final output of the sense amplifier.

In the simulation, from time  $t = 0$  to reference point  $V1$ ,  $wrb = 1$  and the signal “ $V_{dd}$ ” is high. So all the transistors are switched on and the cell is in write mode. During this time, data is constantly toggling and the output signal (state) is following the data. This signifies write 0 and write 1 operations.

Between the time interval  $V1$  and  $V2$ ,  $wrb$  has become 0 but signal “ $V_{dd}$ ” is still high. It means that the cell is in read mode. Since the previously written data is 1, the output signal “state” is high throughout this time interval. This signifies read 1 operation.

Between time intervals  $V2$  and  $V3$ , the signal “ $V_{dd}$ ” is 0. It means the supply voltage to the cell is 0. Hence, all the transistors are turned off. And during this phase, no operation can be performed on the cell. This signifies the “hold” operation of the SRAM cell.

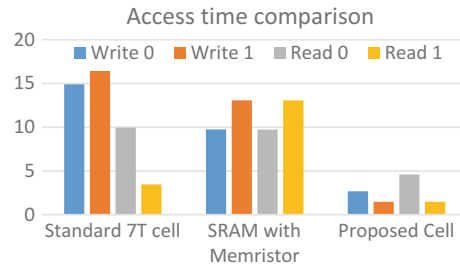
After the time interval  $V3$ , signal “ $V_{dd}$ ” is high. That means supply voltage is switched on and all the transistors are switched on as well. At this point signal  $wrb = 0$  so cell is in read mode and the current data = 0. But the last legal write operation performed on the cell was “write 1” operation (between  $t = 0$  and  $t = V1$ ). This data is retained even when the cell has been turned off (between  $t = V2$  and  $t = V3$ ). So when read operation is performed after  $V3$ , output is high even though the current data is 0. This confirms the nonvolatility of the cell.

It can be observed from Table 1 that leakage current reduces as the upper SVL circuit and lower SVL circuits are integrated with the SRAM cell. Pictorial depiction of the same can be seen in Fig. 9.

The access times of different SRAM cells for read and write operations are shown in Table 2. It shows that proposed SRAM cell can be accessed quicker than other cells. The pictorial representation of access time of SRAM cells is shown in Fig. 10.

**Table 2** Access time of SRAM cells

Access Time	Write 0 (in ps)	Write 1 (in ps)	Read 0 (in ps)	Read 1 (in ps)
Standard 7T SRAM cell	14.9	16.43	9.959	3.463
SRAM with Memristor	9.742	13.08	9.73	13.07
Proposed SRAM Cell	2.68	1.48	4.6	1.48

**Fig. 10** Comparison of access times of SRAM cells

## 5 Conclusion

SRAM cell is designed using Memristor and SVL technique and simulated successfully. SVL technique reduces the leakage power and inclusion of memristors brings in the property of non-volatility to the cell. The radical decrease (90.59%) of leakage power is helpful in portable devices and systems where conserving power is an important criterion. The reduction in access time makes it suitable to be utilized in time-critical missions. Thus, the proposed cell can be utilized in fields like military, industries, etc.

## References

1. P. Upadhyay, Sarthak Ghosh, R. Kar, D. Mandal, S. P. Ghoshal, (2014), Low Static and Dynamic Power MTCMOS Based 12T SRAM Cell for High Speed Memory System. In: 11th IEEE International Joint Conference on Computer Science and Software Engineering, DOI:[10.1109/JCSSE.2014.6841869](https://doi.org/10.1109/JCSSE.2014.6841869)
2. A. Agarwal, C. H. Kim, S. Mukhopadhyay, and K. Roy (2004) Leakage in Nano-Scale Technologies: Mechanisms, Impact and Design Considerations. In: Proc. Of the 41st Design Automation Conference (DAC 04), pp. 6–11, DOI:[10.1145/996566.996571](https://doi.org/10.1145/996566.996571)
3. A. Agarwal, H. Li, and K. Roy, (2002) DRG-Cache: A data retention gatedground cache for low power. In: Proceedings of the 39th Design Automation Conference, pp. 473–478, DOI:[10.1145/513918.514037](https://doi.org/10.1145/513918.514037)
4. Qikai Chen, Saibal Mukhopadhyay, Aditya Bansal and Kaushik Roy (2005) Circuit-aware Device Design Methodology for Nanometer Technologies: A Case Study for Low Power SRAM Design. In: TVLSI, Volume 13, Issue 3, DOI:[10.1109/DATe.2006.243868](https://doi.org/10.1109/DATe.2006.243868)

5. Shyam Akashe, Shishir Rastogi, Sanjay Sharma (2011) Specific power illustration of proposed 7T SRAM with 6T SRAM using 45 nm technology. In: International Conference on Nanoscience, Engineering and Technology (ICONSET), DOI:[10.1109/ICONSET.2011.6167982](https://doi.org/10.1109/ICONSET.2011.6167982)
6. Shyam Akashe, Meenakshi Mishra, Sanjay Sharma (2012) Self-controllable Voltage level Circuit for Low Power High Speed 7T SRAM cell at 45 nm technology. In: IEEE, DOI:[10.1109/SCES.2012.6199024](https://doi.org/10.1109/SCES.2012.6199024)
7. L. O. Chua (1971) Memristor -the missing circuit element. In: IEEE Transactions on circuit theory, vol.18, no.5, pp. 507–519, DOI:[10.1109/TCT.1971.1083337](https://doi.org/10.1109/TCT.1971.1083337)
8. S. Williams (2008) How we found the missing Memristor. In: IEEE SPECTRUM, vol.45, PP.28–35, DOI:[10.1109/MSPEC.2008.4687366](https://doi.org/10.1109/MSPEC.2008.4687366)
9. Thangamani. V (2013) Design of Low Power Resistive Random Access Memory using Memristor. In: International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 9, ISSN: 2278-0181
10. Basavaraj Madiwalar and Dr. Kariyappa B S (2013) Single Bit-line 7T SRAM cell for Low Power and High SNM. In: International Multi Conference on Automation, Computing, Control, Communication and Compressed Sensing, pp. 223–228 DOI:[10.1109/iMac4s.2013.6526412](https://doi.org/10.1109/iMac4s.2013.6526412)

# A Study on Certificate-Based Trust in MANETs

K. Gowri Raghavendra Narayan, T. Srinivasa Rao, P. Pothu Raju and P. Sudhakar

**Abstract** Mobile Ad hoc Network (MANET) is a wireless network of mobile nodes or mobile devices are connected without any specific infrastructure; it has the properties of self-configuration and self-maintenance nature, so it does not have any centralized control. The basic requirement for ad hoc network is to provide security. For providing authentication (trust) to the nodes, secure routing is the primary challenge. For improving the security, the most important thing is to examine the trustworthiness of nodes or mobile devices by identifying selfish and malicious nodes in the network. We are using the concept of trust and applying it on popular routing protocols of MANETs like DSR and AODV for providing secure routing. And also we are using the following algorithms, calculation of trust value (level) algorithm, identification of node misbehavior and modification of trust value algorithm, certificate distribution algorithm, and route innovation and maintenance algorithm for all the nodes in the network. Now, our concept of certificate-based trust and algorithms deals with the malicious nodes and provides a high secure routing.

**Keywords** MANET · Security · Selfish nodes · NS-2

## 1 Introduction

MANETs are wireless network of mobile nodes, in this, the mobile nodes or mobile devices can directly communicate with all the remaining nodes with their radio ranges, if the nodes do not have direct communication range uses intermediate nodes [1]. For the purpose of forwarding packets, the nodes will have to depend on the nodes. Because of the open network and dynamic topology, limited band width and energy constraints make this type of networks vulnerable to network attacks [2]. For the purpose of communication, the mobile nodes use the concept of routing.

---

K.G.R. Narayan (✉) · T.S. Rao · P.P. Raju · P. Sudhakar  
Vasireddy Venkatadri Institute of Technology, Guntur, India  
e-mail: Kgmnarayan9@vvit.net

And the routing is done using the routing protocols. For secure communication between the nodes, the routing protocol must deal with the existence of malicious nodes and also the attack made by malicious nodes. Many trust-based solutions have been proposed but they have the problem of falsified trust values, spoof ID, and sometimes may be the source or destination nodes are deeply involved in authenticating the trust values of the intermediary nodes [3].

Here in this paper, we are proposing the concept of certificate-based trust with a trust certificate and trust value. And also we are proposing algorithms such as calculation of trust value (level) algorithm, identification of node misbehavior, and modification of trust value algorithm, certificate distribution algorithm, and route innovation and maintenance algorithm for all the nodes in the network, so that every node will get a trust value and certificate that states the level of trustworthiness of each node individually, during the path discovery, the communication node will get the certificate details and trust value of a node which it is going to connect. Only if the trust level of that intermediate node is bigger than a fixed value then it starts communication with that particular as intermediate node otherwise, i.e., if the trust level is less than a fixed value then that particular node is avoided in communication and its certificate is revoked. So that it will provides us most reliable and secure end-to-end communication between the nodes.

## 2 Related Work

Amir Pirzada et al. proposed establishing trust in pure ad hoc networks [4], they have used the concept of trust agents and they used parameters precision and acknowledgement of packet, blacklisted nodes, gratuitous route replies. Wei Liu et al. proposed A Study on Certificate Revocation in Mobile Ad Hoc Networks [5], a threshold-based certificate revocation approach which will perform revocation of certificates of the node based on a threshold value so that it will provide the secure communication. Himadri Nath Saha et al. proposed Study of Different Attacks in MANET with Its Detection and Mitigation Schemes [6], it gives the classification of several attacks which are common against the ad hoc network routing protocols, such as mobile versus wired attackers, passive versus active attacks inside versus outside attacks, layered attacks, data versus control traffic attacks.

Frank Kargl et al. proposed Advanced Detection of Selfish or Malicious Nodes in Ad hoc Networks [7], they used various sensors that can detect several kinds of selfish nodes. Suppose any node detected as selfish node multiple times then it is eliminated from the network. Amir Khusru et al. proposed A Novel Methodology to Overcome Routing Misbehaviors in MANET Using Retaliation Model [8], in this model, there are various parameters like number of packets forwarded, number of packets received, packet forwarding ratio which are used to find out bonus and grade points. The bonus point given to the nodes depends on the packet drop and grade is for isolating the selfish nodes.



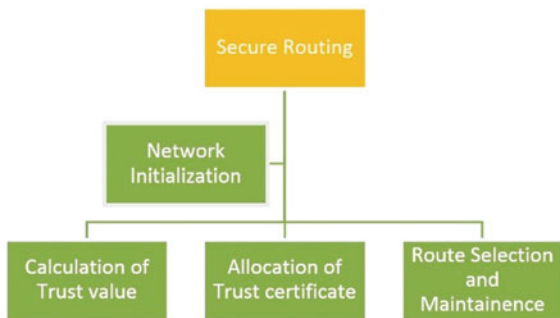
### 3 Methodology

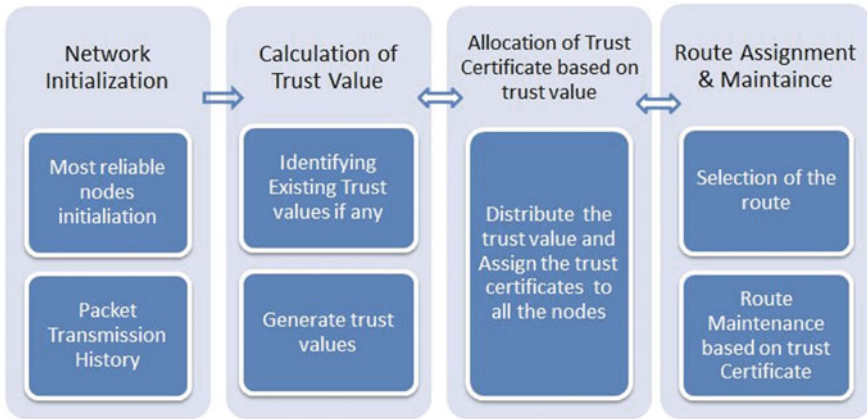
Figure 1 describes the flow of the proposed work, secure routing based on trust value and trust certificate. Figure 2 describes the detailed steps taken in the proposed idea.

The following Algorithm-1 explains the calculation of trust value and Algorithm-2 describes allocation of trust certificate, Algorithm-3 explains certification revocation, Algorithm-4 describes the identification of misbehavior and modification of trust value, and Algorithm-5 will give the route innovations and route maintenance.

```
-----  
Algorithm 1: Calculation of Trust value (level)  
-----  
Initialization: Maxtrustval = N; where N>1  
               Mintrustval=1;  
1: if | Trustval(Nodei) > ΔSigExpTime||Nodei==Newnode |  
   2: Trustval(Nodei) ← Compute(Trustval(Noden))  
   end  
end  
-----
```

**Fig. 1** Description of proposed idea





**Fig. 2** Detailed description of proposed idea

---

**Algorithm 2:** Certificate Allocation

---

Initialization: Maxtrustval = N; where N>1  
 Mintrustval=1;  
 1: if | Trustval(Node<sub>i</sub>) > Mintrustval && !SigExpTime |  
     2: Node<sub>i</sub> ← Allocate(TrustCertificate)  
     End  
 End

---



---

**Algorithm 3:** Certificate Revocation

---

Initialization: Maxtrustval = N; where N>1  
 Mintrustval=1;  
 1: if | Trustval(Node<sub>i</sub>) < Mintrustval || sigExpTime |  
     2: Node<sub>i</sub> ← Revoke(TrustCertificate)  
     End  
 End

---

---

**Algorithm 4:** Identification of misbehaviour and modification of Trust value
 

---

```

Initialization: Maxtrustval = N; where n>1
    Mintrustval=1;
    Count=0;
  for each i , where i is a node
    if | no.of recvpkts<no.ofsendpkts |
      count=count+1;
      Trustval(Nodei) -- ;
    end
  if | no.ofrecvpkts > no.ofsentpkts |
    count=count+2;
    Trustval(Nodei) -- ;
    end
  if | no.ofrecvpkts==no.ofsentpkts |
    count=0;
    Trustval(Nodei)++;
    end
  end
  for each i , where i is a node
    if | pktdelay > ThresholdVal |
      count=count+3;
      Trustval(Nodei)--;
    end
  end
end
end
end
  
```

---

---

**Algorithm 5:** Route Innovation and Maintenance

---

Sender side:

```

1: NodeDestination  $\xrightarrow{\text{Request}}$  Trustval
2: Node  $\epsilon$  (Source, intermediate)  $\xrightarrow{\text{Verify}}$  Trustval(Node $\epsilon$ )
3: for valid Trustval from 1 to n
4: if |Certificate(Node $\epsilon$ ) == TRUE |
    4.1. Node  $\xrightarrow{\text{Checks}}$  Trustval(FromNode, ToNode)
        4.1.1. if | FromTrustval > MinTrustval &&
            ToTrustval > MinTrustval |
        4.1.2 Node  $\xrightarrow{\text{Appends}}$  (Node_Id, Trustval) to RREQ
        4.1.3. Node  $\xrightarrow{\text{Sends RREQ}}$  PathDiscoverer
            end
        end
    4.2: else
        4.2.1: Node  $\xrightarrow{\text{Revoke}}$  TrustCertificate
    end

```

Receiver Side:

```

5: Node  $\xleftarrow{\text{Receives}}$  RREQ
6: Node  $\xleftarrow{\text{Verifies}}$  RREQ
7: if | !NodeDestination |
    7.1. Repeat steps 1 to 4
    end
8: else
    8.1. for each RREQ
        8.1.1. NodeDestination  $\xleftarrow{\text{Computes}}$  Multiplies(Trustval, (id1,id2,...idn))
        8.1.2. NodeDestination  $\xleftarrow{\text{Finds}}$  Maximum (Multiplies(Trustval, (id1,id2,...idn)))
    8.2. NodeDestination  $\xleftarrow{\text{Selects}}$  Routeoptimum
    8.3. NodeDestination  $\xleftarrow{\text{Generate}}$  RREP
    8.4. NodeDestination  $\xrightarrow{\text{Sends RREP}}$  NodeSource
        end
    end
end

```

---

The first level is network initialization, in that network is initialized with required number of nodes. After which network initialization calculation of trust value for each node is calculated using Algorithm-1 and then for each node the trust certificate is issued, for that purpose Algorithm-2 is used. After allocation of certificates to all the nodes, the communication is started using the concept of routing, for that purpose Algorithm-5 is used. After some time, if any node behavior is abnormal then its certificate revocation is done for that particular node and its trust value is also decreased. For the purpose of identification of misbehavior and modification of trust value in the network, Algorithm-4 is used and for the purpose of certificate revocation of a node, Algorithm-3 is used.

### 4 Simulation and Result

The simulations are done by using Network Simulator-2 [9,10]. Figures 3, 4, 5, 6, and 7 show screenshot of the simulation of proposed idea in Network Simulator-2.

Figures 3 and 4 show initialization network with required number of nodes and generation of trust values for every node. Figure 5 shows certificate allocation to every node. Figure 6 shows the identification of malicious node and certificate revocation and route modernization with trusted nodes. Figure 7 shows the modernized route with trusted certificate nodes.

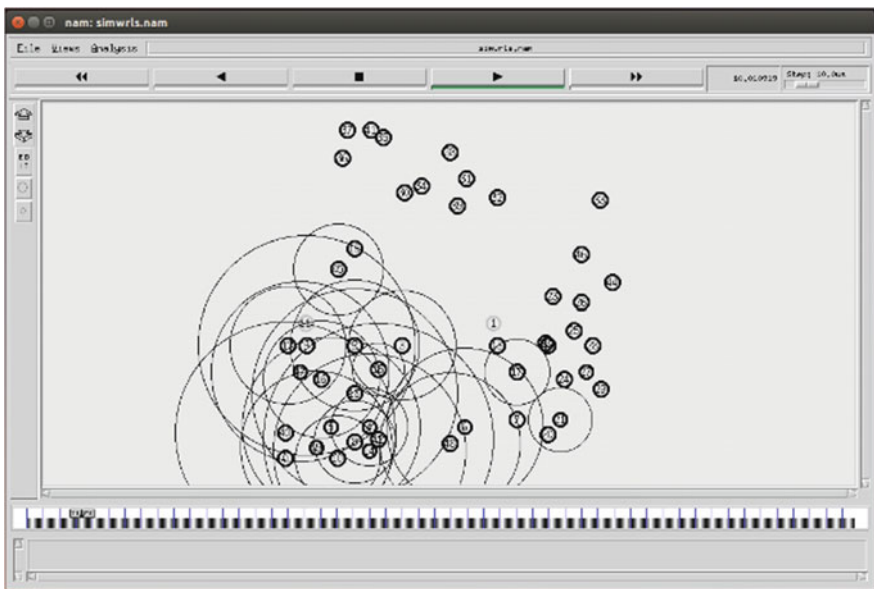


Fig. 3 Screenshot of network initialization in NS-2

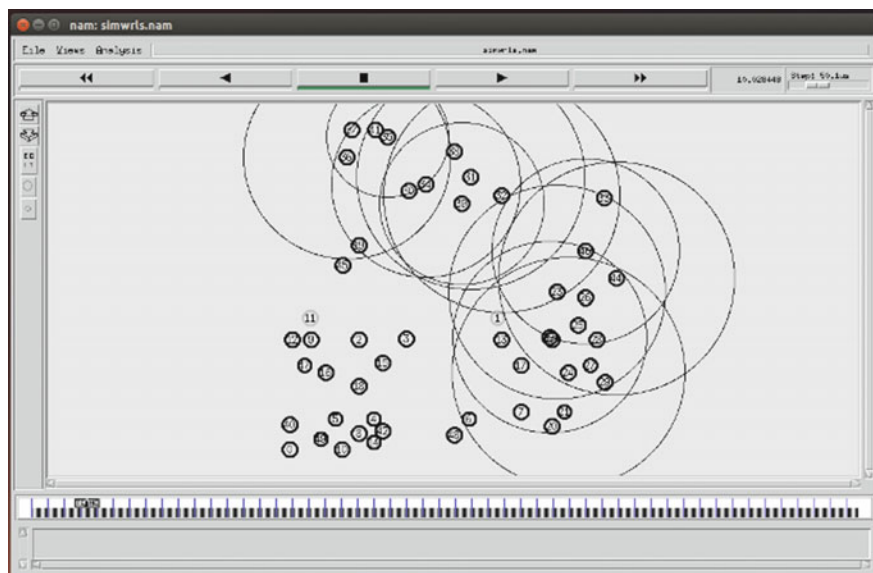


Fig. 4 Screenshot of network initialization in NS-2

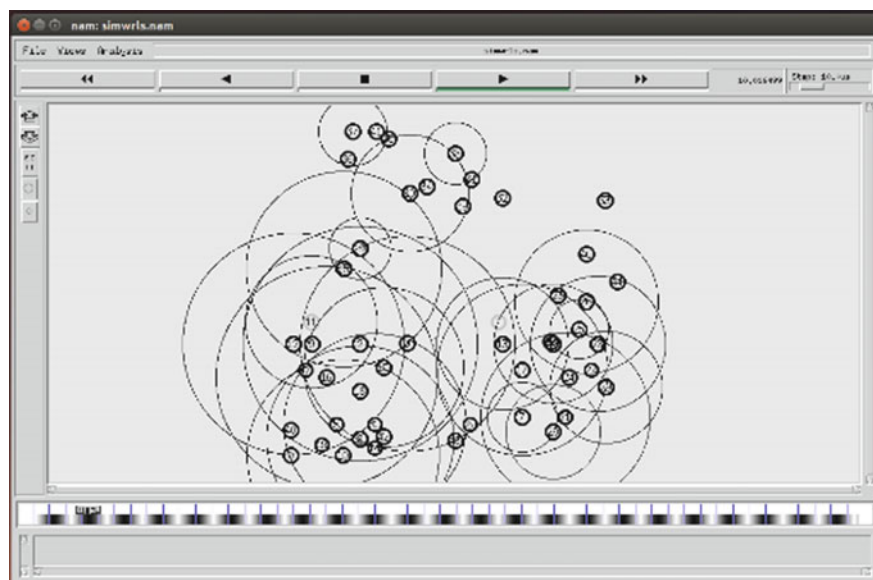
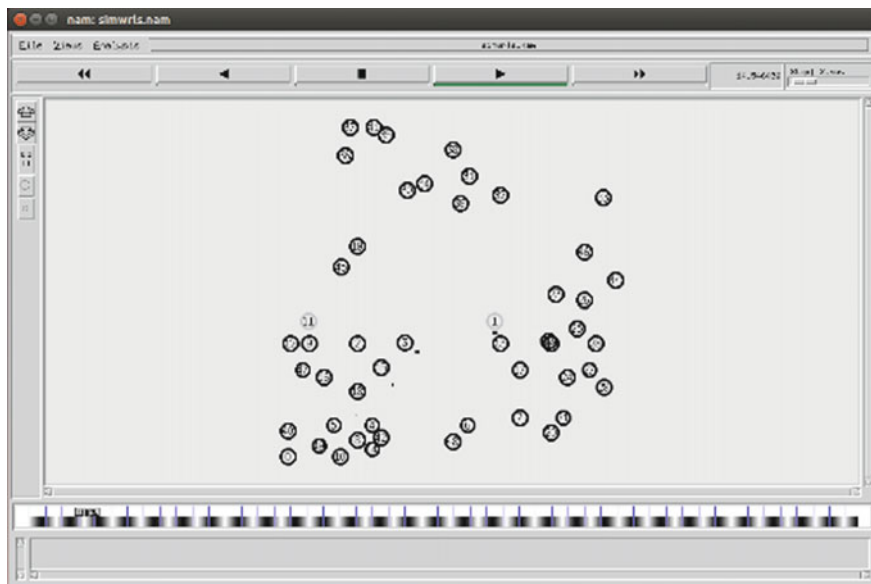
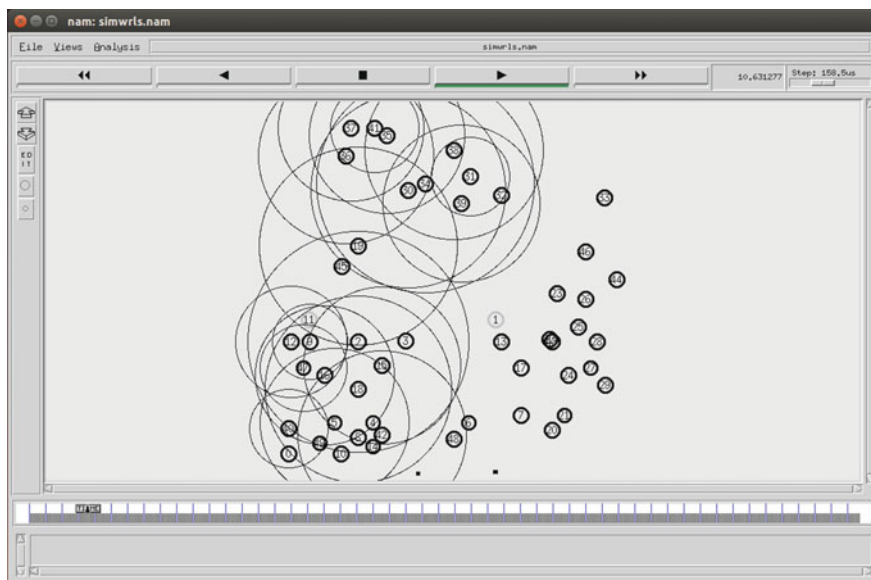


Fig. 5 Distribution of trust values and certificate NS-2



**Fig. 6** Identification of malicious node in NS-2



**Fig. 7** Screenshot of route modification and omitting the malicious node in communication in NS-2

**Table 1** Simulation parameters in NS-2

Parameters	Value
No of nodes	3, 5,10, 20, 30, 40, 50
Maximum speed	10 m/s
Minimum speed	2 m/s
Simulation time	150 s

**Table 2** Simulation parameters in NS-2

Parameters	Value on NS2
Simulation time	150 s
Simulation area	900 × 700 m
Node movement model	Random Waypoint Model
Speed	2–10 m/s
Traffic type	FTP
Packet size	1040 bytes
Bandwidth	2 Mb/s
Packet rate	2 Mb/s
No of nodes	3, 5,10, 20, 30, 40, 50
No of source destination	1,2
Connection	TCP
Propagation	Two ray ground

In this scenario, we have taken the following parameters and values on NS-2. (Table 1 and 2)

Figures 8, 9, 10, 11, and 12 give the results of simulation process with secure routing, the results are drawn from Gnuplot [12]. Figure 8 shows the throughput comparison in the network between undetected malicious nodes in basic routing and detection and omitting the malicious nodes in the network with trust certificate. The red line in graph shows the decrease in throughput with the malicious node and the green line shows omitting the malicious node in the communication path and increase in throughput of the network with secure routing. Figure 9 shows PDR of basic model and secure model. Red line in graph shows the decrease in the PDR with presence of malicious node in basic model and the green line shows increase in PDR even the existence of malicious node by omitting them in the communication path. Figure 10 shows average delay in the network due to attacker. The red line in graph shows delay in the network without any attacker and the green line shows the delay of the network with attacker and its elimination in the network. Figure 11 shows the routing load due to the presence of malicious node, the network is trying to omit the node in the path and selecting another trusted node as intermediate node. For that, RREQ and RREP are depicted in the graph. Red line in graph shows RREQ and green line shows RREP. Figure 12 shows the trust certificate utilization rate between the presence of attacker and absence of attacker. The red line shows attacker traffic and its elimination. During this time, the rate of certificate utilization increases with increase in the number of nodes. The green line shows the certificate utilization rate with absence of attacker in the network.



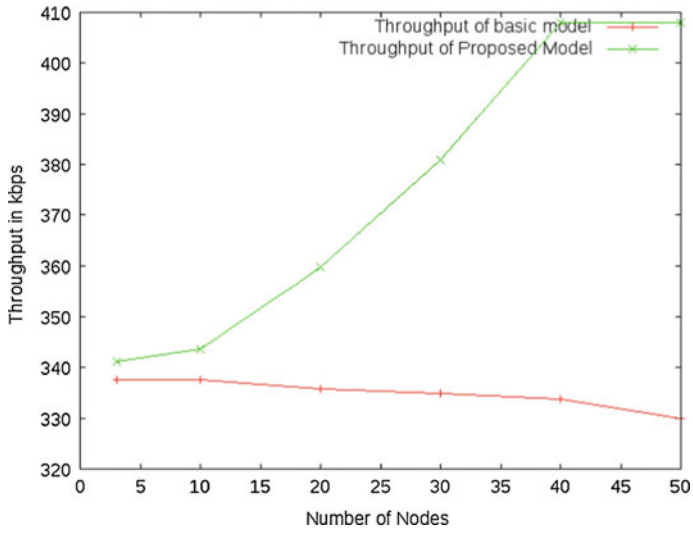


Fig. 8 Throughput comparison between basic routing and secure routing

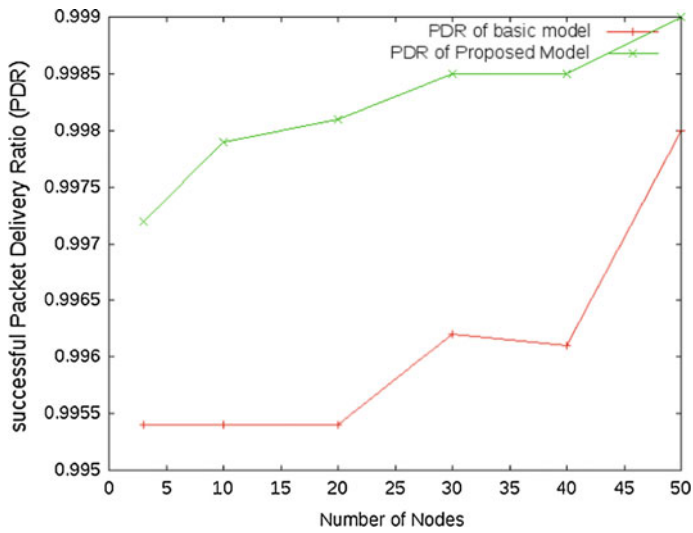


Fig. 9 PDR comparison

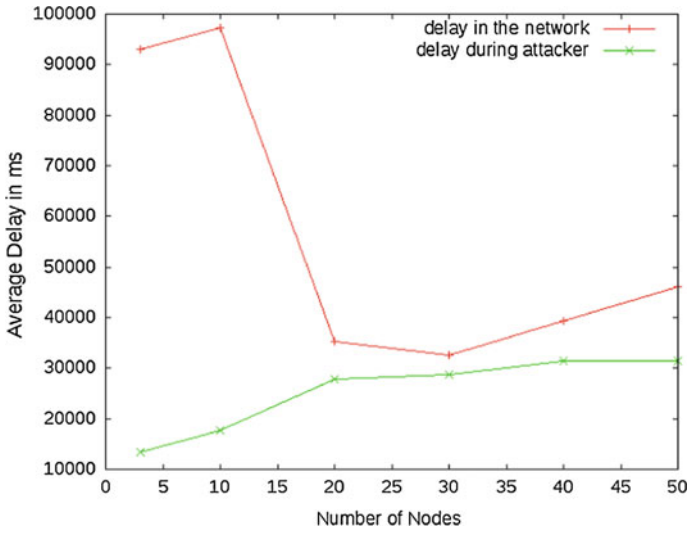


Fig. 10 Delay comparison

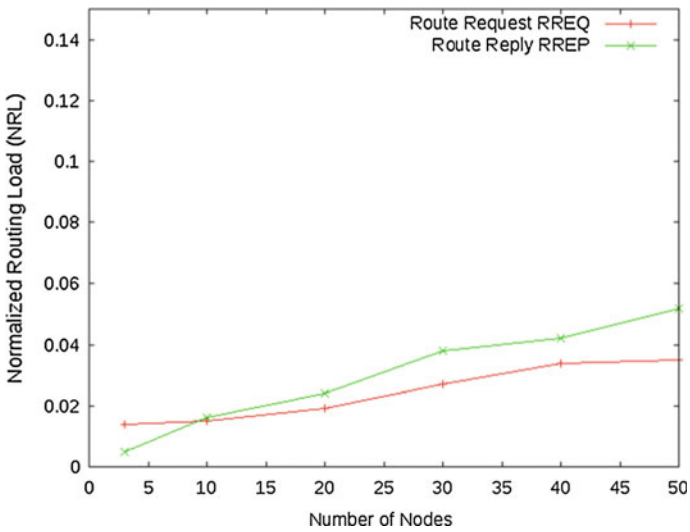


Fig. 11 Routing Load comparison b/w RREQ and RREP

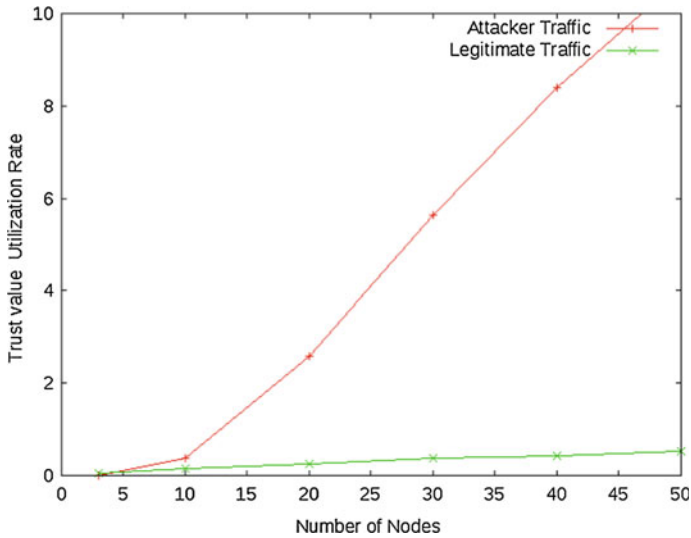


Fig. 12 Trust certificate utilization Rate

## 5 Conclusion

Our proposed idea, a certificate and algorithms based trust provides the selfish node free network and most trustworthy path and communication in MANETs. With this trustworthy path, the communication becomes secure that leads to good throughput, best PDR and less network delay. Because every node is awarded a trust value and trust certificate before forwarding packet to intermediate node. Our proposed certificate-based trust with trust values and certificate distribution algorithm, revocation algorithm, misbehavior identification algorithm and route innovation, maintenance algorithm provides a reliable and a maximum assurance of hidden malicious and selfish node detection as well as prevention and provides a secure and reliable communication in the network.

## References

1. K. Gowri Raghavendra Narayan, N.V. Ramana Gupta, Dr. M.V. Rama Krishna, "A Survey on Performance Ascertainment of MANET Routing Protocols Using NS-2", International Journal of Engineering and Technical Research (IJETR) ISSN: 2321-0869 (O) 2454-4698 (P), Volume-3, Issue-9, September 2015.
2. Wenjia Li, Anupam Joshi: "Security Issues in Mobile Ad Hoc Networks—A Survey". WISE-WA, wireless sensor networks for the city-wide Ambient Intelligence.
3. M. Ali Dorri and Seyed Reza Kamel and Esmail kheyrikhah Security Challenges In Mobile Ad-hoc Networks: A Survey, International Journal of Computer Science & Engineering Survey (IJCSES) Vol. 6, No. 1, February 2015. DOI:[10.5121/ijcses.2015.6102](https://doi.org/10.5121/ijcses.2015.6102).

4. Asad Amir Pirzada, Chris McDonald “Establishing Trust In Pure Ad-hoc Networks”, in Conferences in Research and Practice in Information Technology-2004.
5. Wei Liu, Hiroki Nishiyama, Nirwan Ansari, Nei Kato “A Study on Certificate Revocation in Mobile Ad Hoc Networks” IEEE Communications Society subject matter experts for publication in the IEEE ICC 2011.
6. Himadri Nath Saha, Dr. Debika Bhattacharyya, Dr. P. K.Banerjee Aniruddha Bhattacharyya, Arnab Banerjee, Dipayan Bose, “Study Of Different Attacks In MANET With Its Detection & Mitigation Schemes” International Journal of Advanced Engineering Technology, IJAET/Vol.III/ Issue I/January-March, 2012/383–388 E-ISSN 0976-3945.
7. Frank Kargl, Andreas Klenk, Stefan Schlott, Michael Weber “Advanced Detection of Selfish or Malicious Nodes in Ad Hoc Networks” in Security in Ad-hoc and Sensor Networks chapter, springer, DOI [10.1007/978-3-540-30496-8\\_13](https://doi.org/10.1007/978-3-540-30496-8_13), Print ISBN 978-3-540-24396-0, Online ISBN 978-3-540-30496-8, pages: 152-165.
8. Md. Amir Khusru Akhtar, G. Sahoo, “A Novel Methodology to Overcome Routing Misbehavior in MANET Using Retaliation Model” International Journal of Wireless & Mobile Networks (IJWMN) Vol. 5, No. 4, August 2013, DOI: [10.5121/ijwmn.2013.5414](https://doi.org/10.5121/ijwmn.2013.5414), pages: 187–202.
9. “Network simulator” <http://www.isi.edu/nsnam/ns/>, and <https://www.nsnam.org>.
10. “NS Simulator for Beginners” written by Eitan\_Altman\_Tania\_Jimenez.
11. Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specification, IEEE Std. 802.11, 1997.
12. Gnuplot <http://www.gnuplot.info/>.

# User Interest Modeling from Social Media Network Graph, Enriched with Semantic Web

Bansari Shah, Ananda Prakash Verma and Shailendra Tiwari

**Abstract** This paper intends to model a user's interests from his activities on social media as well as the extended social network he is a part of. Besides that, this paper also intends to reflect the change in one's behaviors over a period by following his activities on social media over time. It is observed that one's interests may vary with time, thus inferring it only from his own activity is perhaps inconsistent. But, if a user's network is also traced for similar and correlated actions, then it is possible to infer one's consistent interest to a certain extent. This paper introduces an approach that will consider a user and his network's activities, about common tags, people, and organizations from their social media posts which prepare a strong ground for the desired inference. Furthermore, to enrich inferred interests, the approach mentioned here also utilizes the semantic web using DBpedia ontology.

**Keywords** Social media · Network graph · Semantic web · Ontology

## 1 Introduction

Applications that aim for personalization of content addressing to the desires or needs of an individual would like to adopt a functionality like maintaining user profile which demands an understanding of his interests. When there is a cold start or no prior information available about a user, observing the user on social medias

---

B. Shah (✉) · A.P. Verma · S. Tiwari  
Coviam Technology and Services Pvt Ltd, Bengaluru, India  
e-mail: bansari.shah@coviam.com  
URL: <http://www.coviam.com/>; <https://ahduni.edu.in/seas/>

A.P. Verma  
e-mail: ananda.verma@coviam.com

S. Tiwari  
e-mail: shailendra.tiwari@coviam.com

B. Shah  
School of Engineering and Applied Science, Ahmedabad University, Ahmedabad, India

like Facebook, Twitter, LinkedIn, and knowledge sharing platforms like Quora and Stack overflow can provide relevant information about the user, e.g., looking at Facebook activity one can find out what is happening in person's life. Similarly, Twitter tells a lot about person's domain of interest and from the LinkedIn profile and activities one can build user's professional profile. User's profile on social media is considered as the main source which gives information like geolocation, interest, and skills. But it is observed that generally user profile might remain static, as it does not reflect continuous changes in user's interest over the period of time, which can be tracked by considering activities of a user along with the network he is a part of on social media.

Various approaches have been introduced in the area of user interest modeling from social media mainly work-around text analysis, topic modeling, topic classification, network analysis, and semantic enrichment of user profiles. Inferred interests of a user will be helpful in recommending articles of their interests from the news as well as other information sharing web portals. Moreover, for an organization, it becomes very effective to understand employees through changes in their interests. It helps in improving employees' efficiency and engaging them in right direction.

In this paper, an attempt has been made to infer a user's interests, which also incorporates the change in interests over the period of time. It mainly works on user's Twitter feed. Twitter is one of the most active public microblogging platforms since 2007 with more than 313 Million users and altogether about 1 Billion posts by 2016.<sup>1</sup> In this paper, user's interest has been inferred by considering his tweets along with the tweets collected from the extended network he is a part of. The inclusion of activities from user's network makes it possible to infer interest even if he is least active, which is the main contribution of this paper. In accordance with that, interests are inferred by considering frequently mentioned and highly weighted hashtags, people, and organizations in tweets, derived using Stanford Named-Entity Recognizer (NER). Further results are generalized through Wikipedia using DBpedia Ontology.

This paper is further organized as follows: Sect. 2 describes related work in the domain of user interest modeling. In Sect. 3, the approach of this paper has been explained with derivation and evaluation of results in Sect. 4 and concluded with a glimpse of future work in Sect. 5.

## 2 Related Work

All prior approaches introduced for discovering interests generally work around the content of tweets. They mainly use text analysis and classification methods. A basic approach for inferring interest from tweets follows Bag-of-Words model.

---

<sup>1</sup><https://about.twitter.com/company>.

It classifies tweets into relevant categories. But studies in [1] argue that it does not work well with tweets as they are small and limited to 140 characters only.

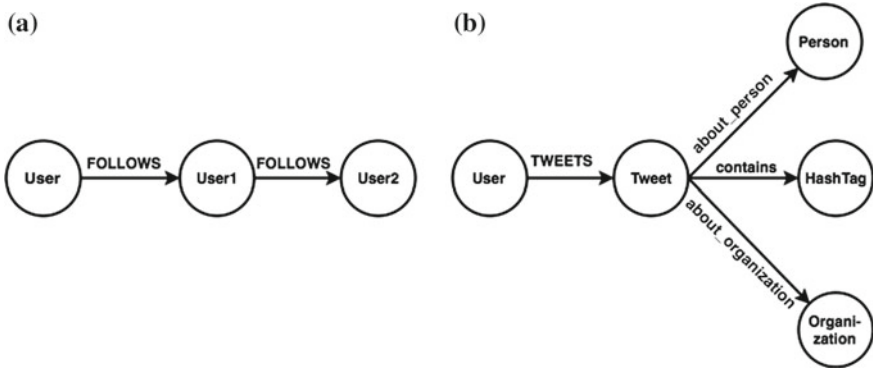
In addition to the bag-of-words model, topic Models [2] like LDA (Latent Dirichlet Allocation) and LLDA (Labeled LDA) have played a major role in direction of inferring user interest in terms of topics by considering all of the user's tweets as a single document. Where each document is a distribution of topics and each topic is a distribution of words. LDA (Unsupervised) [3] is a topic modeling approach based on Bayesian analysis of the original texts, where a topic is defined as a set of co-occurring words based on their probabilistic distribution. Topics generated through LDA can be labeled by referring knowledge bases like Wikipedia [4], where Wikipedia article containing top topical terms can be a label for that topic. LLDA (supervised LDA) works similarly like LDA. But it requires corpus as a labeled set of categories [5, 6]. Topics generated through LLDA maps to these categories only. Both approaches work on tweets posted as well received by a user. Bhattacharya [6] introduced Twitter list based approach which outperformed LLDA. However, experimental results [7] suggest that it is difficult to infer user's interest by considering only tweets/retweets, as it includes user's daily activities as well.

Other approaches introduced for user interest modeling build a semantic profile of a user from his activities on social media. The technique employed by TUMS [8] generates a semantic profile for a user based on entity, hashtag, and topic-based user modeling strategies. The method proposed by Abel [9] constructs semantic user profiles by linking tweets to a relevant news article that represents individual's activities in a better context. As mentioned in [10], topics were discovered by identifying categories from Wikipedia folksonomy of categories sub-tree that maps with disambiguated entities extracted from tweets. Such knowledge bases can be exploited [11], through hierarchical relationship among Wikipedia categories to create a richer user profile.

Unlike above approaches, the approach introduced in this paper not only uses machine learning techniques or does semantic enrichment of a user profile to infer user  $u$ 's interests but also considers other users whom user  $u$  follows (i.e.,  $u$ 's followings) up to certain levels that result into weighted inference. Besides that, the introduced approach also intends to reflect the change in user's interests over a period by considering tweets created over time.

### 3 Approach

In this paper, the approach for inferring user interests is divided into two phases: (1) Inferring niche interest of a user from the network and the activity graph (2) Inferring general interests by enriching niche interests with the semantic web. They are further explained in Sections [3.1 and 3.2] (Fig. 1).



**Fig. 1** **a** Network graph of a user up to depth 2, **b** Activity graph

### 3.1 *Inferring Niche Interest*

Here, niche interest is defined as the interest inferred from activities of user ‘ $u$ ’ and other users he follows. The conceptual flow of deriving niche interest is explained in Fig. 2. In order to do that, network graph of user  $u$  is built from his following list collected using Twitter public API. It follows property graph model implemented with the Neo4j graph database [12]. Built property graph represents a 2 level network of users on social media whom  $u$  follows, each with its property set and with a relation FOLLOWS as shown in Fig. 1. The depth of network can be increased to higher levels.

As the user would follow others from various possible domains, either he is interested in or he is from the same domain. So, considering a user with the network he follows up to some depth as shown in Fig. 1 makes results stronger. The various domains, that other users in  $u$ ’s network belong to, are again inferred using DBpedia as explained in subsection 3.2

With tweets collected from all user nodes in above-created network of a user, an activity graph for each node is built as shown in Fig. 1. Created activity graph forms the network of hashtags, people, and organizations mentioned in tweets. People and organizations that the user talks about in tweets are retrieved using Stanford NER (3 class classifier) from Stanford coreNLP library [13]. These represent main entities about what/whom the posted tweets are. Niche interests are inferred at last as a weighted aggregation of these entities from the whole network.

### 3.2 *Inferring General Interest*

In this phase of inferring general interests of a user, mainly people and organizations from niche interest derived as in Sect. 3.1 are considered. The data says that if a user is talking about certain people/organizations in his tweet, he might be interested in the



domain they belong to. That domain is considered as a general interest of a user. To derive it, DBpedia ontology [14] of person and organization is used. It helps to extract structured information from Wikipedia. Also in the case of other users on Twitter whom user ‘u’ follows, are well-known organizations or are experts of some domain and has information on Wikipedia, their domain of expertise is retrieved as well using a DBpedia. The conceptual flow of Inference of general interest is shown in Fig. 3.

In the case of a person, relevant domains are retrieved from RDF graph of DBpedia ontology using SPARQL query language by referring relationship `dbo:field` and `rdfs:type`. It says what a person is, e.g.—a cricketer or scientist and belongs to which domain. Same can be applicable in the case of organizations through property `dbo:industry`.

For example, if a user is following “Andrew Ng”, based on the results retrieved from DBpedia by executing below SPARQL, it can be inferred that he is interested in “Artificial Intelligence”.

```

PREFIX dbpedia: <http://dbpedia.org/resource/>
PREFIX dbo: <http://dbpedia.org/ontology/>
SELECT ?known_for WHERE{
dbpedia:Andrew_Ng dbo:field ?known_for .
}
    
```

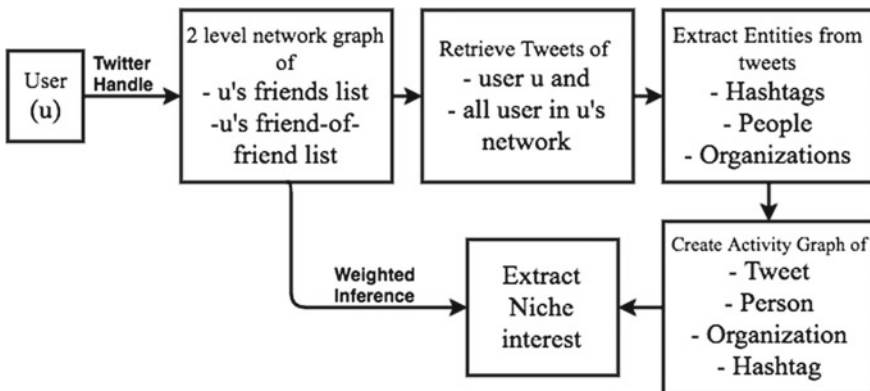


Fig. 2 Conceptual diagram of inferring niche interest

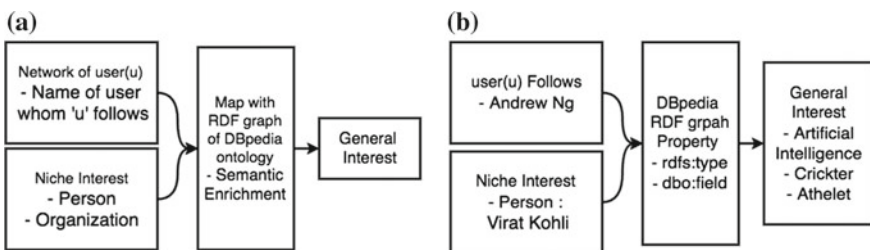


Fig. 3 a Conceptual diagram of deriving general interest b Example of inferred general interest

Similarly if “Virat Kohli” is a niche interest of a user, from the results derived from DBpedia by executing below SPARQL, it can be inferred that user likes cricket and is interested in sports.

```
PREFIX dbpedia: <http://dbpedia.org/resource/>
SELECT ?class_type, ?label WHERE{
dbpedia:Virat_Kohli a ?class_type.
?class_type rdfs:label ?label.
}
```

### 3.3 Calculation of Results

Results are derived as a weighted aggregation of interests common to a user and other users on social media he follows, as it makes this inference more stronger. The network a user follows has more impact in case the user is least active but follows many other users of his interest on social media.

Weight is defined as a property of a relation between each user nodes in user  $u$ 's network and the associated niche interests. It decreases as the depth of the network increases. Besides weight, the occurrence of niche interests is also considered to differentiate daily activities, as its occurrence will be quite less and it will be in direct relation to the user.

For inferring weighted niche interest, in this approach, weights are statically defined as following:

- Weight of interests in direct relation to user  $u$ 's activity: 1.0
- Weight of interests through first level of users in the network  $u$  follows: 0.5
- Weight of interests through second level users in the network  $u$  follows: 0.3.

Normalization of this weight can help to retain same results in different weight allocation.

Niche interests which are common to all user nodes in user  $u$ 's network will increase its weight and that will be inferred as a strong niche interest. In this weighting model, common interest among all user nodes in  $u$ 's network will be weighted with 1.8. Based on this calculation, results are shown in Sect. 4.

## 4 Results and Evaluation

### 4.1 Results from Proposed Method

Based on the calculation of weighted inference of user's niche interest as given in Sect. 3.3, inferred results are as in Table 1.

**Table 1** Inferred niche interest (based on hashtag, person, and organization mentioned in tweets) in order of weight—for Twitter user @rahulvit09

HashTag in tweets	Times occurred	Weight	Person in tweets	Times occurred	Weight	Organization in tweets	Times occurred	Weight
Docker	99	1.8	Shailesh Kumar	11	1.8	IBM	11	1.8
Ubuntu	11	1.8	Hadoop	19	1.8	CERN	7	1.8
Cassandra	45	1.8	Donald Trump	17	1.8	NSA	17	1.8
gaming	10	1.8	Hillary Clinton	1	1.0	Google	53	1.8
Bigdata	248	1.8	Swift	1	1.0	Linux	2	1.5
Xbox	3	1.5	Zayed AI	18	0.8	BitTorrent	1	1.0

Through this approach of building user's network and activity graph with time as a property of a node along with hashtags mentioned in tweets, change in user's interests over the period can also be reflected as shown in Table 2.

## 4.2 Evaluation of Proposed Method

The Interest of a user is best known to himself. Keeping that in the account, introduced approach is evaluated by focusing on Twitter users around us, which includes users of both types—frequent and infrequent user.

**Table 2** Change in interest over a period for Twitter user @rahulvit09

Year	2017	2016	2015
Interest	Docker, Machine learning	fifthel, Akka, Cassandra, Aerospike	Bigdata, Microsoft, Facebook, Machine Learning, Linux

**Table 3** Examples of interests predicted with proposed approach and compared with user's actual interest

Twitter user	User's niche interest	Predicted top niche interest	User's general interest	Predicted general interest
@rahulvit09 – infrequent user – 580 tweets – 187 friends	Cassandra, Hadoop, Docker, big data, Google	Cassandra, Hadoop, Docker, Google, NSA, IBM,	Technology, Travel, Distributed Computing, Cloud, Computing	Distributed Computing, Internet, Software, Cloud, Computing,
@Tiwari_tweets – infrequent user – 423 tweets – 105 friends	Modi, NASA, Trump, Times now, YearInSpace	Modi, Virat Kohli, NASA, Trump, TimesNow	Cricketer, Artist, Newspaper, written work, writer, politics	Athlete, Cricketer, Financial services, Newspaper, Literature
@Mr_Spark – infrequent user – 8 tweets – 26 friends	Nyx, Pudge, Slardar, Newbee, DotaPit	Nyx, Trump, Slardar, Pudge, Dota2 Newbee	Video game, soccer player	Video game, athlete, soccer player
@kinshukkar – frequent user – 1758 tweets – 255 friends	VR, Bot, ISRO, Coviam, Google, Microsoft	Facebook, Messenger, Bot, VR, Rahul Gandhi, Microsoft, ISRO	Software, comedian, Soccer player, business person, Aerospace Eng	Software, artist, monarch, Internet, Soccer player, businessperson

For each user at least 10,000 tweets, both from the ones he posted as well as from his two-level deep network were analyzed by creating the network and an activity graph. As a result of this, niche and general interests were inferred. After the analysis, the authentication of accuracy was done with the user in consideration himself as shown in Table 3. The feedback from the concerned users implied that niche interests were quite accurate in reference to their activities on Twitter, but general interests were too generic. This inference of general interests can be improved by including more properties from the DBpedia ontology.

## 5 Conclusion and Future Work

In this paper, an approach was introduced to infer user's interests from social media, considering activities of both user and the network he is a part of. Derived interests were enriched through the semantic web using DBpedia ontology. Evaluation of method stated that approach was quite effective for inferring interest even if the user is least active. Besides that, change in user's interest is also reflected over a period. In future, inferred interest will be used as a part of user's profile for building recommendation system.

## References

1. Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M.: Short Text Classification in Twitter to Improve Information Filtering. Proc. 33rd Int. ACM SIGIR Conf. Res. Dev. Inf. Retr. SE - SIGIR'10. 841–842 (2010).
2. Blei, D.M., Lafferty, J.D.: Topic Models. Text Min. Classif. Clust. Appl. 71–89 (2009).
3. Blei, D.M., Edu, B.B., Ng, A.Y., Edu, A.S., Jordan, M.I., Edu, J.B.: Latent Dirichlet Allocation. J. Mach. Learn. Res. 3, 993–1022 (2003).
4. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic Labelling of Topic Models. Proc. 49th Annu. Meet. Assoc. Comput. Linguist. 1536–1545 (2011).
5. Ottoni, R., Casas, D. Las, Pesce, J.P., Wagner, M.J., Wilson, C., Mislove, A., Almeida, V.: Of Pins and Tweets: Investigating How Users Behave Across Image-and Text-Based Social Networks. Proc. Eighth Int. AAAI Conf. Weblogs Soc. Media. 386–395 (2014).
6. Bhattacharya, P., Zafar, M.B., Ganguly, N., Ghosh, S., Gummadi, K.P.: Inferring user interests in the Twitter social network. Proc. 8th ACM Conf. Recomm. Syst. - RecSys'14. 357–360 (2014).
7. Wagner, C., Liao, V., Pirolli, P., Nelson, L., Strohmaier, M.: Its not in their tweets: Modeling topical expertise of twitter users. Proc. - 2012 ASE/IEEE Int. Conf. Privacy, Secur. Risk Trust 2012 ASE/IEEE Int. Conf. Soc. Comput. Soc. 2012. 91–100 (2012).
8. Tao, K., Abel, F., Gao, Q., Houben, G.J.: TUMS: Twitter-based user modeling service. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 7117 LNCS, 269–283 (2012).
9. Abel, F., Gao, Q., Houben, G.J., Tao, K.: Semantic enrichment of twitter posts for user profile construction on the social web. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6643 LNCS, 375–389 (2011).

10. Michelson, M., Macskassy, S. a: Discovering users topics of interest on twitter: a first look.'10 Proc. fourth Work. Anal. noisy unstructured text data. 73–80 (2010).
11. Kapanipathi, P., Jain, P., Venkataramani, C., Sheth, A.: User interests identification on Twitter using a hierarchical knowledge base. Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8465 LNCS, 99–113 (2014).
12. Miller, J.: Graph Database Applications and Concepts with Neo4j. Proc. 2013 South. Assoc. 141–147 (2013).
13. Manning, C.D., Bauer, J., Finkel, J., Bethard, S.J., Surdeanu, M., McClosky, D.: The Stanford CoreNLP Natural Language Processing Toolkit. Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr. 55–60 (2014).
14. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R., Hellamnn, S.: DBpedia - A cystallization point for the Web of Data. Web Semant. Sci. Serv. Agents World Wide Web. 7, 154–165 (2009).

# Swarm and Artificial Immune System-Based Intelligence Techniques for Geo-Spatial Feature Extraction

Lavika Goel, Mallikarjun Swamy and Raghav Mantri

**Abstract** The paper deals with the analytical study of nature inspired algorithms with respect to geo-spatial feature extraction. Geo-spatial feature extraction is a very important aspect of remote sensing and lately the procedures have become more and more complex due to the usage of multispectral satellite images which increases the number of dimensions for the classification problem. The paper intends to reinforce the point that latest nature inspired algorithms are better classifiers when it comes to land cover classification of satellite images. As a part of it, this paper focuses upon two techniques, one of which is a hybridized version of two nature inspired algorithms namely Bat Algorithm and Charged System Search and the other being Clonal Selection Algorithm. The first technique presented in the paper combines the advantages of two algorithms to produce better classified images. The dataset that has been worked upon for testing the efficiency of the technique is a multispectral satellite image of Alwar region in Rajasthan, India containing seven bands.

**Keywords** Nature inspired intelligence · Land cover feature extraction  
Charged system search · Bat algorithm · Clonal selection

## 1 Introduction

Geo-Spatial feature extraction is the process of having all the features of interest identified in a spatial or geographic data. This process is a very important aspect in remote sensing. The applications of this procedure are diverse. It can be used in fields like ecology, geology and industries like defence, Natural resources identification, Disaster Risk Management and Climate Change Adaptation.

---

L. Goel (✉) · M. Swamy · R. Mantri  
Department of Computer Science and Information Systems, BITS Pilani, Pilani Campus,  
Pilani, India  
e-mail: lavika.goel@pilani.bits-pilani.ac.in

Land Cover feature extraction is an important division in geo-spatial feature extraction. It generally involves the usage of satellite images. Most of the satellite images that are produced everyday are of no physical significance on their own. The images are nothing but digital number values observed for reflectance of various radiations across electromagnetic spectrum. The different wavelengths at which images were captured are separated by filters and correspond to the increased dimensionality in the dataset. Due to aforementioned reasons, every pixel in a satellite image may be represented using many values and the representation may not be accurate depending upon the resolution of the satellite image. Also, some bands (digital numbers corresponding to a range in the electromagnetic spectrum) in the dataset may be irrelevant to the feature under consideration. Moreover, various atmospheric and topographic corrections render the data slightly inaccurate. These challenges call for a heuristic based classification technique which produces sub optimal solutions in a short span of time, taking into consideration various shortcomings of the dataset.

Nature inspired algorithms facilitate the process of producing quick results at the cost of optimality. The paper highlights two such techniques. First, is a hybrid of two swarm intelligence techniques namely Bat Algorithm and Charged System Search. The best optimizing characteristics of each algorithm are amalgamated in the hybrid version to produce better results. Bats echolocation ability for Bat Algorithm and Coulombs law coupled with Newton's law of motions for Charged System Search are inspirations behind their respective algorithms. Second is Clonal Selection Algorithm (CLONALG), which is inspired by the clonal selection theory of acquired immunity given by Burnet [1]. The best optimizing characteristics of CLONALG are cloning and somatic hyper-mutation of the antibody that has the most affinity with the antigen, so that it can further capture similar antigens even more efficiently.

## 2 Related Work

Extensive research has been conducted for the problem of land cover feature extraction. Feature extraction by creation of segments (or super pixels) is done by Simple Linear Iterative Clustering which is followed by classification using Support Vector Machines and Random Forests [2]. Labour cost in Land cover feature extraction can be greatly reduced by the usage of automatic classification techniques given enough prior knowledge which is followed by classification using fuzzy support vector machines to produce good results [3]. Recently, a lot of attention has been dedicated to support vector machines (SVMs) for the classification of multispectral remote sensing images [4–9]. SVMs have often been found to provide higher classification accuracies than other widely used pattern recognition techniques, such as the maximum likelihood and the multilayer perceptron neural network classifiers [10]. Spectral Index Ratios are also an effective means for land cover feature extraction. The Normalized Difference Class Index (NDCI)



calculated as part of this methodology gives ratios, which produce realistic results because this technique also takes into consideration the common areas of spectral signatures of various classes, the image is classified, into its purview to mitigate misclassification [11, 12]. Usage of pan-sharpening algorithms also enhances the accuracy of results by a decent extent [13].

The choice of Bat Algorithm, Charged System Search and Clonal Selection Algorithm out of a multitude of nature inspired algorithms was made after consulting the proposed taxonomy of nature inspired computational intelligence techniques [14]. Swarm Intelligence (SI), which is a part of Nature Inspired computational Intelligence techniques, is the collective behaviour of decentralized and self-organized systems. The techniques involved in SI are usually inspired from the natural elements like attraction between two charged bodies [15] and echolocation of bats to find prey [16]. Research for SI started in the late 1980 s. The applications for SI have been diversifying ever since. It is not only used for the conventional optimization problems but also various other problems like library materials acquisition, communications, medical dataset classification, dynamic control, heating system planning, moving objects tracking and prediction.

After identifying the algorithms (BA and CSS) to work on, literature survey has been more focussed. In his paper, XS Yang talks about a new meta-heuristic technique inspired by echolocation of bats [16]. This particular paper provided appropriate insights into the origins of BA and facilitated its implementation. The proposed variation of BA was inspired by its data clustering application [17]. Similarly, for CSS two papers were considered helpful, one for the explanation of its origin [15] and other for its application of data clustering [18].

Nature Inspired techniques like Biogeography Based Optimization have already been implemented for geo-spatial feature extraction [19], whose adaptation techniques have been incorporated in the pseudo code. Artificial Immune System (AIS)-based Computational Intelligence Techniques emerged in 1980 s and tries to build models and investigate abstractions of human immune system. Gradually, AIS got divided into four subfields which have emerged prominently. These are Negative Selection Algorithms (NSA), immune network algorithms (INA), danger theory algorithms (DTA) and clonal selection algorithms (CSA) [1].

The Clonal Selection Algorithm was originally proposed by De Castro and Van Zuben [20]. Brownlee, in his book [1] has given the basic Clonal Selection Algorithm pseudo code to optimize a mathematical function. The algorithm in the book is implemented based on the function optimization algorithm proposed by De Castro and Van Zuben [21]. The basic idea about how to adapt the algorithm for land cover feature extraction has been taken from this code. The models built by taking inspiration from the immune system have been used to solve many day-to-day problems, [22–27]. CLONALG, although developed initially as a general machine learning approach, has now found its application into the domain of pattern recognition, function optimization, and combinatorial optimization

problem domains [28]. Since then, it has been worked upon and improved by many researchers. A betterment over Clonal Selection Algorithm, called the Novel Clonal Selection Algorithm, was proposed and tested on travelling salesman problem [29]. To improve efficiency, representation selection and parameterization, Garrett proposed Advanced Clonal Selection (ACS) [30].

### 3 Brief Review of Optimization Techniques Used

Geo-spatial feature extraction is a broad term which is used frequently in the field of remote sensing while analyzing geographic or spatial datasets to extract useful information. Land cover feature extraction, sub division of geo-spatial feature extraction, is considered for the evaluation of efficiency of Nature Inspired Algorithms in question. This process involves classifying various pixels of image into any of the relevant categories based on their land cover, like water, vegetation, etc. The ground truth, i.e. correctly classified pixels are available to us in the form of training data which makes this a supervised classification technique. The spectral signature of each category is calculated using the training data. Ultimately, every land cover feature extraction problem boils down to calculating the closeness of pixels to the spectral signature of each category relevant to the dataset.

#### 3.1 BA and CSS

Charged System Search is inspired by the Coulombs law and Newton's laws of motion which are used for exploitation and exploration of best possible solutions to any optimization problem, respectively. Coulombs law states that the force between any two charged bodies is directly proportional to the product of magnitude of charge carried by the bodies and inversely proportional to the square of the distance between them. The mathematical equation for Coulombs law is

$$F_{ij} = k_e \frac{q_i q_j}{r_{ij}^2}, \quad (1)$$

where  $k_e$  is a constant called the Coulomb constant,  $q_i, q_j$  are the magnitude of the two charges,  $r_{ij}$  is the distance between the two charges.

Newton's laws of motions produce following equations of motion which will be used in the implementation of CSS.

$$a = \frac{v_{\text{new}} - v_{\text{old}}}{\Delta} \quad (2)$$

$$r_{\text{new}} = \frac{1}{2} a \cdot \Delta t^2 + v_{\text{old}} \cdot \Delta t + r_{\text{old}} \quad (3)$$

where,  $a$  is acceleration of the body,  $v_{\text{new}}$  and  $v_{\text{old}}$  are final and initial velocities of the body,  $\Delta t$  is the time interval for which the body has been moving,  $r_{\text{new}}$  and  $r_{\text{old}}$  are the final and initial positions of the body, respectively. The BA is inspired by acoustics of echolocation of bats and the following assumptions are made regarding the behaviour of bats for the sake of BA:

- All bats identify prey (food) using their echolocation ability and they are able to distinguish between the obstacles and prey.
- Bats fly randomly with velocity  $v_i$  at position  $x_i$  with a fixed frequency  $f_{\text{min}}$ , varying wavelength  $\lambda$  and loudness  $A_0$  to search for prey. They can automatically adjust the wavelength (or frequency) of their emitted pulses and adjust the rate of pulse emission  $r \in [0, 1]$ , depending on the proximity of their target.
- The loudness varies from a large value  $A_0$  to a constant minimum value  $A_{\text{min}}$ .

The following formula for the propagation of sounds is used in the implementation of BA.

$$\lambda = \frac{v}{f}, \quad (4)$$

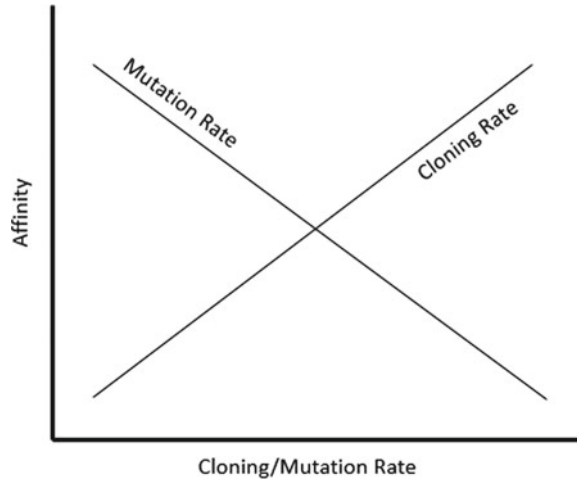
where the  $v$  is the speed of the sound in air which is typically 340 m/s,  $f$  and  $\lambda$  are the frequency and wavelength of sound wave, respectively.

### 3.2 Clonalg

The Clonal Selection theory of acquired immunity by Burnet was inspired by Darwinian natural theory of evolution. The theory proposes that antigens (the foreign material) select for the lymphocytes (B- and T-cells). That lymphocyte is selected which has the most affinity with the antigen. When a particular type of lymphocyte is selected, the cells proliferate and make thousands of clones by itself. The clones made are differentiated into two different cell types, plasma and memory cells. Plasma cells produce large quantities of antibody and have a short lifespan, whereas memory cells live for an extended period in the body of the host anticipating future recognition of the same antibody.

Along with cloning, another important feature of this theory is that when a cell is selected and proliferates, it is subjected to small copying errors, called somatic hyper-mutations, that mutate the shape of the receptors a little bit thus changing the

**Fig. 1** Graph showing variation of mutation and cloning rate w.r.t. Affinity

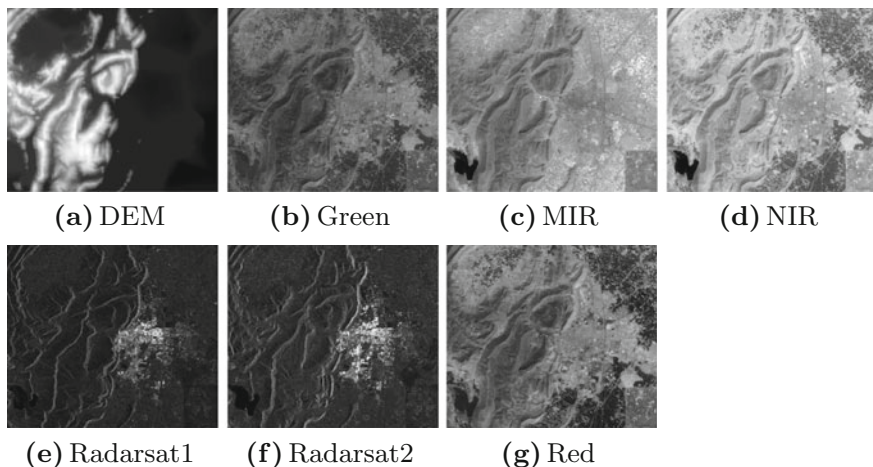


subsequent antibody recognition capabilities of both the antibodies bound to the lymphocyte cell's surface, and the antibodies that plasma cells produce.

In the general CLONALG model, the antibodies are treated as candidate solutions and are selected based on affinity. Affinity can be calculated either by matching against an antigen pattern or via evaluation of a pattern by a cost function [28]. The cloning rate of the selected antibody is proportional to the affinity with the antigen, whereas the hyper-mutation of clones is inversely proportional to affinity with the antigen as shown in Fig. 1. The newly formed set of clones competes with the existent antibody population for membership in the next generation [28].

## 4 Dataset Used

The nature inspired techniques have been adopted for classification of multispectral image of Alwar region in Rajasthan, India. The size of the image is  $472 \times 576$  pixels and the ground resolution is 23.5 m. The image has been captured using LISS-III sensor and has seven different bands namely red, green, near IR, mid IR, radarsat-1, radarsat-2 and digital elevation model. The satellite image is given in Fig. 2.



**Fig. 2** 7-Band Satellite Image Of Alwar Area In Rajasthan, India (Courtesy Of Defence Terrain and Research Lab DTRL), Defence And Research Development Organization (DRDO), India

## 5 Proposed Methodology

The paper proposes two nature inspired techniques namely Hybridized version of Bat Algorithm and Charged System Search and Clonal Selection Algorithm.

### 5.1 Hybrid BA/CSS

The hybridisation of BA and CSS has been done in order to combine the best features of both the algorithms for the sake of classification. The BA produces the spectral signatures (cluster centres) by using the training data which in turn is used in CSS to classify the unclassified pixels in the image.

**Bat Algorithm** The crux of the BA lies in the fact that bats adjust their frequency and loudness after every iteration depending upon their closeness to the prey. The pseudo code mentioned in Algorithm 1 for BA also incorporates the random walk in the algorithm. The BA allows the candidate solutions to move towards the best solution and towards their personal best in every iteration. Following formulas calculate the frequency, velocity and position of bats, loudness, and pulse emission rate after every iteration, respectively,

$$f_i = f_{\min} + (f_{\max} - f_{\min})\beta \quad (5)$$

$$v_i^t = v_i^{t-1} + (x_i^{t-1} - x_*)f_i \quad (6)$$

$$x_i^t = x_i^{t-1} + v_i^t \quad (7)$$

$$A_i^{t+1} = \alpha A_i^t \quad (8)$$

$$r_i^{t+1} = r_i^0 \cdot (1 - e^{-\beta t}), \quad (9)$$

where  $\beta$  is a uniform random number between  $[0, 1]$  and  $x_*$  is the current global best solution, obtained after comparing all the solutions among all the  $n$  bats.

---

**Algorithm 1:** Bat Algorithm Pseudocode

---

Initialize bat population  $x_i$  ( $i=1, \dots, n$ ) and velocity  $v_i$ .

Define frequency  $f_i$  at  $x_i$ .

Initialize loudness  $A_i$  and rate of pulse emission  $r_i$ .

**while** *termination condition is false* **do**

    Generate new solutions by using Eqs. (5), (6) and (7).

**if**  $rand > r_i$  **then**

        select global best solution among all the existing solutions.

        generate solutions using random walk.

**if**  $rand < A_i$  &  $f(x_i) < f(x_*)$  **then**

        accept the new solutions.

        update the loudness  $A_i$  (Eq.8) and rate of pulse emission  $r_i$  (Eq.9).

    sort the bats.

---

**Charged System Search** The formulation of CSS by Kaveh and Talatahari is one among many recent advancements in the field of Nature Inspired meta-heuristic techniques. The technique ensures the exploitation and exploration of search space using various laws of physics. Specifically, Coulomb and Gauss law is used for exploitation while, Newton's second law helps in exploration aspect of the technique. The formulae and details of which have described in the brief review section of the paper.

**Hybridization of BA with CSS** Once the cluster centres are calculated using BA, the CSS in conjunction with window based search is used to classify the image as shown in the pseudocode shown in Algorithm 2. In case of CSS, the charged particles (CP) are analogous to the  $n$ -dimensional vector which represents the cluster centres calculated from the training data. The fitness function used in the pseudocode is Mahalanobis distance function, when implemented in MATLAB. The force between CP and any pixel is calculated using Coulomb law equation mentioned earlier. As evident from the pseudo code of Hybrid BA/CSS, the cluster centres calculated by BA are viewed as charged particles whose charge and mass are initialized to one while velocity to zero before the iterations of CSS begin.

In every iteration of CSS, a square window of  $n \times n$ , containing  $n^2$  pixels is analyzed. Inside the so-called window, every pixel is again examined using the

Coulomb law. The pixel in question is classified after calculating the force exerted on it by the different classes. The force exerted by each CP is calculated using Eq. (1), wherein the  $q_i$  is not taken into consideration because of it representing the pixel which is an uncharged particle,  $q_j$  is the charge of CP and  $r_{ij}$  is the Euclidean distance between the CP and the pixel to be classified. Without any loss of information or generality the constant in Eq. (1) can be set to one. So, the pixel is assigned to the class whose CP exerts the maximum force on it. This step can be visualized as an uncharged particle moving towards charged particle by phenomenon called charging by induction. After the uncharged particle is attracted by the most influential CP, it coalesces with the same leading to some minor changes in the mass and charge of CP. This step is repeated for each pixel in the window leading to a significant change in the mass, charge and position of CPs. This aspect of CSS helps in tapping of geographical proximity of pixels belonging to same class.

---

**Algorithm 2:** Hybrid BA/CSS Pseudocode

---

Initialize cluster centers positions using the BA (Cluster centers are spectral signatures of each class).

Initialize charge (unit charge), mass (unit mass) and velocities (zero initial velocity)

**foreach** *window(square)* **do**

**foreach** *pixel*  $\in$  *window* **do**

        Calculate the force exerted by each CP (Spectral signature) by Coulomb's Law on the unclassified pixel and assign it to the CP exerting maximum force.

        Increment the charge and mass of CP in question by one.

        Change the location (value of spectral signature) of CP using equations of motion.

**foreach** *CP* **do**

**if**  $fit(BA\ value) < fit(new\ value)$  **then**

            Position of CP = BA value.

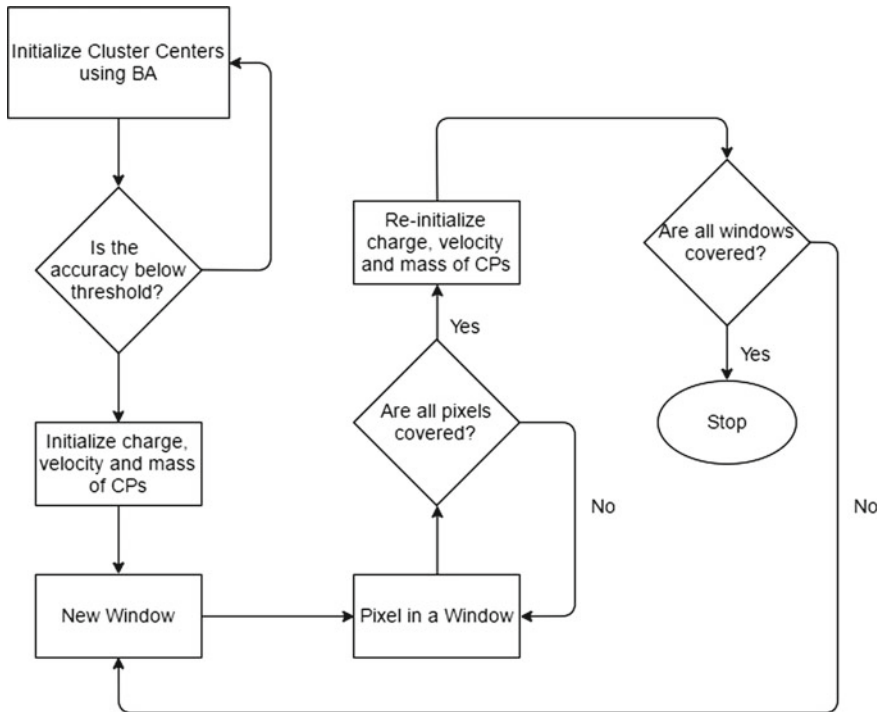
    Re initialize mass and charge of each CP to default.

---

After the window is closed, the Mahalanobis distance of the new CP (cluster centre) is compared with that of old CP (produced by BA), if it is better (lesser) then it is used as the new CP otherwise discarded at the end of window as corrupted CP. The same procedure is continued till all the windows are covered and all the pixels are classified. The procedure has been shown as flowchart in Fig. 3.

**Justification** Due to the hybridization of BA and CSS optimization occurs at following levels:

- Cluster centres are representative of the whole training dataset hence reducing the time complexity.
- Window based search technique exploits the high probability of two features being geographically close.



**Fig. 3** Flowchart for image classification using hybrid BA/CSS

- At the end of every iteration CP positions fitness is re-evaluated so that the effect of misclassified pixel is not carried to the next iteration.
- In every iteration, the charges of CPs are changed which is also accompanied by change in their mass, which reinforces the concept of accumulated charge and counteracts the effect of increasing charge.

## 5.2 Clonalg

The proposed Algorithms for Geo-Spatial feature Extraction is Clonal Selection based Algorithm (CSA). The algorithm generates antibodies corresponding to every class (barren, rocky, urban, vegetation, water) each time for every window/cluster and the unclassified pixel is treated as an antigen. The unclassified pixel is classified in the class whose antibody has maximum affinity with the pixel. The records in the antibody are cloned and mutated based on affinity value (Algorithm 3).

Cloning rate is directly proportional to affinity, while mutation is inversely proportional to affinity (Fig. 1). Instead of calculating the affinity, the algorithm



uses the difference of each antibody with the antigen. So, an antibody having the least difference with the antigen is the one that will be selected to bind (classify the pixel) the antigen.

The Algorithm comprises the the following steps:

**Finding Approximate Maximum and Minimum Sum of Differences possible**

First a rough estimate of the maximum and minimum difference is obtained. This is done by finding the difference between random pixels and antibodies representing the classes. The maximum and minimum differences are stored in the variables *maxdiff* and *mindiff*, respectively. This is done to know the approximate range of differences. This range will be used to determine cloning rate and mutation rate.

---

**Algorithm 3: Main CLONALG function**

---

Calculate approximate maximum and minimum affinity.

**foreach** *window(square)* **do**

    Generate the antibodies.

**foreach** *pixel ∈ window* **do**

        Find antibody having maximum affinity with the pixel.

        Based on affinity, maximum affinity and minimum affinity decide cloning and mutation rates.

        Clone & mutate the selected antibody.

        Remove weakest set of cells.

        Consume pixel in antibodies class.

---

**Making a Window and Initializing Antibodies** The image is parsed using a square window of a predetermined size. For each window, the antibodies are constructed from the original training set. During the classification of all the pixels in the window, the antibodies are mutated, and no new records are taken from the training set. The window based search helps incorporate the concept of memory cells and plasma cells. While cloning, two types of cells are formed, Plasma cells and Memory cells. Plasma cells have a short life span and are used to fight the antigen only during the time of infection. On the other hand, memory cells have longer lifespan and they reside in the immune system and help in identifying the same antigen in case of future attacks. In the algorithm, during the time pixels of a window are being classified, the antibodies keep getting mutated and newly cloned plasma cells keep replacing the weakest of the lot. But once a window is done with classification, another set of antibodies is built from the original training set, which is analogous to the concept of memory cells in the immune system.

**Classifying a Pixel** The difference of a pixel (antigen) is calculated with each and every antibody (representing a particular class). The one giving minimum difference (hence having maximum affinity) is classified as the pixel's class.

**Calculating the Cloning rate and Mutation Rate** The cloning rate is a function of the difference between pixel and affinity and was calculated earlier. Cloning rate also takes into consideration the *maxdiff* and *mindiff*. The cloning rate is inversely

proportional to the difference value. So if the difference is more than maxdiff, then cloning rate should be minimum, and if the difference is less than mindiff, cloning rate should be maximum. Rest of the cloning rates can be determined based on where the difference value lies in the range [mindiff, maxdiff].

Mutation is implemented by replacing a certain number of elements of the cloned cells with that of the antigen's elements. Mutation is directly proportional to the difference between pixel and antibody. When the difference is minimum, mutation rate has to be minimum, and when the difference is more than maxdiff or near it, the mutation rate has to be maximum.

**Cloning and Mutating** Once the antibody (class) with minimum difference is selected, the cells of the same antibody are sorted according to differences with the antigen. The cells are sorted in ascending order. The weakest cells are replaced with the clones of the strongest one. Moreover some elements of the clones are replaced with elements picked from the antigen. This ensures mutation of the cloned cells (Algorithm 4). Thus the cloned and mutated antibody is used for classification purpose of the remaining unclassified pixels in the particular window.

All these steps together form the clonal selection algorithm for image classification. The flowchart (Fig. 4) depicts the algorithm.

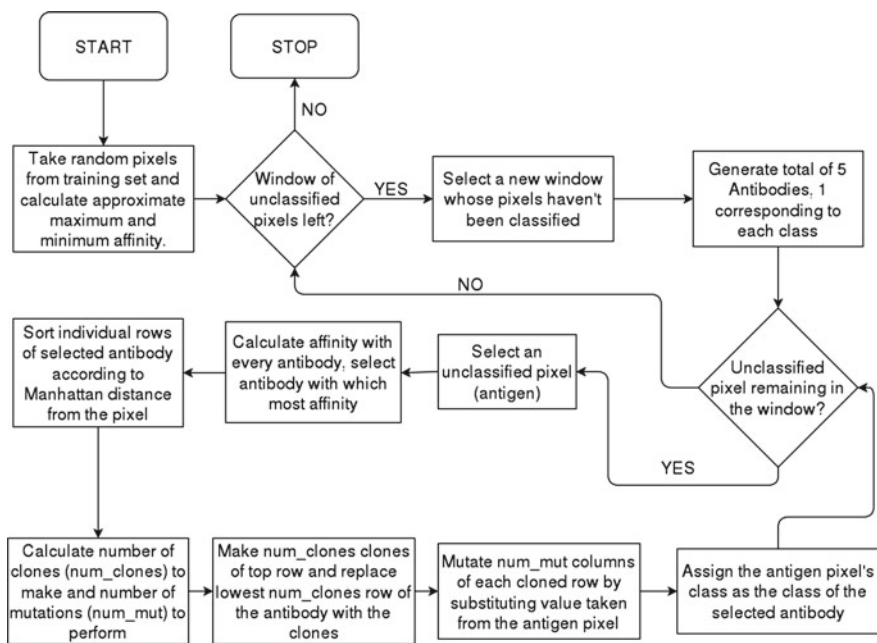


Fig. 4 Flowchart for image classification using CLONALG

---

**Algorithm 4:** Clone\_Mutate Function (*Function that clones and mutates the selected antibody*)

---

Sort the cells of the antibody type according to difference with the pixel.  
 Determine number of clones to make (num\_clones)  
 Determine number of mutations to make (num\_mut)  
 Replace last num\_clones cells with cell having minimum difference  
 (Cloning)  
 Replace num\_mut elements of each clone with elements of the pixel  
 (Mutation)

---

## 6 Accuracy Assessment of Proposed Algorithms

Both the algorithms, BA/CSS and CLONALG, were tested on the Alwar Dataset. The labelled training set pixels were used to determine the accuracy. The label of each training set pixel is compared with the predicted label for that pixel. The labelled data consists of 191 Barren pixels, 288 Rocky pixels, 417 Urban pixels, 329 Vegetation pixels and 206 Water pixels, making a total of 1431 labelled training pixels. The Error matrices (Tables 1 and 5) were plotted for both the algorithms. The  $j$ th column element of the  $i$ th row simply shows how many pixels of the class  $[i]$  were classified into class  $[j]$  by the algorithm. So the diagonal entries show the correctly classified pixels while non-diagonal entries correspond to misclassified pixels. For example, in the first row of Table 1, out of the 198 Barren training pixels, 158 are correctly classified as Barren, 10 are misclassified as Rocky and 23 are misclassified as Urban. The last column of the error matrix gives the actual total count of each class of pixels. And the last row gives the total count of pixels predicted to be of a certain class. The non-diagonal elements represent the omission and commission errors. Omission error refers to the pixels that belonged to a certain class but were misclassified into other classes. Commission error refers to the pixels classified in a certain class, whereas they actually belonged to some other class. The non-diagonal row elements give the omission error, while the non-diagonal column elements give the commission errors (Table 2).

The Producers accuracy is also calculated for each class. Producers accuracy is calculated by dividing the number of correctly classified pixels of a certain class by the total number of pixels in the training set of that class. It simply indicates how well training pixels of a class are classified. Tables 3 and 7 give the producers accuracy for BA/CSS and CLONALG, respectively.

The Users accuracy is calculated by dividing the number of correctly classified pixels of a class by the total number of pixels classified in that class. It is a measure of commission error and indicates the probability that a pixel that has been

**Table 1** Error matrix for Hybrid BA/CSS

	Barren	Rocky	Urban	Vegetation	Water	Total
Barren	158	10	23	0	0	191
Rocky	3	285	0	0	0	288
Urban	66	0	351	0	0	417
Vegetation	0	0	0	329	0	329
Water	0	0	0	0	206	206
Total	227	295	374	329	206	1431

**Table 2** Producer's accuracy for Hybrid BA/CSS

Feature	Accuracy calculation	Producer's accuracy
Barren	158/227	69.6
Rocky	285/295	96.6
Urban	351/374	93.8
Vegetation	329/329	100
Water	206/206	100

**Table 3** User's accuracy for Hybrid BA/CSS

Feature	Accuracy calculation	User's accuracy
Barren	158/191	82.7
Rocky	285/288	98.9
Urban	351/417	84
Vegetation	329/329	100
Water	206/206	100

classified in that class actually belongs to that class. Tables 2 and 3 represent the users accuracy for BA/CSS and CLONALG, respectively.

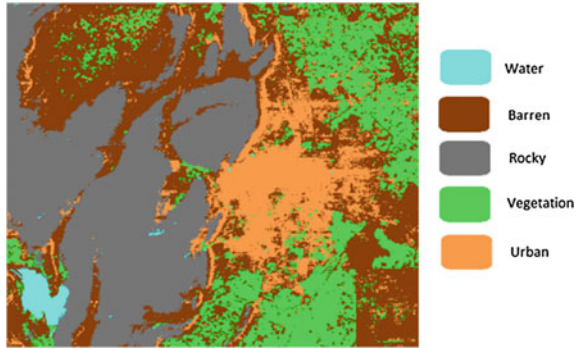
Finally Kappa coefficients were calculated for both the algorithms. It is degree of matching between reference data set and classification. The Kappa coefficient was for the accuracy assessment of BA/CSS and CLONALG and their comparison with other algorithms.

## 6.1 Hybrid BA/CSS

The image classified using Hybrid BA/CSA produces a kappa coefficient of 0.9048.

As evident from the error matrix and producers accuracy the barren feature has not been properly classified. Out of 191 barren pixels only 158 were correctly

**Fig. 5** Classified image of Alwar region using Hybrid BA/CSS



classified. In addition to that many of urban pixels were also misclassified as barren. Vegetation and water features have been perfectly classified and no other feature pixels have been misclassified as them. It can be deduced that the Barren and Urban features have a lot of similarities which has not only led to poor classification but also heavy misclassification amongst themselves. The classified image of Alwar Region using Hybrid BA/CSS is shown in Fig. 5.

## 6.2 Clonalg

For the case of CLONALG, different combinations of antibody\_size and window size produced different values of the kappa coefficients (Table 4). It was observed that the best results were obtained when

$$\text{Window\_size} \approx \text{Antibody}/15 \tag{10}$$

All further calculations were reproduced for the last case, i.e antibody\_size = 45 and window\_size = 3.

**Table 4** Kappa coefficients for different combinations of Antibody\_size and Window\_size

Antibody_size	Window_size	Kappa coefficient
191	400	0.245688
191	191	0.867955
100	50	0.926334
100	10	0.973346
40	10	0.967114
40	4	0.975121
60	4	0.975116
40	2	0.974224
40	1	0.974229
45	3	0.976900

**Table 5** Error matrix for CLONALG

	Barren	Rocky	Urban	Vegetation	Water	Total
Barren	174	0	17	0	0	191
Rocky	3	285	0	0	0	288
Urban	0	0	417	0	0	417
Vegetation	0	0	0	329	0	329
Water	0	0	0	0	206	206
Total	177	285	434	329	206	1431

**Table 6** Producer's accuracy for CLONALG

Feature	Accuracy calculation	Producer's accuracy
Barren	174/177	98.3
Rocky	285/285	100
Urban	417/434	96
Vegetation	329/329	100
Water	206/206	100

**Table 7** User's accuracy for CLONALG

Feature	Accuracy calculation	User's accuracy
Barren	174/191	87.4
Rocky	285/288	98.9
Urban	417/417	100
Vegetation	329/329	100
Water	206/206	100

The error matrix (Table 5) clearly shows that all pixels of urban, vegetation and water have been correctly classified. Out of the 288 rocky pixels, 3 were misclassified as barren. The misclassification is most for the barren pixels, where out of 191, 17 have been misclassified as Urban (Table 6).

From the users accuracy it can be seen that apart from Rocky and Barren pixels, rest all have been classified 100% correctly. Accuracy of classifying barren pixels is quite low compared to the rest. The barren pixels are being misclassified as urban pixels (Table 7).

The classified image of Alwar Region using CLONALG is shown in Fig. 6 along with the colour legend (Fig. 7).

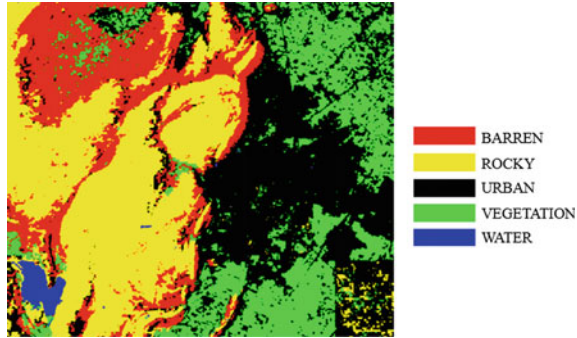


Fig. 6 Classified image of Alwar region using CLONALG

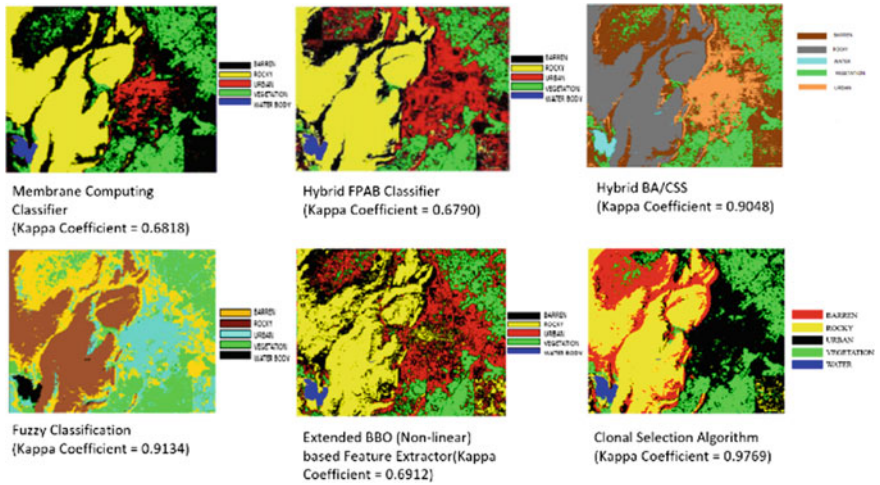
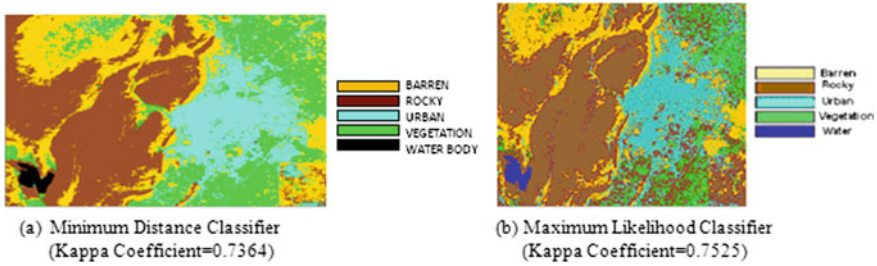


Fig. 7 Classified images of Alwar region after applying various soft computing techniques

Table 8 Comparison of Hybrid BA/CSS and CLONALG with other techniques

Minimum distance to mean	Maximum likelihood	Membrane computing classifier	Hybrid FPAB	Fuzzy classification	PSO
0.7364	0.7525	0.6818	0.679	0.9134	0.970
Extended BBO	Hybrid ACO/BBO	Hybrid ACO2/PSO	CLONALG	Hybrid BA/CSS	
0.6912	0.96699	0.975	0.9769	0.9048	



**Fig. 8** Classified images of Alwar region after applying probabilistic classifiers

## 7 Conclusion and Future Scope

As is evident from the Table 8, Clonal Selection Algorithms performance is very much comparable to the existing soft computing techniques. The accuracies of Hybrid ACO-BBO and Hybrid ACO2/PSO come close while comparing with CLONALG. After analyzing Table 8 more carefully, it is clear that the traditional probabilistic classifiers like Minimum Distance to Mean and Maximum Likelihood fall behind the nature inspired computational techniques by a huge margin when it comes to land cover feature extraction. Hybrid BA/CSS has been introduced for the sake of classification problems and not completely adapted for land cover feature extraction which explains its relatively low kappa coefficient. The evolution of recent nature inspired algorithms provide us with a wide range of flavours for working on land cover feature extraction. The methodology that has been showcased in this paper can also be extended to other applications which can be modelled into a classification problem (Fig. 8).

## References

1. Brownlee, J.: How to escape traps using clonal selection algorithms. In: Technical Report 070209A (2007).
2. Alcocer, R., Zenteno-Jimenez, E., Barrios, J. M.: Automatic Land Use and Land Cover Classification Using RapidEye Imagery in Mexico. In: AAAI Workshop (2015). doi:10.13140/RG.2.1.1685.8081
3. Aitkenhead, M. J., Aalders, I. H.: Automatic Land cover mapping of Scotland using expert system and knowledge integration methods. In: Remote Sensing of Environment, vol. 115(5), pp. 1285–1295. Elsevier (2011). doi:10.1016/j.rse.2011.01.012
4. Hermes, L., Friauff, D., Puzicha, J., Buhmann, J. M.: Support vector machines for land usage classification in landsat TM imagery. In: Proc. IGARSS, Vol. 1, pp. 348–350. IEEE (1999). doi:10.1109/IGARSS.1999.773494
5. Roli, F., Fumera, G.: Support vector machines for remote-sensing image classification. In: Proc. SPIE, vol. 417, pp. 160–166. IEEE (2001).



6. Huang, C., Davis, L. S., Townshend, J. R. G.: An assessment of support vector machines for land cover classification. In: *Int. J. Remote Sens.*, Vol. 23(4), pp. 725–749. Taylor and Francis Group (2002). doi:[10.1080/01431160110040323](https://doi.org/10.1080/01431160110040323)
7. Gualtieri, J. A., Cromp, R. F.: Support vector machines for hyperspectral remote sensing classification. In: *Proc. SPIE*, Vol. 3584, pp. 221–232. DEStech (1998).
8. Gualtieri, J. A., Chettri, S. R., Cromp, R. F., Johnson, L. F.: Support vector machine classifiers as applied to AVIRIS data. In: *Summaries 8th JPL Airborne Earth Science Workshop*, pp. 217–227. Elsevier (1999).
9. Gualtieri, J. A., Chettri, S. R.: Support vector machines for classification of hyperspectral data. In: *Proc. IGARSS*, pp. 813–815. IEEE (2000). doi:[10.1109/IGARSS.2000.861712](https://doi.org/10.1109/IGARSS.2000.861712)
10. Abd, H. A. A. R.: Feature Extraction and Based Pixel Classification for Estimation the Land Cover thematic map using Hyperspectral data. In: *International Journal of Engineering Research and Applications*, vol. 3(3), pp. 686–693 (2013).
11. Jawak, S. D., Luis, A. J.: A spectral index ratio-based Antarctic land-cover mapping using hyperspatial 8-band WorldView-2 imagery. In: *Polar Science*, vol. 7(2), pp. 18–38. Elsevier (2013). doi:[10.1016/j.polar.2012.12.002](https://doi.org/10.1016/j.polar.2012.12.002)
12. Jawak, S. D., Luis, A. J.: Iterative Spectral Index Ratio Exploration for Object-based Image Analysis of Antarctic Coastal Oasis Using High Resolution Satellite Remote Sensing Data. In: *International Conference on Water Resources, Coastal And Ocean Engineering*, vol. 4, pp. 157–164. Elsevier (2015).
13. Alparone, L., Baronti, S., Garzelli, A., Nencini, F.: A global quality measurement of pan-sharpened multispectral imagery. In: *IEEE Geosci. Remote Sens.*, Vol. 1(4), pp. 313–317. IEEE (2004). doi:[10.1109/LGRS.2004.836784](https://doi.org/10.1109/LGRS.2004.836784)
14. Goel, L., Gupta, D., Abraham, A., Panchal, V. K.: Taxonomy of Nature Inspired Computational Intelligence: A Remote Sensing Perspective. In: *Fourth World Congress on Nature and Biologically Inspired Computing*, pp. 200–206. IEEE Press, Mexico City (2012). doi:[10.1109/NaBIC.2012.6402262](https://doi.org/10.1109/NaBIC.2012.6402262)
15. Kaveh, A., Talatahari, S.: A novel heuristic optimization method: charged system search. In: *Acta Mech.*, Vol. 21 (3–4), pp. 267–289. Springer (2010). doi:[10.1007/s00707-009-0270-4](https://doi.org/10.1007/s00707-009-0270-4)
16. Yang, X.S.: A New Metaheuristic Bat-Inspired Algorithm. In: J. R. Gonzalez et al.(eds.) *Nature Inspired Cooperative Strategies for Optimization*. Studies in Computational Intelligence, vol. 284, pp. 65–74. Springer, Berlin (2010).
17. Senthilnath, J., Kulkarni, S., Benediktsson, J. A., Yang, X. S.: A Novel Approach for Multispectral Satellite Image Classification Based on the Bat Algorithm. In: *IEEE Geoscience and Remote Sensing Letters*, vol. 13(4), pp. 599–603. IEEE Press (2016). doi:[10.1109/LGRS.2016.2530724](https://doi.org/10.1109/LGRS.2016.2530724)
18. Kumar, Y., Sahoo, G.: A charged system search approach for data clustering. In: *Progress in Artificial Intelligence*, vol. 2(2), pp. 153–166. Springer, Berlin (2014). doi:[10.1007/s13748-014-0049-2](https://doi.org/10.1007/s13748-014-0049-2)
19. Goel, L., Gupta, D., Panchal, V. K.: Hybrid bio-inspired techniques for land cover feature extraction: A remote sensing perspective. In: *Applied Soft Computing*, vol. 12(2), pp. 832–849, Elsevier (2012). doi:[10.1016/j.asoc.2011.10.006](https://doi.org/10.1016/j.asoc.2011.10.006)
20. De Castro, L., Von Zuben, F. J.: The Clonal selection algorithm with engineering applications. In: *GECCO 2000, Workshop Proceedings, Workshop on Artificial Immune Systems and Their Applications*, Las Vegas, USA, pp. 36–37 (2000).
21. De Castro, L., Von Zuben, F. J.: Learning and optimization using the clonal selection principle. In: *IEEE Transactions on Evolutionary Computation*, vol. 6, n. 3, pp. 239–251. IEEE (2002). doi:[10.1109/TEVC.2002.1011539](https://doi.org/10.1109/TEVC.2002.1011539)
22. Hofmeyr, S. A., Forrest, S.: Immunity by Design: An Artificial Immune System. In: *Proceedings Genetic and Evolutionary Computation Conference*, pp. 1289–1296 (1999).
23. De Castro, L. N., Timmis, J.: *Artificial Immune Systems: A New Computational Intelligence Approach*. In: *Lecture Notes in Computer Science*, London, United Kingdom. Springer-Verlag (1996). doi:[10.1007/978-3-540-73922-7](https://doi.org/10.1007/978-3-540-73922-7)

24. Hong, L.: A Particle Swarm Optimization Based on Immune Mechanism. In: International Joint Conference on Computational Sciences and Optimization, IEEE Computer Society, pp. 670–673. IEEE (2009). doi:[10.1109/CSO.2009.21](https://doi.org/10.1109/CSO.2009.21)
25. Zhang, Z., Xin, T.: Immune Algorithm with Adaptive Sampling in Noisy Environments and Its Application to Stochastic Optimization Problems. In: IEEE Computational Intelligence Magazine, pp. 29–40. IEEE (2007). doi:[10.1109/MCI.2007.906681](https://doi.org/10.1109/MCI.2007.906681)
26. Fang, T., Fu, D., Zhao, Y.: A Hybrid Artificial Immune Algorithm for Feature Selection of Ovarian Cancer Data. In: International Workshop on Education and Training and International Workshop on Geoscience and Remote Sensing, Computer Society, pp. 681–685. IEEE (2008). doi:[10.1109/ETTandGRS.2008.285](https://doi.org/10.1109/ETTandGRS.2008.285)
27. Ulker, E. D., Ulker, S.: Comparison Study for Clonal Selection Algorithm. In: International Journal of Computer Science and Information Technology (IJCSIT), vol 4(4), pp. 107–118 (2012). doi:[10.5121/ijcsit.2012.4410](https://doi.org/10.5121/ijcsit.2012.4410)
28. Beonwlee, J.: Clever Algorithms: Nature Inspired Programming Recipes (2012).
29. Zhao, M., Tang, K., Lu, G., Zhou, M., Fu, C., Yang, F., Zhang, C.: A Novel Clonal Selection Algorithm and its Application. In: International Conference on Apperceiving Computing and Intelligence Analysis, pp. 385–388. IEEE (2008). doi:[dx.doi.org/10.1109/ICACIA.2008.4770049](https://dx.doi.org/10.1109/ICACIA.2008.4770049)
30. Garrett, S. M.: Parameter-free, adaptive clonal selection. In: Congress on Evolutionary Computing, pp. 1052–1058, IEEE (2004).
31. Brownlee, J.: Clonal Selection Algorithms. In: Technical Report 070209A, Complex Intelligent Systems Laboratory (CIS), Centre for Information Technology Research (CITR), Faculty of Information and Communication Technologies (ICT), Swinburne University of Technology (2007).

# Network Intrusion Detection System to Preserve User Privacy

Sireesha Rodda and Uma Shankar Rao Erothi

**Abstract** A wide range of malicious attacks and threats are increasing day by day with the growth and development of internet and network technologies. Enforcing network security is important to protect data or information in the computer network against attacks from intruders. The right of privacy of the user must be respected even on the network-resident data. This paper evaluates the performance of four different classifiers on a standard network intrusion detection dataset. The original values in the dataset are anonymized in order to protect the user's privacy. All the experiments were performed on IBM SPSS Premium Modeler. The effectiveness of the techniques is tested using different evaluation measures.

**Keywords** Network intrusion detection system · Machine learning  
Data classification · Privacy-preserving data mining

## 1 Introduction

With the drastic growth of Internet usage in different fields, security of network traffic is becoming a crucial issue for computer networks. It is necessary to provide security, availability, integrity, and confidentiality to users in spite of ever-evolving malicious entities on the internet [1, 2]. The role of Network Intrusion Detection Systems (NIDS) is to effectively thwart malicious attacks.

Different data mining (DM) techniques have been effectively used for building an efficient NIDS. Different DM techniques such as Neural Networks [3], k-Nearest Neighbor [4], Decision Trees [5], Artificial Immune System [6], Genetic

---

S. Rodda (✉) · U.S.R. Erothi  
Department of CSE, GITAM Institute of Technology, GITAM University,  
Visakhapatnam, India  
e-mail: sireesha@gitam.edu

U.S.R. Erothi  
e-mail: umashankar.erothi3@gmail.com

Programming [7], Naïve Bayes [8] etc. were successfully employed to provide network intrusion detection.

Different NIDS algorithms may be categorized into anomaly detection and misuse/signature detection. Misuse detection identifies intrusions that match to known attacks while anomaly detection attempts to identify unusual activities that deviate from normal behavior. The existing systems, however, have some deficiencies as they are incapable of detecting novel attacks (signature-based) that have never been observed/recorded before. NSL-KDD [9] is a standard benchmark intrusion dataset for identifying network intrusions.

The remainder of the paper is structured as follows: Related work in the area is discussed in Sect. 2. Section 3 provides an overview of the dataset used. Section 4 analyzes the experimental results obtained from different classification techniques. Finally, Sect. 5 concludes the paper.

## 2 Related Work

Many intrusion detection systems (IDS) were designed and implemented using various popular classifiers with each holding different classification capabilities for different intrusions.

Farid et al. [10] used machine learning algorithms for an adaptive network intrusion detection system using decision trees and the naive Bayesian classifier. The classifiers maintained balanced detection rate and false positives for different types of attacks in the network traffic.

Pan et al. [11] presented a hybrid IDS using C 4.5 and neural networks. The performance of neural networks shows higher detection for DOS and probe attacks compared to U2R and R2L attacks, while C 4.5 displayed higher accuracy in detecting U2R and R2L when compared to neural networks.

Ayeiet al. [12] presented a hybrid technique using both anomaly and misuse detection approaches. This approach combined the features of J48, K-NN, and Boyer Moore algorithms.

Peddabachigari et al. [13] proposed a hybrid intelligent system for modeling NIDS. Decision trees and support vector machines are combined as a system model and an ensemble model by combining the constituent classifiers. The individual base classifiers and other hybrid algorithms are combined to minimize computational complexity and maximize detection accuracy.

Rangadurai et al. [14] presented the two-stage architecture for network intrusion detection system. A probabilistic classifier is used in the first stage to detect potential attacks and Hidden Markov Model in the second stage to narrow down the potential attack IP addresses in the network traffic. Depren et al. [15] presented a novel hybrid architecture for IDS utilizing both signature-based and anomaly based detection using Self-Organizing Map and J48 algorithm, respectively.

### 3 Dataset Description

The popular NSL\_KDD benchmark dataset has been employed to validate different network intrusion systems. NSL\_KDD is an improved version of KDD\_CUP'99 [16] dataset used to study the effectiveness of the NIDS system. The NSL\_KDD dataset eliminates these redundant records from the training and testing datasets in order to improve the evaluation results of intrusion detection and different learning techniques in general.

The NSL\_KDD dataset consists of one decision attribute and 42 conditional features; each record is labeled as either attack or as an normal. The training dataset contains of a total of 12,162 instances out of which 6503 are normal and 5659 are attacks. The testing dataset contains 5001 instances out of which 2569 are normal and 2432 are labeled as attacks. The attacks are of 22 types and they are classified into four categories: Dos, Probe, U2R, and R2L. Detailed description of different attacks in the dataset is clearly outlined in Rodda and Erothi [17].

Figures 1 and 2 show the distribution of different types of attacks in the NSL\_KDD train and test datasets, respectively. The description of different attributes present in the dataset is depicted in Fig. 3.

**C 5.0 [18]** is a decision tree model represented as a tree structure. The instances are split based on the attribute that provides maximum information gain. This process repeated until most of the subsamples belong to the same class. Pruning is applied to avoid overfitting.

**Support Vector Machine [18]** “kernel functions” find nonlinear solutions and solve classification problem which is not linearly separable. SVM works by mapping data to high-dimensional feature space and construct optimal hyper plane as

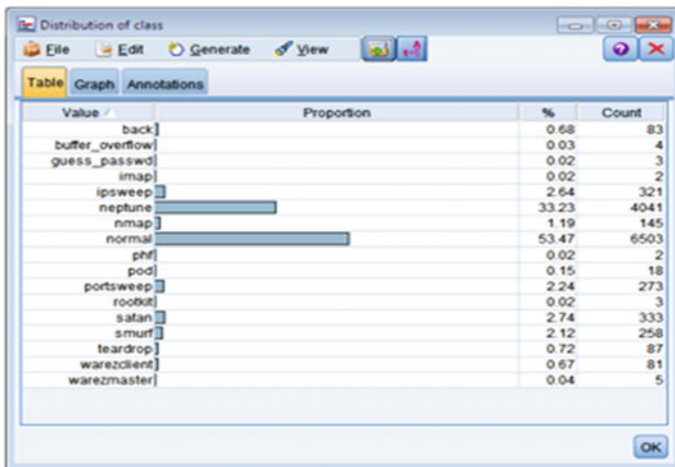


Fig. 1 Distribution of training dataset

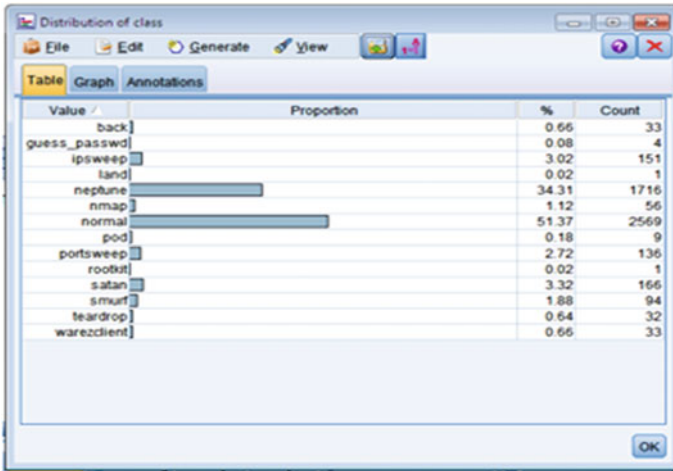


Fig. 2 Distribution of test dataset

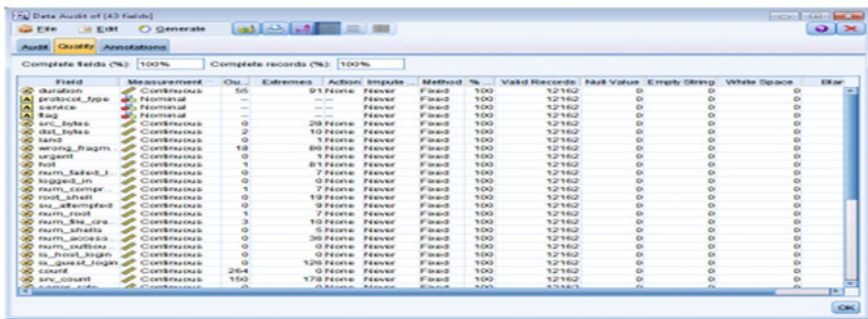


Fig. 3 Description of attributes in the NSL\_KDD dataset

the decision function to categorize positive and negative data points with the maximal margin.

**Neural networks [18]** are nonlinear data modeling tools that can be used to learn the relationship between inputs and outputs, and then generalize the input–output pairs to find patterns in data.

### 4 Experimental Results

The experiments were conducted on Intel Core i5-2400 CPU 3.10 GHz PC with 4 GB of RAM running on a 32-bit operating system. The results reported have been obtained from IBM SPSS Premium Modeler version 17.0.

Experiments were conducted on benchmark NSL-KDD dataset. The performance of three popular classifiers C5.0, SVM, and Neural Network is evaluated against an ensemble classifier. The ensemble model consists of C5.0, SVM, and Neural Network, is applied on the transformed dataset as shown in Fig. 4.

To preserve the identity of the users, the dataset is first transformed using the “Anonymize node”. The flow for obtaining the transformed dataset is shown in Fig. 5. Only the sensitive attributes are anonymize, leaving the non-sensitive attributes unchanged. The sensitive attributes that must be transformed in order to preserve user privacy is performed as shown in Fig. 6.

The ensemble considered uses bagging to combine the predictions of the constituent base classifiers. The efficiency of the four classifiers is evaluated in terms of

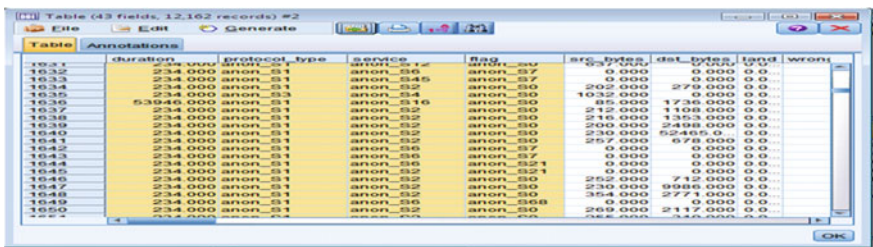


Table	Annotations	duration	protocol_type	service	flag	src_bytes	dst_bytes	land	wrong
10.31		234.000	anon_s1	anon_s1	anon_sav	537.000	0.000	0.000	0.000
10.32		234.000	anon_s1	anon_s6	anon_s7	0.000	0.000	0.000	0.000
10.33		234.000	anon_s1	anon_s45	anon_s7	0.000	0.000	0.000	0.000
10.34		234.000	anon_s1	anon_s15	anon_s0	202.000	278.000	0.000	0.000
10.35		234.000	anon_s3	anon_s4	anon_s0	1032.000	0.000	0.000	0.000
10.36		5346.000	anon_s1	anon_s116	anon_s0	42.000	1738.000	0.000	0.000
10.37		234.000	anon_s1	anon_s2	anon_s0	212.000	1408.000	0.000	0.000
10.38		234.000	anon_s1	anon_s2	anon_s0	210.000	1353.000	0.000	0.000
10.39		234.000	anon_s1	anon_s2	anon_s0	240.000	2488.000	0.000	0.000
10.40		234.000	anon_s1	anon_s2	anon_s0	230.000	62485.0	0.000	0.000
10.41		234.000	anon_s1	anon_s2	anon_s0	257.000	678.000	0.000	0.000
10.42		234.000	anon_s1	anon_s6	anon_s7	0.000	0.000	0.000	0.000
10.43		234.000	anon_s1	anon_s6	anon_s7	0.000	0.000	0.000	0.000
10.44		234.000	anon_s1	anon_s6	anon_s21	0.000	0.000	0.000	0.000
10.45		234.000	anon_s1	anon_s2	anon_s0	230.000	9386.000	0.000	0.000
10.46		234.000	anon_s1	anon_s2	anon_s0	252.000	712.000	0.000	0.000
10.47		234.000	anon_s1	anon_s2	anon_s0	230.000	9386.000	0.000	0.000
10.48		234.000	anon_s1	anon_s2	anon_s0	364.000	2774.000	0.000	0.000
10.49		234.000	anon_s1	anon_s6	anon_s68	0.000	0.000	0.000	0.000
10.50		234.000	anon_s1	anon_s2	anon_s0	269.000	2117.000	0.000	0.000

Fig. 4 Transformed dataset

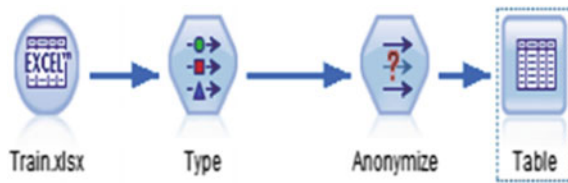
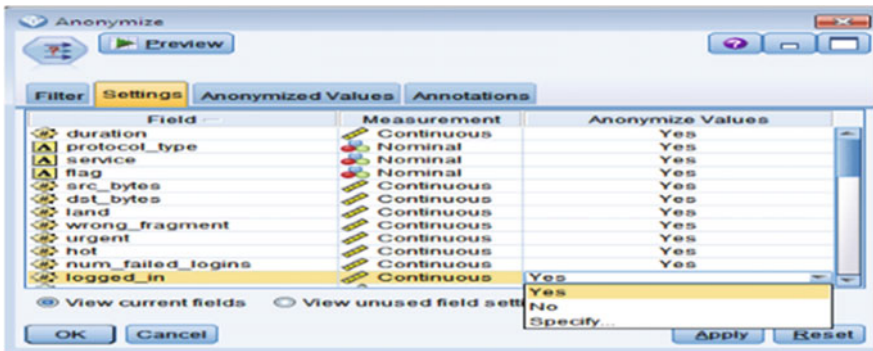


Fig. 5 Flow for Anonymizingnetwork data



Field	Measurement	Anonymize Values
duration	Continuous	Yes
protocol_type	Nominal	Yes
service	Nominal	Yes
flag	Nominal	Yes
src_bytes	Continuous	Yes
dst_bytes	Continuous	Yes
land	Continuous	Yes
wrong_fragment	Continuous	Yes
urgent	Continuous	Yes
not	Continuous	Yes
num_failed_logins	Continuous	Yes
logged_in	Continuous	Yes

Fig. 6 Identifying sensitive attributes

Accuracy, False Alarm Rate (FAR), Detection Rate (DR), Precision, F-Measure, and G-mean. *Accuracy* is the ratio of the total number of correct predictions to the total number of instances. *DR* is the percentage of number of attacks detected to the total number of intrusions present. *Precision* is the proportion of the number of attacks correctly identified to the number of instances identified as attacks by the classifier. *FAR* is the fraction of the number of normal traffic requests falsely identified as attacks to the number of normal class instances. *F-Measure* is the harmonic mean of precision and recall parameters. *G-Mean* is the geometric mean of precision and recall parameters.

Figure 7 shows the flow diagram of Ensemble model for NIDS. Default parameters have been used. The performance of the transformed dataset is evaluated separately by the constituent classifiers.

The percentage of correctly classified instances for the ensemble as well as other classifiers is summarized in Table 1.

From Table 1, it can be observed that the ensemble approach returns the best accuracy when compared with other individual classifiers. The SVM classifier comes close to the best performance.



Fig. 7 Ensemble model for network intrusion detection system

Table 1 Performance of classification models

Classification nodes	Correctly classified (%)	Incorrectly classified (%)
Ensemble	<b>99.56</b>	<b>0.44</b>
C 5.0	99.30	0.7
SVM	99.54	0.46
Neural networks	99.18	0.82
k-NN	99.48	0.52
BayesNet	97.90	2.1
C & R Tree	93.58	6.42
Quest	84.94	15.06
CHAID	91.46	8.54
Discriminant	94.58	5.42
Logistic	59.43	40.57



C 5.0						Support Vector Machine					
	Dos	Probe	U2R	R2L	Normal		Dos	Probe	U2R	R2L	Normal
Dos	1882	2	0	0	1	Dos	1880	0	0	0	5
Probe	2	501	0	3	3	Probe	0	504	0	0	5
U2R	0	0	0	1	0	U2R	0	0	0	0	1
R2L	0	0	0	33	4	R2L	0	0	0	34	3
Normal	3	7	1	5	2553	Normal	0	2	0	1	2556

Neural Networks						Ensemble					
	Dos	Probe	U2R	R2L	Normal		Dos	Probe	U2R	R2L	Normal
Dos	1878	5	0	0	2	Dos	1880	0	0	0	5
Probe	2	500	0	2	5	Probe	1	505	0	0	3
U2R	0	0	1	0	0	U2R	0	0	0	0	1
R2L	0	2	0	33	2	R2L	0	0	0	34	3
Normal	2	7	0	2	2558	Normal	2	2	0	1	2564

Fig. 8 Confusion matrix

C 5.0					Support Vector Machine				
	TP	FN	FP	TN		TP	FN	FP	TN
Dos	1882	3	5	3111	Dos	1880	5	0	3116
Probe	501	8	9	4483	Probe	504	5	2	4490
U2R	0	1	1	4999	U2R	0	1	0	5000
R2L	33	4	9	4955	R2L	34	3	1	4963
Normal	2553	16	8	2424	Normal	2566	3	14	2418

Neural Networks					Ensemble				
	TP	FN	FP	TN		TP	FN	FP	TN
Dos	1878	7	4	3112	Dos	1880	5	3	3113
Probe	500	9	14	4478	Probe	505	4	2	4490
U2R	1	0	0	5000	U2R	0	1	0	5000
R2L	33	4	4	4960	R2L	34	3	1	4963
Normal	2558	11	9	2423	Normal	2564	5	12	2420

Fig. 9 Summary of attacks

The confusion matrix for the ensemble classifier and its constituent classifiers is provided in Fig. 8. On examining the confusion matrices, it can be concluded that Neural Networks were efficient in identifying the lone test instance belonging to the U2R category where other classifiers failed to identify the same. C 5.0 identified highest instances belonging to Dos attacks while SVM returned lower false alarm rate (FAR) when compared with C 5.0, Neural Networks, and Ensemble model. Ensemble model outperforms the others in detecting Probe, R2L and Normal type.

The true positive, false positive, true negative, and false negative values obtained by different classifiers on the test dataset with respect to the four categories of attacks are summarized in Fig. 9. It can be observed that C 5.0 has identified highest number of Dos attacks when compared with other classifiers. Ensemble classifier identified highest number of Probe and R2L attacks. Neural networks and support vector machine work best in detecting attacks belonging to U2R and normal categories.

Figure 10 shows the performance measures of individual base classifiers and ensemble approach. Except neural networks, other classifiers fail to detect U2R

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	99.84	99.84	99.9	99.78
Probe	99.88	99.66	99.86	99.54
U2R	99.98	99.96	99.98	100
R2L	99.92	99.74	99.92	99.84
Normal	99.66	99.52	99.66	99.6
<b>Weighted Average</b>	<b>99.67</b>	<b>99.60</b>	<b>99.71</b>	<b>99.61</b>

**(a) Accuracy**

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	99.73	99.84	99.73	99.62
Probe	99.21	98.42	99.01	98.23
U2R	0	0	0	100
R2L	91.89	89.18	91.89	89.18
Normal	99.80	99.37	99.88	99.57
<b>Weighted Average</b>	<b>99.63</b>	<b>99.30</b>	<b>99.60</b>	<b>99.37</b>

**(c) Detection Rate**

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	99.78	99.78	99.86	99.69
Probe	99.40	98.32	99.30	97.74
U2R	0	0	0	100
R2L	94.44	83.53	94.44	89.19
Normal	99.66	99.52	99.66	99.60
<b>Weighted Average</b>	<b>99.55</b>	<b>99.31</b>	<b>99.58</b>	<b>99.31</b>

**(e) F-Measure**

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	0.09	0.16	0	0.12
Probe	0.04	0.20	0.04	0.31
U2R	0	0.02	0	0
R2L	0.02	0.18	0.02	0.08
Normal	0.49	0.32	0.57	0.37
<b>Weighted Average</b>	<b>0.07</b>	<b>0.15</b>	<b>0.08</b>	<b>0.15</b>

**(b) False Alarm Rate**

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	99.84	99.73	100	99.78
Probe	99.60	98.23	99.60	97.27
U2R	0	0	0	100
R2L	97.14	78.57	97.14	89.18
Normal	99.53	99.68	99.45	99.64
<b>Weighted Average</b>	<b>99.58</b>	<b>99.31</b>	<b>99.45</b>	<b>99.29</b>

**(d) Precision**

Class	Ensemble	C5.0	SVM	Neural Networks
Dos	99.78	99.78	99.86	99.69
Probe	99.40	98.32	99.30	97.74
U2R	0	0	0	100
R2L	94.47	83.70	94.47	89.18
Normal	99.66	99.52	99.66	99.60
<b>Weighted Average</b>	<b>99.55</b>	<b>99.31</b>	<b>99.58</b>	<b>99.31</b>

**(f) G-Mean**

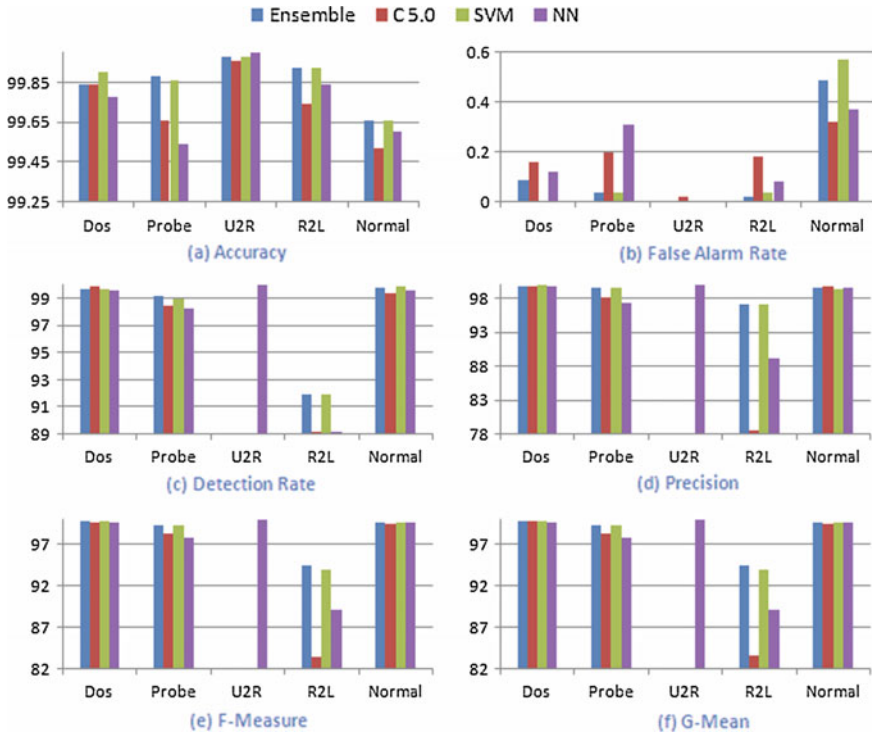
**Fig. 10** Attack classification

attacks. Figure 10a shows that ensemble-based approach returns best values for Probe, U2R, R2L attack types and normal instances whereas SVM returns best accuracy for Dos attacks. Similarly, SVM achieves 0% error rate for Dos category whereas ensemble approach obtains lowest False alarm rate for other categories. From Fig. 10, it can be observed that SVM works well for Dos attack type, neural network for U2R attack type and Ensemble approach for others.

Figure 11 presents the performance measures on the transformed NSL\_KDD dataset. It is observed that the ensemble model achieved good results overall with the exception of U2R attacks, the SVM classifier achieved highest detection rate for normal type, and C 5.0 achieved much higher detection rate for Dos attack as compared to the ensemble model.

## 5 Conclusion and Future Work

This paper explores different data mining based approaches to building network intrusion detection systems. Privacy to sensitive attributes is preserved by anonymizing the corresponding attribute values. The performance of different



**Fig. 11** Performance of ensemble model and base classifiers

approaches is implemented using IBM SPSS Modeler 17.0. It is observed that the ensemble-based approach provides overall better results when compared with other popular classification techniques.

**Acknowledgements** The authors would like to express their gratitude to the Science and Engineering Research Board (SERB), Ministry of Science & Technology, Govt. of India under grant No. SB/FTP/ETA-0180/2014 for providing support.

## References

1. Kotsiantis, Sotiris, Dimitris Kanellopoulos, and Panayiotis Pintelas. "Handling imbalanced datasets: A review." *GESTS International Transactions on Computer Science and Engineering* 30.1 (2006): 25–36. doi:10.1.1.96.9248.
2. McHugh, John, Alan Christie, and Julia Allen. "Defending yourself: The role of intrusion detection systems." *IEEE software* 17.5 (2000): 42–51. doi:10.1109/52.877859.
3. Surana, Shraddha. "Intrusion Detection using Fuzzy Clustering and Artificial Neural Network." *Advances in Neural Networks, Fuzzy Systems and Artificial Intelligence*. (2014): 209–217.

4. Dokas, Paul, LeventErtöz, Vipin Kumar, Aleksandar Lazarevic, Jaideep Srivastava, and Pang-Ning Tan. "Data mining for network intrusion detection." Proc. NSF Workshop on Next Generation Data Mining. (2002): 21–30.
5. Mulay, Snehal A., P. R. Devale, and G. V. Garje. "Intrusion detection system using support vector machine and decision tree." International Journal of Computer Applications 3.3 (2010): 40–43.
6. Zamani, Mahdi, and MahnushMovahedi. "Machine Learning Techniques for Intrusion Detection." arXiv preprint arXiv:1312.2177(2013):1–11.
7. Abraham, Ajith, and Ravi Jain. "Soft computing models for network intrusion detection systems." Classification and clustering for knowledge discovery. Springer Berlin Heidelberg, 2005. 191–207.
8. Panda, Mrutyunjaya, and ManasRanjanPatra. "Network intrusion detection using naive bayes." International journal of computer science and network security 7.12 (2007): 258–263.
9. McHugh, John. "Testing intrusion detection systems: a critique of the 1998 and 1999 darpa intrusion detection system evaluations as performed by lincoln laboratory." ACM Transactions on Information and System Security (TISSEC) 3.4 (2000): 262–294.
10. Farid, DewanMd, NouriaHarbi, EmnaBahri, Mohammad Zahidur Rahman, and Chowdhury Mofizur Rahman. "Attacks classification in adaptive intrusion detection using decision tree." World Academy of Science, Engineering and Technology 63 (2010): 86–90.
11. Pan, Zhi-Song, Song-Can Chen, Gen-Bao Hu, and Dao-Qiang Zhang. "Hybrid neural network and C4. 5 for misuse detection." Machine Learning and Cybernetics, 2003 International Conference on. Vol. 4. IEEE (2003): 2463–2467. doi:10.1109/ICMLC.2003.1259925.
12. Ibor, Ayei E., and Gregory Epiphaniou. "A Hybrid Mitigation Technique for Malicious Network Traffic based on Active Response." International Journal of Security and Its Applications 9.4 (2015): 63–80.
13. Peddabachigari, Sandhya, Ajith Abraham, CrinaGrosan, and Johnson Thomas. "Modeling intrusion detection system using hybrid intelligent systems." Journal of network and computer applications 30.1 (2007): 114–132. doi:10.1.1.74.2371.
14. Karthick, R. Rangadurai, Vipul P. Hattiwale, and BalaramanRavindran. "Adaptive network intrusion detection system using a hybrid approach." 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012). IEEE (2012):1–7. doi:10.1109/COMSNETS.2012.6151345.
15. Depren, Ozgur, Murat Topallar, EminAnarim, and M. Kemal Ciliz. "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks." Expert systems with Applications 29.4 (2005): 713–722.
16. Tavallae, Mahbod, EbrahimBagheri, Wei Lu, and Ali A. "A detailed analysis of the KDD CUP 99 data set." Proceedings of the Second IEEE Symposium on Computational Intelligence for Security and Defence Applications 2009. doi:10.1109/CISDA.2009.5356528.
17. Rodda, Sireesha, Erothi, Uma Shankar Rao. "Class Imbalance Problem in the Network Intrusion Detection Systems." International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) (2016).doi:10.1109/ICEEOT.2016.7755181.
18. IBM, "IBM SPSS Modeler 17 User's Guide," IBM, 1 August 2015. [Online]. Available:<ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.0/en/ModelerUsersGuide.pdf>. [Accessed 1 November 2015].

# Hiding Encrypted Multiple Secret Images in a Cover Image

Prashanti Guttikonda, Hemanthi Cherukuri  
and Nirupama Bhat Mundukur

**Abstract** This work proposes an encryption method and a spatial domain method for hiding multiple binary secret images in a single cover image. Here, the sender and receiver has to share a secret key and this secret key is XORed with the binary secret images resulting in scrambled images. The scrambled binary images are concealed by placing the bits in the LSB of the cover image (either in gray scale or color image).The transmission of images is more secured with this method since it requires extraction as well as decryption with the correct key. The PSNR, MSE values, and histogram of the encryption method and LSB technique are given with MATLAB.

**Keywords** Cryptography · Steganography · Information security  
Least significant bit substitution

## 1 Introduction

To protect the data transferred through network as well with the individual privacy data, several methods have been developed. Encryption is one such method which leads to a noise data that is visible. Steganography is another method where the privacy data is invisible. Steganography can be enhanced by combining it with encryption. Encryption makes the secret data into noisy data and steganography hides the noisy data in a digital media thereby creating a powerful system. The simplest technique of all steganography methods is a LSB method which is a spatial

---

P. Guttikonda (✉) · H. Cherukuri  
Vignan's Lara Institute of Technology & Science, Vadlamudi, Guntur, AP, India  
e-mail: prashantiguttikonda77@gmail.com

H. Cherukuri  
e-mail: hemanthi9666@gmail.com

P. Guttikonda · N.B. Mundukur  
VFSTR University, Vadlamudi, Guntur, AP, India  
e-mail: nirupama\_mca@vignanuniversity.org

domain technique. The encryption methods use symmetric and asymmetric keys. In symmetric key, both sender and receiver share one key where as in asymmetric key both sender and receiver use two different keys.

## 2 Related Work

Work of Nadeem Akhtar and Pragati Johri is concerned with a variation of plain LSB (Least Significant Bit) algorithm. The stego-image quality is improved by using bit inversion technique. In this technique, certain least significant bits of cover image are inverted after LSB steganography that co-occur with some pattern of other bits and that reduces the number of modified LSBs [1]. Approach proposed by Ling Xi, Xijian Ping, and Tao Zhang presents an improved LSB matching steganography, which complementarily modifies the pairs of pixels with adjacent intensity to embed secrete message. In LSB matching steganography, when adding or subtracting one from the cover image pixel, two adjacent bins of the histogram will be altered—the bin value of the modified pixel’s intensity increased by one, and one of its adjacent bin’s value decreased by one. Based on the alteration of histogram caused by LSB matching, the improved algorithm embeds two bits in a pair of pixels with adjacent intensity one time so as to minimize the alteration of histogram [2]. Pallavi Das and Satish Chandra Kushwaha have focused on concealing multiple secret images in a single 24-bit cover image using LSB substitution based image steganography [3]. Amirfarhad Nilizadeh and Ahmad Reza Naghsh Nilchi have proposed a new steganography algorithm that combines two different steganography methods, namely Matrix Pattern (MP) and Least Significant Bit (LSB), which are presented for RGB images. The MP method is an algorithm which, first, divides the “Cover-Image” into nonoverlapping  $B \times B$  blocks. Then, it hides the data in the 4th through 7th bit layers of the blue layer of the “Cover-Image”, by generating unique tixt2 matrix patterns for each character in each block. The LSB method is an algorithm that hides data in the least significant bit of the “Cover-Image” pixels, which has the least visible effect on the transparency of the “Stego-Image” [4]. M. S. Sutaone and M. V. Khandare have proposed a steganography system that is designed for encoding and decoding a secret file embedded into an image file using random LSB insertion method in which the secret data are spread out among the image data in a seemingly random manner. This can be achieved using a secret key [5].

## 3 Proposed Method

The encryption method takes a secret key and is XORed with each binary image. The results obtained are then concealed in the image using LSB steganography.

### 3.1 Encryption Algorithm

This is a simple stream cipher in which the key is transferred into binary format and is then XORed with the bit of binary image (Fig. 1).

### 3.2 Decryption Algorithm

The decryption procedure is shown in Fig. 2 where scrambled image is XORed with the key.

### 3.3 Hiding in Grayscale Image

- (1) Read the cover image and its size should be equal to size of  $N$ \*size of binary image, where  $N$  is number of binary images.

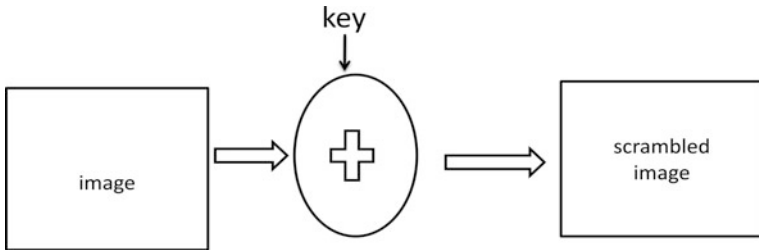


Fig. 1 Encryption algorithm

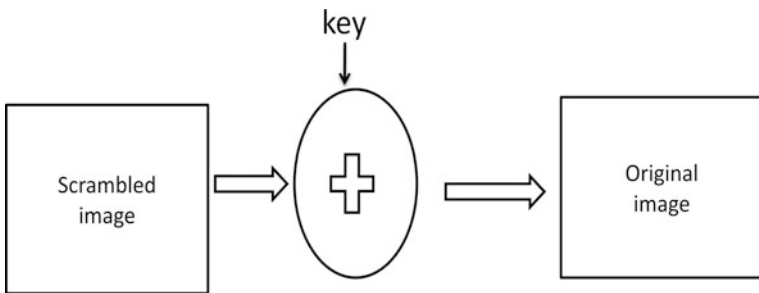


Fig. 2 Decryption algorithm

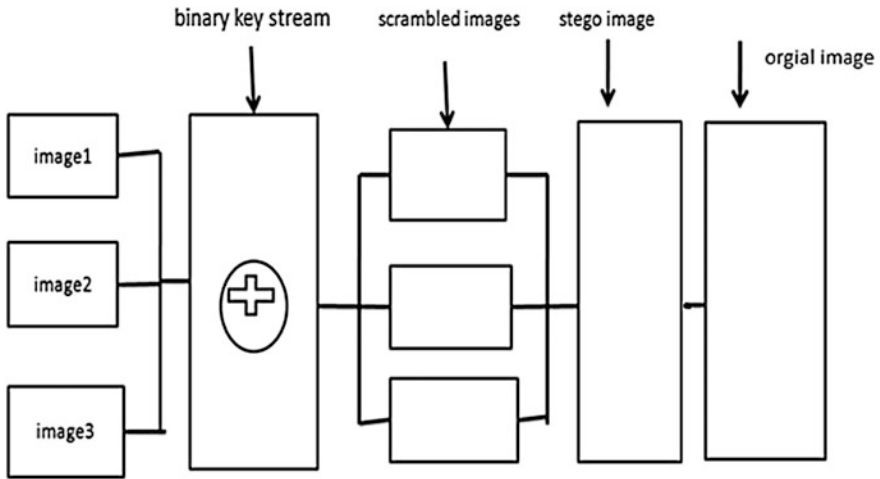


Fig. 3 Hiding in gray scale image

- (2) Read the multiple binary images and scramble them using encryption algorithm.
- (3) The scrambled images are the placed one after the other in the cover image, i.e., first image is hidden in the  $N$  pixels and second image is placed from  $N + 1$  pixel and so on (Fig. 3).

### 3.4 Extract in Gray Scale Image

Input: Encoded message, grayscale image

Output: Stego image

Steps:

1. Extract the encrypted secret image1 from LSB of the stego image1.
2. XOR the extracted image1 with the binary key stream.
3. Secret image1 is obtained.
4. Repeat the Steps 1–3 until the  $N$  images are extracted.
5. End.

### 3.5 Color Images

The scrambled images are hidden in RGB components of color image, i.e., first image is hidden in red component, second image is in gray and third image is in blue components (Fig. 4).



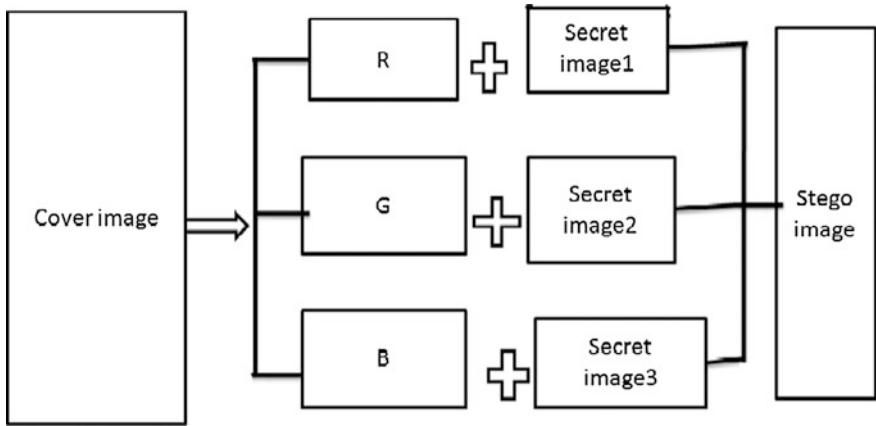


Fig. 4 Hiding in color image

### 3.6 Extract the Color Image

Input: Encoded message, color image

Output: Stego image

Steps:

1. Read the cover image.
2. Extract the RGB components from the cover image.
3. Retrieve the encrypted secret image 1 from R component and XOR with the binary key stream. Thus secret image 1 is obtained.
4. Retrieve the encrypted secret image 2 from G component and XOR with the binary key stream. Thus, secret image 2 is obtained.
5. Retrieve the encrypted secret image 3 from B component and XOR with the binary key stream, Thus image 3 is obtained.
6. End.

## 4 Results and Discussion

### 4.1 Secret Images

Figure 5 shows the three secret images which are to be XORed with the secret key. The size of the three secret images is 256\*256.



**Fig. 5** a first secret image, b second secret image c third secret image

## 4.2 Encrypted Images

The three secret images in Fig. 5 are XORed with the key stream “hello world” and the corresponding encrypted images are shown in Fig. 6. The histograms of the each encrypted images are also shown.

## 4.3 Decrypted Images

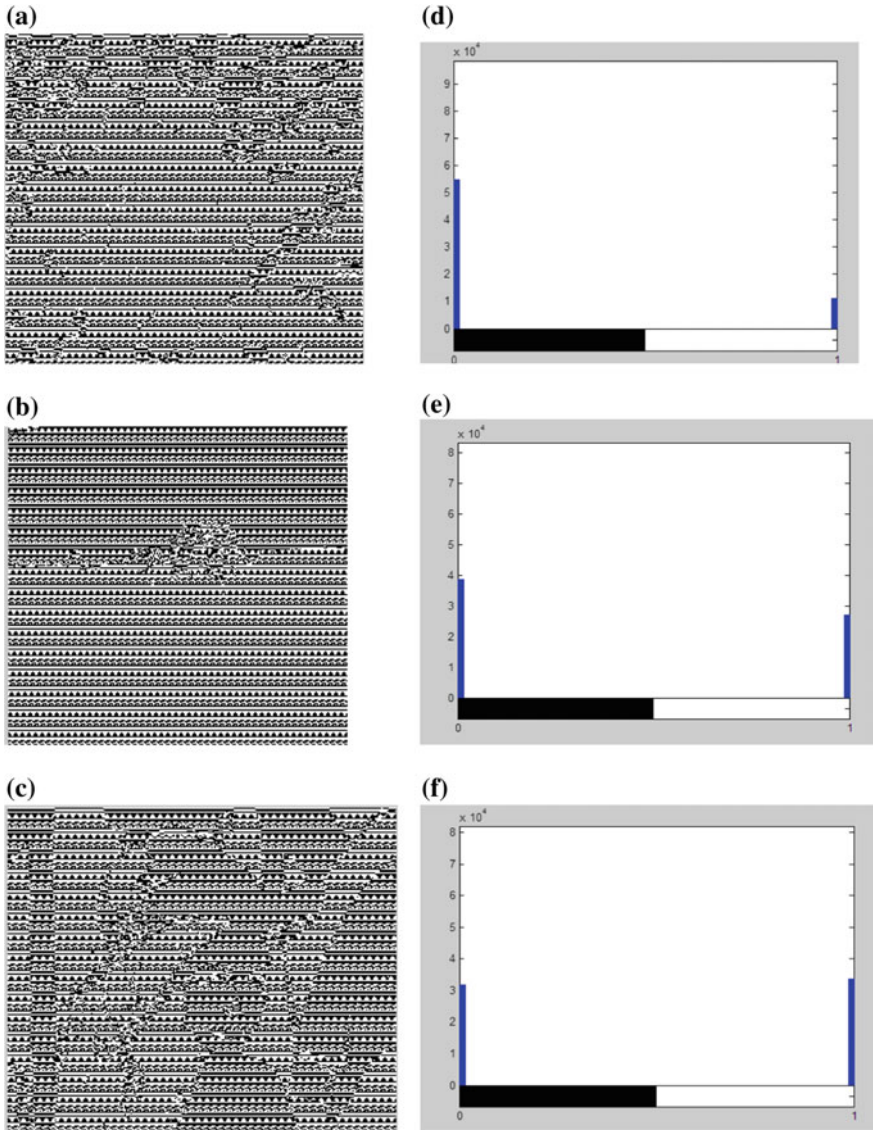
Figure 7 shows the decrypted images with the correct key that is hello world. Figure 8 shows the decrypted images with the wrong key meet me at toga party. This shows that images can be extracted only with the key that is used during encryption.

## 4.4 Stego Images

Figure 9 shows the two stego images along their MSE and PSNR values. The size of the two images is 512\*384. The MSE and PSNR values of images are shown in Table 1

## 4.5 Visual Quality and Histograms of Color Images

Figure 10 shows the two stego images along their histograms. The histograms of the stego images show definite amount of change from the histogram of cover images, but the visual quality of stego image shows no major changes. The cover images are of size 256 \* 256 \* 3 and the three secret images are of size 256 \* 256. Table 2 shows the MSE values and PSNR values.



**Fig. 6** Encrypted images of **a** first secret image, **b** second secret image, **c** third secret image and their corresponding histograms



Fig. 7 Decrypted images with key hello world



Fig. 8 Decrypted images with wrong key meet me at toga party

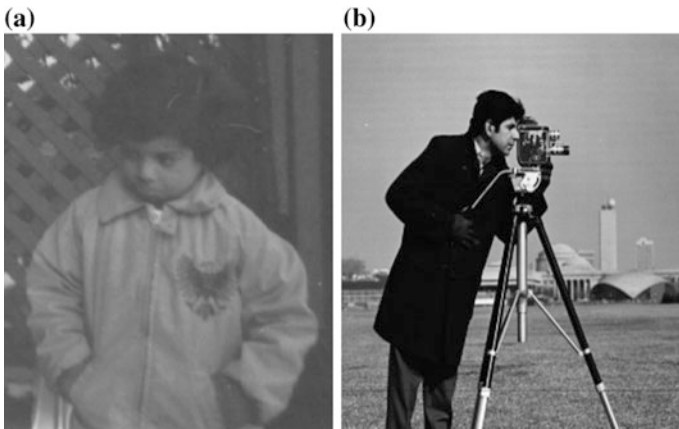


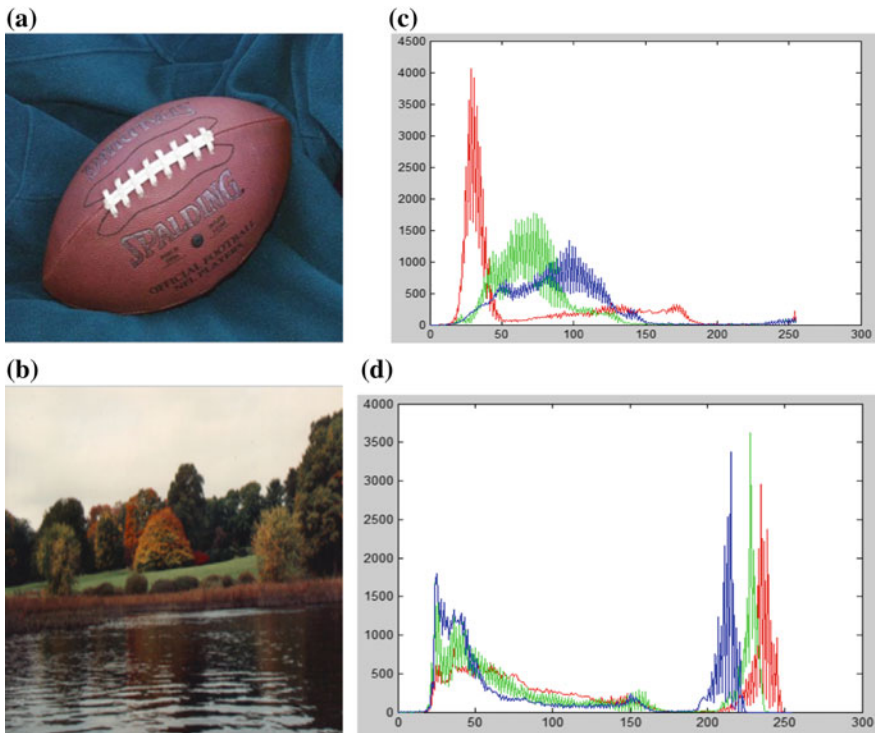
Fig. 9 Stego images

**Table 1** Quality parameters of different stego images

S. No.	Cover image	MSE of cover image to stego-image	PSNR of cover image to stego-image (db)
1	Pout.tif (512 * 384)	1.4949	46.3847
2	Cameraman (512 * 384)	1.5036	46.3596

**Table 2** Quality parameters of different color images

S. No.	Cover image	MSE of cover image to stego-image	PSNR of cover image to stego-image (db)
1	Football (256 * 256)	0.4988	51.15147
2	Autumn (256 * 256)	0.5008667	51.13407



**Fig. 10** Stego images of **a** football with 3 binary images hidden, **b** autumn with 3 binary images hidden, **c** histogram of stego image, **d** histogram of stego image

## 5 Conclusion

The simplest encryption technique is XORing the key with the secret image and the LSB is the simplest method for steganography. Combination of these techniques provides a better security system rather than using them individually. The quality parameter for both encryption and lsb method gives satisfactory results. In future, the embedded data can be encrypted by using other encryption techniques like DES, AES which are somewhat complicated encryption algorithms when compared to the proposed one.

## References

1. Pallavi Das, Satish Chandra Kushwaha, Madhuparna Chakraborty (2015) Data hiding using randomization and multiple encrypted secret images. International conference on communications and signal processing (ICCSP) pp 298–302. doi:[10.1109/ICCSP.2015.7322892](https://doi.org/10.1109/ICCSP.2015.7322892).
2. Khalid A, Al-Afandy, Osama S. Faragallah (2016) High security data hiding using image cropping and LSB least significant bit steganography. Information Science and Technology (CiSt) 4th IEEE International Colloquium.
3. Aakaash jois, Tejaswini L (2016) Survey on LSB Data Hiding Techniques. International Conference on Wireless communication. Signal processing and networking (WiSPNET) pp 656–660. doi:[10.1109/WiSPNET.2016.7566214](https://doi.org/10.1109/WiSPNET.2016.7566214).
4. B G Priyanka, S.V,Sathyanarayana (2014) A stegnographic system for embedding image and encrypted text. International conference on contemporary computing and informatics (IC3I) pp 1351–1355. doi:[10.1109/IC3I.2014.7019666](https://doi.org/10.1109/IC3I.2014.7019666).
5. Nadeem Akhtar, Pragati Johri,Shahbaaz Khan (2013) Enhancing the Security and Quality of LSB Based Image Stegnography. 5th International Conference and Computational Intelligence and Communication Networks pp 385–390. doi:[10.1109/CICN.2013.85](https://doi.org/10.1109/CICN.2013.85).

# Achieving Higher Ranking to Webpages Through Search Engine Optimization

B. Swapna and T. Anuradha

**Abstract** Search engine optimization is a process to maximize the number of visitors to a particular website or a webpage and to display the page on the initial search results page. It follows some methods and techniques to give ranking to the webpages. Major search engines like Google, Yahoo, Bing, etc., list the webpages on the search results page based on their ranking only. Search engines are like a bridge to connect to the Internet. In recent days, users are getting required information through search engines only. Based on the user's query, search engines show the quality webpages on the results page. So, search engines are responsible for providing ranking to the efficient webpages. This paper proposes some new optimization techniques to improve page ranking.

**Keywords** LSI keywords · Off-page SEO · On-page SEO · SERP  
User traffic · Webpage ranking

## 1 Introduction

The basic process of searching specific data is same in all famous search engines. Users enter a query for searching information about a particular content. The webpages containing data related to search query are shown on the Search Engine Results Page (SERP). With the help of the webpage ranking algorithms, search engines determine the position of the webpages which need to be shown on the SERP [1].

A huge amount of content is available in online, but finding the related content is a major key factor now a days. Search engines like Google, Yahoo, Bing etc., are the primary tools which provide search facility to the users for finding information

---

B. Swapna (✉) · T. Anuradha  
Velagapudi Ramakrishna Siddhartha Engineering College, Kanuru, Vijayawada, India  
e-mail: swapna9282@gmail.com

T. Anuradha  
e-mail: atadiparty@gmail.com

on the web. Users' queries are searched based on the keywords. When user enters a query or phrase, search engines start finding content matching webpages and list them on SERP [2]. Search engines are gaining popularity day by day as they help the users in quickly finding and filtering the information as per user's requirement. For that reason, search engines are giving rankings to the webpages. The webpages which have good ranking will be listed in the initial SERPs and will automatically be accessed by more number of users. The sites or pages which doesn't have better ranking will go to last SERPs and sometimes may not be accessed at all by the users. Search Engine Optimization (SEO) techniques will suggest the site owners to follow some techniques which will be useful to get better ranking and more user traffic [3, 4]. This paper introduces some new SEO techniques to achieve high page ranking and improve user traffic to the particular website.

## **2 Guide Lines to Achieve Ranking to Webpages or Websites**

Search engines are specific tools to do indexing of the webpages in an efficient manner and provide quality content to the users by ranking to webpages or web-blogs. Based on the usage of particular keywords in the webpages, the search engines rank them within a short duration after they were placed in the World Wide Web [5]. To get the top rank, some new techniques are required a part from keywords. When the user enters a query in the search engine, the related webpages will be displayed on the SERP in an order based on the ranking of the page. Search engines follow ranking algorithms for giving ranking to the webpages and for every 6 months, they do minor changes to that algorithms for avoiding fraud websites and duplicate content.

The website owners use SEO tools and techniques like keyword generation tools which suggest related keywords to be used in the content and position checker which checks the keyword position, etc., to improve their page ranking. Most of the website owners use SEO techniques like on-page and off-page SEO techniques for achieving good webpage ranking within less duration after the site is placed in the web. Google, Yahoo, Bing, etc., give ranks to efficient and quality content webpages or websites when they follow particular rules and regulations without violating the instructions.

### ***2.1 LSI Keywords to Improve Webpage Ranking***

Search engines like Google always try to do better service to the users by understanding the search query and display the pages related to users query. For doing this, they focus on semantic search and use Latent Semantic Indexing (LSI)



keywords [6]. An LSI keyword is a phrase that contains the words similar to the main keyword—it is often a synonym. For example, when a search engine discovers a page with a word Apple, it decides whether the term relates to a fruit or a brand based on LSI keywords. If the page also contains other words like taste, fruit, etc., related to the main keyword Apple, it decides the page is about the fruit Apple. So, the usage of LSI keywords will improve the ranking of the page.

## 2.2 *The Basic Search Engine Algorithms*

There are three pieces of software that form the components of a search engine [7]. They are

1. Spider Software which crawl the Internet in search of fresh webpages. It concentrates only on the text matter of the page.
2. Index Software which works by making sense of the heap of links, URLs, and text paragraphs.
3. Query Software which is at the frontend of a search engine. It presents the outcome of the hardworking Spider and Index software.

## 2.3 *Traditional SEO Techniques*

Once the website is placed in the web, it should get good ranking by the search engines, then only they will be visible in the initial SERPs and get more user traffic. To get proper ranking, SEO techniques should be used [8, 9]. Majorly Search Engine Optimization techniques are divided into two categories. They are on-page search engine optimization (On-Page SEO) and off-page search engine optimization (Off-Page SEO).

**On-Page SEO:** Mainly on-page SEO focuses on the following parameters to get better ranking by the search engines [10]. And it also suggests some changes to the webpages at development side for getting good ranking.

(1) URL of the Webpage or Website (2) Meta Tags (3) Focus Keyword Setting (4) Header Tags (5) Alt Tags (6) Title of the Webpage (7) Keyword Placement in the content (8) Description (9) Spelling and Grammar (10) Application of Bold and Italic styles to the Keywords in the content (11) Page loading (12) Putting Keywords in the Header Tags and Alt Tags (13) Keyword Density (14) Uniqueness of the webpage content

**Off-Page SEO:** The off-page SEO's main work is to interact with other websites and search engines to turn the traffic to the website. Not only Google, most of the search engines assign strength to the websites in the form of giving ranking to its pages based on off-page SEO. Some of the major off-page SEO techniques are provided below.

(1) Social Media (2) Facebook (3) Twitter (4) Header Tags (5) WhatsApp (6) Blog Submission (7) Internal Linking (8) Link Building (9) Directory Submission (10) Press Release Submission (11) Article Submissions

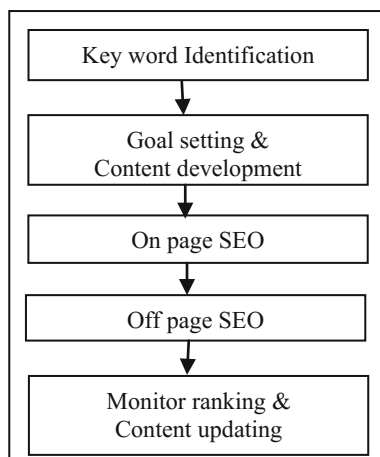
## 2.4 *SEO Process Flow*

Figure 1 shows the flow diagram of the SEO process. The first step in SEO process is Keyword Identification/Searching. The process here is, whenever a user enters a search query, the search engine will display different webpages in the order of their ranking. This is done by search engine based on the keywords used in the pages of the website. To get proper position in the search, the website owners can choose long tail keywords with less competition when compared to short keywords with high competition, because there is a chance to obtain good ranks with the long tail keywords within less time. The second step is goal setting which is nothing but website content development and optimization by analyzing the content of competitor websites. The third step is to use on-page SEO techniques and off-page SEO techniques for getting higher ranking in the search engines. And the last step is monitoring ranking and try to keep updating the content. Updating should be done at regular intervals, otherwise search engines will think that the website is not in active condition.

## 2.5 *Latest SEO Techniques*

Along with the on-page SEO techniques and off-page SEO techniques, the website developers need to concentrate on the following factors also to get top rank in search results [11]. They are mainly

**Fig. 1** SEO process flow diagram



- **Header Tags:** Most of the web content developers know the importance and usage of the Header tags but they concentrate only on H1, H2, H3, and H4 tags and forget the remaining tags. That is not the correct way to approach the search engine. Importance should be given to remaining tags like H5, H6, etc., also.
- **Power Point Template:** PPTs should be included in the content to keep the users in the website for longer time. So, attractive Power Point Presentations with straight points including the keywords need to be implemented.
- **You Tube Videos:** You Tube videos need to be attached to the webpages to make the search engine think that the webpage has connection with even you tube browser also. So, there is a chance of giving priority to the website by search engines.
- **Bold Keywords:** If the webpage has more than three lines of content, then long tail keywords need to be included and they should be made bold for easy identification. Search engines when doing indexing process automatically give some value to bold keywords, because they think that those keywords are most important ones related to the content.
- **Comment Backlinks:** Some popular websites provide links to an unpopular or newly hosted website based on their requirements. They may ask questions relevant to their websites and expect answers from other sites. If other sites are giving proper solutions, then automatically the other sites will get a chance to get a link to be attached to that particular popular website. When a search engine runs its ranking algorithm, it automatically finds the websites which have their profile backlinks. Based on this, it assumes that the website contains quality content which is useful to the users. This is also one of the methods in off-page SEO.
- **Convert PPT to PDF form:** Many PDF websites are there to provide information in PDF format for easy downloading and reading purpose. So, the website PPT can be converted into PDF to get PDF link of that site.

### 3 Experimental Work

The experimental work is done in collaboration with design team of [www.compareegg.in](http://www.compareegg.in). The website's main purpose is to compare the product prices at various online stores like Amazon, Flipkart, Snapdeal, eBay, etc., and display them online. Generally most of the people, before they buy any product, search for price details of a particular product which has same features and specifications from different companies. There are many websites that already existed to compare the prices of products. So, to achieve high rank to our website within less time, Google ranking algorithms were analyzed as well as some high ranking websites' content patterns along with keyword stuffing and usage of keywords in the content were observed. It was identified that they were mostly using latest SEO techniques mentioned in

Sect. 2.5. All these techniques were incorporated in the website design of compareegg.in and also the following three new SEO techniques were proposed:

1. In using the headers in the page, a hierarchy is followed. And the content that is not seen by most of the users about the product is given a H6 header.
2. Uniqueness of the content which is not matching with any other website of this type.
3. Google’s Keyword planner will list all the keywords related to the content and the volume of the keyword. Based on this, importance was given to other keywords which are in the next position with respect to volume of usage along with focus keyword. And it was seen that these keywords appear at least three times in every webpage.

Figure 2 shows the ranking position of compareegg.in website in Google SERP when the keyword moto gold mobile is given as search keyword. Figure 3 shows the user traffic of compareegg.in website without using latest SEO techniques or proposed techniques. Here only on-page SEO and off-page SEO techniques were used. Figure 4 shows the user traffic of compareegg.in website after using latest SEO.

The comparison of Fig. 3 and Fig. 4 shows that after applying latest SEO techniques, users count increased from 1299 to 2317 and page views of the site increased from 3845 to 9881.

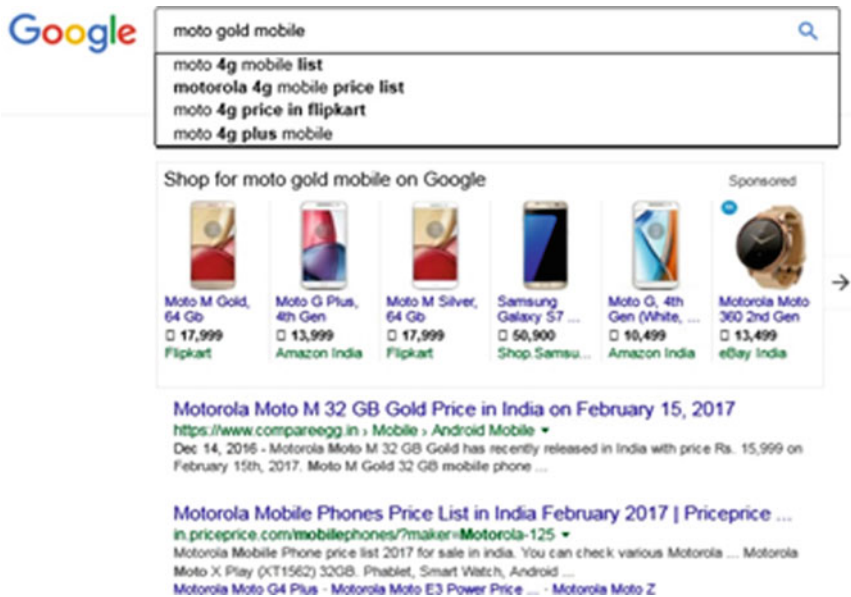


Fig. 2 Motorola Moto M 32 GB Gold Price Webpage position in Google Results page

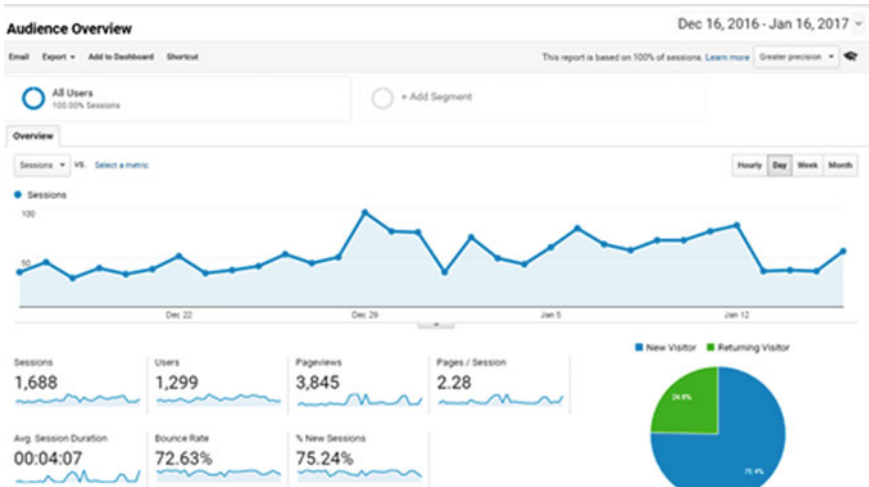


Fig. 3 Compareegg.in website traffic results without using new SEO techniques

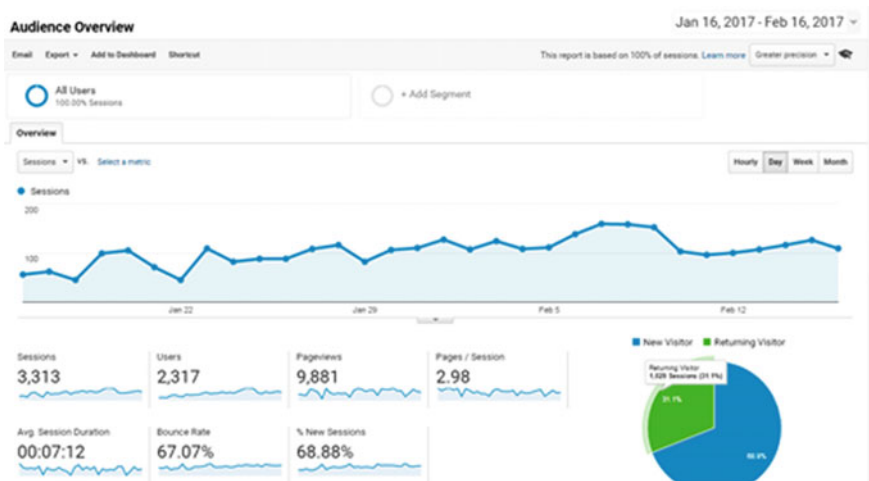


Fig. 4 Compareegg.in website traffic analysis results after using new SEO techniques

## 4 Conclusions

Search engine optimization is an efficient method of getting high page ranking for a website. It can also improve the site traffic. It is very much needed for the commercial websites to attract the number of users to their sites as the correct usage of SEO techniques will make the site identified by search engines.

**Acknowledgements** The authors thank compareegg.in website management team for allowing the authors to work along with site development team and for providing the result screenshots.

## References

1. Cui M, Hu S (2011) Search engine optimization research for website promotion. In: Information Technology, Computer Engineering and Management Sciences (ICM), International Conference on. IEEE 4: 100–103. doi: [10.1109/ICM.2011.308](https://doi.org/10.1109/ICM.2011.308).
2. Usha M, Nagadeepa,N(2015)On-Page and Off-Page Optimization Techniques for Search Engine Results Page(SERP). International Journal of Advanced Research in Computer Science and Software Engineering. 5(1): 1042–1046.
3. urRehman K., Khan MA (2013) The foremost guidelines for achieving higher ranking in search results through Search Engine Optimization. International Journal of Advanced Science and Technology. 52: 101–110.
4. Pabitha C, (2015). Search Engine Optimization by Eliminating Duplicate Links. International Journal of Engineering Research in Management & Technology. 4(1): 53–58.
5. Chengling Zhao, Jiaojiao Lu, FengfengDuan. (2009) Application and Research of SEO in the Development of Web2.0 Site. Knowledge Acquisition and Modeling, International Symposium on, 01: 236–238. doi: [10.1109/KAM.2009.69](https://doi.org/10.1109/KAM.2009.69).
6. Almpandis G, Kotropoulos C, Pitas I (2005) Focused crawling using latent semantic indexing—An application for vertical search engines. In: International Conference on Theory and Practice of Digital Libraries 2005 pp. 402–413. Springer Berlin Heidelberg. doi: [10.1007/11551362\\_36](https://doi.org/10.1007/11551362_36).
7. Hussien AS (2014) Factors Affect Search Engine Optimization. International Journal of ComputerScience and Network Security 14(9): 28–33.
8. Jain A (2013) The Role and Importance of Search Engine and Search Engine Optimization. International Journal of Emerging Trends & Technology in ComputerScience 2(3): 99–102.
9. Su AJ, Hu YC, Kuzmanovic A, KohCK(2014) How to improve your search engine ranking: Myths and reality. ACM Transactions on the Web (TWEB). 2014 8(2): 8. doi: [10.1145/2579990](https://doi.org/10.1145/2579990).
10. Bansal M, Sharma D. Improving Webpage Visibility in Search Engines by Enhancing Keyword Density Using Improved On-Page Optimization Technique. International Journal of Computer Science and Information Technologies 6(6): 5347–5352.
11. SEO Techniques and Strategies for 2017 (Infographic) <https://mytasker.com/blog/seo-strategies-and-techniques/>.

# Parallel Computing Algorithms for Big Data Frequent Pattern Mining

Subhani Shaik, Shaik Subhani, Nagaraju Devarakonda  
and Ch. Nagamani

**Abstract** Frequent Pattern Mining (FPM) is a focused research area with a goal of identifying the patterns that appear in the dataset most frequently. Due to huge increase in data volume and large search space, it is necessary to study the parallel computing algorithms for mining the frequent patterns. In the last two decades, many sequential algorithms have been implemented for solving FPM problem. Yet no more efficient algorithm exist for today's large data volumes called big data. In this paper, we presented a scalable parallel algorithm for big data frequent patterns mining. Three key challenges are identified to parallel algorithmic design: load balancing, work partitioning, and memory scalability. The experimental results are carried out using different datasets such as chess, census, mushroom, Kosarak, pumsb, connect and a comparison is made with existing parallel approaches. The experimental results show scalable performance and yield significant gains over different machines.

**Keywords** Big data · Parallel frequent pattern · Map reduce · Pattern growth  
IDD · Hybrid models

## 1 Introduction

Frequent pattern mining is a focused research area in data mining, with a goal of finding the patterns that appear in a dataset most frequently. The large volumes of data in present time required streaming frameworks for big data frequent pattern

---

S. Shaik (✉)

Acharya Nagarjuna University, Guntur, Andhra Pradesh, India  
e-mail: subhnicse@gmail.com

S. Subhani

St. Mary's Women's Engineering College, Guntur, Andhra Pradesh, India

N. Devarakonda

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

Ch. Nagamani

Andhra Loyola Institute of Engineering & Technology, Vijayawada, India

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_10](https://doi.org/10.1007/978-981-10-6319-0_10)

mining. Extraction of frequent patterns in transaction-oriented database is an essential data mining task to several mining processes such as classification, association rule generation and time series analysis. Most of these mining tasks require multiple passes over the database and if the database size is big, then scalable solutions connecting multiple processors are required. In the last two decades, many efficient pattern mining algorithms are implemented for solving FPM problem. Unlike most of the earlier parallel approaches based on different variants of a priori algorithm, FPM does not explicitly result in duplication of entire counting data structure on each processor. FPM has many applications in various domains, such as Intrusion detection, software bug recognition, drug discovery, credit card fraud prevention, spatiotemporal data, and market basket analysis. Scalable parallel algorithms have significant gain for recognizing pattern mining in the background of Big Data. This paper presents latest advanced techniques for work out the Big Data FPM in parallel and analyzing them over the lens. Three key challenges are identified to parallel algorithmic design: load balancing, work partitioning and memory scalability. These key challenges as a structure of reference, we extract and express algorithmic design patterns. We start our discussion by presenting a brief literature review on frequent pattern mining algorithms in Sect. 2. Section 3 and 4 describes regarding scalable challenges and big data computation. Section 5 gives parallel computing algorithms for big data frequent pattern mining. Finally Sects. 6 and 7 describe analytical results and conclusion.

## 2 Related Work

The identification of patterns that appear in a dataset most frequently is an important task to several data mining issues related clustering and classification. The value and importance of long patterns are gaining increasing recognition in a wide range of domains including social network analysis, business intelligence, bioinformatics, and software engineering. Yet, identifying the long pattern remains a challenging task due to the prohibitively large number of smaller patterns which often need to be generated first. In this section, we presented a brief review on parallel computing algorithms for FPM. Jiawei Han and Jian Pei et al. presented a novel approach called FP growth for mining the frequent patterns. Divide and conquers mechanism is used to project on division databases. Further, an efficient data structure has been implemented for database squeezing and quick memory transfer. Such as methodology may eliminate or significantly reduce the total number of candidate sets to be generated and reduce the volume of the database [1]. Jilles Vreeken and Nikolaj Tatti et al. introduced a classic model for closed and non-derivable item sets that are used to prune unwanted item sets. This model can either be dynamic or static and we can repeatedly revise the model for finding the novel patterns. Siegfried Nijssen, Albrecht Zimmermann proposed a restraint-based pattern mining systems for discovering dissimilar types of patterns fulfilling constraints. It accomplishes high-level languages in which programmers can simply specify



restraints, general search algorithms discover patterns for several jobs expressed in the necessity language [1]. Association rules are mainly used to count the transactions where the items occurred together. Yet, counting absenteeism of items is extreme if the more number of likely items is extremely large. In some applications, it is difficult for knowing the association between the absence and the presence of items. These rules are called negative association rules. In the next sections, we presented the parallel frequent pattern mining and big data computation.

### 3 Parallel Frequent Pattern Mining

#### 3.1 Frequent Pattern Mining—a Generic View

Let  $D$  be a given transaction database for identifying the frequent patterns, which consists of a set of transactions  $\{q_1, q_2, \dots, q_n\}$  and a user-specified minimum support  $\emptyset$  (where  $\emptyset$  lies between 0 and 1). It discovers the complete set of all patterns that contained at least a specified percentage of transactions in the transaction database. Based on given problem, the input pattern may be a graph, series, item set, or a tree. A variety of algorithms are implemented for solving the FPM problem such as Tree projection, FP growth, Apriori algorithm, and Éclat. The following pseudocode represents a generic view of frequent pattern mining. (Fig. 1)

The above algorithm represents a transaction database with  $Q$  and a user-defined support with  $m$  as an input. Frequent pattern data store  $FP$  is initializes to empty. Insert all the frequent patterns from data store frequent pattern, and generate a candidate key. If the generated candidate key is above the threshold level, then pattern is stored in  $FP$ . The process is iterated until all the  $FP$ 's from the database are finished.

```

Algorithm
  Input: Database Q, Minimum Support m
  Begin
    FP = {}; // Frequent pattern data-store FP is
              Initializes to empty
    Insert length-one frequent pattern in FP
    Until all frequent pattern in FP
    Begin
      Generate a candidate pattern P from one(or more)frequent
      Pattern (m) in FP
      If support (P, Q) ≥ m
        Add P to frequent pattern set FP;
    End End
  
```

**Fig. 1** A generic view of FPM

### 3.2 Scalable Challenges

The ability to gather huge amount of data has significantly improved because of approaches in hardware and software platforms. The streaming of big data analysis is slightly dissimilar and poses different challenges for the mining procedure. Three key problems related to frequent pattern mining are work partitioning, dynamic load balancing, and memory scalability. These key challenges as a reference frame, we extracted key algorithmic design patterns. Among these algorithms, memory scalability design is difficult to overcome the problem of parallel algorithms. Load balancing design challenge is also critical for efficient parallel execution of the frequent patterns. Work partition is a major issue for parallel frequent pattern in data mining. These three design issues continue to open much more problems for FPM. So, improved methods are estimating the cost of solving issues at each level of parallel frequent pattern. This process improves the efficiency of the system in these design issues.

### 3.3 Basic Mining Methodologies

Apriori algorithm is the basic frequent pattern method for candidate generation and it can be described recursively in hierarchy fashion. This algorithm contains three steps that are repetitive, for various values of  $k$ , where  $k$  is the cost of the pattern generated in the present iteration. These three steps are (i)  $C_{k+1}$  candidate patterns generation joins the patterns in  $F_k$  (ii) the  $C_{k+1}$  pruning of candidates, for all the subsets not lying in  $F_k$  (iii) the validation of pattern in  $C_{k+1}$  against the transaction database  $Q$ , to determine the subset of  $C_{k+1}$  which is truly frequent. When the set of frequent  $k$ -patterns in  $F_k$ , a given iteration is empty and corresponding algorithm is terminated. The pseudocode of the complete procedure is presented in Fig. 2.

Apriori algorithm computes the candidate's similar BFS method, decomposing the item sets pattern into level-wise depends on similarity classes:  $k$  item sets are computed earlier than  $(k + 1)$  item sets. Imagine lexicographic order of item sets, the search space can be divided into prefix- and postfix-based classes. Figure 2 illustrates similarity classes for prefix and postfix item sets of length one, in sequential order, for sample dataset. To discover frequent item sets of size one, immediately mined similarity classes separately. This algorithm follows lexicographic order, patterns increases by placing the correct items that track the previous parent item. (Fig. 3)

The FP-growth algorithm transforms the problem of finding lengthy frequent patterns to penetrating for smaller ones recursively and then combines the suffix. It uses the smallest frequent items as a suffix, giving fine selection. The procedure substantially minimizes the search costs and straight extracts the frequent patterns. The FP-growth algorithm is presented in Fig. 2.

**Algorithm** Apriori (Database  $Q$ , Support  $m$ )

```

Begin
  Compute F1 and F2 patterns using specialized
  counting methods;
  k =: 2;
  While Fk is not empty do
    Begin
      Generate Ck+1 by using joins on Fk;
      Prune Ck+1 with Apriori subset pruning trick;
      Generate Fk+1 by counting candidates in
      Ck+1 with respect to Q at support m;
      k =+ 1;
    End of the frequent patterns level wise;
    Return all frequent item set;
  End

```

**Fig. 2** Apriori algorithm**Algorithm** FP- growth (FPT, S, P)

```

// FPT - Tree on Frequent Items
// S-Minimum Support and P-Current Item set Suffix.
Begin
  If FPT is a single path do
    For every C of nodes in path do
      Inform all patterns C U P;
    Else
      For every item i in FPT do
        Begin
          Produce pattern  $P_i = \text{set } i \cup P$ ;
          Inform pattern  $P_i$  as frequent;
          Use pointer to extract condition prefix
            paths for item one;
          Construct conditional Frequent Pattern Tree
             $FPT_i$  from condition;
          From prefix paths after eliminating
            infrequent items;
          If ( $FPT_i \neq \emptyset$ ) FP- growth ( $FPT_i, P_i, S$ )
        End
      End
    End

```

**Fig. 3** Frequent pattern growth algorithms

## 4 Big Data Computation

### 4.1 Design Fundamentals of Parallel Algorithms

Parallel algorithm is a formula that tells us how to resolve a known problem using many processors. However, specifying a parallel algorithm involves more than just specifying the added dimension of concurrency, and the algorithm designer must specify few principles. Parallel algorithm that yields the performance commensurate with the computational and storage resources is employed to resolve the problem. Often, dissimilar choices yield the finest performance on dissimilar parallel architectures or under diverse parallel programming paradigms. In practice, parallel algorithms include: work separation that can be dispersed to the processors, the concurrent pieces of work mapped to many processes, supervise memory scalability data distributed by all processors, and regulate the processors at different levels of the parallel load balancing for execution of the program.

### 4.2 Shared Memory Systems

At the time of design parallel algorithms, we must be aware of memory model for controlling and deciding how data will be accessed and stored, which in turn plays a straight activity of parallel algorithm in the presence of design and performance. Shared memory systems are parallel computing machines that share a separate memory address locations. In recent years, common memory machines have become steady as to everywhere multi-core systems are present everywhere.

### 4.3 Distributed Memory Systems

In distributed systems [2], the process has accessing privileges to internal private memory. During inter-process communication, task information, and sharing the input data must be done explicitly. Through network transmissions, processes exchange the data by writing and reading records on a shared file system. There are two procedures usually used for distributed memory systems. In practice, distributed memory systems may include message passing and map reduce

The message passing model developed the actor model of calculation [3], which is featured by inherent processing parallel within system. In this model, processes are by actors, and they interrelate through direct message passing. All messages must appeal to transmit and get a message; the request must endeavor to reduce network traffic, so message passing is acutely moderate than accessing local memory in the network traffic. Normally message passing based systems are designed based on parallel algorithms for the partition of data. In local memory,

the designing of the message passing activity holds an exact data of the input. For distributed memory systems, MapReduce [2] is an advanced programming scheme that occurs for data-intensive processing. It provides an uncomplicated technique of writing parallel programs and Hadoop is most popular open source system that has been developed in present days [4]. There is a problem of representing the input as a set of value pairs. The MapReduce processes the key, due to processing which generates one more key value pair. Generated value pairs are merged together by MapReduce framework and key values. Entire map process is allocated individual ID and mapping process repeats over every data in its allocated text and generates the key value pair.

## 5 Parallel Computing Algorithms

In frequent pattern mining, many challenging issues are faced by the big data analytics. A major issue arises when the data is large enough to be stored in distributed systems. The intermediate results of data are acquired to shuffling the mining process throughout the distributed nodes. The cost of data results is referred as transfer cost of distributed nodes connecting in the network. If the size of dataset is extremely high, then the design paradigm requires careful measures in the presence of data transfer costs and disk access constraint.

### 5.1 *Count Distribution Algorithm*

In CD algorithm, each site [5] has consecutive association mining activities and broadcasts the support counts of candidate item sets. Support counts of first item sets from every site are calculated using Apriori approach [6]. It broadcasts these item sets to additional sites and detects the universal frequent of first item sets. Accordingly, every site generates candidate 2—item sets and calculates their support counts. It minimizes the transaction length and detects extra similar transactions and stores in main memory. The dataset initially consists of frequent and infrequent items. Due to the total transactions could go above the main memory limit. The aim of this problem is to propose a method that partitions the large volume of data into different horizontal segments.

### 5.2 *Data Distribution Algorithm*

Agrawal and Shafer and Han et al. [7, 8] designed a data distribution and instead of splitting data horizontally, each computational node processes only a part of the set of candidates. In the first step of each iteration  $k$ , the candidates of length  $k$  are

distributed among nodes. Each node computes support of all its candidates and sends the information to the other nodes. Subsequently, pruning is performed and frequent patterns are processed in the next iteration. Good workload balance is achieved by assigning an equal number of candidates to each node.

### ***5.3 Intelligent Data Distribution Methods***

Han et al. [8] proposed IDD which addresses the main problems of DD. The local storage part of the database can send every one of the other PEs by means of the linear-time ring-based broadcast. Although DD partitions the candidates uniformly, amid the processors, it fails to split the job done on each transaction. Once candidates fit in the memory, IDD algorithms switch to CD algorithm. IDD performs a prefix-based partitioning and single item as a substitute of a round-robin candidate partitioning.

### ***5.4 Hybrid Methods***

Han et al. [8] developed a hybrid distribution algorithm which associates strategies, distributing data, and candidates. Nodes are dividing into different segments and the database is dispersed over these segments. The candidates are separated and refined by one node in a group. (Figure 4)

The parallel processing algorithm partitions large volumes of data into different segments. Then every segment, it prunes sporadic items and places each transaction in the main memory. While placating the transactions, we need to check whether they are in the main memory or not. If yes, then increase the transactions pointer by 1. Else, place the same transaction in the main memory. Finally, it places all main memory openings for this separator into a temp file. Each local site generates support counts and broadcasts them to every one of other sites to let each site compute globally frequent item sets parts for that pass. The HD algorithm attains more improved performance than clean IDD and CD algorithms. Hybrid distribution algorithm combines CD and IDD.

## **6 Results and Discussion**

In the section, the implementation of scalable parallel frequent pattern mining algorithm for big data frequent patterns mining is presented. The experimental results are carried out using different datasets such as chess, census, mush room,

```

Algorithm: Parallel Computing Algorithm
// FN is Non-frequent item set
//GN is global -frequent item set
// C2 is Global Frequent support counts from receiver
Begin
  For each transactions t =1 to N do
  Begin
    For each two subsets s of t
    If (s ∈ global frequent item set GN) then
    Begin
      Support=support+1;
    End
    t = Remove FN (t);
    Add t to output table;
  End
  Source to destination (C2);
  F2 =destination to destination (FG);
  C3=Candidate item set;
  Q=obtain the transactions from table t;
  While (Ck ≠ ∅) do
  For each transaction t in transaction table Q
  For each k subsets m of transaction set t
  If (m ∈ Ck)
  m.support=m.support+1;
  k=K+1;
  Source to destination (Ck); //generating candidate Item set
  of k+1 passes
  Ck + 1= Candidate item set
End

```

**Fig. 4** Parallel computing algorithms

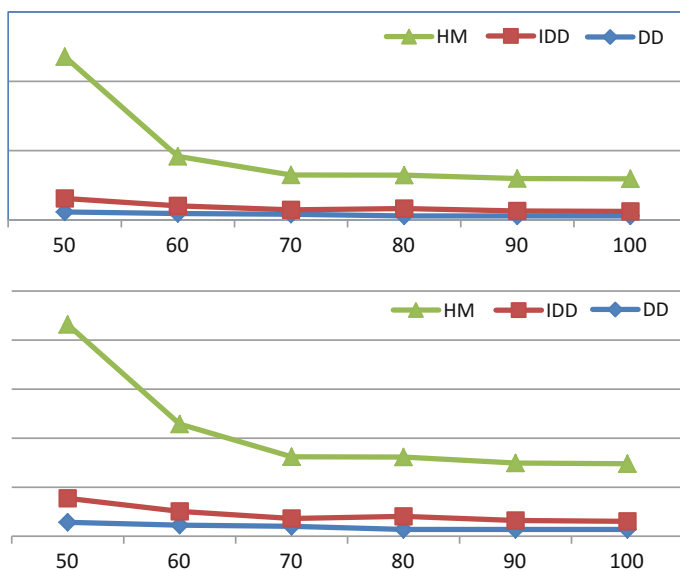
Kosarak, pumsb and connect as shown in Fig. 3. The majority of the parallel FPM algorithms are designed for well-established DMM (distributed memory) or SMP (shared memory) systems. Few algorithms have been implemented for inventing patterns on heterogeneous environments like GRID, hybrid systems like Clumps and even on novel processor architecture like SMT (simultaneous multithreading) and CMP [9]. (Table 1) (Fig. 5)

However, additional work is required to parallelize refined FPM algorithms on these promising elevated performance processing architectures. The above graph represents the least support on the decimal logarithm and horizontal axis of the finishing time in seconds on the vertical axis. The datasets basic for diagrams on the left side are rather density; those basics of the diagrams on the right side are rather sparse. The statistical result shows scalable performance on distinct machines and scope of the problems. A comparison is made with traditional parallel approaches and our proposed parallel computing algorithm shows significant gains [10].

**Table 1** Experimental results on six different datasets

S. No	Dataset	Support Count	FPM Computing Algorithm					
			APA	FPG	CD	DD	IDD	HM
1	Census	50	1.76	2.54	37.255	0.563	0.41	7.086
2	Chess	60	1.119	1.4	19.601	0.453	0.36	3.565
3	Mushroom	70	0.76	0.79	8.322	0.406	0.34	2.516
4	Kosarak	80	0.595	0.43	4.039	0.281	0.34	2.418
5	pumsb	90	0.455	0.3	1.844	0.281	0.34	2.345
6	Connect	100	0.17	0.24	1.156	0.281	0.33	2.345

Where APA-Apriori algorithm, FPG-FP Growth, CD-Count Distribution, DD-Data Distribution, IDD-Intelligence Data Distribution, M-Hybrid Models



**Fig. 5** Experimental results on six different datasets

## 7 Conclusion

Frequent pattern mining (FPM) is an imperative, focused research area in data mining, with a goal of identifying the patterns that appear in a dataset most frequently. Due to enormous increase in data volume and large search space, it is necessary to study the parallel computing algorithms for mining frequent patterns. In the last two decades, many sequential algorithms have been implemented for solving FPM problem. Yet no more efficient algorithm erects for today’s large volumes called “Big Data”. More research is needed for validating big data analytics using extant models. In this paper, an implementation of adequate scalable



parallel algorithm for big data FPM is presented. For mining the frequent patterns, three key challenging threats are identified to parallel algorithmic design: load balancing, work partitioning, and memory scalability. The experimental results are carried out using different datasets such as chess, census, mushroom, Kosarak, pumsb, connect and a comparison is made with existing parallel approaches. The analytical results show scalable performance and yields significant gains over different machines. The future enhancement to this work may include some updating such as using optimization to make the database proficient.

**Acknowledgements** Authors are grateful to Dr. D. Nagaraju, HOD & Professor, Dept. of Information Technology, Lakireddy Bali Reddy College of engineering, Vijayawada, India, for giving valuable suggestions and continued support to carry out this work. I am truly thankful to all the people who directly or indirectly helped me in this research work.

## References

1. <http://www.springer.com/in/book/9783319078205>.
2. Jeffrey Dean and Sanjay Ghemawat. Map reduce: Simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, January 2008.
3. Carl Hewitt, Peter Bishop, and Richard Steiger. Universal modular actor formalism for artificial intelligence. In *Proceedings of the third International Joint Conference on Artificial intelligence, IJCAI-73*, pages 235–245. Morgan Kaufmann Publishers Inc., 1973.
4. Foster, I., Kesselman, C., Nick, J., Tuecke, S.: *The physiology of the grid: an open grid services architecture for distributed systems integration*. Technical report, Global Grid Forum (2002).
5. M. Zaman, Ashrafi, T. David, K. Smith, “ODAM: An Optimized Distributed Association Rule Mining Algorithm”, In *IEEE Distributed Systems Online*, Los Alamitos 2004.
6. A. Inokuchi, T. Washio, H. Motoda, “An Apriori-Based Algorithm for Mining Frequent Substructures from Graph Data” In *Lecture Notes in Computer Science*, pg. 13–23, 2000.
7. Agrawal, R.—Shafer, J.C.: *Parallel Mining of Association Rules*. *IEEE Transactions on Knowledge and Data Engineering*, Vol. 8, 1996, No. 6, pp. 962–969.
8. Han, E.-H. Karypis, G. Kumar, V.: *Scalable Parallel Data Mining for Association Rules*. In *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data*, May 13–15, 1997, Tucson, Arizona, USA, 1997, J. Peckham, Ed., ACM Press, pp. 277–288.
9. J.R. Jeba and Dr. S.P. Victor, “Comparison of Frequent Item Set Mining Algorithms”, *J.R. Jeba et al, / (IJCSIT) International Journal of Computer Science and Information Technologies*, Vol. 2 (6), 2011, 2838–2841.
10. Amit Mittal, Ashutosh Nagar, Kartik Gupta and Rishi Nahar, “Comparative Study of Various Frequent Pattern Mining Algorithms”, *International Journal of Advanced Research in Computer and Communication Engineering*, Vol. 4, Issue 4, April 2015.

# Writer Identification System for Handwritten Gurmukhi Characters: Study of Different Feature-Classifier Combinations

Sakshi, Naresh Kumar Garg and Munish Kumar

**Abstract** In this paper, we are exploring various features and classifiers for writer identification in light of Gurmukhi text handwriting. The identification of the writers based on a piece of handwriting is a challenging task for pattern recognition. The writer identification framework proposed in this paper includes diverse stages like image preprocessing, feature extraction, training, and classification. The framework first prepares a skeleton of the character so that meaningful data about the handwriting of writers can be extracted. The feature extraction stage incorporates various plans, namely, zoning, diagonal, transition, intersection and open end points, centroid, the horizontal peak extent, the vertical peak extent, parabola curve fitting, and power curve fitting based features. In order to assess the prominence of these features, we have used four classification techniques, namely, Naive Bayes, Decision Tree, Random Forest and AdaBoostM1. For experimental results, we have collected 49,000 samples from 70 different writers. In this work, maximum accuracy of 81.75% has been obtained with centroid features and AdaBoostM1 classifier.

**Keywords** Feature extraction · Classification · Naive bayes · Decision tree  
Random forest · AdaBoostM1

---

Sakshi · N.K. Garg

Department of Computer Science & Engineering, GZS Campus College  
of Engineering & Technology (Maharaja Ranjit Singh Punjab Technical University),  
Bathinda, Punjab, India  
e-mail: me.sakshi51@gmail.com

N.K. Garg

e-mail: naresh2834@rediffmail.com

M. Kumar (✉)

Department of Computer Applications, GZS Campus College of Engineering & Technology  
(Maharaja Ranjit Singh Punjab Technical University), Bathinda, Punjab, India  
e-mail: munishcse@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_11](https://doi.org/10.1007/978-981-10-6319-0_11)

125

## 1 Introduction

Writing is implied by the representation of language in a textual medium using an arrangement of ‘signs or symbols’. Every individual has his own writing style which relies on a considerable measure of variables like particular shape of letters, spacing between letters, etc. Handwriting of a person is also subject to the mental condition of the individual like his level of inspiration, anger, joy, and others. In any case, it is found that handwriting of a person is generally steady however might be influenced gradually with age. Writer identification, in general plays an important role in forensic, writer verification schemes, and related branches of science. The parameters are typically considered for writer identification is comprehensiveness uniqueness, aging, accessibility, etc. The same is applicable in Gurmukhi script too with an extra essentialness of finding the advancement and development of the Gurmukhi script. It is clear that the significance of writer identification has turned out to be more vital nowadays. So, the number of researchers engaged with this challenging issue, is going on top of these opportunities. There are many languages all through the world. Each language represents an alternate danger to the writer identification issue contingent upon the characteristics of the language. It is clear that the identification problem differs across various languages. The handwriting-based writer identification is an active research area. A comprehensive survey of writer identification till 1989 was presented by Lorette and Plamondon [8, 11]. Zois and Anastassopoulos [14] have presented a writer identification framework based on English and Greek writers. They have achieved an accuracy of 95% for writer identification of both English and Greek writers. Leeham et al. [9] proposed a technique to recognize the writers based on numerals written by them. Schlapbach and Bunke [12, 13] presented a Hidden Markov Model (HMM) based writer identification and verification strategy. They prepared an individual HMM for each writer’s handwriting. The identification technique was tested with data set collected from 650 scholars. Gazzah and Amara [2] have proposed an approach for writer identification framework based on offline handwritten Arabic script documents. The proposed technique depends on combining the global and neighborhood feature sets by using the genetic algorithm so as to take out the redundant and irrelevant features. They have considered two classifiers, SVM and MLP for recognition. They have noticed that MLP performs preferable outcomes over SVM and got precision of around 94%. Ghiasi and Safabakhsh [3] have proposed an effective technique for writer recognition using a code book. They have utilized Farsi database which incorporates short, medium, and extensive messages and results demonstrate that the productivity of short messages are more effective. Maadeed [10] has proposed a writer identification framework based on Arabic handwritten text. He has also compared the edge direction distribution features with other features of Arabic text and used  $k$ -NN classifier for recognition. Writer identification framework is divided into two categories, namely, text dependent and text independent. Depending on the text content, text-dependent methods match the same characters and hence require the

writer to write the same text. Text-independent methods are able to identify writers independent of the text content and it does not require comparison of the same characters. A very few studies in Indian languages have been documented so far. Currently, writer identification of handwritten Gurmukhi script documents is done manually. In this paper, we have presented a study of different features and classifiers combinations for text dependent writer identification model based on Gurmukhi text handwriting. This paper is divided into six sections. The introduction and related work have been presented in Sect. 1. Section 2 illustrates the Gurmukhi script and data collection phase. Section 3 portrays the feature extraction techniques. In Sect. 4, we have briefly discussed about classification techniques. Section 5 incorporates experimental results using these features and classifiers. At long last, conclusion and future works are exhibited in Sect. 6.

## 2 Gurmukhi Script and Data Set

*Gurmukhi* script is the script used for writing the *Punjabi* language. In Gurmukhi script, there are 35 basic character constants out of which the initial three are vowel bearers. In this work, we have considered all these 35 fundamental Gurmukhi characters. Twenty samples of each character written by each writer are taken. In this manner, we have collected 49,000 samples from 70 different writers. 70% data of 49,000 samples has been taken as training dataset and rest of data is considered as testing dataset.

## 3 Feature Extraction Techniques

The performance of writer identification system, basically, is dependent on the features that are being extracted. The extracted features ought to have the capacity to classify a writer in a unique way. In this work, we have explored various features like, zoning, diagonal, transition, intersection and open end points, centroid, horizontal peak extent, vertical peak extent, parabola curve fitting, and power curve fitting based features. These features are extracted by using a hierarchical technique presented by Kumar et al. [6]. Numerous scientists have been utilized zoning, diagonal, transitions and intersection point's, etc., based features for printed or handwritten character recognition work. Parabola curve fitting and power curve fitting based features are provided by Kumar et al. [7]. They have used these features for offline handwritten Gurmukhi character recognition. They have also proposed peak extent based feature extraction techniques for handwritten Gurmukhi character recognition [5]. In this work, we have considered these strategies for proposing a text dependent writer identification framework based on Gurmukhi text handwriting.

## 4 Classification Techniques

Classification is the last stage of the writer identification framework which is used to classify the writers based on the features extracted in the previous phase. In this stage, just a final decision is taken about the unknown writer of character, to which class it belongs to find the identity of a writer. In this paper, we have considered four different classification methods, namely, Naive Bayes, Decision Tree, Random Forest, and AdaBoostM1. The Naive Bayes [4] classifier is a basic method which has a very clear semantics representing a probabilistic knowledge. It assumes that in a given class, predicative attributes are conditionally independent. Attributes in decision tree are nodes and each leaf node is representing a class of the writer. Decision tree classifiers are used to classify various subsamples of dataset. Random forest is one of the most popular and powerful machine learning algorithms [1]. It is a type of ensemble machine learning algorithm. Random forest removed the over-fitting crisis of decision tree. The *meta* estimator that fits the number of decision tree classifiers for such purpose is called random forest. The random forest uses averaging to help in getting better predictive accuracy and control over-fitting. Random forest is unexcelled in accuracy among existing supervised learning algorithms for classification and runs efficiently on large data bases [12]. AdaBoostM1 is a machine learning *meta* algorithm and can be used in conjunction with many other types of learning algorithms to improve their performance. The output of the other learning algorithms is combined into a weighted sum that represents the final output of the boosted classifier. AdaBoostM1 is adaptive in the sense that subsequent weak learners are tweaked in favor of those instances misclassified by previous classifiers. In this work, we have considered AdaBoostM1 with random forest classifier for improve the proposed writer identification accuracy.

## 5 Experimental Results and Discussion

In this section, we have presented experimental results of various features and classifiers considered in this work. For experimental results, we have used a dataset of 49,000 samples collected from 70 different writers of isolated handwritten characters of Gurmukhi script. 70% data from 49,000 samples is taken as training dataset and rest of data is considered as testing dataset. Experimental results are derived using distinctive feature extraction methods and classification procedures. These outcomes are graphically portrayed in Fig. 1. We have accomplished most extreme identification accuracy of 81.75% with centroid features and AdaBoostM1 ensemble classification technique as depicted in Table 1. True Positive Rate (TPR) and False Positive Rate (FPR) of each writer for this case (Centroid features and AdaBoostM1 classification) are presented in Figs. 2 and 3.

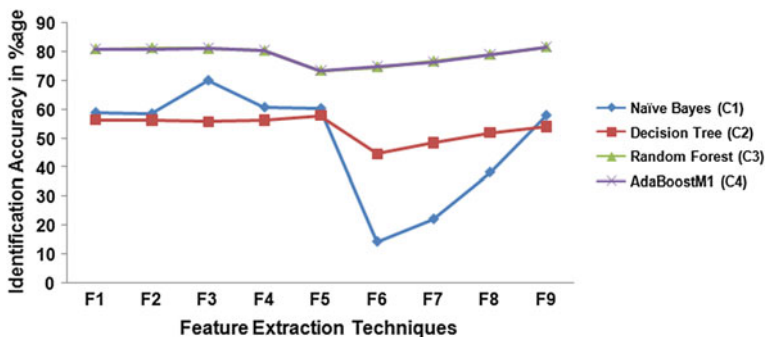


Fig. 1 Classifier-wise writer identification accuracy for different features

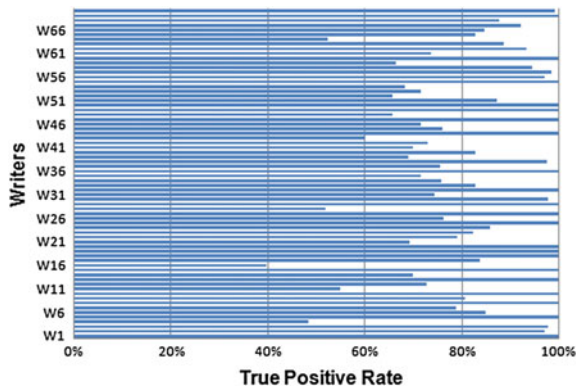
Table 1 Writer identification accuracy with different features and classifiers

	Classification Technique			
	Naive Bayes (C <sub>1</sub> )	Decision Tree (C <sub>2</sub> )	Random Forest (C <sub>3</sub> )	AdaBoostM1 (C <sub>4</sub> )
Zoning features (F <sub>1</sub> )	58.93%	56.45%	81.04%	81.16%
Diagonal features (F <sub>2</sub> )	58.54%	56.25%	81.46%	80.96%
Transition features (F <sub>3</sub> )	70.10%	55.8%	81.19%	81.37%
Intersection and open end points based features (F <sub>4</sub> )	60.75%	56.21%	80.49%	80.68%
Parabola curve fitting based features (F <sub>5</sub> )	60.37%	57.84%	73.51%	73.57%
Power curve fitting based features (F <sub>6</sub> )	14.40%	44.65%	74.78%	74.89%
Horizontally peak extent based features (F <sub>7</sub> )	22.24%	48.45%	76.77%	76.59%
Vertically peak extent based features (F <sub>8</sub> )	38.29%	51.88%	79.07%	79.21%
Centroid features (F <sub>9</sub> )	57.97%	54.06%	81.70%	81.75%

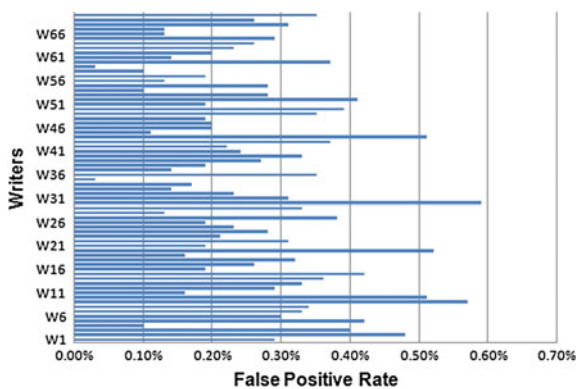
## 6 Conclusion and Future Scope

In this paper, we have presented a study of different features and classification techniques for text-dependent writer identification system. Maximum identification accuracy of 81.75% has been accomplished with centroid features and AdaBoostM1 ensemble classifier. This accuracy may be improved either by increasing the span of training dataset or by using the various optimal feature selection techniques like PCA, Correlation Feature Set (CFS), etc. This work can be employed to other scripts like Devanagari, Bengali, and Tamil and so forth which are similar to the Gurmukhi script after building the training dataset of these scripts.

**Fig. 2** True Positive Rate with Centroid Features and AdaBoostM1 Classifier



**Fig. 3** False Positive Rate with Centroid Features and AdaBoostM1 Classifier



## 7 Declaration

In this manuscript, we have used 49,000 samples of Gurmukhi characters collected from 70 different writers. These all individuals who participated in this work have given their consent for publish this dataset in this manuscript.

## References

1. Breiman L 2001 Random Forests, *Machine Learning*, 45(1):5–32.
2. Gazzah S and Amara N B 2008 Neural networks and support vector machines classifiers for writer identification using Arabic script, *The International Arab Journal of Information Technology*, 5(1): 92–101.
3. Ghiasi G and Safabakhsh R 2010 An efficient method for offline text independent writer identification, *In Proceedings of the 20th International Conference on Pattern Recognition*, 1245–1248.

4. John G H and Langley P 1995 Estimating Continuous Distributions in Bayesian Classifiers, *In Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*, 338–345.
5. Kumar M, Sharma R K and Jindal M K 2013 A Novel Feature Extraction Technique for Offline Handwritten Gurmukhi Character Recognition, *IETE Journal of Research*, 59(6): 687–692.
6. Kumar M, Jindal M K and Sharma R K 2014a A Novel Hierarchical Techniques for Offline Handwritten Gurmukhi Character Recognition, *National Academy Science Letters*, 37(6): 567–572.
7. Kumar M, Sharma R K and Jindal M K 2014b Efficient Feature Extraction Techniques for Offline Handwritten Gurmukhi Character Recognition, *National Academy Science Letters*, 37(4):381–391.
8. Leclerc F, Plamondon R 1994 Automatic signature verification: the state of the art 1989–1993, *International Journal of Pattern Recognition and Artificial Intelligence*, 8(3):643–660.
9. Leeham G, Chachra S 2003 Writer identification using innovative binarised features of handwriting numerals, *In the Proceedings of the 7<sup>th</sup> International Conference on Document Analysis and Recognition (ICDAR)*.
10. Maadeed S A 2012 Text-dependent writer identification for Arabic Handwriting, *Journal of Electrical and Computer Engineering*, 13: 1–8.
11. Plamondon R, Lorette G 1989 Automatic Signature Verification and Writer Identification The State of the Art, *Pattern Recognition*, 22(2):107–131.
12. Schlapbach A, Bunke H 2005 Writer identification using an HMM based hand writing recognition system: to normalize the input or not?, *In the Proceedings of 12th Conference of the International Graphonomics Society*, Salerno, Italy.
13. Schlapbach A, Bunke H 2007 A writer identification and verification system using HMM based recognizers, *Pattern Analysis and Application*, 10(1):33–43.
14. Zois E, Anastassopoulos V 2000 Morphological Waveform Coding for Writer Identification, *Pattern Recognition*, 33(3):385–398.



# Modified CSM for FIR Filter Realization

N. Udaya Kumar, K. Durga Teja, K. Bala Sindhuri and P. Rakesh

**Abstract** Digital filters are used in communication and digital signal processing applications and they are classified into two types: Finite Impulse Response (FIR) filter and Infinite Impulse response (IIR) filter. For high-speed- and area-efficient applications, FIR filter realization with less delay and less area including reconfigurability is needed. In this paper, modified constant shift method (MCSM) architecture is proposed for the realization of FIR filter with reconfigurability using Verilog HDL. This architecture is implemented in Xilinx Virtex-4 FPGA device (xc4v11x200-11ff1513).

**Keywords** Modified constant shift method (MCSM) · Constant shift method (CSM) · Reconfigurable FIR filter · Shifter unit · Adder unit

## 1 Introduction

Reconfigurability in FIR filter means changing of filter coefficients during runtime [1] and such filter is called reconfigurable FIR filter. It plays an important role in software-defined radio (SDR) systems [2]. Multiple Constant Multiplication (MCM) technique [3, 4] cannot be used for the realization of such type of filters when the filter coefficients are dynamically changes. Constant Shift Method (CSM) and Programmable Shift Method (PSM) architectures [5] can be used for the realization of reconfigurable FIR filters without any modification in the hardware.

---

N. Udaya Kumar (✉) · K. Durga Teja · K. Bala Sindhuri · P. Rakesh  
S.R.K.R. Engineering College, Bhimavaram, India  
e-mail: n\_uk2010@yahoo.com

K. Durga Teja  
e-mail: kumilidurgateja@gmail.com

K. Bala Sindhuri  
e-mail: k.b.sindhuri@gmail.com

P. Rakesh  
e-mail: rakeshpulletikurthi.rp@gmail.com

For the realization of reconfigurable FIR filters, the major hardware depends on coefficients multiplication with input sequence. To perform this multiplication, the processing elements (PEs) of CSM and PSM architectures are used. As a result, the area of reconfigurable FIR filter using PSM architecture is less compared to CSM architecture whereas delay is increased [5]. For high-speed- and area-efficient applications, the delay and area should be less.

In this paper, the modified constant shift method (MCSM) architecture for the implementation of reconfigurable FIR filter is proposed in order to reduce the delay and area compared to CSM architecture.

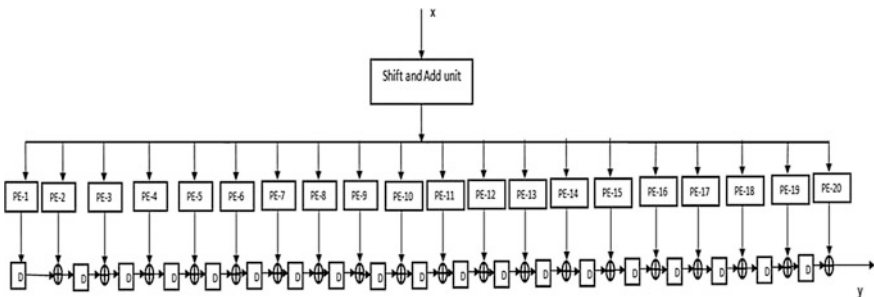
The organization of this paper is as follows: Section 2 describes reconfigurable FIR filter. Section 3 explains BCSE algorithm. Section 4 describes CSM architecture. Section 5 explains MCSM architecture. Section 6 shows the simulation results of reconfigurable FIR filter using CSM and MCSM architectures. Section 7 shows design summary and synthesis report of reconfigurable FIR filter using CSM and MCSM architectures. Finally this work is concluded in Sect. 8.

## 2 20 Tap Reconfigurable FIR Filter

The block diagram of 20 tap reconfigurable FIR filter in transposed direct form is shown in Fig. 1. It consists of processing elements (PEs) that implements coefficient multiplication with the input sequence ( $x$ ) and then delayed version of the output of each PE is added to obtain filter output ( $y_{fir}$ ) as shown in Fig. 1. The single PE and single adder is called a Tap. For an  $N$ -tap filter, there is  $N$  number of PEs. The block diagram of processing element (PE) for both CSM and MCSM is shown in Fig. 2.

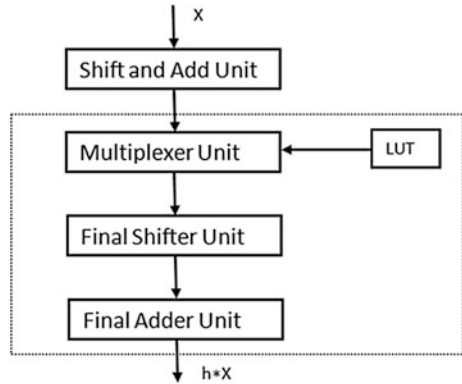
The functions of different blocks of PE are as follows:

- (1) Shift and Add unit: The architecture of shift and add unit of PE for both CSM and MCSM is shown in Fig. 3. This unit generates the  $m$ -bit binary common subexpressions (BCSs) based on coefficient partitioning which are used to



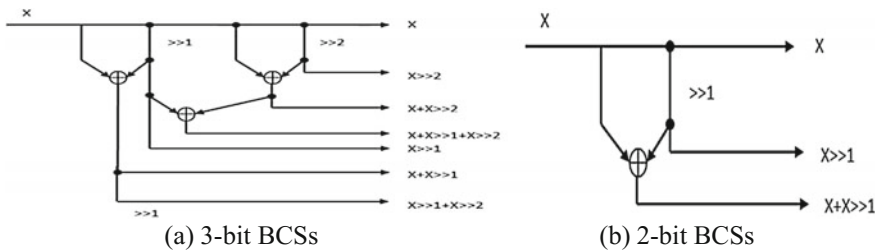
**Fig. 1** Block diagram of 20 tap reconfigurable FIR filter in transposed direct form

**Fig. 2** Block diagram of processing element (PE) of CSM and MCSM



reduce the complexity of multiplication by eliminating the common subexpressions (CSs) [6] and also by reducing number of adders using BCSE algorithm explained in Sect. 3. In CSM architecture, shift and add unit generates both 2-bit BCSs and 3-bit BCSs whereas in MCSM architecture generates only 3-bit BCSs. In Fig. 3 ( $\gg 1$ ) and ( $\gg 2$ ) represents constant shift values of  $2^{-1}$  and  $2^{-2}$  respectively.

- (2) Multiplexer unit: This unit consists of multiplexers that selects BCSs from shift and add unit based on input coefficient to PE. In CSM architecture for  $n$ -bit word length coefficient, there are  $\lceil n/3 \rceil$  multiplexers whereas in MCSM architecture the multiplexer unit is divided into  $\lceil n/16 \rceil$  multiplexer groups of 6 multiplexers (in which five are 8:1 mux and one is 2:1 mux). E.g. for 32 bit coefficient, there are total  $\lceil 32/3 \rceil = 11$  multiplexers in CSM whereas in MCSM there are  $\lceil 32/16 \rceil = 2$  multiplexer groups. The output of multiplexer unit is called intermediate additions.
- (3) Final Shifter unit: This unit performs the constant shift operations to the obtained, intermediate additions from multiplexer unit and the output of final shifter unit is called final intermediate additions. This can be explained using output expression of PE



**Fig. 3** Architecture of shift and add unit for CSM and MCSM

$$y1 = 2^{-4}x + 2^{-6}x + 2^{-15}x + 2^{-16}x \quad (1)$$

By coefficient partitioning [7], we obtain

$$y1 = 2^{-4}(x + 2^{-2}x) + 2^{-15}(x + 2^{-1}x), \quad (2)$$

where  $(x + 2^{-2}x)$  and  $(x + 2^{-1}x)$  are called intermediate additions obtained from multiplexer unit and the constant shift values  $2^{-4}$ ,  $2^{-15}$  in (2) are performed by final shifter unit to obtain final intermediate additions such as  $2^{-4}(x + 2^{-2}x)$  and  $2^{-15}(x + 2^{-1}x)$ . Both CSM and MCSM architectures have same final shifter unit except  $2^{-16}$  constant shift value present in MCSM.

- 4) Final adder unit: This unit computes the sum of all final intermediate additions  $2^{-4}(x + 2^{-2}x)$  and  $2^{-15}(x + 2^{-1}x)$  as in (2).

### 3 Binary Common Subexpression Elimination Algorithm (BCSE)

The multiplication operation can be implemented by shift and add operations. BCSE algorithm [8] is more efficient in reducing the number of adders needed to realize the multipliers that are using shift and add operations. The BCSE algorithm also deals with elimination of redundant binary common subexpressions (BCSs) that occur within the coefficients.

In this algorithm, the 2-bit binary representation can form one BCS of  $[1 \ 1]$  and it can be expressed as

$$[1 \ 1] = x + 2^{-1}x, \quad (3)$$

where  $x$  is input signal. In remaining 2-bit BCSs such as  $[0 \ 1]$   $[1 \ 0]$  there are only one nonzero bit so they do not require any adder. Similarly, the 3-bit binary representation can form four BCSs, which are  $[0 \ 1 \ 1]$ ,  $[1 \ 0 \ 1]$ ,  $[1 \ 1 \ 0]$  and  $[1 \ 1 \ 1]$ . These BCSs can be expressed as

$$[0 \ 1 \ 1] = X1 = 2^{-1}x + 2^{-2}x \quad (4)$$

$$[1 \ 0 \ 1] = X2 = x + 2^{-2}x \quad (5)$$

$$[1 \ 1 \ 0] = X3 = x + 2^{-1}x \quad (6)$$

$$[1 \ 1 \ 1] = X4 = x + 2^{-1}x + 2^{-2}x, \quad (7)$$

where  $x$  is input signal. In remaining 3-bit BCSs such as  $[0\ 0\ 1]$ ,  $[0\ 1\ 0]$  and  $[1\ 0\ 0]$  there are only one nonzero bit, so they do not require any adder. A straightforward realization of above BCSs would require five adders. However  $X1$  can be obtained from  $X3$  by a right shift operation without using any extra adders as

$$X1 = 2^{-1}x + 2^{-2}x = 2^{-1}(x + 2^{-1}x) = 2^{-1}X3 \quad (8)$$

Also,  $X4$  can be obtained from  $X3$  by using an addition operation to  $X3$  as

$$X4 = x + 2^{-1}x + 2^{-2}x = X3 + 2^{-2}x \quad (9)$$

However, only three adders are needed to realize the BCSs  $X1$  to  $X4$ . Thus the adders and redundant binary common subexpressions (BCSs) are reduced using BCSE.

## 4 CSM Architecture

The block diagram for processing element of CSM architecture is shown in Fig. 4. In CSM architecture [9], the sign-bit (that tells whether coefficient is positive or negative valued) and  $n$ -bit coefficient are stored in the Look up Table (LUT) and then this  $n$ -bit coefficient is partitioned into groups of 3-bits from the most significant bit (MSB) where the number of 3-bit groups is equal to  $\lceil n/3 \rceil$  and these 3-bit groups are given as selection inputs for the multiplexers. These multiplexers select the BCSs obtained from shift and add unit based on selection input from LUT and the obtained outputs from multiplexers are called intermediate additions. The intermediate additions are shifted by constant shift values of final shifter to obtain an output called final intermediate additions, and these are finally added using final adder unit and the resultant is shifted by constant shift value of  $2^{-1}$  to obtain output of PE. This output of PE and its complemented form are given as input to a 2:1 multiplexer with selection input of sign-bit stored in LUT to obtain final output. If sign-bit = 0 (means the coefficient is positive valued), then take output of PE as final output, otherwise take its complemented form.

In this work, the CSM architecture for 32-bit word length coefficient is explained, i.e., for  $h = "0.11111111111111111111111111111111"$ . Here sign-bit = 0, since coefficient ( $h$ ) is a positive value stored in LUT. By partitioning this coefficient ( $h$ ) from MSB, ten 3-bit groups and one 2-bit group are obtained which acts as selection inputs for Mux1 to Mux10 (8:1 mux) and Mux11 (4:1 mux) respectively as shown in Fig. 4. The output of PE of CSM i.e.  $y1 = h * x$  can be implemented using shift and add operations as

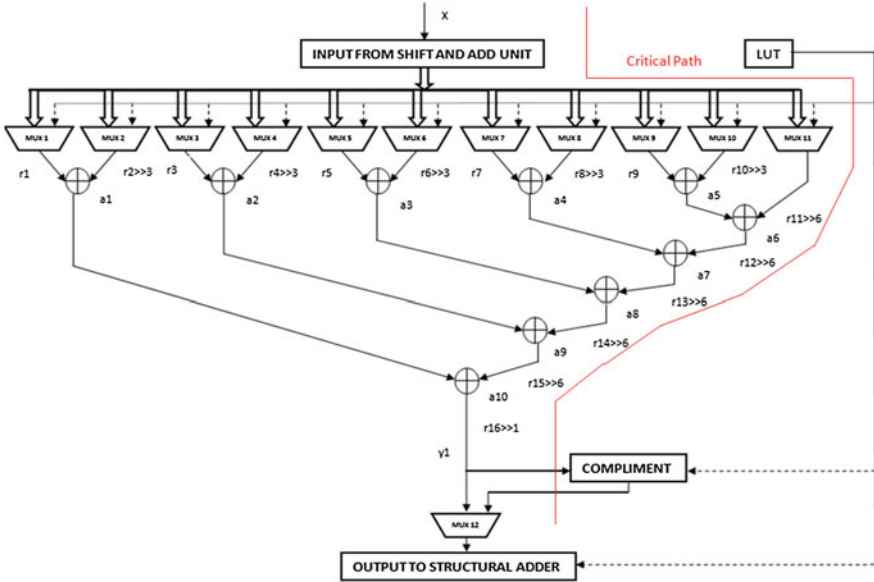


Fig. 4 Architecture of processing element (PE) of CSM

$$\begin{aligned}
 y1 = & 2^{-1}x + 2^{-2}x + 2^{-3}x + 2^{-4}x + 2^{-5}x + 2^{-6}x + 2^{-7}x + 2^{-8}x + 2^{-9}x + 2^{-10}x \\
 & + 2^{-11}x + 2^{-12}x + 2^{-13}x + 2^{-14}x + 2^{-15}x + 2^{-16}x + 2^{-17}x + 2^{-18}x + 2^{-19}x \\
 & + 2^{-20}x + 2^{-21}x + 2^{-22}x + 2^{-23}x + 2^{-24}x + 2^{-25}x + 2^{-26}x + 2^{-27}x \\
 & + 2^{-28}x + 2^{-29}x + 2^{-30}x + 2^{-31}x + 2^{-32}x
 \end{aligned}
 \tag{10}$$

By partitioning coefficient ( $h$ ) in (10) into groups from MSB such that to form BCSSs, we obtain as

$$\begin{aligned}
 y1 = & 2^{-1}[x + 2^{-1}x + 2^{-2}x + 2^{-3}x + 2^{-4}x + 2^{-5}x + 2^{-6}x + 2^{-7}x + 2^{-8}x + 2^{-9}x \\
 & + 2^{-10}x + 2^{-11}x + 2^{-12}x + 2^{-13}x + 2^{-14}x + 2^{-15}x + 2^{-16}x + 2^{-17}x + 2^{-18}x \\
 & + 2^{-19}x + 2^{-20}x + 2^{-21}x + 2^{-22}x + 2^{-23}x + 2^{-24}x + 2^{-25}x + 2^{-26}x + 2^{-27}x \\
 & + 2^{-28}x + 2^{-29}x + 2^{-30}x + 2^{-31}x]
 \end{aligned}
 \tag{11}$$

$$\begin{aligned}
 y1 = & 2^{-1}[(x + 2^{-1}x + 2^{-2}x) + 2^{-3}(x + 2^{-1}x + 2^{-2}x) + 2^{-6}[(x + 2^{-1}x + 2^{-2}x) + 2^{-3}(x + 2^{-1}x + 2^{-2}x)) \\
 & + 2^{-6}[(x + 2^{-1}x + 2^{-2}x) + 2^{-3}(x + 2^{-1}x + 2^{-2}x)) + 2^{-6}[(x + 2^{-1}x + 2^{-2}x) + 2^{-3}(x + 2^{-1}x + 2^{-2}x)) \\
 & + 2^{-6}[(x + 2^{-1}x + 2^{-2}x) + 2^{-3}(x + 2^{-1}x + 2^{-2}x)) + 2^{-6}(x + 2^{-1}x)]
 \end{aligned}
 \tag{12}$$

Finally, from (12) the output of PE of CSM architecture can be obtained as

$$y1 = 2^{-1}[(r1 + 2^{-3}r2) + 2^{-6}[(r3 + 2^{-3}r4) + 2^{-6}[(r5 + 2^{-3}r6) + 2^{-6}[(r7 + 2^{-3}r8) + [(r9 + 2^{-3}r10) + 2^{-6}r11]]]]]]] 2^{-6} = 2^{-1}[r16] \quad (13)$$

In (13)  $r1, r2, r3, r4, r5, r6, r7, r8, r9, r10$  are 3-bit BCSs equals to  $(x + 2^{-1}x + 2^{-2}x)$  and  $r11$  is 2-bit BCS equals to  $(x + 2^{-1}x)$  that are selected by using Mux1 to Mux10 (8:1 muxes) and Mux11 (4:1 mux) respectively as shown in Fig. 4 and these are called intermediate additions which are further shifted and added by final shifter unit (such as  $\gg 1, \gg 3, \gg 6$ ) and final add unit (such as  $a1, a2, a3, a4, a5, a6, a7, a8, a9, a10$ ) respectively to obtain  $r16$  and then  $r16$  is shifted by constant shift value of  $2^{-1}$  (i.e.,  $r16 \gg 1$  as shown in Fig. 4) to obtain output of PE ( $y1$ ). This output of PE ( $y1$ ) is the final output since sign-bit = 0.

Due to 4:1 mux that selects 2-bit BCSs in PE of CSM architecture the area is increased and also delay is increased due to large critical path [10] (stages between input and output) in PE as shown in Fig. 4. Similarly, CSM architecture for higher word length coefficient such as 64-bit, 128-bit and soon the delay and area largely increased and also the complexity of coefficient partitioning is increased as in (13). All these factors lead to large delay and large area for the realization of reconfigurable FIR filter using CSM architecture. To overcome these drawbacks, the modified constant shift method (MCSM) architecture for the realization of reconfigurable FIR filter is proposed.

## 5 MCSM Architecture

In MCSM architecture, the  $n$ -bit coefficients are partitioned into groups of 16 bits from the most significant bit (MSB) and are stored in LUT and the number of 16-bit groups is equal to  $\lceil n/16 \rceil$ . The sign-bit of coefficient is also stored in LUT. In this architecture, the multiplexer unit is also divided into  $\lceil n/16 \rceil$  multiplexer groups of six multiplexers (in which five are 8:1 mux and one is 2:1 mux) with 16-bit selection input from LUT. Each 16-bit selection input is divided into five 3-bit groups and one 1-bit from MSB that are used as selection inputs for five 8:1 mux which selects 3-bit BCSs and one 2:1 mux which selects input  $x$ , respectively. The output of these six multiplexers of each multiplexer group are called intermediate additions and they are further shifted and added by final shifter unit and final adder unit to obtain final output of each multiplexer group. Now the final output of each multiplexer group (except the first multiplexer group) are shifted by constant shift value of  $2^{-16}$  and all are added together along with the final output of first multiplexer group to obtain final stage output. This final stage output is then shifted by constant shift value of  $2^{-1}$  to obtain output of PE. This output of PE and its complemented form are given as input to a 2:1 multiplexer with selection input of

sign-bit stored in LUT to obtain final output. If sign-bit = 0 (means the coefficient is positive valued), then take output of PE as final output, otherwise take its complemented form.

In this work, the MCSM architecture for 32-bit coefficient is explained, i.e., for “ $h = 11111111111111111111111111111111$ ”. Here sign-bit = 0 since coefficient ( $h$ ) is a positive value stored in LUT. This 32-bit coefficient is partitioned into two 16 bits groups as “1111111111111111” and “1111111111111111” and is stored in LUT and the multiplexer unit is divided into two multiplexer groups of six multiplexers (i.e., Mux1 to Mux6) with 16-bit selection input as shown in Fig. 5. The Mux1 to Mux6 generates intermediate additions which are further shifted and added by final shifter unit (such as  $\gg 1, \gg 3, \gg 6$ ) and final add unit (such as  $a1, a2, a3, a4, a5, a6, a7, a8, a9, a10$ ) respectively to obtain final outputs of two multiplexer groups such as  $g1$  and  $g2$  and can be expressed as

$$g1 = [(r1 + 2^{-3}r2) + 2^{-6}[(r3 + 2^{-3}r4) + 2^{-6}[(r5 + 2^{-3}r6)]]] \quad (14)$$

$$g2 = [(r7 + 2^{-3}r8) + 2^{-6}[(r9 + 2^{-3}r10) + 2^{-6}[(r11 + 2^{-3}r12)]]] \quad (15)$$

where  $r1, r2, r3, r4, r5,$  and  $r6$  are intermediate additions of first multiplexer group and  $r7, r8, r9, r10, r11,$  and  $r12$  are intermediate additions of second multiplexer group and  $g1$  and  $g2$  are final outputs of first and second multiplexer groups. In this

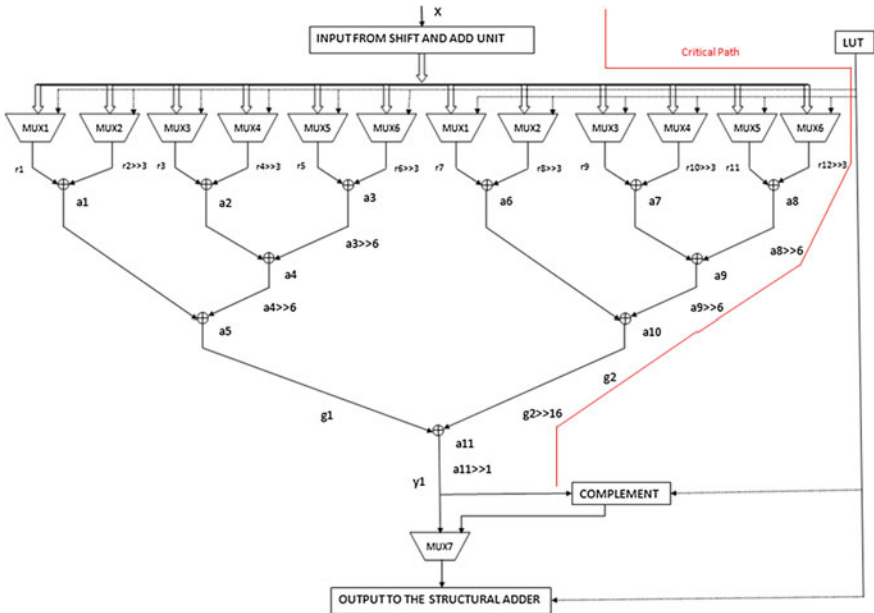


Fig. 5 Architecture of processing element (PE) of MCSM



32-bit coefficient case, the  $r_1, r_2, r_3, r_4, r_5, r_7, r_8, r_9, r_{10}, r_{11}$  are equal to  $(x + 2^{-1}x + 2^{-2}x)$  and  $r_6, r_{12}$  are equal to  $x$ . Now the final output of second multiplexer group is shifted by constant shift value of  $2^{-16}$  (i.e.,  $g_2 \gg 16$  as shown in Fig. 5.) and is added to final output of first multiplexer group to obtain final stage output as  $[g_1 + 2^{-16}g_2]$ . This final stage output is shifted by constant shift value of  $2^{-1}$  to obtain output of PE. Thus output of PE of MCSM architecture can be obtained as

$$y_1 = 2^{-1} [g_1 + 2^{-16}g_2] \tag{16}$$

Thus the complexity of coefficient partitioning is decreased as in (16). Due to 2:1 mux in PE of MCSM instead of 4:1 mux in PE of CSM architecture, the area is decreased and also due to smaller critical path [10] (stages between input and output) in PE of MCSM compared to PE the delay is reduced. Similarly, MCSM can yield less complexity in coefficient partitioning and less delay and less area compared to CSM for higher word length coefficients. Thus the reconfigurable FIR filter can be implemented using modified constant shift method (MCSM) architecture with less delay and less area and less complexity.

### 6 Simulation Results

The 20 tap reconfigurable FIR filter shown in Fig. 1 is implemented using PEs of CSM and MCSM architectures in Xilinx 12.2 and the resulted simulation figures are shown in Figs. 6 and 7 respectively in which “yfir” represents the filter output and “xin” represents the input sequence and  $h_0, h_1, h_2, h_3, h_4, h_5, h_6, h_7, h_8, h_9, h_{10}, h_{11}, h_{12}, h_{13}, h_{14}, h_{15}, h_{16}, h_{17}, h_{18}, h_{19}$  represents 20 32-bit coefficients for 20 tap reconfigurable FIR filter.

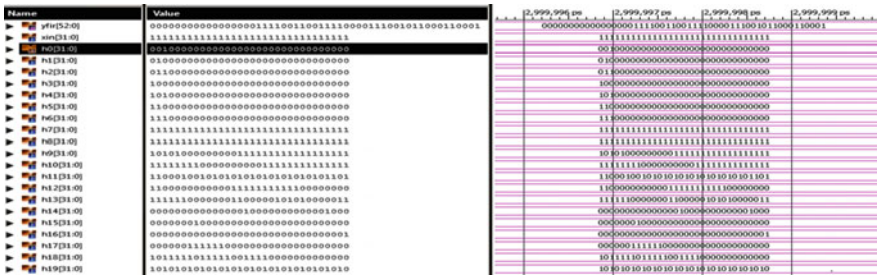


Fig. 6 Simulation result of 20 tap reconfigurable FIR filter using CSM architecture

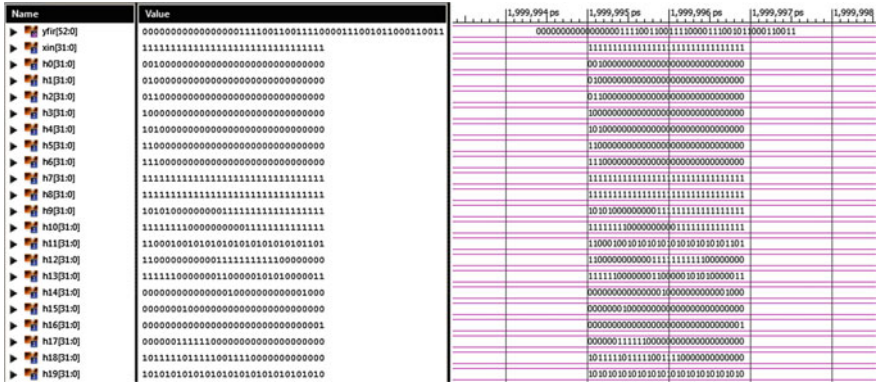


Fig. 7 Simulation result of 20 tap reconfigurable FIR filter using MCSM architecture

(a) Device Utilization Summary				(b) Device Utilization Summary			
Logic Utilization	Used	Available	Utilization	Logic Utilization	Used	Available	Utilization
Number of 4-input LUTs	57,818	178,176	32%	Number of 4-input LUTs	56,778	178,176	31%
Number of occupied Slices	31,501	89,088	35%	Number of occupied Slices	30,940	89,088	34%
Number of Slices containing only related logic	31,501	31,501	100%	Number of Slices containing only related logic	30,940	30,940	100%
Number of Slices containing unrelated logic	0	31,501	0%	Number of Slices containing unrelated logic	0	30,940	0%
Total Number of 4-input LUTs	59,601	178,176	33%	Total Number of 4-input LUTs	58,881	178,176	33%
Number used as logic	57,818			Number used as logic	56,778		
Number used as a route-thru	1,783			Number used as a route-thru	2,103		
Number of bonded I/Os	745	960	77%	Number of bonded I/Os	745	960	77%
Average Fanout of Non-Clock Nets	2.86			Average Fanout of Non-Clock Nets	2.80		

Fig. 8 Design summary of 20 tap reconfigurable FIR filter using a CSM b MCSM

## 7 Design Summary and Synthesis Report

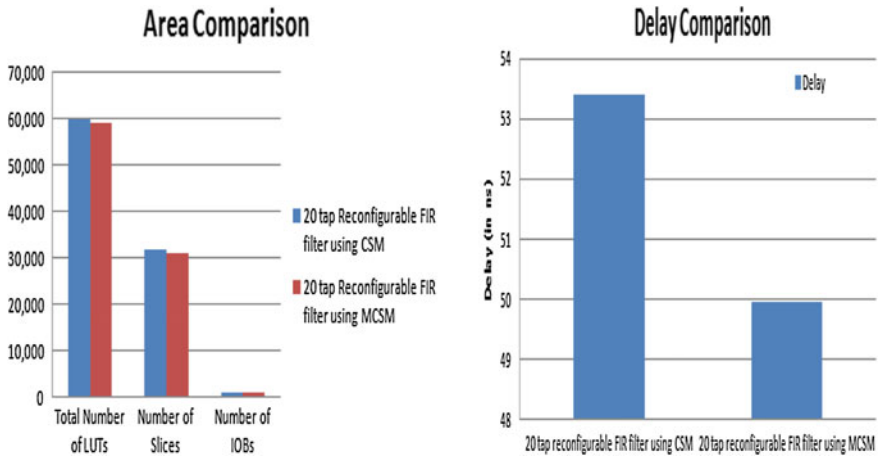
The Design Summary for 20 tap reconfigurable FIR filter using CSM and MCSM architectures are shown in Fig. 8.

From the design summary, it is observed that total number of LUTs is decreased by 1.21% and number of slices is decreased by 1.78% and the average fan-out is decreased by 2.10% for 20 tap reconfigurable FIR filter using MCSM compared to CSM. The synthesis report of 20 tap reconfigurable FIR filter using CSM and MCSM architectures is tabulated in Table 1.

From Table 1, it is observed that the delay is decreased by 6.50% for 20 tap reconfigurable FIR filter using MCSM compared to CSM. The graphical representation of area and delay comparisons of reconfigurable FIR filter implementation using CSM and MCSM architectures are shown in Fig. 9.

**Table 1** Synthesis report

Delay (in ns)	20 tap reconfigurable FIR filter using CSM	20 tap reconfigurable FIR filter using MCSM
	53.413	49.947



**Fig. 9** Area and delay comparisons of 20 tap reconfigurable FIR filter using CSM and MCSM

## 8 Conclusion

The modified constant shift method (MCSM) architecture is proposed for implementing reconfigurable FIR higher order filters with low complexity. The MCSM architecture yields less area, less delay, and low complexity compared to CSM architecture and can be used in high-speed- and area-efficient applications. This work is implemented in virtex-4 xc4vllx200-11ff1513 FPGA for 32-bit word length coefficients.

## References

1. K.H. Chen and T.D. Chiueh, “A low-power digit-based reconfigurable FIR filter”, *IEEE Trans. Circuits sus. II*, vol. 53, no. 8, pp. 617–621, Aug. 2006.
2. T. Hentschel, M. Henker and G. Fettweis, “The digital front-end of software radio terminals”, *IEEE Personal communication Mag.* Vol. 6, no. 4, pp. 40–46, Aug. 1999.
3. L. Ming and Y. Chao, “The multiplexed structure of multi-channel FIR filter and its resources evaluation”, *International conference on computer distributed control and intelligent environmental monitoring (CDCIEM)*, March 2012.
4. Pramod Kumar Meher and Yu Pan “MCM based implementation of block FIR filters for high speed and low power application”, *19th International conference on VLSI and System-on-chip*, IEEE, 2011.

5. P. Mahesh and A.P. Vinod, "New Reconfigurable architectures for implementing FIR filters with low complexity" IEEE transactions on computer-aided design of integrated circuits and systems, vol. 29, no. 2, February 2010.
6. Richard I Hartley "Subexpression sharing in filters using Canonical signed digit multiplier" IEEE transactions on circuits and systems-II Analog and Digital Signal Processing, vol. 43, no. 10, oct. 1996.
7. A.P. Vinod and E.M.K. Lai, "An efficient coefficient-partitioning algorithm for realizing low complexity digital filters", IEEE transactions Computer aided design Integr. Circuits sys., vol. 24, no. 12, pp. 1936–1946, Dec. 2005.
8. R. Mahesh and A.P. Vinod, "A New common subexpression elimination algorithm for realizing low complexity higher order digital filters" IEEE transactions on computer-aided design of integrated circuits and systems, vol. 27, no. 2, February 2008.
9. V. Sandhiya, S. Kathick, M. Valarmathy, "A survey on New Reconfigurable architectures for implementing FIR filters with low complexity" International conference on computer communications and Informatics (ICCCI), Jan. 3–5, 2014.
10. Xin Lou, Yan Jun YU, "Lower Bound Analysis and Perturbation of Critical Path for Area-Time Efficient Multiple Constant Multiplications" IEEE transactions on computer aided design of integrated circuits and systems, vol. 36, Feb. 2017.

# Toward Design and Enhancement of Emotion Recognition System Through Speech Signals of Autism Spectrum Disorder Children for Tamil Language Using Multi-Support Vector Machine

C. Sunitha Ram and R. Ponnusamy

**Abstract** This paper presented a methodology of an enhanced system for recognizing and classifying emotion from the speech signals of Autism Spectrum Disorder children for Tamil Language using Multi-Support Vector Machine. In this effort, a real database is recorded from the ASD children speech from a special school. The recorded database is categorized and named as primary emotion: Anger, Neutral, Happiness, Sadness, and Fear. It is speaker-independent. Likewise, speaker-dependent Tamil emotional database (Tamil\_DB) and Telugu emotional database (Telugu\_DB) are collected from movies and categorized as primary emotions. A standard Berlin Corpus (EMO-DB) is used as a training dataset. With these datasets, good and suitable features are extracted from the samples by Mel Frequency Cepstral coefficients and global parameters. Further, these databases are trained by these features and classified primary emotions using multi-support vector machine to measure the effectiveness accurately. Results indicate that speech corpus exhibits the values are closer.

**Keywords** Mel frequency cepstral coefficient • Global parameters  
Multi-support vector machine • Autism spectrum disorder • Databases  
Computational intelligence

---

C. Sunitha Ram (✉)  
Department of CSE, SCSVMV University, Kanchipuram, India  
e-mail: sunithabasha@gmail.com

R. Ponnusamy  
Department of CSE, SriLakshmiAmmal College of Engineering, Chennai, India  
e-mail: r\_ponnusamy@hotmail.com

## 1 Introduction

Human–Computer Interaction (HCI) involves the knowledge, improvement, and design of the system between users and computers. Human–Computer Interaction is updated and enhanced by different technology in real world. HCI is the study of bidirectional communication between human and computer technologies. HCI is a regulation that attracts reformation and creativeness. HCI is related with multi-sided computer technologies.

A current study of HCI interface techniques are virtual, augmented, and mixed reality and machine vision learning with humans that can be later extent to emotions.

To design HCI user needs physical, cognitive, and affective which determines the mechanics of interaction, which determines in which way human can understand the system and interrelate with it, which deals with human emotions respectively. Based on enhancement in technology, HCI system architecture is alienated by two different modalities as unimodal and multimodal HCI.

Still, research is going to incorporate human emotions with HCI technology by analyzing, recognizing, and classifying emotion through audio signals. Touch GUI, Web User Interface, Visual systems, Virtual and augmented reality, specialized interfaces like robotic systems, Wearable technology and Mobile Apps speech emotion recognition are in the mounting stage. Emotion plays important role between human beings and computers. Researchers are still developing technology to intermingle emotion in multimodal HCI. Human can express their emotions by verbal and nonverbal communication. Speech is the immediate communication to utter emotion through interactive devices. Affective computing area is used to focusing the design and growth of devices which synthesize human emotions [1].

Human speech has promising quality for realize emotions which is relevant to design Automatic Speech Emotion Recognition. Human–robots interactions have been implemented via many advance technology in the world such as playing toys, speech therapists, and e-learning for special children. In real life, machine alone understands human emotion by speech and to provide appropriate response is in development. Autism spectrum disorder has problem in brain growth by various difficulties in social interaction, redundant action, and communication problem either in verbal or nonverbal for children. They show different types of expression than normal children. Only parents and therapists can recognize their emotion. Without their attention normal people is hard to recognize their emotion. Our research is paying attention to design, develop, and automate Speech Emotion Recognition system to recognize primary emotion from ASD children speech.

To achieve accuracy above-mentioned points are demonstrated by Computational Intelligence (CI). It is a computational methodology to solve complex real-world problems. Features of computational intelligence are clustering, categorization, and abilities to improve the systems. A complicated problem in HCI such as speech emotion recognition is solved by different methodologies and techniques are implemented via CI.

First, we discuss about utmost work completed in the area of emotional speech recognition in different languages in different sources all over the world. Later, we discuss about ASD children emotion in multimodal systems. The growing responsiveness on autism has recent research for speech.

To extract good features researchers tried to identify and implement by many methods such as Discrete Wavelet Transform (DWT), Spectral features like Mel Frequency Cepstral Coefficients (MFCC), Linear Prediction Coefficients (LPC), global features like Fundamental frequency (F0), Mean, Median, Standard deviation, Minimum, Maximum and Range, Voice quality features like amplitude, energy, pitch and duration, nonlinear-based features like Teager energy Operator (TEO). Various classification methods have been implemented for the completion of speech emotion recognition such as Gaussian Mixture Model(GMM), Multi Layer Perceptron (MLP), Hidden Markov model (HMM), k-nearest Neighbors (KNN), Granular SVM (GSVM), Support Vector Machines (SVM), Radial Basis Function (RBF), Back Propagation Neural Network (BPNN), and Multi-Support Vector Machines (MSVM).

- Case studies of speech emotion recognition are
  - HCI tutoring system
  - Lie detector used in CBI
  - Car board system to recognize and response of driver's emotion
  - Detect and giving priority to customer angry state in call center
  - Diagnostic tool by speech therapists
  - Recognize partner emotional state when they are in long distance
- Real-time application of speech emotion recognition
  - A wearable EMG AUBADE is a platform used in healthcare applications
  - Kismet recognize emotion based on speech and visual
  - ERMIS recognize emotion by MPEG 4 standard input framework

In this work, Tamil, Telugu, and ASD children databases are collected to train and classify primary emotion and the results are tested with standard Berlin database. Tamil and Telugu databases are called as speaker-independent collected from movies, according to the speech utterances emotions are segregated and labeled as primary emotion. ASD children database called as speaker-dependent recorded from special school and labeled as primary emotion. Next standard Berlin database is freely accessible downloaded from the Internet. These databases are initially extracted good features, trained, and classify primary emotion from speech to build SER system. As mentioned above, researchers use many spectral features in speech emotion recognition which produces better result in feature extraction. In this work, MFCC and global parameters have been used to extract good features from speech signal and remove bad features like noise, space, etc. MFCC and global parameters are the finest approach to examine signal from speech.

After extracting good features, multi-SVM is used for training, testing, and classifying the primary emotion of all databases. A free toolkit for speech emotion

recognition presents various mechanism and components which is usable to recognize emotion from the speech signals. Most of the toolkits like MATLAB, WEKA are targeted to implement feature extraction methods for speech recognition, speaker recognition, speech emotion recognition, and speaker emotion recognition processing. This paper is ordered as follows: Segment 2 summaries literature review of ASD and speech emotion recognition other languages, Segment 3 illustrates the design of the set out work, Segment 4 exhibits the results and discussions and Segment 5 shows the conclusion and upcoming directions.

## 2 Literature Review

From the period of the study, initially utmost effort was finished in the region of the world in the part of emotional recognition from speech in different languages. Afterward research about ASD children emotion recognition in different sources was investigated. In literature review the remarks observed of ASD children emotional speech for different languages around the world has inspired (Table 1).

## 3 Methodology of ASER

The emotional speech recognition system has three modules in the proposed system: preprocessing, feature extraction, and classification. Figure 1 shows proposed automatic speech emotion recognition system. All the modules are described in detail in the subsequent sections.

### 3.1 Preprocessing Steps

Extracting the good features from the speech signals is fundamental process in recognition of primary emotion is called preprocessing steps. Exclusion of noise and unspoken seconds or minutes are the important parameters of preprocessing steps. Initially it is constructing by speech databases. It is divided into two types namely speaker dependent (Natural or Elicited) and speaker-independent (simulated or induced). Speaker-dependent is recorded from actor or real speakers. Speaker independent is collected from cinemas (.mp3 format). Table 2 gives the details description of training and recording information of ASDDB, Tamil-DB, Telugu-DB and Berlin (EMO\_DB) databases. Figure 2 shows training speech signals are preprocessed into any file format and latter it is converted into.wav format.

Initially, to design and develop the automatic speech emotion recognition system four databases are used.



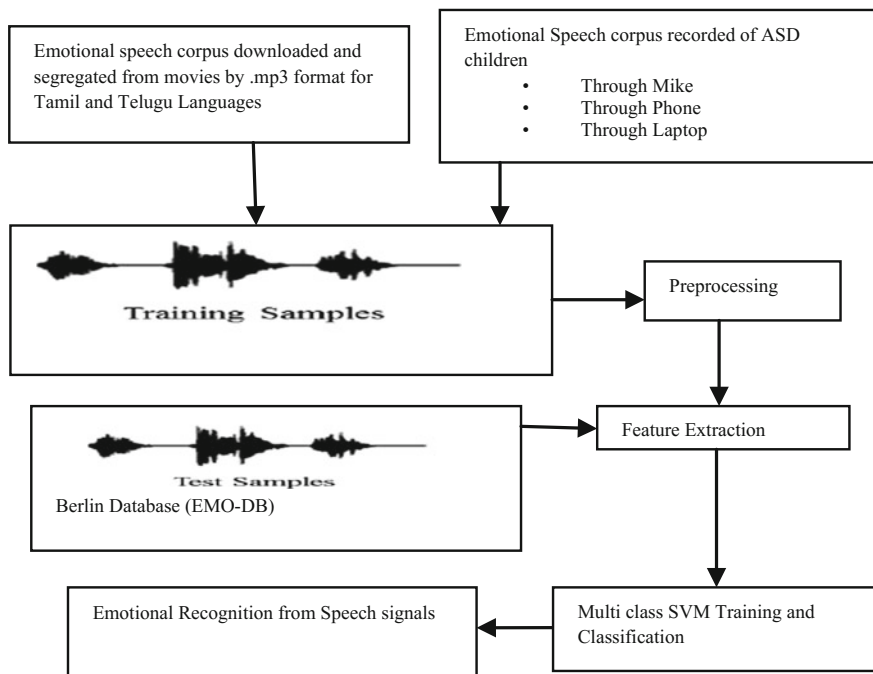
**Table 1** Outline on related work on other different languages and ASD emotional speech

Authors	Speech parameters	Primary Emotions	Participants	Feature extraction & Classification Methods
Rajoo et al. [2]	Gestures, Facial expression, Speech signals and body movements	Happy, Sad, Neutral and Anger	10 Drama professional speakers (5 Male, 5 Female) with 120 sentences (Malayalam)	MFCC & kNN
			10 native Mandarin speakers(5 Female and 5 Male) with 80 sentences (Mandarin)	
Poorna et al. [3]	Speech signals	Happy, Anger, Sad	8 native Chinese speakers(4 Female and 4 Male) with 40 short sentences	Energy, pitch contour, quefrequency coefficient & Hybrid rule based K-means and Multiclass SVM
Ladde et al. [4]	Speech signals	Anger, Happy, Sad and Neutral	10 samples	HMM & SVM
Palo et al. [5]	Speech signals	Angry, Happy, Sad and Surprise	Native children database	Radial Basis Function Network
Zhang et al. [6]	Speech signals	Angry, Fear, Happy, Neutral, Sad and Surprise	Chinese database	Penalty factor, Kernel function & Multiclass SVM
Meddeb et al. [7]	Speech signals	Neutral, Sad, Fear, Anger and Happy	REGIM_TES-Arabic database	Pitch, Energy, MFCC, Formant, LPC, Spectrogram & RBF kernel Multiclass SVM
Sinith, M. S et al. [8]	Speech signals	Happy, Sad, Anger and Neutral	EMO-DB and SAVEE database (Malayalam)	Pitch, energy and MFCC & SVM
Mathew D Lerner et al. [9]	Face, speech and EEG	Happy, Sad, Fear and Anger	40 English speakers (14 Female and 26 Male) with 40 short sentences	Event related potential
Catherine et al. [10]	Face and speech(verb al and nonverbal)	Happy, sad, Fear, Surprise, Disgust and Anger	99 English participants(53 ASD and 46 others)	Structural Equation modeling
I-Fan Lin I et al. [11]	Speech signals	Gender classification	14 Japanese Peoples (20–47 years, 3 females)	Two-way mixed design ANOVA

(continued)

**Table 1** (continued)

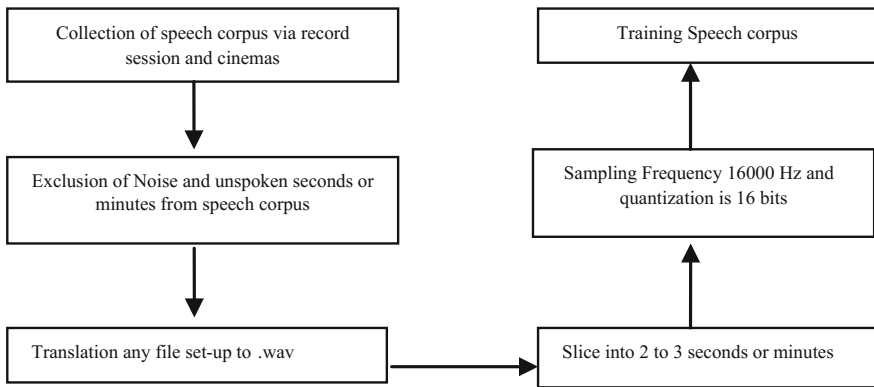
Authors	Speech parameters	Primary Emotions	Participants	Feature extraction & Classification Methods
Erik Marchi et al. [12]	Face, voice and body gestures	Happy, Anger, Sad, Surprise, Afraid and Neutral	10 Hebrew ASD children(5 Male and 5 Female)	Energy, pitch, duration and MFCC & SVM
Jean Xavier et al. [13]	Face and voice	Joy, Sad, Fear, Disgust, Neutral and Anger	19 Germany ASD children(14 boys and 5 girls)	GLMM
Laurence Chaby et al. [14]	Face, speech and postures	Happy, Sad, Neutral and Anger	6–12 years of ASD children(French)	k-NN and Dynamic HMM



**Fig. 1** Proposed automatic speech emotion recognition system

**Table 2** Details of training prerequisite method illustration

SI. No	Process description	ASD Database	Tamil Database	Telugu Database	Berlin Database
1.	Speaker	14 Boys, 15 Girls	7 Female, 8 Male	6 Female, 8 Male	10 Female, 10 Male
2.	Speech corpus	35 phrases of boys and 26 phrases of girls	15 phrases of male and 13 phrases of female	15 phrases of male and 13 phrases of female	500 phrases of both male and female
3.	Age group	7–12 years	30–50 years	40–50 years	21–35 years
4.	Recording Environment	Special School, Class Room	Laboratory	Laboratory	Laboratory
5.	Sampling Frequency, fs	16,000 Hz	16,000 Hz	16,000 Hz	16,000 Hz
6.	File format	.wav(laptop), .ogg(mobile)	Convert.mp4 to .wav	Convert.mp4 to .wav	.wav
7.	Duration	2 to 3 min	2 to 3 min	2 to 3 min	2 to 3 min
8.	Quantization	16 bits	16 bits	16 bits	16 bits
9.	Types of database	Speaker dependent	Speaker independent	Speaker independent	Speaker independent
10.	Samples Type	Training	Training	Training	Testing



**Fig. 2** Preprocessing the speech corpus

- Speaker-independent database
  - Tamil speech emotion corpus (TAMIL\_DB)
  - Telugu speech emotion corpus (TELUGU\_DB)
  - Berlin speech emotion corpus (EMO\_DB)

- Speaker-dependent database
  - Autism children speech emotion corpus(ASD\_DB)

### 3.2 Feature Extraction

Extracting and selecting the good features should be efficient to improve the accurate values to separate primary emotions is the initial step of ASER. The parameters of feature extraction are MFCC, Pitch F0, SD\_Mean, Maximum and Minimum to characterize the signals. In this work, it will assess the primary emotion like happiness, sadness, anger, fear and neutral from the speech based on 39 double-delta MFCC and global parameters. Steps involved in the extracting the features are

Step 1: MFCC feature extraction steps are implemented by the following steps and the results are stored in.xls format.

$$\text{Preemphasis} - Y(n) = X(n) - a * X(n - 1) \quad (1)$$

$$\text{Frame} - \text{floor}((1 - N) / M) + 1 \quad (2)$$

$$\text{Hamming windowing} - W(n, \alpha) = (1 - \alpha) - \alpha \cos(2\pi n / (N - 1)) \quad (3)$$

$$\text{Fast fourier transform} - Y(w) = FFT[h(t) * X(t)] = (w) * X(w) \quad (4)$$

$$\text{Triangular Bandpass Filters} - mel(f) = 2595 * \ln(1 + f / 700) \quad (5)$$

- Logarithm—Taking log of bandpass filter
- Discrete Cosine Transform- Convert log into cos transformation
- 39 Double-Delta MFCC

The steps involved in 39 double-delta MFCC are

- 12 MFCC

$$1 \text{ energy feature - Energy} = \sum x^2(t), \quad t \text{ varies from } t1 \text{ to } t2 \quad (6)$$

$$12 \text{ delta MFCC features} - \Delta d(t) = (c(t+1) - c(t-1))/2 \quad (7)$$

- 12 double-delta MFCC features
- 1 delta energy feature
- 1 double-delta energy feature

The cepstral features extraction are calculated by 12 delta MFCC features ( $\Delta$ ) i.e., vector and acceleration is calculated by 12 double-delta MFCC features ( $\Delta\Delta$ ) to obtain 39 dimensional MFCC feature extraction for automatic emotional speech recognition system.

Step 2: Training samples like Tamil-DB, Telugu-DB and ASD-DB and testing samples like EMO-DB databases are used for implementation.

Step 3: The framework values are stored in matrix form. Using global features like pitch frequency F0, maximum and minimum and standard deviation is calculated from matrix values using formulas.

- Pitch F0

$$\text{Pitch-period} = (20 + \text{samplerate}) * (1/Fs) \quad (8)$$

$$\text{Pitch - frequency (F0)} = 1 / \text{pitch - period} \quad (9)$$

$$\text{Mean value, } \mu = \sum F0_i / N, \quad i \text{ varies from } 1 \text{ to } N \quad (10)$$

$$\text{Standard deviation, } \sigma = \sqrt{1/N \sum (X_i - \mu)^2}, \quad i \text{ varies from } 1 \text{ to } N \quad (11)$$

$$\text{Maximum value} = \max (|i_{F0_{\max}}|) / N \quad (12)$$

$$\text{Minimum value} = \min (|i_{F0_{\min}}|) / N \quad (13)$$

Comparison was done by Tamil, Telugu, and ASD Tamil with standard database EMO-DB. For classification multiclass SVM method is implemented.

Implementation work is done by MATLAB R2012a. Table 3 shows feature extraction steps involved after preprocessing steps.

### 3.3 Multiclass SVM Classification

Feature extraction processes are calculated and it is used as inputs to Multiclass SVM. Multiclass SVM database categorization setback can be calculated into

$$S = \{(x_i, l_i) | x_i = R^m, l_i \in \{1, \dots, C\}, i = 1, 2, \dots, N\} \quad (14)$$

The multiclass SVM technique is trained and classified the databases based on the parameters as shown below:

- If (Pitch F0 =  $\alpha_1 \pm 0.03$ ) and (SD-Mean =  $\beta_1 \pm 0.02$ ) and (max value =  $\gamma_1 + 0.04$ ) and (min value =  $\gamma_1 - 0.04$ )  $\Rightarrow$  **Neutral**
- Else if (Pitch F0 =  $\alpha_2 \pm 0.05$ ) and (SD-Mean =  $\beta_2 \pm 0.02$ ) and (max value =  $\gamma_2 + 0.04$ ) and (min value =  $\gamma_2 - 0.04$ )  $\Rightarrow$  **Anger**
- Else if (Pitch F0 =  $\alpha_3 \pm 0.05$ ) and (SD-Mean =  $\beta_3 \pm 0.02$ ) and (max value =  $\gamma_3 + 0.04$ ) and (min value =  $\gamma_3 - 0.04$ )  $\Rightarrow$  **Sadness**
- Else if (Pitch F0 =  $\alpha_4 \pm 0.05$ ) and (SD-Mean =  $\beta_4 \pm 0.02$ ) and (max value =  $\gamma_4 + 0.04$ ) and (min value =  $\gamma_4 - 0.04$ )  $\Rightarrow$  **Fear**
- Else  $\Rightarrow$  **Happiness**

where  $\alpha_i$ ,  $\beta_i$  and  $\gamma_i$  are possibility values of global parameters of and 0.02, 0.03, 0.04 and 0.05 are the threshold values of emotions.

## 4 Experiments Results Using Multiclass SVM

The necessary features extraction and classification methods are described and calculated by above steps. Results obtained using MFCC, global parameters and Multiclass SVM are given below.

**Table 3** Feature extraction steps involved in ASER

Steps involved in Feature extraction	Requirements
Training speech corpus	Tamil-DB, Telugu-DB, ASD-DB
Testing speech corpus	EMO-DB (Berlin)
Mel Frequency Cepstral Coefficients	Preemphasis, Framing, Hamming windowing, Fast Fourier Transform, Triangular Band pass filter, Logarithm, Discrete Cosine Transform (DCT), 39 Double Delta MFCC
Global parameters	Pitch frequency (F0), Standard deviation (Mean), Maximum, Minimum
Known primaryemotion	Happiness, Sadness, Neutral, Anger and Fear

## 4.1 Feature Extraction

Feature extraction values are calculated frame by frame.

- 12 MFCC (m1-m12)
- 1 energy feature (e1)
- 12 delta MFCC features (dm1-dm12)
- 12 double-delta MFCC features (ddm1-ddm12)
- 1 delta energy feature (de1)
- 1 double-delta energy feature (dde1)

Figure 3 shows results of 39 dimensional MFCC feature extraction values of ASD children for anger emotion method. The presentation study of MFCC feature extraction is achieved by showing preemphasis method. Frame values represented by columns and MFCC parameters values (m1-m12, e1, dm1-dm12, ddm1-ddm12, de1, dde1) are represented by rows.

Figure 4 shows 39 dimensional of MFCC after preemphasis of sample Tamil language of anger primary emotion and it was preemphasized by all types of datasets for primary emotion.

## 4.2 Multiclass SVM

The multiclass SVM is implemented by formulas and the system is trained and classified based on the parameters for training and testing samples. The threshold values are calculated after some permutation. One-vs-all classification method is used to classify the N-primary emotion. During the comparison Multiclass SVM gives minimum difference between training and testing speech samples. Figure 5 shows experimental results for Multiclass SVM for Tamil, Telugu, ASD Tamil, and Berlin databases. The overall fusion values of Tamil, Telugu, ASD children and Berlin datasets are shown in Table 4.

These feature set processes 65 to 75.4% classification precision for Fear, Anger, Happiness, Sadness, and Neutral using MFCC and global features extraction and multiclass SVM for classification. Figure 6 shows sample screen of ASER system for ASD children with above-mentioned languages.

## 5 Conclusion

In this work, an Automatic Speech Emotion Recognition System for ASD children is designed from signals for Tamil language. Databases contain emotional speech by practicing ASD children by the practitioners which is recorded and converted into standard format. The observations are made with Telugu-DB, Tamil-DB, and

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	
1	M FCC	Frame 1				Frame 2				Frame 3				Frame 4			
2	m1	7.691479	8.813917	9.427308	9.027773	7.906897	8.039592	10.27296	12.18078	9.949951	5.501337	-36.0586	-28.4246	-23.4996	-18.7266	-14.2	
3	m2	9.382166	9.462713	9.163184	9.026264	8.799677	8.362695	8.160762	6.45883	6.392522	6.483248	5.242628	4.032166	1.560751	-1.58296	-3.07	
4	m3	-2.74947	-4.50614	-5.58525	-5.66421	-6.01906	-6.02184	-5.32738	-6.00541	-6.04742	-3.65023	-4.17545	-4.50342	-4.75775	-6.04582	-6.29	
5	m4	-2.98065	-3.68088	-4.16588	-4.44026	-5.67266	-6.24476	-4.92616	-3.59516	-1.16729	0.926761	-0.72756	-0.67914	-0.45358	-0.88794	-1.75	
6	m5	-3.14289	-2.87988	-3.53457	-2.97914	-1.43765	-1.78278	-1.80702	-2.9398	-4.46407	-2.47679	0.332092	-1.82104	-0.82872	-2.39653	-1.28	
7	m6	0.315666	1.028813	0.571206	-0.51893	-1.7848	-1.87536	-2.16733	-3.40375	-3.74236	-4.4739	2.195092	1.677719	3.014358	0.964403	1.894	
8	m7	-0.25396	-1.52322	-2.4202	-2.89608	-3.75928	-4.92484	-4.10568	-4.29196	-5.06804	-5.54695	2.5989	1.607638	1.028813	1.240926	2.045	
9	m8	-0.33979	-1.08971	-0.94657	-1.00769	-2.11177	-2.53175	-1.60019	-1.60761	-1.91548	-1.18945	2.201118	2.10604	0.009966	0.476482	2.277	
10	m9	-1.45922	-1.27904	-1.73719	-1.56475	-1.63373	-0.46197	0.983595	1.461174	2.156082	2.138208	1.843351	1.633982	0.865205	0.699361	1.894	
11	m10	-0.35419	-0.53332	-0.83635	-0.64614	-0.3268	-0.23619	0.63793	-0.32321	-1.48211	-0.3748	1.833236	1.299502	0.979514	0.365362	0.803	
12	m11	0.144241	-0.38109	-0.62854	-0.74522	-0.14224	-0.00908	0.283761	0.410808	-0.03018	-1.31421	-0.08716	0.886013	1.648898	0.373736	0.441	
13	m12	-1.78476	-1.62433	-1.40411	-1.28776	-1.82945	-2.23502	-1.37939	-1.09789	-1.55259	-1.62609	-0.98647	-0.14944	-0.62865	-1.1046	-1.04	
14	e1	-1.75261	-2.04694	-2.7318	-2.73887	-2.14358	-2.76185	-2.57559	-2.97914	-3.39208	-2.81948	-0.26889	-0.0002	0.292743	-0.391	-0.16	
15	dm1	-0.53947	-0.28745	-0.1108	0.381317	-0.48325	-0.08698	0.532421	0.246756	-0.0661	-1.37833	-0.90555	-0.607	0.273896	0.566037	0.434	
16	dm2	-0.63173	-0.34501	-0.14097	-0.69061	-1.05536	-1.01216	-0.70079	-0.72978	-0.62008	0.397261	-0.57396	-0.06525	0.321428	-0.10291	-0.32	
17	dm3	-0.56126	-0.3419	-0.56582	-1.21121	-1.52281	-1.3412	-0.48313	-0.41624	-0.32648	1.127792	0.863987	-0.20744	0.025729	-0.06441	-0.67	
18	dm4	0.054334	0.276517	-0.24045	-0.28384	-0.37858	-0.42171	0.304803	0.8834	1.290939	1.871828	0.403353	-0.12699	0.842409	0.776928	0.070	
19	dm5	-0.47181	-0.65872	-0.68358	0.167612	0.676863	0.76392	0.746395	0.713727	0.514929	0.462331	-0.26929	-0.3088	-0.39871	-0.52856	0.174	
20	dm6	-0.42472	-0.97422	-0.73795	-0.58254	-0.15407	-0.18866	-0.15148	-1.11667	-1.44428	-1.45317	-0.11883	-0.5024	0.035458	-0.66985	-0.70	
21	dm7	1.028813	0.571206	-0.51893	-1.7848	-1.87536	-2.16733	-3.40375	-3.74236	-4.4739	-6.31154	0.378916	-0.10334	0.723852	0.140262	-0.16	
22	dm8	-1.52322	-2.4202	-2.89608	-3.75928	-4.92484	-4.10568	-4.29196	-5.06804	-5.54695	-6.25224	-19.341	-18.8411	-21.319	-24.4859	-24.6	
23	dm9	-1.08971	-0.94657	-1.00769	-2.11177	-2.53175	-1.60019	-1.60761	-1.91548	-1.18945	-0.4387	13.44084	12.33435	11.699059	10.06854	8.995	
24	dm10	-1.27904	-1.73719	-1.56475	-1.63373	-0.46197	0.983595	1.461174	2.156082	2.138208	1.181207	4.715959	4.611461	5.280659	6.277004	6.214	
25	dm11	-0.53332	-0.83635	-0.64614	-0.3268	-0.23619	0.63793	-0.32321	-1.48211	-0.3748	-0.65267	1.58685	1.816495	2.709779	3.113456	3.663	

Fig. 3 39 dimensional MFCC feature extraction values of ASD children for anger emotion

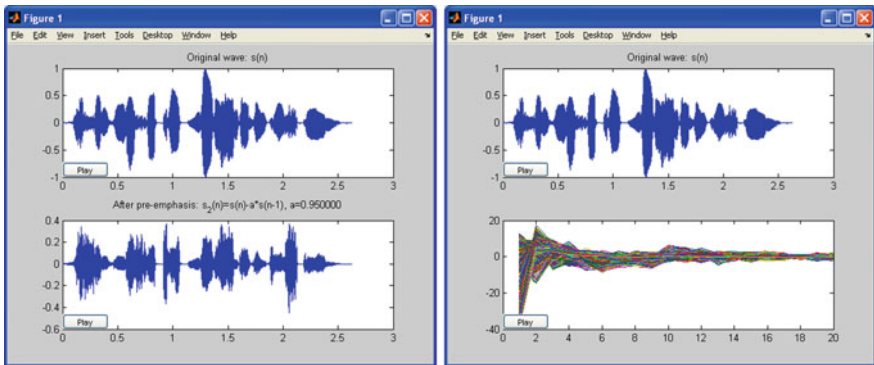


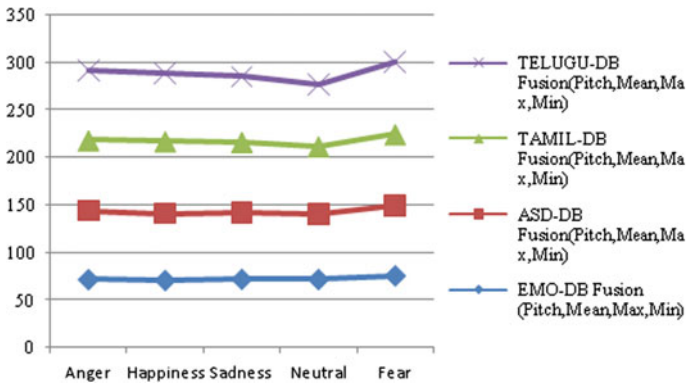
Fig. 4 39 dimensional MFCC after preemphasis for Tamil language of anger emotion

EMO-DB. A proportional study of feature extraction methods such as MFCC and global parameters are performed to extract good feature from the speech signals. These methods are combined with multiclass support vector machine for classification and the result is shown in confusion matrix. The performance of these techniques are tested and evaluated and it is found to be efficient in recognizing and classifying primary emotion of ASD children from speech signal. Future direction

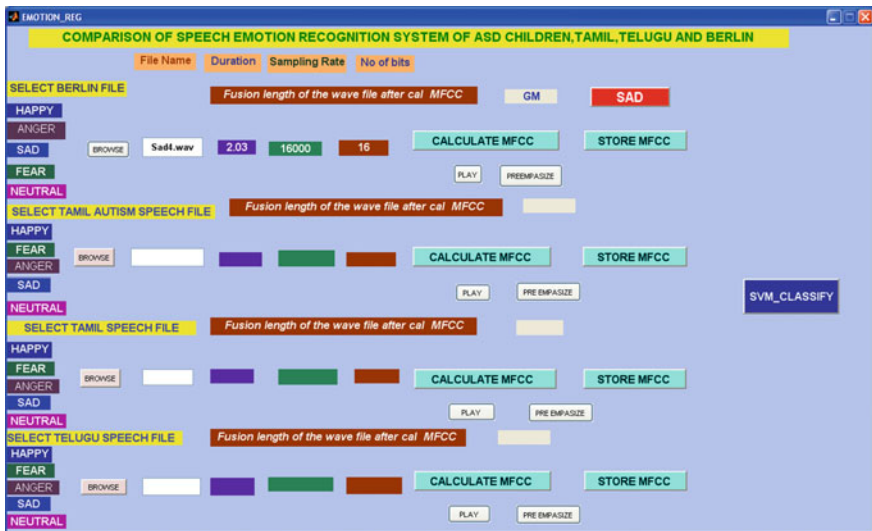


**Table 4** Confusion matrix of the Multiclass SVM

Type of Databases	Features	Anger	Happiness	Sadness	Neutral	Fear
EMO-DB	Fusion (Pitch,Mean, Max,Min)	71.25	70.41	71.61	71.33	75.23
ASD-DB	Fusion (Pitch,Mean, Max, Min)	73.57	71.33	70.78	69.67	75.02
TAMIL-DB	Fusion (Pitch,Mean, Max,Min)	73.49	75.47	74.08	70.67	74.28
TELUGU-DB	Fusion (Pitch,Mean, Max,Min)	72.87	71.91	68.68	65.45	75.92



**Fig. 5** Inference of all databases using Multiclass SVM



**Fig. 6** Sample screen of ASER system for ASD children with other datasets

hybrid classification and multimodal emotion can be explored. It may be supportive to identify primary emotion of ASD adults, finding secondary emotion from continuous speech.

**Acknowledgements** The authors would like to thank SCSVMV University for supporting this work.

## References

1. Picard, Rosalind W., and Roalind Picard. *Affective computing*. Vol. 252. Cambridge: MIT press, 1997.
2. Rajoo, Rajesvary, and Ching Chee Aun. "Influences of languages in speech emotion recognition: A comparative study using Malay, English and Mandarin languages." *Computer Applications & Industrial Electronics (ISCAIE)*, 2016 IEEE Symposium on. IEEE, 2016.
3. Poorna, S. S., et al. "Emotion recognition using multi-parameter speech feature classification." *Computers, Communications, and Systems (ICCCS)*, International Conference on. IEEE, 2015.
4. Ladde, Pravina P., and Vaishali S. Deshmukh. "Use of Multiple Classifier System for Gender Driven Speech Emotion Recognition." *Computational Intelligence and Communication Networks (CICN)*, 2015 International Conference on. IEEE, 2015.
5. Palo, Hemanta Kumar, Mihir Narayan Mohanty, and Mahesh Chandra. "Statistical feature based child emotion analysis." (2015): 53–6.
6. Zhang, Weishan, et al. "Emotion Recognition in Speech Using Multi-classification SVM." *Ubiquitous Intelligence and Computing and 2015 IEEE 12th Intl Conf on Autonomic and Trusted Computing and 2015 IEEE 15th Intl Conf on Scalable Computing and Communications and Its Associated Workshops (UIC-ATC-ScalCom)*, 2015 IEEE 12th Intl Conf on. IEEE, 2015.
7. Meddeb, Mohamed, Hichem Karray, and Adel M. Alimi. "Speech emotion recognition based on arabic features." *Intelligent Systems Design and Applications (ISDA)*, 2015 15th International Conference on. IEEE, 2015.
8. Sinith, M. S., et al. "Emotion recognition from audio signals using Support Vector Machine." *Intelligent Computational Systems (RAICS)*, 2015 IEEE Recent Advances in. IEEE, 2015.
9. Lerner, Matthew D., James C. McPartland, and James P. Morris. "Multimodal emotion processing in autism spectrum disorders: an event-related potential study." *Developmental cognitive neuroscience* 3 (2013): 11–21.
10. Jones, Catherine RG, et al. "A multimodal approach to emotion recognition ability in autism spectrum disorders." *Journal of Child Psychology and Psychiatry* 52.3 (2011): 275–285.
11. Lin, I-Fan, et al. "Vocal identity recognition in autism spectrum disorder." *PLoS one* 10.6 (2015): e0129451.
12. Marchi, Erik, et al. "Typicality and emotion in the voice of children with autism spectrum condition: evidence across three languages." *INTERSPEECH*. 2015.
13. Xavier, Jean, et al. "A multidimensional approach to the study of emotion recognition in autism spectrum disorders." *Frontiers in psychology* 6 (2015).
14. Chaby, Laurence, et al. "Exploring multimodal social-emotional behaviors in autism spectrum disorders: an interface between social signal processing and psychopathology." *Privacy, Security, Risk and Trust (PASSAT)*, 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom). IEEE, 2012.

# Aadhaar Card Voting System

Lingamallu Naga Srinivasu and Kolakaluri Srinivasa Rao

**Abstract** Nowadays a reliable voting system is needed for our society to eliminate the fake voting and to improve flexibility, transparency, and reliability. This paper presents a novel voting system by using QR code of Aadhaar card. The Aadhaar QR code is decrypted by using binarization and Reed–Soloman error correction. Based on the decrypted Aadhaar number in the QR code, it provides a citizen’s name and fingerprint details in our database. This Aadhaar voting system (AVS) accepts a citizen fingerprint. If the authentication is success, then it allows the citizen to vote otherwise it does not allow the citizen to vote. This paper provides three document files by using DIARY technique. First document file provides voter’s name and voting time. Second document file provides voter’s address and information. Third document provides individual party votes and total number of votes. These three document files can be generated and deleted by given specific password through special officer. This AVS machine ON/OFF can be controlled by specific password through special officer. By using this type of system, it will eliminate fake voting and provides more transparency and reliability.

**Keywords** Aadhaar card · Binarization · Reed–Soloman error correction  
Diary technique · Minutiae · Ridges · Fingerprint

## 1 Introduction

So far in India two types of voting systems are used. First one is ballet voting system and second one is Electronic voting machine (EVM). In EVM voting system, the machine does not recognize the authorized person but this is being done

---

L.N. Srinivasu (✉) · K.S. Rao  
Department of ECE, Kallam Haranadha Reddy Institute of Technology,  
Guntur 522019, Andhra Pradesh, India  
e-mail: lingamallusrinivas@gmail.com

K.S. Rao  
e-mail: ksrmtchrao@gmail.com

manually which can be overcome in this proposed Aadhaar card Voting System (AVS).

Nowadays, Aadhaar card utilization is increasing day by day in India. Aadhaar card is used in electronic mobiles, money transactions, to identify the authorized person, etc. This paper develops a novel voting system using Aadhaar card. Aadhaar card contains a citizen information, Aadhaar number, QR code. In that, Aadhaar QR code contains a valid Aadhaar number. By decoding the QR code, the Aadhaar number is obtained. The citizen information can be accessed by using the Aadhaar number. The citizen information contains an iris data, fingerprint data, address, etc. Based on the Aadhaar QR code, a virtual voting System using diary technique is developed.

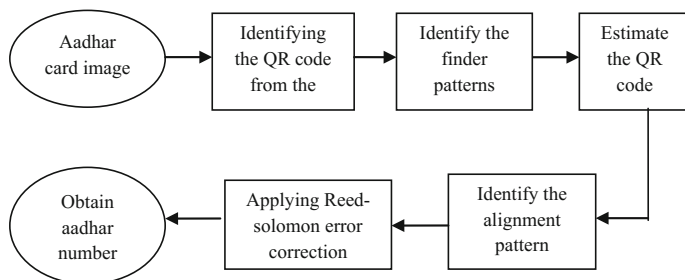
The AVS allows the citizen Aadhaar QR code. The Aadhaar number is extracted by the decoding of QR code. Extract the citizen information and fingerprint from the database based on the Aadhaar number.

Next AVS allows the citizen finger print in run time. If database fingerprint and runtime fingerprint matches, then it allows for voting, otherwise it does not accept the citizen to vote.

The AVS generates three document files. First document file contains the voter's name and voting time. Second document file contains complete voter information like voter's name, address, and voting time. Third document file produced by specific officer with specific password. This file contains a data of the voter or individual party data. It also provides the total number of votes and percentage of the voting.

## 2 Decoding the QR Code

This stage consists of several stages. First stage is identifying the QR code from the Aadhaar card. Second stage is finding the finder patterns and QR code version. Third stage is identifying the alignment pattern. Final stage is performing the Reed-Solomon error correction for decoding the text in QR code (Fig. 1).



**Fig. 1** Block diagram of QR code decoding

### 2.1 Identification of QR Code

A new binarization technique is applied to the Aadhaar card for identifying the black and white module (QR code) in Aadhaar card image. The Aadhaar card image is shown in Fig. 2.

The Aadhaar image is divided into several blocks to achieve the binarization. Compute the threshold  $C_1$  to every block by using the below formula.

$$C_1(X, Y) = \text{mean}(X, Y) * \left[ 1 + \left( \frac{\text{standard deviation}(X, Y)}{P} \right) \right], \tag{1}$$

where the conventional value for  $P$  is 1250.  $X, Y$  are the correspondent row and column of the block. Compute another threshold  $C_2$  by applying convolution between the block and kernel  $K$ .

where

$$K = \frac{1}{10} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 1 \end{bmatrix} \tag{2}$$

Identify the QR code block by using the threshold  $C_2$ . The Fig. 3 shows the QR code obtained after the binarization.

### 2.2 Identify the Finder Patterns and QR Code Version

Finder patterns of QR code are obtained by the detecting the top left, top right, and bottom left finder patterns in the QR code. This corner finder patterns are detected by the collecting all points in image with matching ratio of 1:1:3:1:1, both horizontal and vertical. Finder pattern of QR code can be obtained by merging the collecting points. The ratio 1:1:3:1:1 will specify the module widths of finder pattern.

Fig. 2 Aadhaar card



**Fig. 3** QR code image using binarization



QR code version can be estimated by using the module size (MS) and distance between upper left and upper right finder pattern. The module size can be calculated by using the formula given below

$$\text{Module Size (MS)} = \frac{\text{Width of the finder pattern}}{7} \quad (3)$$

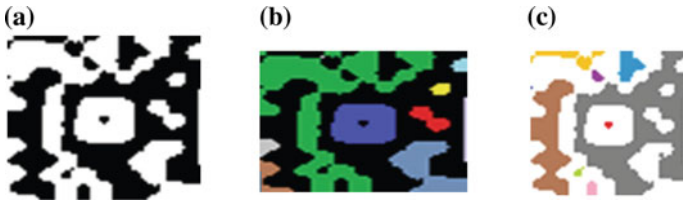
The formula for calculating the QR code version is shown below

$$\text{Version} = \frac{((\text{Euclidean distance (fp}_1, \text{fp}_2)/\text{MS}) - 10)}{4}, \quad (4)$$

where  $\text{fp}_1$  and  $\text{fp}_2$  are top positioned right indent finder pattern and top positioned left indent finder pattern, respectively and MS represents the module size.

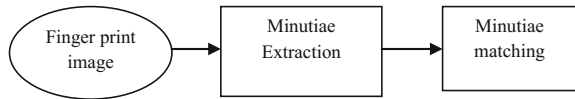
### ***2.3 Identify the Appropriate Alignment Patterns***

Low-resolution images and inappropriate binarization make the deformation of alignment patterns. To find out the alignment patterns, first roughly locate and extract one alignment pattern with specified radius by using finder patterns. This alignment pattern can be taken as a sub-image. Next the need is to calculate the connected component of black pixel and white pixel of the sub-image. The connected component of black pixels and the connected component of white pixels are shown in Fig. 4. For each connected component of black pixel  $C_i$  in Fig. 4b, check its border pixels. If all border pixels are adjacent to the same connected component of white pixel  $C_j$  in Fig. 4c, then check all border pixels in  $C_j$ . Once all border pixels in  $C_j$  are adjacent to the same black pixel connected component  $C_k$  in Fig. 4b, these components possibly contain alignment pattern. Finally, calculate the centroid  $C_i$  as the centroid of alignment pattern.



**Fig. 4** **a** Extracted image with radius = 30 pixels, **b** connected component of black pixels, **c** connected component of white pixels

**Fig. 5** Block diagram of fingerprint design description



### 2.4 Error Correction

This paper uses Reed–Soloman code for error correction. In this error correction Peterson–Gorenstein–Zierler algorithm is used.

## 3 Authentication of Fingerprint

In authentication of fingerprint, stored template fingerprint is compared with the input fingerprint. In authentication of fingerprint consists of two stages. First stage is minutiae extraction and second stage is minutiae matching (Fig. 5).

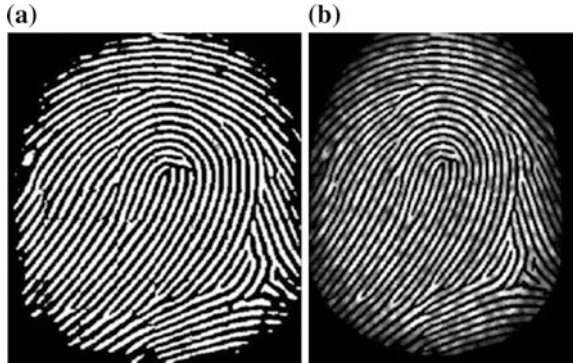
### 3.1 Minutiae Extraction

In this stage consists of two steps. First step is image enhancement and image segmentation and second step is final extraction.

#### 3.1.1 Image Enhancement and Segmentation

Image enhancement is used to improve the contrast of ridges and furrows in fingerprint. It also connects the false break points of the ridges. Adaptive histogram equalization technique and adaptive binarization method comes under the image enhancement.

**Fig. 6** **a** Finger print image after binarization, **b** Finger print image before binarization



Adaptive histogram equalization method improves the global contrast of an image. It increases the gain of lower contrast pixels without affecting the global contrast.

Adaptive binarization method performs the binarization to the image. Binarization converts gray image to binary image. It assigns logical 1 to furrows and logic 0 to ridges. Due to the binarization, fingerprint ridges are highlighted is shown in Fig. 6.

### **Image Segmentation:**

Image segmentation can be achieved by identifying ROI. ROI is used to extract only ridges and furrows part and eliminates the area which does not contain ridges and furrows. It can be achieved by ROI extraction by morphological methods and block direction estimation.

### **Block Direction Estimation:**

In block direction estimation, the binarized image is divided into  $16 \times 16$  blocks. After dividing the blocks, the following algorithm is applied to each block.

- (a) Determine the gradient values along  $x$ -direction ( $gdx$ ) and  $y$ -direction ( $gdy$ ) for each pixel of the block.
- (b) The following formula is applied to each block and then obtain the least square approximation of block direction.

$$\beta = \frac{1}{2} \tan^{-1} \left( \frac{2 * (gdx * gdy)}{gdx^2 - gdy^2} \right), \quad (5)$$

where  $gdx$ ,  $gdy$  represents the gradient values along  $x$ -direction and  $y$ -direction respectively.  $\beta$  is used for least square approximation of block direction. After completing the estimation of each block direction, apply the following formula to each block which are used to remove the information on ridges and furrows.



$$E = \frac{(gdx * gdy) + (gdx^2 - gdy^2)}{(w \times w) * (gdx^2 - gdy^2)}, \tag{6}$$

where  $w \times w$  represents the block size. In each block, if the value of  $E$  falls below the threshold value then the block can be considered as background block. The direction map image of binarized image is shown in Fig. 7.

**Extraction of ROI by Morphological Operations:**

Extraction of ROI can be achieved by using two operations. The two operations are open and close. Close operation can be used to shrink the image and reduce the small cavities. Open operation can be used to expand the image and reduce the peaks introduced by background noise.

**3.1.2 Final Minutiae Extraction**

The final minutiae extraction consists of four operations. The four operations are ridge thinning, minutiae marking, removal of false minutiae, and finally minutiae representation.

**Ridge Thinning:**

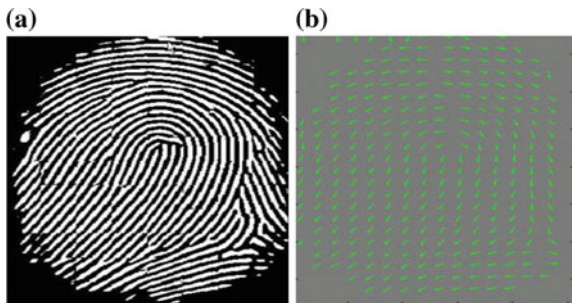
The ridge thinning process reduces the redundant pixels of ridges. It is repeated until the obtained ridges are only one pixel wide. This can be achieved by using the following MATLAB function. The ridge thinning image is shown in Fig. 8.

```
bwmorph(binaryImage, 'thin', Inf);
```

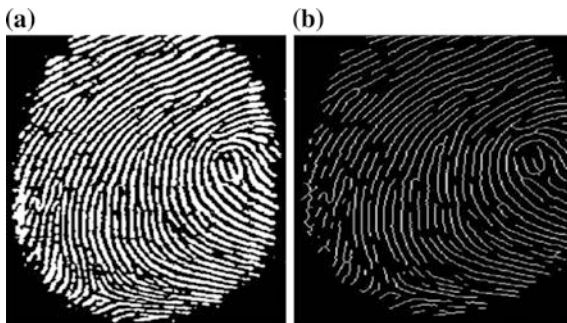
**Minutiae Marking:**

Minutiae marking done to  $3 \times 3$  block are as follows:

**Fig. 7** **a** Binarized image, **b** Direction map of binarized image



**Fig. 8** **a** Image before, **b** Image after thinning



**Fig. 9** Representation of ridge branch

0	1	0
0	1	0
1	0	1

**Fig. 10** Representation of ridge ending

0	0	0
0	1	0
0	0	1

- (a) If the central pixel is 1 and has exactly three one-value neighbors, then the central pixel will be a ridge branch as shown in Fig. 9.
- (b) If the central pixel is 1 and has only one one-value neighbor, then the central pixel will be a ridge ending is shown in Fig. 10.

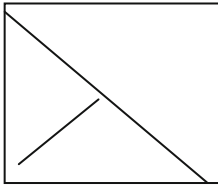
**Removal of False Minutiae:**

This stage removes the false minutiae. The false minutiae occurs due to the over amount of inking. Before going to the removal of false minutiae, first calculate the Inter-ridge Distance (ID). It is the average distance between two ridges. Calculate the inter-ridge distance to each row by using the following formula:

$$\text{Inter ridge distance} = \frac{\text{sum all pixels with value 1}}{\text{row length}} \tag{7}$$

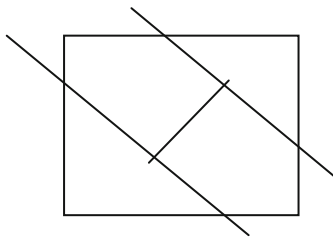
The following steps are used to remove the false minutiae.

1. If the value of  $d(\text{bifurcation, termination})$  is less than ID and the two minutia will exist in the same ridge, then discard the both (case  $m_1$ )

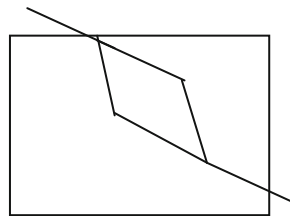


$m_1$

2. If the value of  $d(\text{bifurcation, termination})$  is less than ID and the two minutia exist in the same ridge then will discard the both (case  $m_2$  and  $m_3$ )

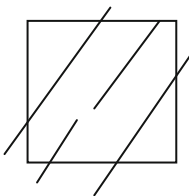


$m_2$

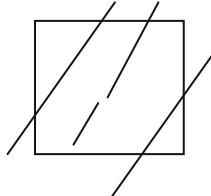


$m_3$

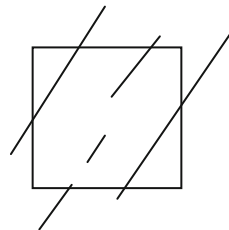
3. If the value of  $d(\text{bifurcation, termination})$  is approximately equal to ID and their directions are coincident with a small angle variation and not with any other termination located between the two terminations, then will discard both of them (case  $m_4, m_5, m_6$ )



$m_4$

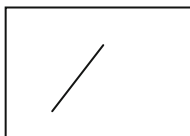


$m_5$



$m_6$

4. If the value of  $d(\text{bifurcation, termination})$  is less than ID and the two minutia are exist in the same ridge then will discard the both (case  $m_7$ )



$m_7$

where  $d(X, Y)$  is the distance between the two minutia points.

### Minutiae Representation:

In minutiae representation, each minutiae is characterized by the following parameters:

- (1)  $x$ -coordinate
- (2)  $y$ -coordinate
- (3) Orientation
- (4) Ridge associated with it.

A bifurcation can be split into three terminations such as  $x$ - $y$  coordinates (pixel adjacent to the bifurcating pixel), orientation and an associated ridge.

The orientation of each termination ( $t_x, t_y$ ) can evaluated by following method:

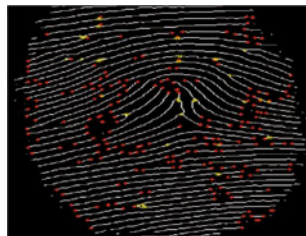
1. Extract a ridge segment whose starting point is the termination and length is  $D$ .
2. Sum up all  $x$ -coordinates of points in the ridge segment.
3. Divide above summation with Inter-ridge Distance (ID) value to get  $s_x$ . To get  $s_y$  use the same procedure.

The result of minutiae extraction is shown in below Fig. 11.

## 3.2 Minutiae Matching

In minutiae matching, compare the database fingerprint image and runtime processed image. For minutiae matching, use the elastic match algorithm. Elastic algorithm is used to count the number of minutiae matched pairs.

**Fig. 11** Image of minutiae extraction



An elastic string  $(x, y, \theta)$  match algorithm is used to find number of matched minutia pairs in  $I_1$  and  $I_2$  sets.  $I_1$  and  $I_2$  are the minutiae sets given by

$$I_1 = \{m_1, m_2 \dots m_i\} \text{ where } m_i = (x_i, y_i, \theta_i) \quad (8)$$

$$I_2 = \{m_1, m_2 \dots m_j\} \text{ where } m_j = (x_j, y_j, \theta_j) \quad (9)$$

Elastic string match algorithm calculates the spatial distance between the minutiae  $m_i$  in  $I_1$  and minutiae  $m_j$ . If spatial distance is similar to a given tolerance  $r_0$  between the two minutiae sets, then it is considered as matching, otherwise it is considered as not matching.

## 4 Diary Technique

It is a command in MATLAB software. This function creates a word file with a specific file name. It saves the word file to a particular path, specified in current folder of the MATLAB. If you do not specify the file name to the word file then MATLAB creates a file named as DIARY in the current folder. DIARY command makes the word file to hold the output data.

‘Diary(‘filename’)’ writes a copy of all subsequent keyboard inputs and the resulting output to the named file where filename is the full path name or file name in the current MATLAB folder but it does not hold the figure result.

## 5 Authentication

Authentication is a very important tool to every application. There are several methods existing to provide the authentication. This paper provides the authentication through the password. This password system is used to keep the machine ON, generate the voting information files, delete the voting information file, and switch the machine OFF.

## 6 Results

The machine will be activated through the specific password. If password is correct, it will show “WELCOME TO THE AADHAR VOTING SYSTEM” and allows the voter to vote. The machine will ask for Aadhaar card and scans it to generate the QR code.

Machine performs QR decoding operation and shows the Aadhaar number. Aadhaar number will be checked with database and will generate the citizen

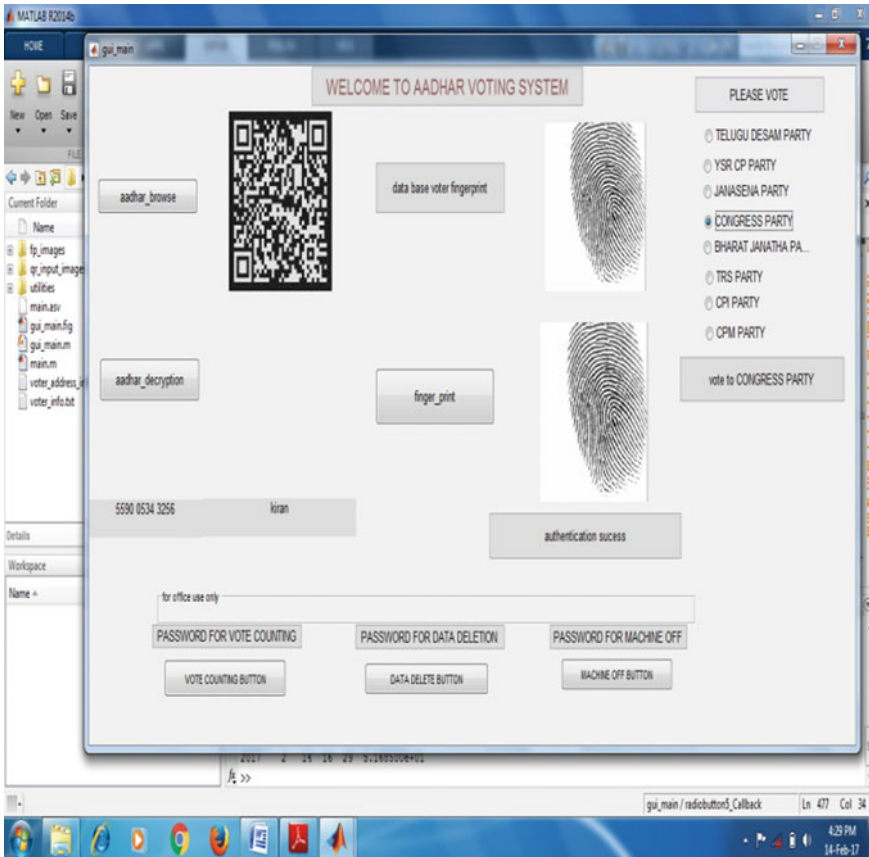


Fig. 12 Voting success

fingerprint. It asks the runtime fingerprint and compares with database fingerprint. If it matches then it will show “AUTHENTICATION SUCCESS” and allows the voter to vote. After voting it shows the status is shown in Fig. 12.

When the runtime fingerprint does not match with the database fingerprint, then it shows the “AUTHENTICATION FAILED” and does not allow for voting. Even if voted, it does not show the vote status. Result is shown in Fig. 13.

After completion of voting, enter the password to generate the voting file. If password matches, then it will show the “PASSWORD MATCH” and will generate the voting file, otherwise it will not do it. Result is shown in Fig. 14.

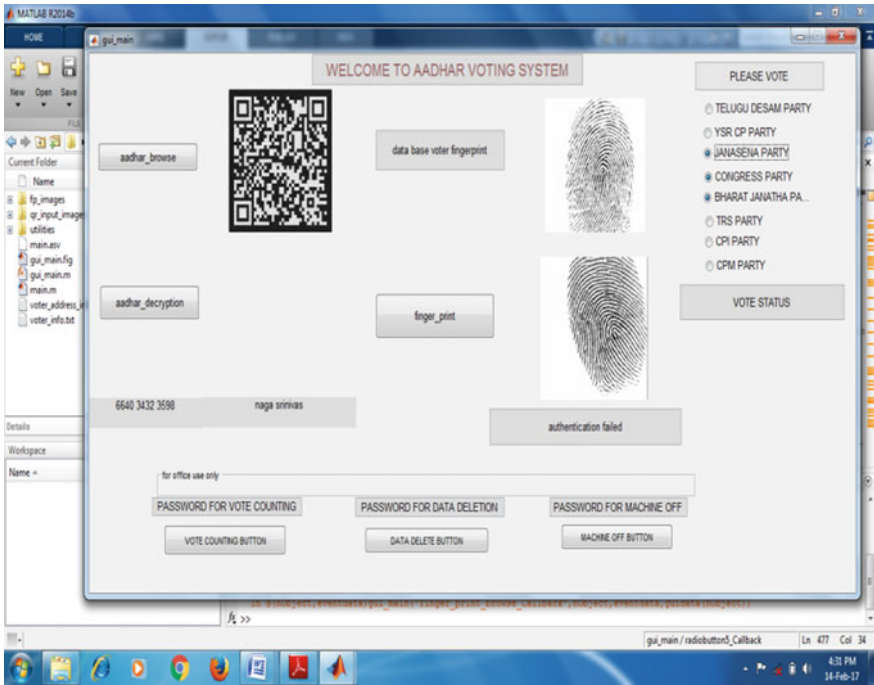


Fig. 13 Voting failure

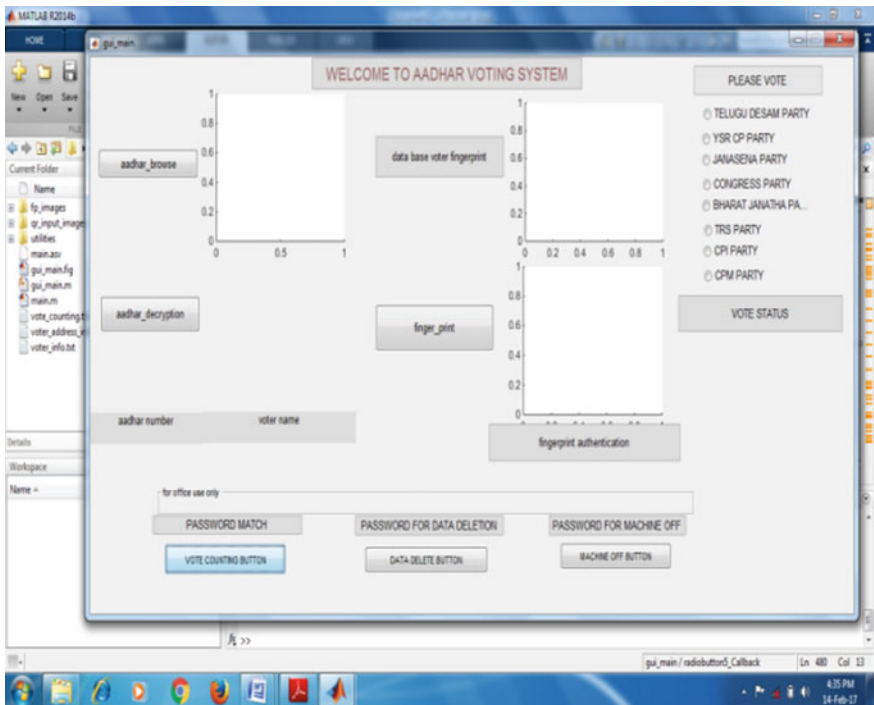


Fig. 14 Voting file generation

## 7 Conclusion

In this paper, it is effectively designed the QR decoding procedure and developed fingerprint matching process to identify the authorized voter to allow for voting. This paper also developed the method to generate the voting files through password which increases reliability and transparency.

## References

1. D. Ashok Kumar and T. Ummal Sariba Begum, "Electronic Voting Machine—A Review" Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, March 21–23, 2012
2. Jeng-An Lin and Chiou-Shann Fuh, "2D Barcode Image Decoding" Hindawi Publishing Corporation Mathematical Problems in Engineering Volume 2013, Article ID 848276, 10 pages
3. Sangram Bana and Dr. Davinder Kaur, "Fingerprint Recognition using Image Segmentation" Sangram Bana, et al./ (ijaest) international journal of advanced engineering sciences and technologies vol no. 5, issue no. 1, 012–023
4. L. F. F. Belussi and N. S. T. Hirata, "Fast component-based QR code detection in arbitrarily acquired images," *Journal of Mathematical Imaging and Vision*, vol. 45, no. 3, pp. 277–292, 2013
5. L. F. F. Belussi and N. S. T. Hirata, "Fast QR code detection in arbitrarily acquired images," in Proceedings of the Conference on Graphics, Patterns and Images (SIBGRAPI '11), pp. 281–288, Macei'o, Brazil, August 2011
6. E. Ohbuchi, H. Hanaizumi, and L. A. Hock, "Barcode readers using the camera device in mobile phones," in Proceedings of the International Conference on Cyberworlds (CW'04), pp. 260–265, Tokyo, Japan, November 2004
7. Y. Liu, J. Yang, and M. Liu, "Recognition of QR Code with mobile phones," in Proceedings of the Chinese Control and Decision Conference (CCDC '08), pp. 203–206, Yantai, China, July 2008
8. Handbook of Fingerprint Recognition by Davide Maltoni, Dario Maio, Anil K. Jain & Salil Prabhakar
9. Fingerprint Recognition, Paper by WUZHILI (Department of Computer Science & Engineering, Hong Kong Baptist University) 2002
10. Fingerprint Classification and Matching by Anil Jain (Department of Computer Science & Engineering, Michigan State University) & Sharath Pankanti (Exploratory Computer Vision Group IBM T. J. Watson Research Centre) 2000
11. C. Y. Lai and M. S. Chen, Extracting QR Code from a nonuniform background image in embedded mobile phones [M.S. thesis], Department of Electrical Engineering, National Taiwan University, 2007
12. Mathworks computer software company <http://www.mathworks.com>



# An Innovative Security Model to Handle Blackhole Attack in MANET

MD. Sirajuddin, Ch. Rupa and A. Prasad

**Abstract** Mobile Adhoc Network (MANET) is a collection of mobile nodes communicating with each other without using any fixed infrastructure. Security is one of the most important issues in the MANET. In this paper we address the blackhole attack and proposed an innovative security model to handle this attack. Our proposed model supports Ad hoc on Demand Distance Vector (AODV) routing protocol which is the most widely used protocol in the MANET. The proposed model not only provides node authentication but also message authentication with low overhead.

**Keywords** MANET · Security · AODV · Authentication · Blackhole attack  
Session key

## 1 Introduction

Mobile Adhoc Network (MANET) is a collection of mobile nodes which communicates with each other without using any fixed infrastructure. It is a multihop and self-organized wireless network. The main strength of MANET is its dynamic topology. Because of this interesting feature MANETs are used in various domains like Military applications, Disaster Management, Environmental monitoring, etc., with flexibility. This dynamic topology leads to various security threats, because any node can become the part of the network. Sometimes malicious node enters the

---

MD. Sirajuddin (✉)  
JNTUK, Kakinada, India  
e-mail: siraj.cs@gmail.com

Ch. Rupa  
VRSEC, Vijayawada, India  
e-mail: rupa.mtech@gmail.com

A. Prasad  
VSU, Nellore, India  
e-mail: prasadjkc@yahoo.co.in

MANET and receives the packets which are not intended for it and even modifies the packet content which is transmitted in the MANET. This malicious node provokes the security risks in the network. The routing and security protocols designed for wired networks cannot be applied on MANETs because all the nodes in ad hoc networks are mobile. These security threats degrade the performance of the MANET. In this paper, we address one of the security threats known as blackhole attack. The existing routing protocols do not handle this attack by default. We need to modify the existing ad hoc network routing protocols to handle this blackhole attack. Due to this reason, the security is still the research issue in the MANET.

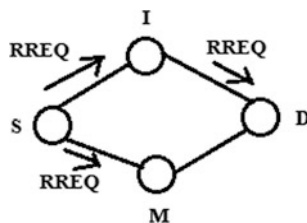
**Blackhole Attack:** It is a type of security attack in which a malicious node waits for RREQ packets from the sender. On receiving RREQ packet, it sends RREP immediately by claiming that it has a shortest route to the destination. As the sender is not aware of this malicious node, it sends all its data to the destination through this malicious node. This malicious node will receive the data which is originally meant for the destination. The presence of this blackhole attack degrades the performance of the MANET.

AODV is most widely used routing protocol in the MANET. It is reactive protocol, i.e., it finds the route whenever it is needed by the source node to transmit data to the destination. In this routing protocol whenever a node wants to transmit the data to the destination, it initiates route discovery process by flooding the RREQ packets. On receiving these RREQ packets, the intermediate nodes forward them towards the destination node. When destination node receives the RREQ packets, it sends RREP packet back to the source node. The source node considers the RREP packet which has the highest sequence number. It then forwards the data packets to the destination by using the route recorded in the RREP packet.

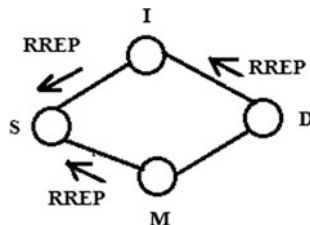
AODV protocol suffers from blackhole attack. Whenever the route discovery procedure is initiated, the blackhole attack comes into picture. Existing AODV protocol do not employ any security measures to resolve this security threat. This concept is depicted in the Fig. 1.

In above figure, *S* is the Source Node, *I* is the intermediate Node, *D* is the destination node, and *M* is the Malicious node. The above figure depicts the route discovery process initiated by the node *S*. It floods RREQ packets to all its neighbors to know the address of node *D*. The neighbors of *S*, *I* and *M* receive the RREQ packets. Since node *I* is the genuine node it forwards the RREQ packet to node *D*. Whereas node *M* is the malicious node; it will not forwards the RREQ packet to its neighbor. Instead of forwarding this RREQ packet to the node *D*, it

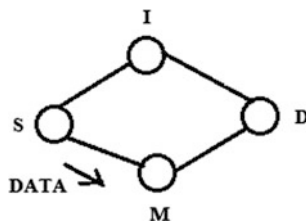
**Fig. 1** Route discovery in AODV with black hole attack



**Fig. 2** Transmission of RREP packet from malicious node



**Fig. 3** Transmission of data to destination through malicious node



sends RREP packet immediately to the node *S* claiming that it is having shortest distance to the destination node *D*. It does this by inserting highest sequence number in the RREP packet. On receiving this RREP packet from malicious node *M*, the source node *S* sends data packet to the destination through this malicious node and thinks that the data would reach the node *D* safely, because the node *S* is unaware that node *M* is the malicious node. This is shown in (Figs. 2 and 3).

Many researchers have proposed various flavors of AODV protocol to handle this blackhole attack. In Sect. 2 we focused on the existing techniques along with their limitations.

## 2 Literature Survey

In this section, we focused on the existing techniques proposed by the researchers to resolve the blackhole attack.

Ahemad et al. [1] used Encryption Verification Method (EVM) to detect blackhole attack and to securely transmit the data in MANET. In our study, we found that this technique requires more memory space at each node because it uses DCT (Data Collection Table) along with the routing tables to detect blackhole attack. However, verification process provides the authentication and confidentiality.

Kumar et al. [2] proposed promiscuous mode based technique to detect blackhole attack and propagates this information to other nodes so that the malicious node is removed from the MANET. But this technique requires extra packets to detect this attack.

Thanikaivel et al. [3] proposed fast and secure data transmission technique in MANET. They used reactive protocol and proactive protocol for routing the packets. They proposed a security model based on central agent and process algorithm. This technique seems to be expensive in terms of processing time and memory since it uses both types of routing protocols together along with a security model to securely transmit the data in the network.

Sharma et al. [4] implemented two solutions to handle blackhole attack. In the first solution, after sending RREQ packets, it waits for RREP packet to arrive from more than two nodes. After receiving two RREPs it checks for shared hops in both RREP packets. If both received RREP packets have shared hops, then the sender ensures itself that the route is secure and selects the route from one of the RREP packets. If no route contains shared hops then the sender waits for other RREPs until a route with shared hops is found. Second solution is based on packet sequence number. This technique detects blackhole attack based on the unique sequence number. If a sender does not find RREP packets with shared hops, then it will never transmit the data.

Krishna et al. [5] proposed trust-based AODV protocol to improve the QOS of the MANET. They used trust value to detect the malicious nodes. It is an extra overhead for a node in MANET. It diminishes the battery life of a node.

Alnumay and Ghosh [6] used ID-based protocol which secures AODV and TCP protocol in MANET. This is based on RSA algorithm. The limitation with this approach is it involves expensive computations to securely transmit the data.

Rivila and Putta [7] proposed Zone Routing Protocol based on Hash Algorithm to provide the security in the MANET.

All the aforementioned schemes handle the blackhole attack efficiently but they are expensive in terms of computation and memory space. In the next section, we proposed a novel security model which resolves blackhole attack with less complexity.

### 3 Proposed Technique

To overcome the limitations of existing techniques we proposed centralized node based security model to securely transmit the data in MANET. In our security model, we used one node as a master node. It is responsible for tracking out the activities of all nodes in a network. It assigns time stamp for each pair of nodes for communication. All the nodes in MANET must register themselves with the master node. It assigns ID's to the registered nodes. On receiving the RREP packet, the source node determines the route which is to be followed to reach the destination. Before sending the data to the destination, it sends a request to master node for session key which is to be used for communication. It is represented as follows:

$$S \rightarrow M : Req || ID_S || ID_R \quad (1)$$

The Eq. (1) represents the request from Sender to the Master node.  $S$  represents the sender;  $M$  is the Master node;  $Req$  indicates request packet;  $ID_S$  is the ID of Sender;  $ID_R$  is the Receiver's ID.

On receiving the request from the source node the master node responds with the session key, which is shown in the Eq. (2).

$$M : T_{SR} || ID_S || T \quad (2)$$

$M$  is the Master node;  $T_{SR}$  is the session key and  $T$  is the time stamp.  $T_{SR}$  is determined by the master node by considering various parameters like traffic, channel bandwidth, etc. After receiving the shared session key, the sender now formulates the message which it wants to send to the destination node. The message is hashed by using the MD5 algorithm. The format of message which is sent from sender node  $S$  to the receiver node  $R$  is as follows:

$$ID_S || ID_R || msg || H(msg) || T_{SR} \quad (3)$$

$ID_S$  is the ID of Sender;  $ID_R$  is the Receiver's ID,  $msg$  is the message;  $H(msg)$ : is the hash value of the message formed by using MD5;  $T_{SR}$  is the session key value generated by the master node. When intermediate node receives this data packet, it forwards the packet towards the destination. This process continues until the data packet reaches the destination. After receiving this data packet within the time stamp duration it regenerates the original message. If the message is valid then the destination node sends the acknowledgement. Before sending the acknowledgement, the destination node requests the master node for the session key. The request from destination node to the master node is as follows:

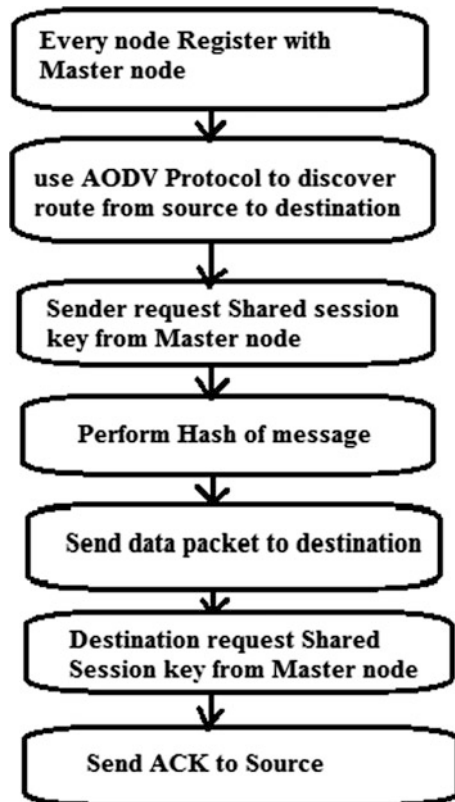
$$R \rightarrow M : Req || ID_R || ID_S \quad (4)$$

$R$ : is the destination node;  $M$  is the master node;  $Req$  indicates request message;  $ID_R$  is the ID of the receiver;  $ID_S$  is the ID of the sender node. The message from the master node to the destination consists of Shared session key, ID of receiver and Time stamp. It is represented in Eq. 5.

$$M : T_{RS} || ID_R || T \quad (5)$$

Once the destination node gets the session key from the master node, it sends the acknowledgement back to the sender. Since the message is encrypted, it cannot be interpreted by the malicious node. This security model resolves the blackhole attack. This concept is implemented over the AODV protocol. Once the route is determined by the route discovery process, this security protocol comes into work. If a malicious node takes the packets from the source node it is of no use for it. If a packet reaches the destination within the specified time stamp then only it is

**Fig. 4** Flow diagram of proposed system



considered for processing otherwise it is rejected. The flow chart for this entire process is depicted in Fig. 4.

## 4 Conclusion

Our proposed security model improves the performance of the MANET by resolving the blackhole attack. It uses hash function to provide the authentication and data integrity. It runs over the AODV routing protocol. This novel security model provides security to the MANET efficiently with low overhead. In future we implement our proposed scheme over modified AODV to increase the QOS of the MANET.

## References

1. Firoz Ahmed, Seok Hoon Yoon and Hoon Oh, "An Efficient Black Hole Detection Method using an Encrypted Verification Message in Mobile Ad Hoc Networks" *International Journal of Security and its Applications*, Vol 6. No. 2. April 2012. pp. 179–184.
2. Pramod Kumar Singh and Govind Sharma, "An Efficient Prevention of Black Hole Problem in AODV Routing Protocol in MANET", *IEEE 11th International Conference on Trust, Security and Privacy in Computing and Communications*. pp. 902–906.
3. B. Thanikaivel and B. Pranisa, "Fast and Secure Data Transmission in MANET", *International Conference on Computer Communications and Informatics ICCCI-2012*.
4. Ms. Nidhi Sharma and Mr. Alok Sharma, "The Black-hole node attack in MANET", *IEEE Second International Conference on Advanced Computing & Communication Technologies*, 2012. pp. 546–550.
5. Radha Krishna Bar, Jyotsna Kumar Mandal and Moirangthem Margit Singh, "QOS of MANET through Trust Based AODV Routing Protocol by Exclusion of Black Hole Attack", *International Conference on Computational Intelligence: Modelling Techniques and Applications CIMTA 2013*. pp. 530–537.
6. Waleed S. Alnumay and Uttam Ghosh, "Secure Routing and Data Transmission in Mobile Ad Hoc Networks", *International Journal of Computer Networks and Communications*, Vol. 6. 2014.
7. Dilli Rivila and Chandra Shekar Reddy Putta, "Enhancing the Security of MANET Using Hash Algorithms", *Eleventh International Multi-Conference on Information Processing (IMCIP-2015)*. pp. 196–206.

# Knowledge Discovery via SVM Aggregation for Spatio-temporal Air Pollution Analysis

Shahid Ali

**Abstract** Air quality information has drawn a lot of attention in every part of the world. People nowadays are more concerned about their health, among them children are at great risk as their lungs are developing at young age and increase in air pollutants will deteriorate their health. Therefore, air quality monitoring stations are placed to examine the air quality and to predict future air quality. In this regard, our research is focused on air quality monitoring, examination and prediction. As we know that air pollution is not a static problem, rather it is spatio-temporal problem as it changes from time to time and location to location. In this regard, a new computational technique named SVM aggregation is proposed for spatio-temporal air pollution analysis. Through knowledge fusion and with the help of SVM aggregation air pollution problem will be addressed systematically from monitoring to examination and future air quality prediction.

**Keywords** Knowledge discovery · Air pollution analysis · Support vector machines

## 1 Introduction

Man made pollutants through combustion, construction, mining, agriculture and warfare are the key contributors towards air pollution in today's environment [1, 2]. The common gaseous pollutants include carbon monoxide, sulphur dioxide, chlorofluorocarbons and nitrogen oxide produced by industry and motor vehicles [3]. Unpleasant air quality can kill many organisms including humans. The emissions of various gases through various resources in air can cause respiratory disease, cardiovascular disease, throat inflammation, chest pain and congestion [4]. In this context, air pollution caused to humans and other living organisms has suggested the formation of a computational technique for detecting, examining, monitoring

---

S. Ali (✉)

Unitec Institute of Technology, Auckland, New Zealand  
e-mail: alis12@unitec.ac.nz

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_16](https://doi.org/10.1007/978-981-10-6319-0_16)

181



and predicting air quality for future. It is noticeable that air pollution analysis often involves processing huge size of spatial image data, and long term historical/temporal data of air parameters. This in practice poses the challenge of not only spatial and temporal knowledge discovery, but also knowledge fusion over spatial and temporal dimensions. Addressing computational air pollutions analysis, the proposed research focuses on constructing spatial and temporal computing environment by SVM aggregation, where a set of individual support vector machines (SVMs) are aggregated for very purpose of data mining, and where the advantage of SVM aggregation computing is explored having (1) outstanding generalization capability of SVM aggregation; (2) speedy parallel SVM aggregation computing; and (3) effective knowledge fusion with the same SVM representation over the spatial and temporal dimensions.

The key questions for undertaking this research along with particular method are as follows:

1. Why spatial and temporal pollution?

Regardless of the previous work in air pollution, there is no evidence in their research and articles focusing on spatial and temporal air pollution analysis integration and interaction deploying SVM aggregation computing.

2. How SVM aggregation computing and knowledge fusion over spatial and temporal dimensions can be applied?

To study the results of huge size image and long-term historic data particularly of spatial and temporal, is a tedious and time-consuming task. Spatial and temporal dimensions fusion via same SVM representation is a complex task.

3. How to make decision making easy and robust?

It is very difficult to response to a smaller change in data analysis mainly of huge size spatial image data and long term historical/temporal data of air parameters that make decision-making tough.

The proposed research is supported by Auckland Regional Council by providing spatio-temporal image and air parameter data of Auckland City and professional advices for air and environmental study.

This paper is organized as follows: Sect. 2 introduces related work and critical analysis for computational air and environmental studies. Section 3 provides the motivation for the research. Section 4 is devoted towards the research methodology of the research. Section 5 provides the research implications to real life case studies applications. The paper concludes with conclusion as in Sect. 6.

## 2 Related Work

### 2.1 Literature of Air and Environmental Studies

In the past, a variety of computational studies have been done for monitoring different aspects of air and environmental pollution. About smoke, water and forest pollution, Carlotto, Lazaroff and Brennan, in 1993 proposed to monitor environmental anomalies using Multispectral Imagery (MSI). In this research, variables causing problems to smoke, water and forest pollution were successfully identified [5]. On fire smoke detection, Li, Khananian, Fraser and Cihlar, studied neural network in 2001 to classify a scene into smoke, cloud or clear background, and generated continuous outputs to represent the mixture portions of these objects [6]. For investigating oil spill in ships, Solberg, Storvik, Solberg and Volden, in 1999 computed a set of features for each dark spot, and authenticated a spot as either an oil slick or a lookalike [7]. For environmental decision support related to urban traffic, Benvenuto and Marani et al. in 2001 conducted a study with an objective to provide direction for the design of an environmental monitoring system to estimate and predict the pollution values [8]. For predicting air temperature up to 12 h ahead, Smith and McClendon et al. in 2006 collected parameters on air temperature, solar radiation, wind speed, rainfall and relative humidity and then applied Artificial Neural Network (ANN) computing to obtain their results [9]. To investigate modern trends in monitoring and analysis of environmental pollutants, Namiesnik et al. in 2001 conducted a study to provide information required for a reliable evaluation of the state of the environmental pollution and the changes taking place [10]. For detecting and monitoring environmental anomalies and changes, Yang, Chenghai and Odvody, et al. in 2015 used spectral classification method, to identify specific areas in image for anomaly and change detection followed by knowledge based techniques to identify general categories based on spectral shape [11]. Dong, Huijuan and Dai introduced two models for immediate response to air pollutant problems, namely Asia-Pacific Integrated Assessment Model/Computational General Equilibrium (AIM/CGE CGE) model and Greenhouse Gas and Air Pollution Interactions and Synergies (GAINS) model. Both these models were combined together for future prediction of CO<sub>2</sub> and other air pollutantants [12].

### 2.2 Review of Computational Methodologies for Air and Environmental Pollution Analysis

In terms of computational methodology, a number of methods have been developed to explore various components of air pollutions. Among them, the neural network method is popularly used in this area. For example, Li and Khananian et al. developed neural networks and threshold classifier methods to identify potential areas covered by smoke and further used texture analysis and spatial filtration to

remove false classified pixels. Ando, Graziani and Pitrone proposed a black box approach consisting of linear, nonlinear and neural network models, for air pollution modelling, where air pollution concentration is predicted as a function of the expected causes, based on meteorological forecasts [13]. For time series prediction of air pollution, Castro, Castillo, Melin and Diazapplied Interval Type-2 Fuzzy Neural Network (IT2FNN) hybrid method to predict the impact of meteorological pollutants as ozone ( $O_3$ ) over an urban area [14]. Zito, Chen and Bell used several NNs, such as multilayer perceptron (MLP), radial basis function (RBF) and modular network (MN) to estimate real-time roadside CO and  $NO_2$  concentration [15]. For complicated environmental data processing, Osowski and Garanty used Support Vector Machine (SVM) plus wavelet decomposition methods for daily air pollution forecasting on  $NO_2$ , CO and  $SO_2$  dust pollutants [16]. Roadknight, Balls, Mills and Palmer-Brown used principal components analysis (PCA) together with standard multilayer perceptron (MLP) and multiple regression analysis for decision support in determining critical levels of ozone pollution [17]. It is noticeable that fuzzy computing is also being used for environmental study. Benvenuto and Marani et al. studied Sophisticated Fuzzy Interference Systems (FISs) to provide directions for the design of an environmental monitoring system to estimate and predict the pollutant values [8]. Arhami, Mohammad and Kamali used ANNs for the prediction of hourly air pollutants nitrogen oxides  $NO_x$ , carbon monoxide (CO), particulate matter  $PM_{10}$ , nitrogen monoxide (NO), nitrogen dioxide  $NO_2$  and ozone ( $O_3$ ) [18]. Computing prediction intervals (PI) and probability of exceeding air quality thresholds were developed based on ANNs and Monto Carlos Simulations (MCSs) for the prediction of hourly air pollutants.

### 3 Motivation of the Proposed Research

It is noticeable that all above researches are just pieces of isolated environmental research. However, air and environmental problems in nature is air spatio-temporal problem involving huge size static as well as timeline dataset. On such huge size spatio-temporal data, traditional methods reviewed above often confronts the difficulty of computational complexity. Further, decision making based on only spatial or temporal data analysis is not sufficient. Similarly, decision making to a small change in huge size data is difficult and slow. Hence, considering the above concerns, we believe to study the air pollution problem from a spatio-temporal perspective, SVM aggregation computation is the best solution.

## 4 Research Methodology

### 4.1 Part 1: A General Methodology for Pollution Pattern Analysis by SVM

The following flow chart in Fig. 1 will explain the methodology part of our research.

In the beginning of this research a database will be constructed based on our prior knowledge that will contain different types of pollution images, for example, water pollution, soil pollution and air pollution, etc. For a specific type of pollution a timeline can be drawn directly from the database. The basic principal of detection part is to find a change that is difficult to judge. Air pollution detection comprises of two tasks, first to find a minor change in air polluted pattern that is difficult to identify and secondly to monitor the size of air polluted image area. Air polluted image based on prior knowledge via classification of images from a large datasets will be autonomously selected. Further size of air polluted area will be monitored through image segmentation to locate boundaries of affected air polluted area. This will indicate how air pollution is affecting life around that area. The residual model will be applied to detect change in air pollution images. The main objective of the examination part is to get information from the referral database to know how good or bad the image is to present grade to consider it as a polluted image. To find the appropriate grade for each air polluted image processed, the image classification method will be utilized. Further this air pollution examination part will explore how serious air pollution problem is currently? and what the current causes of air pollution are? and how we can avoid current air pollution causes to obtain the normal

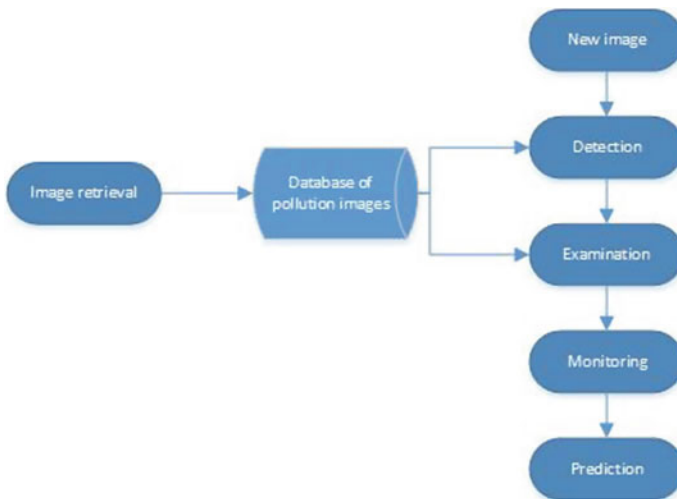


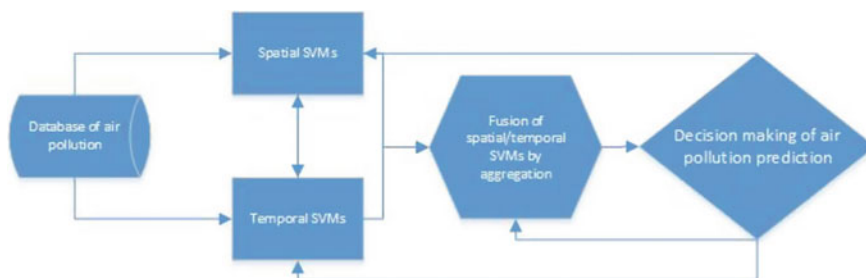
Fig. 1 Methodology for research

results? Further in this research, air polluted image will be monitored. The basic principal of monitoring is to detect small changes in the air pollution image. For monitoring air pollution images a timeline will be drawn from 2001 to 2006. Further, air pollution images will be processed, once a change is detected in the image pattern, it will highlight that there is a need to monitor that polluted region to prevent it from getting further polluted.

#### 4.2 Part 2: A Methodology for Integrated Spatial-Temporal Analysis by SVMs Aggregation

Spatial and temporal data will be obtained from air pollution database. Individual SVMs will be modelled on the spatial and temporal dimension, respectively. For decision making of air pollution problem in future, knowledge from different aspects of SVMs will be integrated. Feedback will be derived for spatial and temporal SVM's and to their fusion respectively based on decision making of air pollution prediction. The following Fig. 2 is an overview of our proposed computational approach for decision making of air pollution quality for future.

The basic principal of prediction is to forecast air pollution in future, with and without conditions. With conditions forecasting means that we can do something for air pollution to get things better. However, with no conditions forecasting means, we are not doing anything for air pollution control but it will give us early warning for air pollution. In the prediction part, air polluted images from years 2001 to 2006 and current year 2010 will be processed, then two predictions will be generated. In the first prediction, the conditions and factors contributing to air pollution are identified, and the forecast will assume that these conditions and factors contributing to air pollution are going to be controlled. In the second prediction, the forecast will assume that the conditions and factors contributing to air pollution are not controlled.



**Fig. 2** Proposed computational approach for decision-making of air pollution prediction

## 5 Case Studies of Real Applications

### 5.1 *Vehicle's Emission Analysis*

Excess of carbon emission from vehicles results in severe air pollution. In order to deal with this problem the data for this study is obtained from two sources Motor Industry Association (MIA), and Auckland Regional Council (ARC). The data from MIA constitutes of vehicle's models with their emission's rate of Carbon dioxide CO<sub>2</sub>. The statistics of emission of Carbon Monoxide (CO) and Nitrogen per Dioxide NO<sub>2</sub> in air on hourly basis each day is obtained from ARC for the year 2008. The research of spatial and temporal SVM aggregation will be conducted to analyze the emissions of carbon dioxide, carbon monoxide and nitrogen dioxide from vehicles. In this study, real engine performances, such as real fuel efficiency and emissions will be investigated against the optimal engine output; and a database of vehicle's models that releases CO<sub>2</sub> will be created based on the information provided by MIA. A vehicle under investigation will be evaluated on the basis of the constructed pollution knowledge database, where the CO and NO<sub>2</sub> emission of the vehicle will be calculated based on the ARC statistics. Once the vehicle is demonstrated to be low in performance and high in emissions against the predicted values, then the vehicle will be recognized as high emission of CO, CO<sub>2</sub> and NO<sub>2</sub>. (i.e. polluted vehicle towards air pollution).

### 5.2 *Auckland City Centre Air Pollution*

Air pollution is a major concern of municipality on harm and discomfort to humans. This study addresses the air quality of Auckland city centre. A database is constructed by ARC, where a series of air pollution images (captured by Advanced Very High-Resolution Radiometer over an area of 45 × 24 km region) focusing Skycity tower are captured from two different viewpoints: Takapuna and Arataki, and statistical values of CO, NO<sub>2</sub> and O<sub>3</sub> are calculated for each hour every day, whole years during 2001–2006. The research of spatial and temporal SVM aggregation will be conducted to analyse the air pollution problem of Auckland. In this regard, over a specific air pollution parameter, individual SVMs will be modelled on spatial and temporal dimension, respectively. Spatial SVMs address static pollution patterns. For example, a pollution pattern extracted from the Skycity tower image. Temporal SVMs summarizes the time line from 2001 to 2006 by support of vector regression. As decision making, knowledge from different aspects SVMs will be integrated to compose a future prediction. The fusion of SVMs will inform us of the seriousness of the air pollution problem in coming periods of time (weeks, months).

## 6 Conclusion

The proposed research poses two aspects research benefits, (1) a series of advanced new methods will be developed for air pollution analysis on the track of spatial and temporal SVM aggregation. The developed computational prototype software will present a clear picture to environmental monitoring authorities as a prediction tool for air pollution investigation; (2) air pollution problem is addressed systematically from detection to examination, monitoring and future prediction. This benefits eventually environmental monitoring authorities to identify polluted regions and to control the level of air pollution.

## References

1. S. H. Linder, D. Marko, and K. Sexton, "Cumulative cancer risk from air pollution in houston: disparities in risk burden and social disadvantage," *Environmental science & technology*, vol. 42, no. 12, pp. 4312–4322, 2008.
2. D. Whelpdale and R. Munn, "Global sources, sinks and transport of air pollution," *Air Pollution*, vol. 1, pp. 289–324, 2015.
3. A. S. Venkataramani and B. J. Fried, "Effect of worldwide oil price fluctuations on biomass fuel use and child respiratory health: evidence from guatemala," *American journal of public health*, vol. 101, no. 9, pp. 1668–1674, 2011.
4. L. Triolo, A. Binazzi, P. Cagnetti, P. Carconi, A. Correnti, E. De Luca, R. Di Bonito, G. Grandoni, M. Mastrantonio, S. Rosa *et al.*, "Air pollution impact assessment on agroecosystem and human health characterisation in the area surrounding the industrial settlement of milazzo (italy): a multidisciplinary approach," *Environmental monitoring and assessment*, vol. 140, no. 1–3, pp. 191–209, 2008.
5. M. J. Carlotto, M. B. Lazaroff, and M. W. Brennan, "Multispectral image processing for environmental monitoring," in *Applications in Optical Science and Engineering*. International Society for Optics and Photonics, 1993, pp. 113–124.
6. L. F. di Vito, "Neuro-fuzzy techniques to estimate and predict atmospheric pollutant levels," in *Neural Nets WIRN Vietri-01: Proceedings of the 12th Italian Workshop on Neural Nets, Vietri sul Mare, Salerno, Italy, 17–19 May 2001*. Springer Science & Business Media, 2012, p. 260.
7. A. S. Solberg, G. Storvik, R. Solberg, and E. Volden, "Automatic detection of oil spills in ers sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, no. 4, pp. 1916–1924, 1999.
8. M. Versaci, "Neuro-fuzzy techniques to estimate and predict atmospheric pollutant levels," in *Neural Nets WIRN Vietri-01*. Springer, 2002, pp. 260–265.
9. B. A. Smith, R. W. McClendon, and G. Hoogenboom, "Improving air temperature prediction with artificial neural networks," *International Journal of Computational Intelligence*, vol. 3, no. 3, pp. 179–186, 2006.
10. J. Namies'nik, "Modern trends in monitoring and analysis of environmental pollutants," *Pol. J. Environ. Stud.*, vol. 10, no. 3, p. 127, 2001.
11. C. Yang, G. N. Odvody, C. J. Fernandez, J. A. Landivar, R. R. Minzenmayer, and R. L. Nichols, "Evaluating unsupervised and supervised image classification methods for mapping cotton root rot," *Precision Agriculture*, vol. 16, no. 2, pp. 201–215, 2015.

12. H. Dong, H. Dai, L. Dong, T. Fujita, Y. Geng, Z. Klimont, T. Inoue, S. Bunya, M. Fujii, and T. Masui, "Pursuing air pollutant co-benefits of co 2 mitigation in china: a provincial leveled analysis," *Applied Energy*, vol. 144, pp. 165–174, 2015.
13. B. Ando, S. Baglio, S. Graziani, and N. Pitrone, "Models for air quality management and assessment," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 30, no. 3, pp. 358–363, 2000.
14. J. Castro, O. Castillo, P. Melin, and A. Rodriguez-Diaz, "A hybrid learning algorithm for interval type-2 fuzzy neural networks in time series prediction for the case of air pollution," in *Fuzzy Information Processing Society, 2008. NAFIPS 2008. Annual Meeting of the North American*. IEEE, 2008, pp. 1–6.
15. P. Zito, H. Chen, and M. C. Bell, "Predicting real-time roadside co and concentrations using neural networks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 9, no. 3, pp. 514–522, 2008.
16. S. Osowski and K. Garanty, "Wavelets and support vector machine for forecasting the meteorological pollution," in *Proceedings of the 7th Nordic Signal Processing Symposium-NORSIG 2006*. IEEE, 2006, pp. 158–161.
17. C. M. Roadknight, G. Balls, G. Mills, and D. Palmer-Brown, "Modeling complex environmental data ieee transactions of neural networks vol. 8," *Month July*, 1997.
18. M. Arhami, N. Kamali, and M. M. Rajabi, "Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte carlo simulations," *Environmental Science and Pollution Research*, vol. 20, no. 7, pp. 4777–4789, 2013.



# A Case Study in R to Recognize Human Activity Using Smartphones

Kella Bhanu Jyothi and K. Hima Bindu

**Abstract** Physical activity recognition is a growing area of research with many applications in medical, surveillance systems and manufacturing industry. We perform a case study to classify the human activity into six categories—Standing, Walking, Walking\_upstairs, Walking\_downstairs, Sitting and Lying using Random Forest algorithm in R. The dataset is of high dimensional numeric data, this report focuses on two Preprocessing methods—Principal Component Analysis and Near zero variance with removal of correlated predictors to identify the best suitable data reduction techniques. This study highlights the Influence of preprocessing, the procedure to fine tune the model.

**Keywords** Physical activity recognition · Surveillance systems  
Categories · Manufacturing industry · Random forest · PreProcessing

## 1 Introduction

Human activity recognition is used to recognize the activities of a human by analyzing the gestures/movements of the persons. Accurate activity recognition is challenging because human activity is complex and highly diverse [1]. The activity recognition is an automated interpretation of on-going events from video data.

Recognition can be accomplished by exploiting the information retrieved from various sources such as environmental [2] or body-worn sensors [2, 3]. Some approaches have used dedicated motion sensors in different body parts such as the wrist, waist, thighs, and chest for achieving good classification performance [4]. These sensors are usually uncomfortable to wear by the humans for long period.

---

K. Bhanu Jyothi (✉) · K. Hima Bindu  
Department of Computer Science and Engineering, Vishnu Institute of Technology,  
Bhimavaram, India  
e-mail: bhanu.kella@gmail.com

K. Hima Bindu  
e-mail: himagopal@gmail.com

Smartphones are bringing up new research opportunities, which are very popular and becoming the central computer and communication device in people's lives and these are emerging technologies for human-centered applications where the user is a rich source of context information and the phone is the first-hand sensing tool [5]. Latest devices come with embedded built-in sensors such as accelerometer, digital compass, GPS, dual cameras, gyroscope, etc. [5, 6].

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations [1]. It provides typical information about the identity of a person, their personality, and psychological state.

Many different applications have been studied by researchers in activity recognition [7]; Explained few applications of it-

- Surveillance Systems—The activity recognition is essential for surveillance and other monitoring systems in public. The present Surveillance Systems are mainly used for recording and it is monitoring by a human which is cost effective. In future the system, will auto recognize the suspicious activities performed by the humans [8].
- Ambient Assisted Living—To monitor the old people who are sick, children and patients human activity recognition is useful. Based on their activities, the system is recognizing what they are doing [7].
- Monitoring the daily activities of elderly people, so that they can live independent.
- Sports Play Analysis—Based on the player's action, the system identifies how the players are playing in the field [8].
- Car Assisting System—By Image and Location processing, the car management system is used to assist driver by providing model for outer environment with support of cameras, GPS and other sensors [7, 8].
- Industry Manufacturing, Assisting—The activity recognition techniques could also assist workers in their daily work. Wearable computing is a kind of extension of the body which allows a worker to perform extraordinary tasks [7].

## 2 Related Work

### 2.1 Description of HAR Dataset

Publicly available Human Activity Recognition dataset is taken from the UCI Machine Learning Repository [9] for experimental validation. The dataset was recorded with 30 subjects, performing six different activities: walking, ascending stairs, descending stairs, sitting, standing, and lying down. The waist mounted smartphone (Samsung Galaxy S II) embedded with 3D-accelerometer and 3D-gyroscope was used to collect data at 50 Hz. More information about the dataset can be found in [10].

The dataset has been partitioned into two sets, where 70% of the volunteers, i.e., 21 subjects were selected to generate the training data and 30% means 9 volunteers for the test data. As well the dataset includes Activity labels and 561 features data. Overall the HAR (Human Activity Recognition) dataset includes 10,299 records and the train data contains 7352 records and the test contains 2947 records.

Each record consists of following attributes [9]:

- Triaxial acceleration from the accelerometer and the estimated body acceleration
- Triaxial Angular velocity captured from the gyroscope
- Each vector in the dataset has 561-features with time and frequency domain variables and activity label.
- Each vector is identified by the identifier of the subject who carried out the experiment

## ***2.2 Existing Models (Algorithms Previously Used to Build the Model on HAR Dataset)***

Many supervised and unsupervised machine learning algorithms are used to recognize the activities—Walking, Sitting and Standing [11] of human by using wearable sensors. Namely k-Nearest Neighbor (k-NN), Gaussian Mixture Models (GMM), Support Vector Machines (SVM), and Random Forest (RF), k-Means, Hidden Markov Model (HMM), Decision Tree and Naïve Bayes Algorithms are applied on the HAR dataset to calculate the accuracy but depends on the features selected the accuracy rate varies. In this case, k-NN—90.1% and HMM—90.2% methods gave the better accuracy compared to other Supervised and Unsupervised algorithms [11].

The Human activity recognition is performed by using multi-class support vector machine method [12] on the same dataset what we used in this paper to build the model and observed the accuracy as 89.3% for 789 samples.

## ***2.3 Proposed Model—Random Forest***

In this paper, we have presented the human activity recognition using Random Forest. Random Forest is one of the classification techniques. Classification is used to assigns items in a collection to target categories or classes [13]. The aim of the classification is to accurately predict the target class for each case in the data.

Random forest is an ensemble classifier that presents of many decision trees and outputs the class which is the mode of the class's output by individual trees [13]. It is one of the most accurate learning algorithms available for many datasets; it produces a highly accurate classifier and runs efficiently on large databases. It can

handle thousands of input variables without variable deletion also; it provides estimates of which variables are important during the classification. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

### 3 Methodology

#### 3.1 Preprocessing

The dataset is of two parts—one part of the dataset is used to develop a model, i.e., Train Dataset (70% of the dataset) and the other is Test Dataset (30% of the dataset) which is used to evaluate the model's performance. HAR data has 561 numeric attributes. The accuracy calculation for this large dataset of 561 attributes is quite typical. In order to get proper accuracy, preprocessing of the data is necessary. Data Preprocessing is a data mining technique that involves transforming the raw data into understandable format [13].

This study observes which data reduction technique suits for this high dimensional numeric data. PCA and zero & near zero variance with the correlation predictors methods were studied to identify the best suitable data reductions technique. The preprocessed dataset are separately used to build the random forest models to check on which model the high accuracy is.

**Near Zero Variance with Correlated Predictors.** Zero Variance removes the attributes which have same unique value across samples [14]. Here, the zero Variance method is applied on the train dataset by using 'preProcess' function in Caret Package [15]. After that the dataset is applied to remove the correlated predictors from the train dataset to get the quality data, using the method 'findCorrelation' in which the predictors which are correlated at the rate 0.95 are removed. The generated dataset contains 277 variables and the correlated predictors are 284 out of 561 attributes. The code we used to preprocess the train dataset is

```
zscaleTrain = preProcess(train[, 1:numPredictors],
method = "zv")
scaledX = predict(zscaleTrain, train[, 1:numPredictor])
correlatedPredictors = findCorrelation(cor(scaledX),
cutoff = 0.95)
reducedCorrelationX = scaledX[, -correlatedPredictors]
```

**Principal Component Analysis.** This is a method of extracting important variables (in form of components) from a large set of variables available in a data set. It taken out low-dimensional set of features from a high dimensional data set with a motive to capture as much information as possible. PCA is more useful when

dealing with three or higher dimensional data. Principal Component Analysis is applied by using the method ‘pca’ in preProcess function [15] with the threshold 0.99 which keeps the components above the variance. The code we used to preProcess the train dataset is

```
pcaTrain = preProcess(scaledX, method = "pca", thresh =
0.99)
pcaX = predict(pcaTrain, scaledX)
```

By this method, the dataset is generated with 179 principal components from 561 attributes of the train dataset.

### 3.2 *Build the Model*

Now the dataset is preprocessed by above two techniques the obtained train datasets is applied to Random Forest algorithm.

The Random Forest algorithm is applied on the dataset by using k-fold cross-validation method [15]. In k-fold cross-validation, the dataset is divided into ‘k’ sub parts and the single sub-sample is retained as a test part and the rest k-1 sub-samples are treated as train data. The cross-validation process is then repeated k times (the folds), and each observation should be used in validation exactly once. The k results from the folds are averaged to produce a single estimation [15]. The ‘k’ value chosen is 10 here.

The Random Forest model creating for the HAR train dataset is by using train() function in Caret package reference [16] of R programming Language. The caret package stands for Classification And Regression Training, contains set of functions that attempt to streamline the process for creating predictive models. The train() function includes many machine learning methods to build the models on a dataset [17]. Here, we used ‘rf’ method in that to develop a model for the dataset. ‘rf’ stands for Random Forest. Before to apply train() method on the dataset, trainControl() function is applied, it is used to generate parameters that further control how models are created.

The accuracy is how often the model trained is correct, which is represented by using Confusion Matrix. A confusion matrix is a summary of prediction results on a classification problem [18]. The number of correct and incorrect predictions are summarized with count values and broken down by each class.

**Tuning the classifier.** We build the random forest model using different mtry (the number of variables randomly sampled at each split) values—2, 5, 10, 15, 20 by train() function using ‘rf’ method on the two preprocessed train datasets.

*Accuracy of the Random Forest model on preprocessed train dataset by method —a.* The accuracy values of the Random Forest model for different mtry values on the train is represented in the Fig. 1.

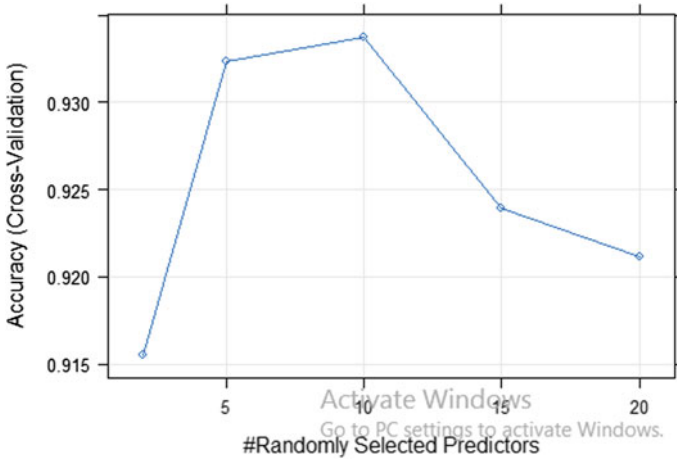


Fig. 1 Accuracy graph for different mtry values

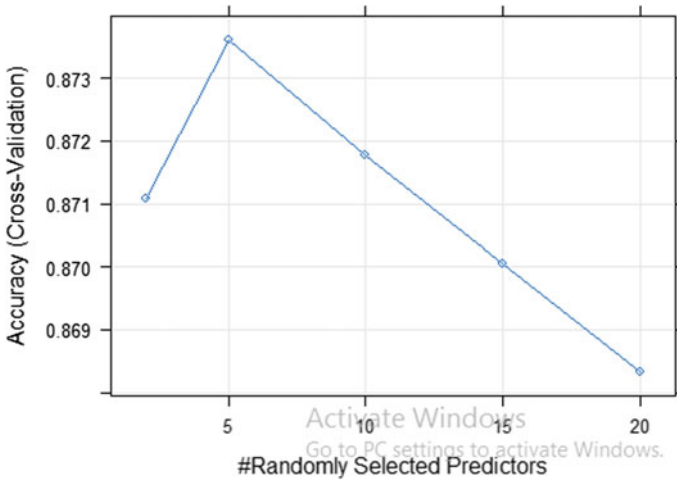


Fig. 2 Accuracy graph for different mtry values

At the mtry value 10, the accuracy of the model is high, i.e., 93.66% which is the highest in all trails and it is used to build the final model.

Accuracy of the Random Forest model on preprocessed train dataset by method—b. The accuracy values of the Random Forest model for different mtry values on the preprocessed train dataset using Principal Component Analysis is represented in the Fig. 2.

**Fig. 3** Work-flow diagram to build the model



At the mtry value 5, the accuracy of the model is high, i.e., 87.36% which is the highest in all trails and but we used mtry = 10 to build the final model to compare with the model-1 performance.

The work-flow will be takes place to analyze Human Activity Recognition dataset using Random Forest algorithm is presented in Fig. 3.

## 4 Results

Using tenfold cross-validation the accuracy we received for the Random Forest model at mtry value 10 is 93.66% on the train dataset which is preprocessed using zero and near zero variance with Correlation Predictors removal and 87.17% with the Principal Component Analysis preprocessed train dataset. Now the test dataset is predicted to preprocess by both the methods what we used into train a model and applied the ‘rf’ method on the preprocessed test datasets to predict the model. By using confusion matrix, the predicted classes are compared with the test dataset labels which we already with us in the dataset.

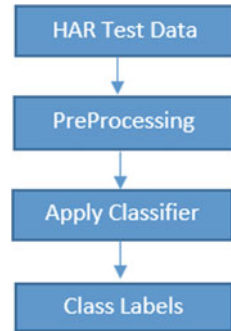
A confusion matrix is a table that is used to describe the performance of a classification model on a set of test data for which the true values are described [19]. By ‘rf’ method, the accuracy of the test dataset preprocessed using ZV with the removal of Correlation Predictors is determined as 94.91% and 89.72% with the PCA preprocessed test dataset.

**Table 1** Comparison of model performances

PreProcessing	Principal component analysis										Zero variance with correlated predictors							
Method	Random forest																	
Activity	Walking	Upstairs	Downstairs	Standing	Sitting	Lying	Walking	Upstairs	Downstairs	Standing	Sitting	Lying	Walking	Upstairs	Downstairs	Standing	Sitting	Lying
Walking	479	30	59	0	0	0	493	23	6	0	0	0	493	23	6	0	0	0
Upstairs	1	431	36	0	1	0	0	443	43	0	1	0	0	443	43	0	1	0
Downstairs	16	10	325	0	0	0	3	5	371	0	0	0	3	5	371	0	0	0
Standing	0	0	0	501	105	2	0	0	0	519	54	0	0	0	0	519	54	0
Sitting	0	0	0	31	383	10	0	0	0	13	434	0	0	0	0	13	434	0
Lying	0	0	0	0	2	525	0	0	0	0	2	537	0	0	0	0	2	537
Accuracy (%)	89.72										94.91							



**Fig. 4** Work-flow diagram on test dataset



The comparison of two random forest models is shown in Table 1 and the work-flow to evaluate the model on the test dataset using the classifier build on train dataset is shown in Fig. 4.

## 5 Conclusion

The goal of the activity recognition is to analyze the common human activities in daily life. Accurate activity recognition is difficult because human activity is complex and highly diverse. The Smart phones played key role in activity recognition because of its features like accelerometers, gyroscopes, GPS, barometers, etc., the smart phones easily detecting the body motion and providing the data of human gestures in tri axial. The data we received is analyzed by the machine learning algorithms to recognize the activity of a human like what he is doing. Here we used the Random Forest algorithm to develop the Activity Recognition System on the provided Human Activity Recognition dataset which is publicly available; the dataset contains the 6 activities of 30 subjects, namely—Sitting, Standing, Walking, Walking\_upstairs, Walking\_downstairs and Lying. The dataset is pre-processed by two techniques here—zero variance with the correlated predictors and principal component analysis and the model performance is compared. The accuracy of the models evaluated on the preprocessed data using Random Forest is 94.91% and 89.72%. From the classification results, we conclude that near zero variance method is preferable than PCA for data reduction.

## References

1. Ronao, Charissa Ann, and Sung-Bae Cho. "Human activity recognition using smartphone sensors with two-stage continuous hidden Markov models." *Natural Computation (ICNC), 2014 10th International Conference on*. IEEE, 2014.

2. R. Poppe. Vision-based human motion analysis: An overview. *Computer Vision and Image Under-standing*, 108(1–2):4–18, 2007.
3. P. Lukowicz, J.A. Ward, H. Junker, M. Stager, G. Troster, A. Atrash, and T. Starner. Recognizing workshop activity using body worn microphones and accelerometers. *Proceedings of the 2nd Int Conference Pervasive Computing*, pages 18–22, 2004.
4. D.M. Karantonis, M.R. Narayanan, M. Mathie, N.H. Lovell, and B.G. Celler. Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring. *IEEE Transactions on Information Technology in Biomedicine*, 10(1):156–167, 2006.
5. R. Nishkam, D. Nikhil, M. Preetham and M.L. Littman. Activity recognition from accelerometer data. In *Proceedings of the Seventeenth Conference on Innovative Applications of Artificial Intelligence*, pages 1541–1546, 2005.
6. Yang, Rong, and Baowei Wang. “PACP: A Position-Independent Activity Recognition Method Using Smartphone Sensors.” *Information* 7, no. 4 (2016): 72.
7. Sunny, J.T., George, S.M., Kizhakkethottam, J.J., Sunny, J.T., George, S.M., & Kizhakkethottam, J.J. Applications and Challenges of Human Activity Recognition using Sensors in a Smart Environment. *International Journal*, 2, 50–57.
8. *Frontiers of Human Activity Analysis*. [http://michaelryoo.com/cvpr2011tutorial/tutorial\\_cvpr2011\\_intro.pdf](http://michaelryoo.com/cvpr2011tutorial/tutorial_cvpr2011_intro.pdf).
9. Human Activity Recognition using SmartPhones Dataset. <https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>.
10. Human Activity Recognition using SmartPhones Dataset. [http://rstudio-pubstatic.s3.amazonaws.com/24009\\_c068b79c74ae4fec8913fc0bf7a8b451.html](http://rstudio-pubstatic.s3.amazonaws.com/24009_c068b79c74ae4fec8913fc0bf7a8b451.html).
11. Physical Human Activity Recognition using Wearable Sensors. <http://www.mdpi.com/1424-8220/15/12/29858/pdf>.
12. Anguita, Davide, et al. “Human Activity Recognition on Smartphones using a Multiclass Hardware-Friendly Support Vector Machine”.
13. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
14. Building Predictive Models in R using the Caret Package. [http://web.ipac.caltech.edu/staff/fmasci/home/astro\\_refs/BuildingPredictiveModelsR\\_caret.pdf](http://web.ipac.caltech.edu/staff/fmasci/home/astro_refs/BuildingPredictiveModelsR_caret.pdf).
15. Data PreProcessing. <https://www.rdocumentation.org/packages/caret/versions/6.0-73/topics/preProcess>.
16. Arlot, Sylvain, and Alain Celisse. “A survey of cross-validation procedures for model selection.” *Statistics surveys* 4 (2010): 40–79.
17. Kuhn, Max. “Predictive Modeling with R and the caret Package”.
18. Data PreProcessing. <https://topepo.github.io/caret/available-models.html>.
19. Visa, S., Ramsay, B., Ralescu, A.L. and Van Der Knaap, E., 2011, April. Confusion Matrix-based Feature Selection. In *MAICS* (pp. 120–127).

# Big Data Collection and Correlation Analysis of Wireless Sensor Networks Yielding to Target Detection and Classification

M. Giri and S. Jyothi

**Abstract** Wireless Sensor Networks (WSNs) is an overwhelming computing field. The sensors are responsible for sensing the data in the field in which the sensors are deployed. For each category of sensor, large numbers of sensors are deployed. These sensors transmit the data along with certain other data like energy, life time and all that. In this context, it is intended to collect data in a big way and find the correlation between specified categories of data in target detection. A new protocol is developed which determines the correlation analysis of parameters like velocity, speed, and distance fields. Using these data the target can be detected. On detection, it is intended to control the missiles for further security. Thus, the wireless sensor networks enhance the security features of the communication field. Further it is also intended to classify the targets based on the data received, using the novel algorithm, as attacking, losing, strengthening, alarming, and winning. The life time of the sensors is also sent along with the data which is used to increase the life time by enabling the sensor recharging and enhance big data sensing in a wider manner all through the life time.

**Keywords** Wireless sensor networks • Big data • Correlation analysis  
Target detection • Life time • Protocol

## 1 Introduction

Big data processing is quite an interesting and increasingly popular domain in the computer field. Big data is characterized by its volume, which is huge in terms of data collection. Wireless Sensor Networks employ sensors like velocity sensors,

---

M. Giri (✉)

Department of CSE, Rayalaseema University, Kurnool, Andhra Pradesh, India  
e-mail: m.giri@gmail.com

S. Jyothi

Department of Computer Science, SPMVV University, Tirupati, Andhra Pradesh, India  
e-mail: jyothi.spmvv@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_18](https://doi.org/10.1007/978-981-10-6319-0_18)

acceleration sensors, speed sensors, and distance sensors for target detection and classification domain. These sensors are considered in an enormous amount in the geo network and they are deployed by various strategies like dropping, random deployment, spreading, independent planting, and so on. These sensors are intelligent on its nature. They sense the target in the geo synchronous domain and transmit the data at a distance not more than 1000 km from the sensor field. The moving target moves across the sensors randomly and it is abundant in nature. The sensors pertaining to the category of sensing senses the data along with the target id, distance, nature of the target, source and direction of motion. These targets are sensed by the corresponding sensor periodically may be one in one second and then transmitted wirelessly to the control center. These data are gathered and organized as big data for the corresponding category of sensing.

The big data that is gathered is received, processed for errors, ignored for errors, refined for storage and stored in a designed data base. The storage is large and the big data is preprocessed. The preprocessing step includes ignoring null values and very large values. The values within the admissible range are detected and stored as big data. Here the application that is considered is target detection and classification. To detect and classify the target, it is intended to determine the correlation between acceleration, speed, velocity and distance. This correlation determines the movement of the target in the wireless zone. Thus, the overall correlation yields the detection of the target. The protocol which is designed anew classifies the target thereon.

## 2 Related Work

In [1], authors have worked on Real-Time Processing Algorithms for Target Detection and Classification in Hyper spectral Imagery in 2001 and it deals with novel algorithms but it deals with images. In [2], authors have worked on supervised target detection and classification by training on augmented reality data but it is a Neural Network based concept which is not pertaining to our current context. In [3], authors have worked on An Energy-Efficient Big Data Gathering Algorithm for WSN and it deals with preliminary data gathering algorithm. In 2011, in [4], authors have worked on Symbolic Dynamic Filtering of Seismic Sensors for Target Detection and Classification. It is an inferior method. Hence, we switch to the proposed method. In [5], authors have done a research work on Drawing Dominant Dataset from Big Sensory Data in Wireless Sensor Networks. It deals with Big Data but not target detection and classification.

In [6], authors have done a work on Extracting Kernel Dataset from Big Sensory Data in Wireless Sensor Networks. Again it does not deal with target detection and classification. In [7], authors have has worked on A Real-Time Big Data Gathering Algorithm Based on Indoor Wireless Sensor Networks for Risk Analysis of Industrial Operations. This work deals with big data gathering algorithm but not correlated to the current work. In [8], authors have has worked on Data and

Energy-Integrated Communication Networks for Wireless Big Data. It does not again deal with our current work. In [9], authors have worked on Improving Constrained Single and Group Operator Placement Using Evictions in Big Data Environments. This deals with operator placement but not target detection and classification. In [10], authors have worked on ELDC: An Artificial Neural Network based Energy-Efficient and Robust Routing Scheme for Pollution Monitoring in WSNs. It is an ANN based method which is a very old method.

In [11], authors have carried out a work on toward a CDS-based Intrusion Detection Deployment Scheme for Securing Industrial Wireless Sensor Networks. It deals with security issues which are not the current topic of discussion. In [12], authors have done a research work on Recent Development in Big Data Analytics for Operations and Risk Management. It deals with the operational statistics of big data. But we expect the discussion to be based on target detection and classification. In [13], authors have done a research work on Big Data Behavioral Analytics Meet Graph Theory: On Effective Botnet Takedowns. This deals with graph theory, whereas the current context is WSN. In [14], authors have worked on A Mechanism Filling Sensing Holes for Detecting the Boundary of Continuous Objects in Hybrid Sparse Wireless Sensor Networks. It deals with boundary detection which is not our current work. In [15], authors have done a work on A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud. It deals with detecting errors alone in big data processing. Hence, it is intended in the current context to introduce the proposed research on formulating a WSN with speed, acceleration, distance and velocity sensors, allowing targets to have mobility in the vicinity of WSN. Then sensing, transmission of data and storage occurs. BDCCA and TDACA are two algorithms of current context which deals with correlation analysis of big data, target detection, and classification. Then a detailed analysis of the results is carried out and our method is proved to be the best method for the undertaken research.

### 3 The WSN Target Detection and Classification Model

WSN is the heart of the Target Detection and Classification Model. In real time the motes on acceleration, speed, velocity and distance are deployed in the field of sensing. The mote senses the data and transmits intelligent data to the control center. The control center receives the data and processes the data and decision are taken for target detection and classification. Figure 1 depicts the scenario in the WSN world.

The category of motes under consideration is Speed, Velocity, Acceleration, and Distance. These motes are deployed in a random manner. These motes interact and sense the data and forward it to the control center. The structure of motes transmission is shown in the Fig. 2. The structure encompasses the fields like node id

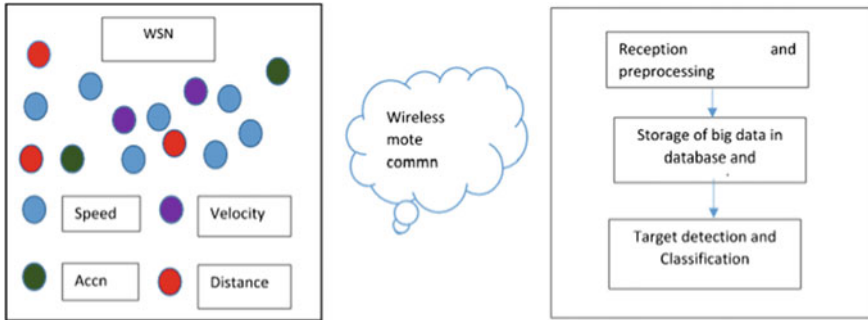


Fig. 1 WSN target detection model

Node id	Cat. of mote	periodicity	life time	Distance of Deployment	wireless media	Data
---------	--------------	-------------	-----------	------------------------	----------------	------

Fig. 2 Motes data transmission structure

which indicates the unique id of the transmitting mote. The next field is the type of mote. This field indicates the mote category like speed, velocity and so on. Next field is the periodicity field which indicates the cyclic period for data transmission which may be set to one sec, two sec, and so on. Another field in the transmission structure is the life time of the motes. This indicates the residual energy in the mote. If the life time is closure to finish condition the mote will be put in the recharge mode. Next field in the mote structure is the distance of deployment from the center to which the data is transmitted. Another field is the wireless transmission media field like ZIGBEE, Bluetooth and so on. Finally, the data itself is transmitted.

In the current context, the control center storage is also discussed. The data that is collected from the motes is saved in the form a database. Few mandatory fields are discussed here. Figure 3 shows the view of the database structure in the control center for the velocity mote. The database view encompasses the fields like category of mote, unique node id, time of sensing, life time of the sensor, distance of the sensor from the target, target identifier, and speed of the target. Also it includes the location of the target, height of the target above the ground level, direction of the motion of the target, and the dynamic orientation of the target pertaining to the current sensing. Apart from this, the big data collects several other data types. Here only the mandatory fields are taken into account.

Cat. Of Mote	Node id.	Time Periodicity	Life Time	Distance of Deployment	Target id	speed	location	Height	Direction	Orientation	...
--------------	----------	------------------	-----------	------------------------	-----------	-------	----------	--------	-----------	-------------	-----

Fig. 3 Sample view of the big data collected for the velocity sensor

### 3.1 *Big Data Collection\_Corr\_Analysis\_Algorithm\_WSN (BDCCAAW)*

- Step 1: Start
- Step 2: for i = 1 to 4 {
- Step 3: for j = 1 to no\_of\_sensors {
- Step 4: random\_deploy (i, j) } };
- Step 5: for i = 1 to 4 {
- Step 6: for j = 1 to no\_of\_sensors {
- Step 7: capture\_data (i, j)
- Step 8: encapsulate (cat, j, time, periodicity, life\_time, dist, target\_id, i, loc, height, direction, orientation, target\_type, network\_topology, packet\_id, seq\_no, protocol)
- Step 9: transmit (packet\_id, send\_id, ccenter\_id, receive\_id, database)
- Step 10: preprocess (packet\_id, database)
- Step 11: Store (packet\_id, database)
- Step 12: corr\_analysis (packet\_id, database) } };
- Step 13: End

#### 3.1.1 Random Deploy Function

In this function, it is taken into account the four categories of sensors like velocity, speed, acceleration, distance. There may be many WSNs for target detection and classification. Each WSN encompasses a random number of sensors for each category of sensors specified. Depending upon the nature of target the deployment of WSN is designed. The distance and location of WSN are suitably designed. The deployment of sensors in WSN is in the open field for the targets and random, spreading, dropping, throwing, and so on. Thus, the sensors are suitably deployed in the field and construct a WSN.

#### 3.1.2 Capture Data Function

The sensors are initialized with random values. Then the sensors are invoked for sensing. The sensor senses the data from the environment.

### 3.1.3 Encapsulate Function

The data that is gathered is constructed in the form of the packet. Such a concept is called encapsulation. The packet consists of header field as well as the big data fields. The header fields include predominantly the category, node id, center id, protocol, and so on. The data fields include time, periodicity, life\_time, dist, target\_id, loc, height, direction, orientation, target\_type, network\_topology and son. The category and node id are variables. Thus, the packet formulation is called Encapsulation.

### 3.1.4 Transmit and Preprocess

The data that is collected and encapsulated as packet is transmitted using the wireless protocol like ZIGBEE or Bluetooth or GSM or GPS is transmitted to the control center where it is received with identity recognized. The received packet is checked for errors in the received fields. If errors are detected the packets are ignored and next packet is collected and processed or other sensors data is encountered currently. Else if there is no error, data preprocessing takes place. In the preprocessing mode, the data is checked for extreme values present. If it is then it is ignored. If the data is in the admissible format then it is stored. The null fields in the packet are ignored and next data is collected. The database stores only the meaningful and useful data about the WSN. This is the significance of data preprocessing.

### 3.1.5 Storing Data

A suitable database is chosen for storing the data. The database category ranges from simple excel sheet, Cassandra, MongoDB, OrientDB, Apache HBase, Neo4j, CouchDB, Terrastore, FlockDB, Hibari, Riak, Hypertable, Blazegraph, Hive, ICE, Infinispan to Redis. The nature and selection of database is dependent on the application chosen. For WSN ICE is the suitable database. The other databases can also be adopted. After choosing the database, the attributes are designed and the packet is converted into the database. The big data database is stored in the cloud for enhanced scalability and virtualization. No\_Sql is the database query language dealing with big data.

### 3.1.6 Correlation Analysis

The category of data considered in the current context is speed, velocity, acceleration, and distance. The targets operation is based on these data. The speed reflects the moving target speed and it has its effect on other parameters. The other attributes are acceleration of the moving target, velocity of the moving target, and the



distance from the sensor of the corresponding target. As the target is a high speed one, the sensors detect the targets in a faster manner as it moves by. If the speed is high, then the target is at long distance from the goal. Else if the speed is low and the acceleration is high, the target is closer to goal. If the acceleration is high then the velocity is high and hence the target is far off from the goal and it is closer if the velocity is considerably reduced. The distance field indicates the target distance from the sensor. Thus, the issues like correlation are discussed.

### 3.2 Target\_Detect\_And\_Classify\_Algorithm (TDACA)

1. Start
2. For i = 1 to 4 {
3. For j = 1 to no\_of\_sensors {
4. Read\_data (i, j) } }
5. For i = 1 to 4 {
6. For j = 1 to no\_of\_sensors {
7. If (velocity<=Vmin, Accn>Amax, Dist<=Dmin, Speed>Smax) then
8. The target is closer to goal and it is classified as Attacking Target
9. Else if (velocity<=Vmax, Accn<=Amin, Dist>=Dmax, Speed<=Smin) then
10. The target is away from goal and it is classified as Losing Target
11. Else if (Vmin<=velocity>=Vmax, Amin<=Accn>=Amax, Dmin<=Dist>=Dmax, Smin<=Speed>=Smax) then
12. The target is half way to goal and it is classified as Strengthening Target
13. Else if (velocity>=Vmax, Accn>=Amax, Dmin<=Dist>=Dmax, Speed>=Smax, height=Hmax, orientation=Omax) then
14. The target is almost toward the goal and it is classified as Alarming Target
15. Else if (velocity<=Vmin, Accn>Amax, Dist<=Dmin, Speed>Smax then
16. For q = 1 to m attackers {
17. The target is closer to goal and it is classified as Winning Target}
18. Else Target type is Unknown } }
19. End

#### 3.2.1 Representation of Classification Algorithm

Target detection and classification is an important issue in the WSN. After detecting the target classification of the target is performed and it is mandatory. The target classes are Attacking, Losing, Strengthening, Alarming and Winning. Figure 4 depicts the classification scenario. The attributes considered for classification typically is Velocity, Acceleration, Speed, Distance, Height, and Orientation. The rules for the target classification are written in the algorithm.

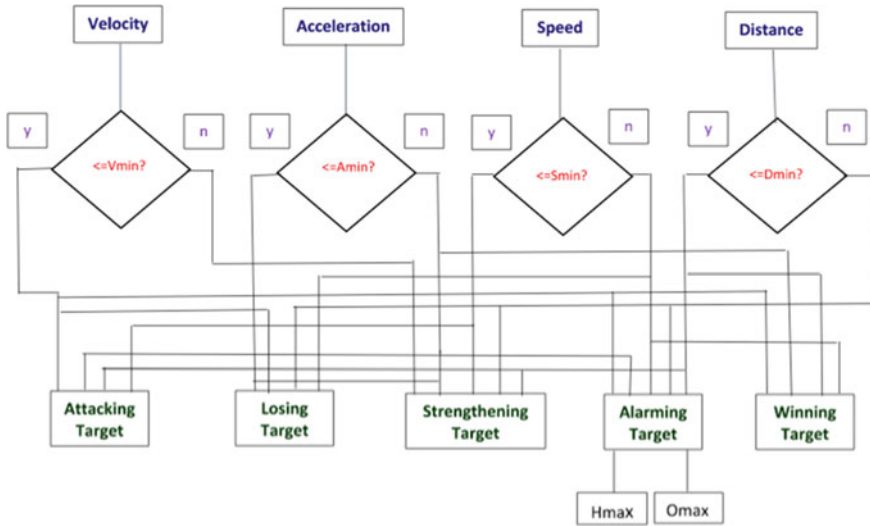


Fig. 4 Target detection and classification system

The first rule specifies if the velocity is less than or equal to minimum velocity and acceleration is greater than the maximum value for the sensor and the Distance is less than or equal to minimum distance of the target from the sensor and Speed is greater than the specified maximum then the target is closer to goal and it is classified as Attacking Target. If the velocity is less than the maximum value and acceleration is less than or equal to minimum value and distance is greater than or equal to the maximum value and speed is less than or equal to the minimum speed then the target is away from goal and it is classified as Losing Target. If the velocity lies between maximum and minimum values and acceleration lies between minimum and maximum values and the distance lies between the minimum distance and the maximum distance and Speed lies between the maximum and minimum value then the target is half way to goal and it is classified as Strengthening Target.

The next rule specifies that if the velocity is greater than or equal to the maximum value and acceleration is greater than or equal to the maximum value and Distance lies between the maximum and minimum values and Speed is greater than or equal to the maximum value and height is equal to maximum value and orientation is equal to the maximum orientation value, then the target is almost toward the goal and it is classified as Alarming Target. The rule for winning target classification is encompassing a loop which runs for one to m number of attackers where the velocity is less than or equal to the minimum value and acceleration is greater than maximum value and Distance is less than equal to minimum distance and Speed is greater than the maximum value then the target is closer to goal and it is classified as Winning Target. Thus, the TDACA description is provided in the text above.

### 4 Results and Analysis

The WSN model is constructed using the sensors deployed in real time in the field ranging from speed sensors, acceleration sensors, distance computing sensors and velocity sensors. These sensors are deployed in road side near toll centers and traffic signaling centers and other nomadic area as preferred. The sensors encompass the GSM motes and they sense the data pertaining to the target and transmit this data to the data collection centers. At the other end, the software is written to monitor and receive the motes data as set periodically as big data. This big data is preprocessed as specified earlier, refined and stored in excel spread sheet in various columns. Cloud storage is also supported for bulkiness, scalability, and virtualization. The data is processed and control decision is taken on the moving target. Then the classification of the goal target is done using this model. Figure 5 depicts the graph for collection of time, periodicity, life time and distance against the number of sensors. These parameters are mandatory for mote packet transmission. Figure 6 displays the big data on location, height, direction, and orientation of target against the number of sensors which is sensed and transmitted by GSM. Figure 7 shows the plot of number of sensors against various mote parameters like node\_id, target\_id, target\_type, packet\_id, Sequence\_no and Protocol.

The protocol varies the number of sensors and records these parameters and plots the graph. As discussed earlier the four categories of sensors which are speed, distance, acceleration, and velocity. These data are flowing in abundance from various sensors. The correlation can be found from the graph which is depicted in Fig. 8. Figure 9 depicts the target classification against the various classifiers. Before classification, the targets are perfectly detected. The target classes are

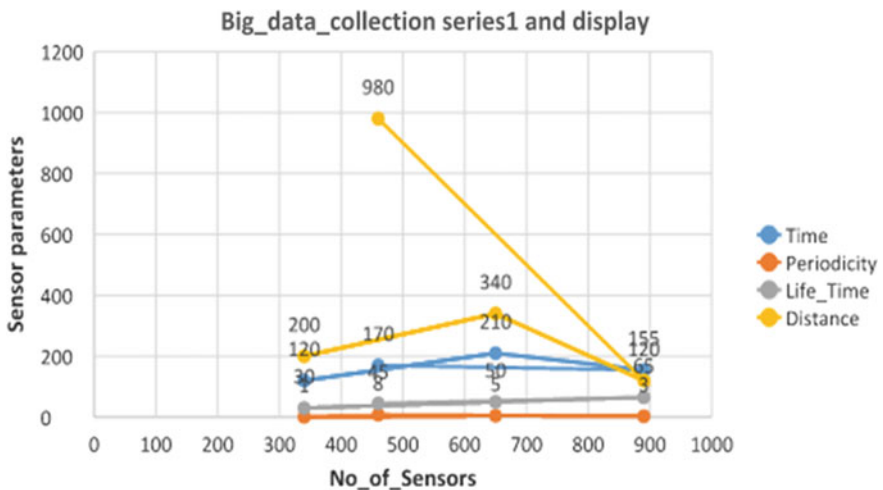


Fig. 5 Big data collection series1 and display

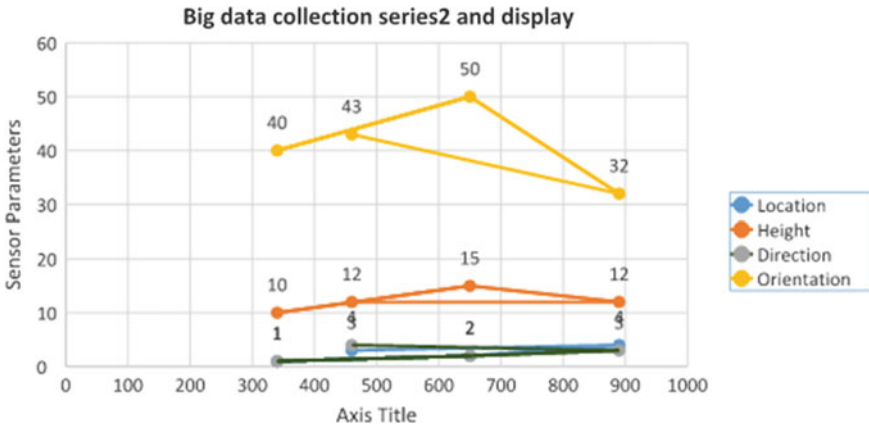


Fig. 6 Big data collection series2 and display

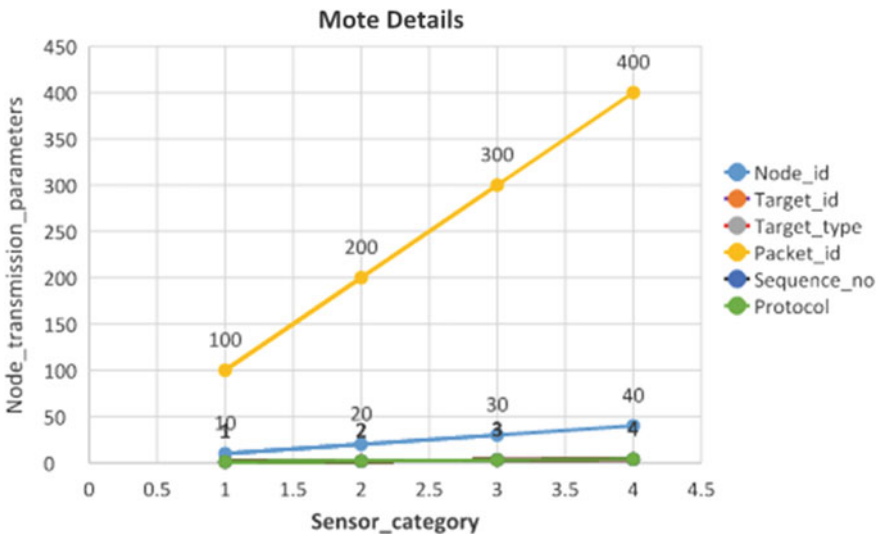


Fig. 7 Mote details collected from the field of deployment

Attacking, Losing, Strengthening, Alarming and Winning. The number of targets is varied and the response is plotted in Fig. 9. Figure 10 depicts the plot of proposed TDACA algorithm against the existing techniques like linear classifier and SVM. The response in terms of percentage is represented in Table 1. The linear classifier correctly classifies 3601 targets out of 4100 targets with efficiency 87.82. The other existing algorithm Support Vector Machine (SVM) classifies 3735 targets correctly

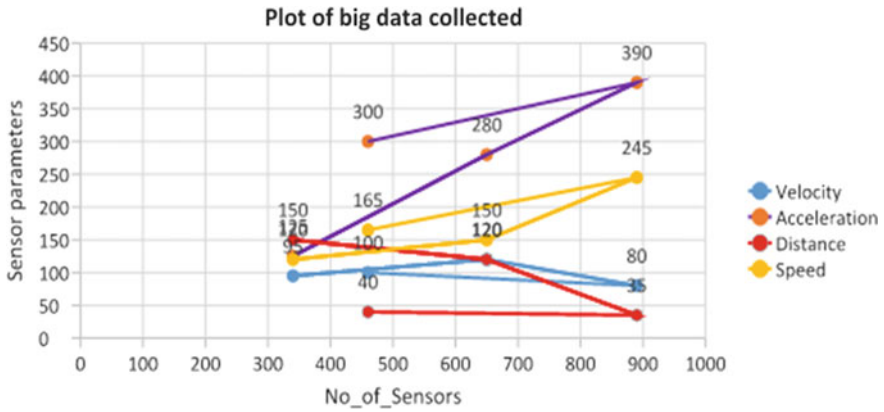


Fig. 8 Graph representing big data collected from sensors

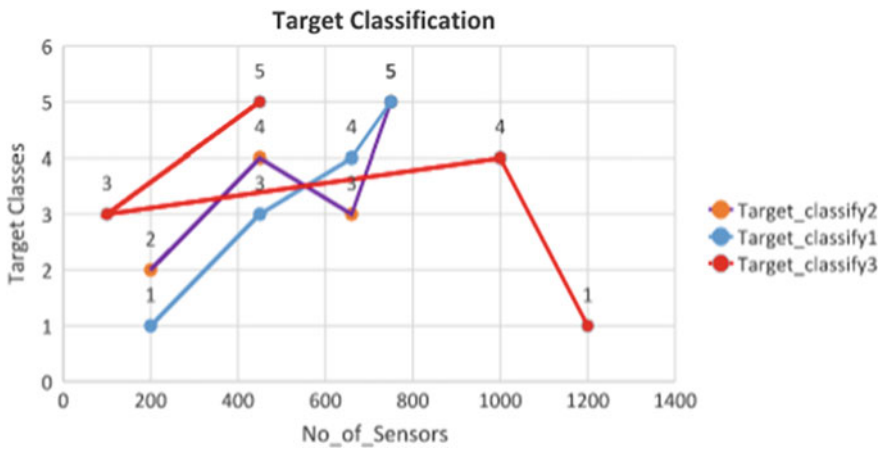
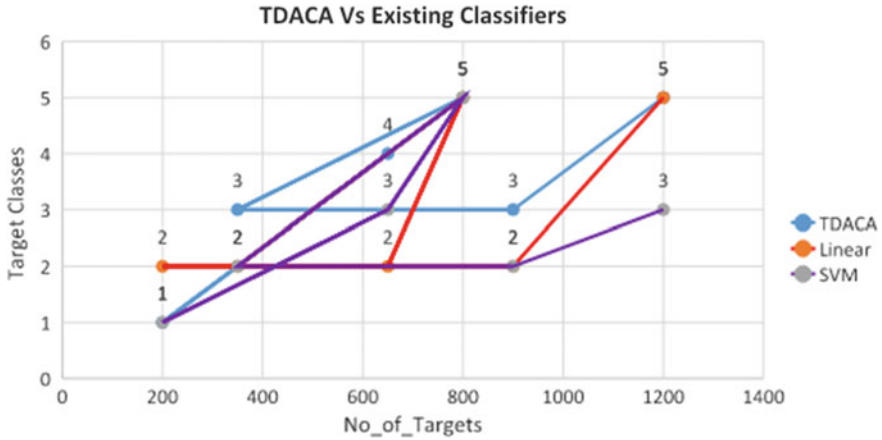


Fig. 9 Target Classification against No\_of\_targets

out of 4100 targets with the classification efficiency of 91.1% and found to be moderate. The proposed TDACA algorithm classifies 4078 targets out of 4100 targets and found to be efficient with efficiency 99.46. Almost all targets are classified correctly after getting detected Thus, it is found to be the best when compared to other classifiers.



**Fig. 10** Comparison of TDACA with existing classifiers

**Table 1** Confusion matrix of various classifiers considered

S. No	No_of_targets	TDACA	Linear	SVM
1	200	196	160	164
2	650	643	613	634
3	800	800	704	732
4	350	347	312	310
5	900	895	834	852
6	1200	1197	978	1043
	Efficiency (%)	4078/4100 = 99.46	3601/4100 = 87.82	3735/4100 = 91.1

## 5 Conclusion

WSN for Target Detection and Classification encompasses the deployment of sensors like speed, velocity, distance, and acceleration. The number and nature of these nodes are varied and the big data is collected from various sensors, received, preprocessed and stored in the cloud in the form of a big data database. The data flows in abundance and hence big data. BDCCA algorithm for WSN receives and finds the correlation of various data received. This correlation helps in grouping the sensors as intelligent nodes for transmission. The next step is to detect and classify the targets. The target classes are Attacking, Losing, Strengthening, Alarming and Winning. The proposed TDACA algorithm is used to classify the targets correctly. The response is plotted. It is compared with other existing algorithms like Linear Classifier and SVM. The existing classifiers have low classification efficiency and have their own drawbacks. Almost all targets are perfectly

classified in this algorithm. Hence it is concluded from the statistics gathered and observed, recorded and plotted that the TDACA algorithm is the best algorithm for moving target detection and classification in the WSN world.

## References

1. Chang, C., I., Ren, H., Chiang, S., S., "Real-Time Processing Algorithms for Target Detection and Classification in Hyper spectral Imagery", *IEEE Transactions on Geo science and Remote Sensing*, Vol. 39, No. 4, 2001.
2. Coiras, E., Mignotte, P., Y., Petillot, Y., Bell, J., Lebart, K., "Supervised target detection and Classification by training on augmented reality data", *IET Radar Sonar Navigation*, Vol.1, Issue 1, pp. 83–90, 2007.
3. Rani, S., Ahmed, S., H., Talwar, R., Malhotra, J., "Can Sensors Collect Big Data: An Energy Efficient Big Data Gathering Algorithm for WSN", *IEEE Transactions on Industrial Informatics*, Vol. Xx, No. X, 2009.
4. Jin, X., Gupta, S., Asok, R., Thyagaraju, D., "Symbolic Dynamic Filtering of Seismic Sensors for Target Detection and Classification", *American Control Conference*, 2011.
5. Cheng, S., Zhipeng, C., Li, J., J., Fang, X., "Drawing Dominant Dataset from Big Sensory Data in Wireless Sensor Networks", *IEEE Conference on Computer Communications*, 2015.
6. Cheng, S., Zhipeng, C., Li, J., J., Hong, G., "Extracting Kernel Dataset from Big Sensory Data in Wireless Sensor Networks", *IEEE Transactions on Knowledge and Data Engineering*, 2016.
7. Ding, X., Yong, T., Yan, Y., "A Real-Time Big Data Gathering Algorithm Based on Indoor Wireless Sensor Networks for Risk Analysis of Industrial Operations", *IEEE Transactions On Industrial Informatics*, Vol. 12, No. 3, June 2016.
8. Yang, K., Qin, Y., Leng, S., Fan, B., Fan, W., "Data and Energy Integrated Communication Networks for Wireless Big Data", *DEINs for Wireless Big Data*, Vol. 4, 2016.
9. Nikos, T., Thanasis, L., Khan, S., U., Xu, C., Z., Zomaya, A., Y., "On Improving Constrained Single and Group Operator Placement Using Evictions in Big Data Environments", *IEEE Transactions on Services Computing*, Vol. 9, NO. 5, 2016.
10. Amjad, M., Zhihan, L., Jaime, L., Umar, M., M., "ELDC: An Artificial Neural Network based Energy-Efficient and Robust Routing Scheme for Pollution Monitoring in WSNs", *IEEE Transactions on Emerging Topics in Computing*, 2016.
11. Bayou, L., Nora, C., B., David, E., Cuppen, f., "Towards a CDS-based Intrusion Detection Deployment Scheme for Securing Industrial Wireless Sensor Networks", 2016 11th International Conference on Availability, Reliability and Security (ARES), pp. 157–166, 2016.
12. Choi, T., M., Chan, H., K., Yue, X., "Recent Development in Big Data Analytics for Operations and Risk Management", *IEEE Transactions On Cybernetics*, Vol. 47, No. 1, 2017.
13. Elias, B., H., Mourad, D., Chadi, A., "Big Data Behavioral Analytics Meet Graph Theory: On Effective Botnet Takedowns", *IEEE Network*, Vol. 31, Issue: 1, 2017.
14. Xiang, J., M., Zhou, Z., Lei, S., Taj, S., Wang, Q., "A Mechanism Filling Sensing Holes for Detecting the Boundary of Continuous Objects in Hybrid Sparse Wireless Sensor Networks", *IEEE Access*, Volume: PP, Issue: 99, 2017.
15. Yang, C., Liu, C., Zhang, X., Surya, N., Chen, J., "A Time Efficient Approach for Detecting Errors in Big Sensor Data on Cloud", *IEEE Transactions on Parallel and Distributed Systems*, Vol.28, Issue 2, 2017.

# Time- and Cost-Aware Scheduling Method for Workflows in Cloud Computing Systems

G. Narendrababu Reddy and S. Phani Kumar

**Abstract** Cloud computing systems provide different options to customers to compute the tasks' based on their choice. Cloud systems provide services to customers as a utility. The customers are focused on the availability of service at low cost and minimum execution time. The performance of cloud systems depends on scheduling of tasks. The groups of tasks which are interdependent are referred as workflows. Workflow tasks scheduling plays an important role to estimate cloud system performance. If we want to reduce the execution time (make span), the cost involved in it will increase. Here, we proposed a novel method which minimizes the cost and time to schedule tasks of workflows. This algorithm schedules the tasks of a workflow to complete the execution in shortest feasible time so as to minimize the price for the services provided to customers. The experimental results show that proposed scheduling algorithm minimizes the make span and cost of workflows when compared with other existing algorithms.

**Keywords** Workflows • Dependent tasks • Task height • Cloud computing  
Make span minimization

## 1 Introduction

A disruptive computing model in which services are provided to customers over the Internet on demand based on pay—as you consume is called as cloud computing [1]. Different types of resources like infrastructure, platform and software are provided by cloud service provider to different types of customers on demand as a service utility. Cloud computing systems can assign, expand and withdraw its

---

G. Narendrababu Reddy (✉)  
CSE Department, GNITS, Hyderabad, India  
e-mail: gnbreddy25@gmail.com

S. Phani Kumar  
CSE Department, GITAM University, Hyderabad, India  
e-mail: phanikumar.s@gitam.edu



services dynamically at any point in time. Cloud environment leverages the virtualization features, and provide scalable resources and software services on demand over the internet to the customers [2]. Applications and files are thronged on a remote cloud which is a combination of multiple systems connected together and is accessible over the Internet in cloud computing. Cloud computing environment can provide adequate resources on demand to the users to process their tasks. In general, there is an inverse relation between the pricing of task computation time to its processing time [3]. Customers always look for the best service at least cost. Minimizing the make span of tasks in a workflow will increase the cost because more resources would be required to minimize the completion time. This is referred as time, cost tradeoff [4].

Workflows representation is adopted in many scientific and general purpose applications and projects in diverse disciplines. The interdependency relations between different tasks in a complex collection can be easily represented by popular workflow models [5]. The workflow represents the interactions and dependencies of tasks in an application. Workflow processing is a tiresome job for distributed and multiprocessor systems like cloud computing, where interdependent relations between tasks have to be considered by processing units. Scheduling of tasks in a cloud environment is an accomplice with different costs based on utilization of resources like processor and its time, memory, I/O, etc. [6]. We need to apply side-by-side processing mechanism to execute interdependent tasks. So one has to take the following points into consideration while scheduling [7]: Allocation of resources to tasks, the order in which tasks are executed by virtual machines (VMs) in the cloud, time and cost involved in the scheduling of these tasks.

The two main QOS constraints, time and cost can be optimized by finding the solutions to the above points. Cost and time are market-driven scheduling strategies rather than performance driven. If we try to reduce the time to process tasks, the scheduling cost would increase.

In this paper, we proposed a method to minimize the overall completion time of tasks in a workflow within a reasonable cost range. This algorithm minimizes the overall completion time of a workflow as minimum as possible and also scale down the number of VMs used in task execution. We focused mainly on minimizing the overall completion time of tasks in a workflow at the same time cost factor minimization to fulfill the important quality of service factors.

The overall completion time of a workflow can be minimized by reducing the completion time of parallel tasks at the same level. Virtual machines with high capacity would be allotted to time-critical tasks to minimize the execution time. The completion time of critical tasks becomes a deadline for noncritical tasks, i.e., remaining noncritical tasks must complete the execution on or before the completion of critical tasks. Cloud computing system pricing model is based on pay for what you use. So the number of virtual machines used to complete the execution of a workflow should be minimum to reduce the overall cost involved in task completion.

The remainder of the paper is organized as follows: Discussion of related works in Sect. 2. In Sect. 3, the workflow scheduling problem is explained. The proposed

approach is discussed in Sect. 4. Section 5 illustrates the experimental results and findings. Conclusion and future enhancements are discussed in Sect. 6.

## 2 Literature Survey

Most of the literature papers on interdependent task scheduling algorithms consider a set of resources and minimizing the make span [8–10]. But we focused on a distinct approach when compared with conventional workflow scheduling algorithms. Here, we proposed a unique algorithm that curtails the number virtual machines used to complete the execution of a workflow. Byun et al. [11] proposed BTS (Balance Time Scheduling) heuristic algorithm for workflows scheduling. For a workflow to complete its execution within a stipulated deadline, BTS estimates the minimum number of computing resources required. BTS is economical, elastic, extensible and universal, but due to its static allocation strategy, potential resource wastage takes place. In [12] Sudarsanam et al. introduced a technique using critical path analysis for estimating the number of resources required. This algorithm is based on the reinforced partial critical path technique. A heuristic dynamic programming algorithm called DCA for a dual criterion workflows scheduling was proposed by Wieczorek et al. [13]. DCA generates and checks tasks schedule repetitively and selects the best among them based on sliding constraint. Minimum resources required to complete a workflow execution within the minimum executable time was presented by Huang et al. in [14]. By varying DAG parameters like size, communication–computation ratio, etc., requirements of resources would be determined based on experimental data gathered from many sample workflows. Here we are proposing a unique method for scheduling dependent tasks of a workflow in cloud computing systems.

## 3 Workflow Scheduling Problem

A Directed Acyclic Graph (DAG) notion is used to represent workflow applications  $W = (t, e)$ , where  $t = \{t_1, t_2, t_3, \dots, t_n\}$  is a set of ‘ $n$ ’ tasks and ‘ $e$ ’ is set of directed edges. If there is a data dependency between ‘ $t_i$ ’ and ‘ $t_j$ ’, an edge ‘ $e_{ij}$ ’ of the form  $(t_i, t_j)$  exists between them, in which ‘ $t_i$ ’ is said to be the parent task of ‘ $t_j$ ’ and ‘ $t_j$ ’ is said to be child task of ‘ $t_i$ ’. So a child task has to wait for execution until its entire parent tasks are completed execution. Figure 1 shows a sample workflow. Customers define deadline ‘ $D$ ’ and budget ‘ $B$ ’ for workflows as constraints when they submit these applications to cloud systems. Scheduling the time-critical tasks of workflows with execution time as major constraint under limited budget is the main objective.

The completion time of a task  $T_i$  is denoted by  $CT_i$  and cost involved in the process as  $fi(CT_i)$ .

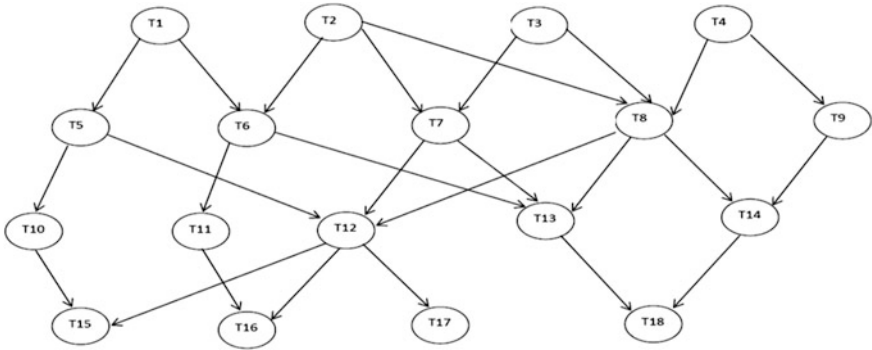


Fig. 1 A sample workflow

Our objective is to minimize the overall completion time of all tasks in a workflow, represented by  $CT_{max}$  and cost involved as  $f(CT)$ .

$P_{ij}$  is a task  $T_i$  processing time on a Virtual Machine  $VM_j$ .

$$P_{ij} \geq 0, \tag{1}$$

where  $i = 1, 2, 3, \dots, n$  and  $j = 1, 2, 3, \dots, m$

Now all tasks processing time in a  $VM_j$  is

$$P_j = \sum_{i=1}^n P_{ij} \tag{2}$$

Where,  $j = 1, 2, 3, \dots, m$

Minimizing  $CT_{max}$

$$\sum_{i=1}^n P_{ij} \leq CT_{max} \quad j = 1, 2, \dots, m \tag{3}$$

$$P_j \leq CT_{max} \quad j = 1, 2, \dots, m \tag{4}$$

Optimally,

$$CT_{max} = \{ \max_{i=1 \text{ to } n} CT_i, \max_{j=1 \text{ to } m} \sum_{i=1}^n P_{ij} \} \tag{5}$$

Minimizing  $f(CT)$

$$f(CT) = \sum_{i=1}^n f_i(CT_i) \tag{6}$$

$f_i(CT_i)$  is the cost involved in the task  $T_i$  execution and  $CT_i$  is the time for  $T_i$  execution.

The main objective of proposed scheduling method is the minimization of  $CT_i$  which in turn reduces  $f(CT)$  value.

## 4 Proposed Approach

The proposed scheduling method is following ‘ $M/M/m$ ’ queuing model. Distinct virtual machines ( $M$ ) process distinct tasks ( $M$ ). ‘ $m$ ’ is the finite server’s notion, but the tasks queue length is perpetual, so any number of tasks can possession in a queue. Poisson distribution can be used to represent workflows inflow to the system for scheduling.

Each workflow consists of several tasks in it. The different tasks in a workflow are allocated to different Virtual Machines. Tasks would be assigned to Virtual Machines in a predefined manner.

$VM = \{VM1, VM2, \dots, VMm\}$  a group of Virtual Machines, which have to process ‘ $n$ ’ jobs denoted by  $J = \{J1, J2, \dots, Jk\}$ . A job  $J_i$  is usually made of several tasks  $\{T1, \dots, T_n\}$ .

All the finite Virtual Machines provide self-sufficient and epidemic scattered services to the tasks of a workflow. Distinct VMs have distinct capacities to serve different tasks. Each virtual machine has a set of tasks in its queue. The task dispatch proportion is a function of a number of tasks present in the queue. The service rate of the scheduling system is the sum of all virtual machines service rates. The scheduling algorithm should allocate all tasks of a workflow to cloud computing resources in order to minimize the overall completion time which in turn reduces the computing cost. So the proposed method is devised to satisfy both time and cost constraints.

The main objective of proposed strategy is to identify a resource provisioning method that should minimize the overall completion time of a workflow. Scheduling of tasks done at different levels/heights with the objective of minimizing the completion time of tasks at that level/height. So the algorithm should decide minimum possible make span for the workflow that optimizes the number of virtual machines used and the task schedule for each height.

Figure 2 [11] shows a model for scheduling workflows in cloud computing systems. Here the customer submits the inputs like workflows, specifications of resources, and quality of service requirements to the workflow management system. Workflow management system consists of resource capacity estimator, resource acquisition and provisioning module, scheduling module, and execution manager. Scheduling of workflow to virtual machines will be taken care by the management system. The capacities of virtual machines those are ready to carry out a task would be estimated by resource capacity estimator. The resource acquisition module is the mediator between resource estimator and resource provisioning. The selection of virtual machines to tasks is said to be resource provisioning. Tasks execution order would be decided by scheduling module and execution manager dispatches tasks to the virtual machines.

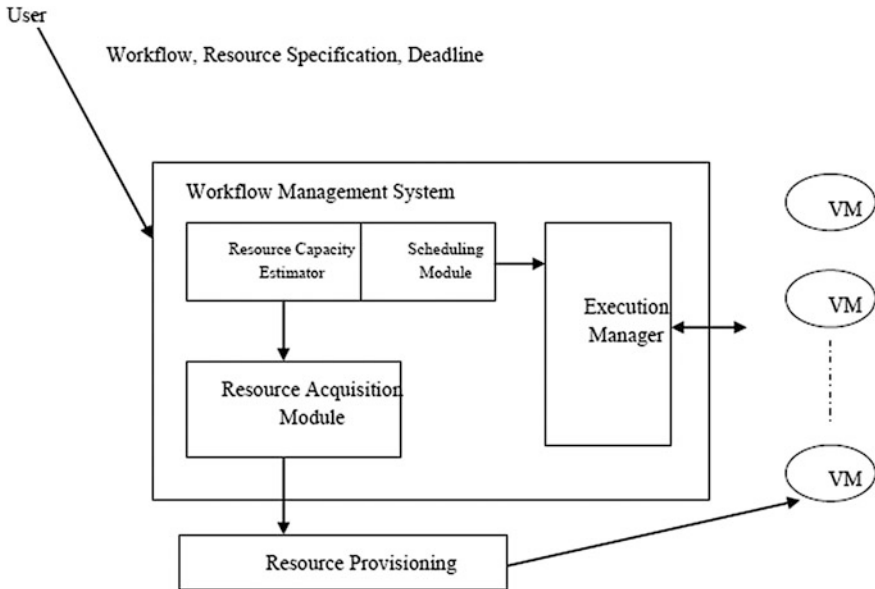


Fig. 2 Model for cloud computing workflows scheduling

The algorithm takes inputs as follows:

- (I) The profile of workflow denoted as a directed acyclic graph  $W = (t, e)$ , where ‘ $t$ ’ is the set of tasks in the workflow ‘ $W$ ’ and ‘ $e$ ’ represents a set of all dependencies among the tasks in the workflow.
- (II) Resources, i.e., Virtual machines.

#### 4.1 Task Properties

1. *Independent task group*: The independent tasks are those which are of the same height. The tasks can be grouped based on their heights so that the tasks of the same height are placed in the same group.
2. *The length of the task*: Basically, it is the number of instructions that have to be executed by a virtual machine represented with million instructions.
3. *Critical and noncritical tasks*: The task which has the maximum culmination time on all the virtual machines of the group is said to be a critical one in that task group. The remaining tasks in the group are said to be noncritical tasks.

### 4.2 Calculation of Height

If we can sort out tasks based on their height hierarchically the scheduling process would become easier. Hou et al. [15] first proposed the method for hierarchical sorting of the tasks for multiprocessor scheduling using a genetic algorithm and was later used by Qu et al. [16] and several others. The height of a particular task is strictly higher than its all predecessors. So all the available tasks at a particular height are not dependent on each other, i.e., all tasks at a given height can be treated as independent.

*Definition of task height:*

$H(Ti)$ —height of a node  $Ti$  can be defined as:

$$\begin{aligned}
 H(Ti) &= 0, \text{ if predecessor of } Ti = 0(\text{null}) \\
 &\text{(OR)} \\
 H(Ti) &= 1 + \text{maximum}(H(\text{predecessor}(Ti))), \\
 &\text{if predecessor of } Ti \neq 0(\text{null})
 \end{aligned}
 \tag{7}$$

Where Predecessor of  $(Ti)$  is all the predecessor of task  $Ti$ .

### 4.3 Algorithm

*Input:*  $W = (t, e)$  where

$t = \{t_1, t_2, \dots, t_n\}$  And

$t_i \rightarrow t_j \in e$  when predecessor of  $(t_j) = (t_i)$

A group of virtual machines ready to accept and carry out tasks execution.

*Output:* All tasks schedule.

*Step1:* Task height calculation.

From equation no: 7,

$$\begin{aligned}
 H(T_i) &= 0, \text{ if predecessor of } T_i = 0(\text{null}) \\
 &\text{(OR)}
 \end{aligned}$$

$$\begin{aligned}
 H(T_i) &= 1 + \text{maximum}(H(\text{predecessor}(T_i))), \\
 &\text{if predecessor of } T_i \neq 0(\text{null})
 \end{aligned}$$

*Step 2:*

$TG_h \leftarrow$  grouping of tasks that are ready to execute after its predecessor execution for each 'h' in 'H'.

If  $TG_h$  is not empty Do

Determine:  $MinimumTime(Ti) = Min\ of\ \{C(Ti, VM\ j) \mid Ti \in TG_h\}$

Calculate:  $Tc = \{Tc \mid Max\ of\ \{MinimumTime(Ti)\}\}$

Assign  $T_c \rightarrow$  a virtual machine which has the high capacity so that it can finish the task in less time.

Expel ' $T_c$ ' from  $TG_h$  group

*Step3:*

For all the tasks in group ' $TG_h$ ', select a task ' $T_i$ ' with minimum latest completion time (lct).

If ( $T_i, T_j \in \text{Min}(lct)$ )

Take  $T_i$  among  $T_i$  &  $T_j$ : Priority ( $T_i$ ) > Priority ( $T_j$ )

For all virtual machines,

If ( $VM_j \in \text{Minimum (probabilistic function)}$ )

Assign  $T_i$  to  $VM_j$

Remove  $T_i$  from  $TG_h$

Select a task ' $T_i$ ' with minimum (lct).

Amend  $VM_j$  available time

Else

Go - on

If  $TG_h \neq \text{Null}$

Appeal for a virtual machine with higher capacity that is assigned for  $T_c$ .

Go back to 3<sup>rd</sup> step.

## 5 Experimental Results

A workflow with 12 tasks as put on display in Fig. 3 was taken for the experimental purpose. The assumption we made that all virtual machines have the same capacity. Different task assignments based on time are shown in Fig. 4a, b. Figure 4a presents assignment of tasks based on proposed algorithm. Figure 4b presents tasks assignment based on the basic 'max-min' scheduling algorithm. The proposed scheduling algorithm also follows the same max-min scheduling for critical tasks, so the overall finishing time is identical for both methods. But the total number of virtual machines used is very less in the proposed method, which in turn leads to an enormous contraction in total cost.

### 5.1 Task Allocation

CloudSim [17] used for simulating the proposed scheduling algorithm and other job scheduling policies. We have correlated proposed scheduling policy with other well-known scheduling algorithms such as 'min-min' and fair 'max-min'. The results for these three algorithms are shown here as below.

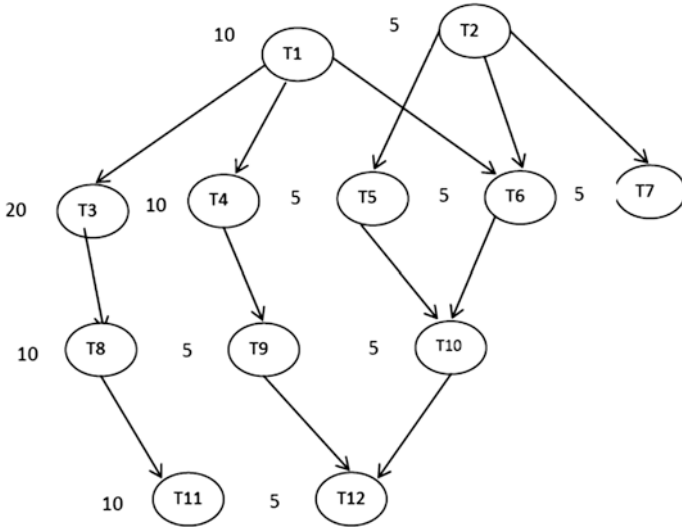


Fig. 3 Sample workflow with 12 tasks

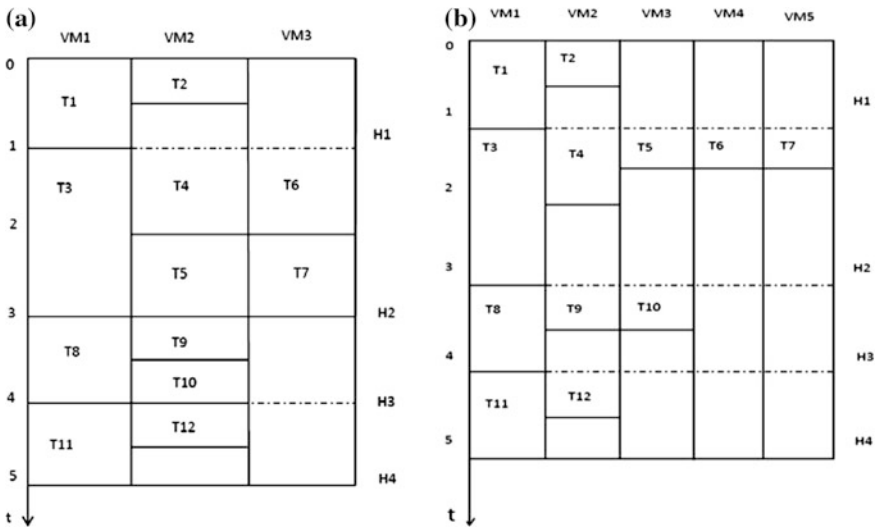


Fig. 4 a Tasks allocation to virtual machines based on proposed method (Only three VMs are used here), b Tasks allocation to virtual machines using max-min scheduling algorithm (Five VMs are used here)

Figure 5 shows the overall completion time of tasks at each height for workflow shown in Fig. 3. Tasks height on X-axis and time on Y-axis. It is apparent that the proposed method achieves superior result than other two methods.



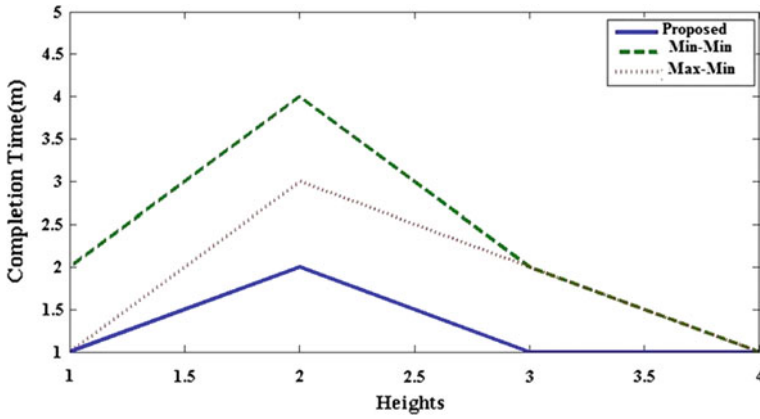


Fig. 5 Task heights versus completion time

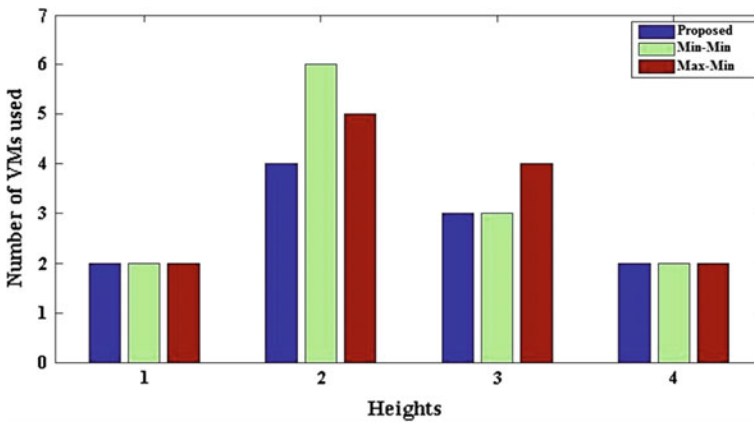


Fig. 6 Task heights versus number of VMs used

Figure 6 shows a correlation of the number of VMs used at each height for assignment of workflow as shown in Fig. 3. It is clearly apparent that proposed method accomplishes superior results. The cost of scheduling is directly related to the number of virtual machines used in that scheduling system, i.e., as the number of VMs increases the cost for the scheduling also will increase. The proposed scheduling method utilizes the less number of virtual machines which will directly brunt on the contraction of price for scheduling.

Figure 7 Depicts correlation of comprehensive completion time of workflow shown in Fig. 3. The proposed algorithm takes plenty of inferior overall completion time taken when compared with existing ‘min-min and max-min’ algorithms.

Fig. 7 Completion time

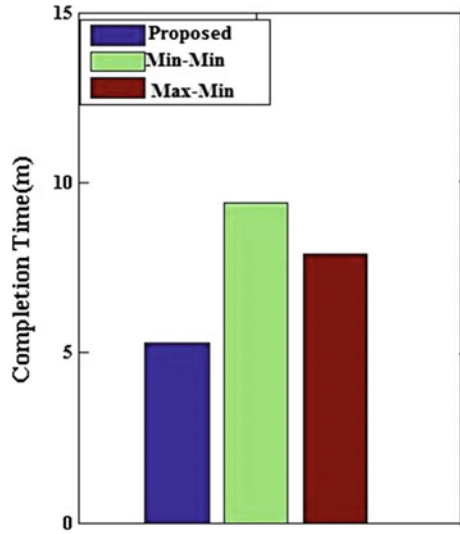


Fig. 8 Price chart

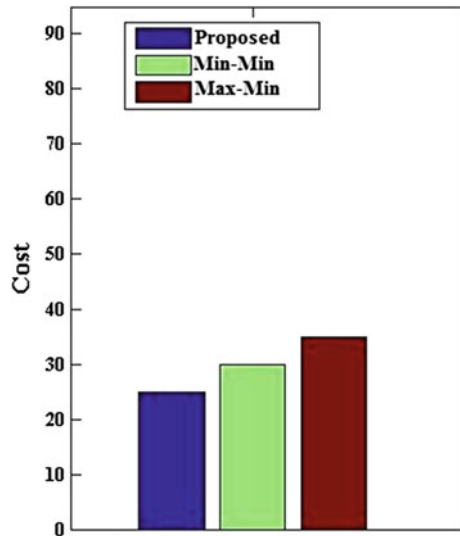


Figure 8 is the total price chart for the comparison of all the three scheduling algorithms of workflow shown in Fig. 3. It shows that the price is considerably minimized when proposed method is used for scheduling the tasks of the workflow.

## 6 Conclusion and Future Work

Workflow scheduling in cloud computing systems is one of the main challenges for service providers to meet Quality of Service (QoS) requirements. The customers are focused on the quick response from the service provider in terms of overall completion time a workflow with minimal cost. Here we proposed a unique scheduling strategy for workflows in the cloud environment to address the issues related to time and cost. The proposed algorithm minimizes the execution time of tasks at different stages based on the height of tasks by forming task groups. The completion time is reduced to a possible minimal level in each height, so the allocation of additional resources takes very less time and hence the overall completion time of workflow is also reduced. CloudSim is used to simulate the proposed algorithm and the outcomes illustrate that it achieves superior results in allocation and processing of workflows. The makespan of workflows is significantly reduced in comparison with other existing algorithms and it also diminishes the total price for execution of a task by catching an extra resource only when it is needed. In the future work, we want to enhance the capabilities of this algorithm in order to reduce the power consumption by systems in data station to reduce the carbon emission which will become a closer step toward green computing.

## References

1. Buyya, R, Pandey, S. and Vecchiola, C. (2009) 'Cloudbus toolkit for market-oriented cloud computing', *CloudCom'09: Proceedings of the 1st International Conference on Cloud Computing*, December 2009, Vol. 5931 of LNCS, Springer, Germany, pp. 24–44.
2. Pandey, S., Wu, L., Guru, S. and Buyya, R. (2011) '*Workflow engine for clouds, Cloud computing: Principles and Paradigms*'. February 2011, pp 321–344, Buya, R., Broberg, ISBN-13:978-0470887998, Wiley Press, New York, USA.
3. Yu, J and Buyya, R 'A taxonomy of workflow management systems for grid computing', *Journal of Grid Computing*, September, Vol. 3 Nos 3–4 (2005), pp 171–200, Springer B.V New York, USA.
4. Yu.J, and Buyya R, (2006) 'A budget-constrained scheduling of workflow applications on utility grids using genetic algorithms', *Workshop on Workflows in Support of Large-Scale Science, Proceedings of the 15th IEEE International Symposium on High Performance Distributed Computing (HPDC 2006, IEEE CS Press, Los Alamitos, CA, USA)*, 19–23 June, Paris, France.
5. Plale, B. et al 'CASA and LEAD: adaptive cyberinfrastructure for real-time multiscale weather forecasting', *IEEE Computer*, Vol. 39, No.11 (2006), pp:56–64.
6. Cao, Q., Wei, Z-B. and Gong, W-M. (2009) 'An optimized algorithm for task scheduling based on activity based costing in cloud computing', *3rd International Conference on Bioinformatics and Biomedical Engineering, 2009, ICBBE 2009*, 11–13 June, pp. 1–3.
7. Li, J., Qiu, M., Niu, J., Gao, W., Zong, Z. and Qin, X. (2010) 'Feedback dynamic algorithms for preemptable job scheduling in cloud systems', *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, 31 August 2010 to September 3, pp. 561–564.

8. Yuan, Y., Li, X. and Wang, Q. (2006) 'Time-cost tradeoff dynamic scheduling algorithm for workflows in grids', *CSCWD'06, 10th International Conference on Computer Supported Cooperative Work in Design, 2006*, 3–5 May, pp. 1–6.
9. Yu, J., Buyya, R. and Ramanohanarao, K. (2008) *Metaheuristics for Scheduling in Distributed Computing Environments*, Springer, Berlin, Germany.
10. Dong, F., and Akl, S.G. (2006) *Scheduling Algorithms for Grid Computing: State of the Art and Open Problems*, January, Tech. rep., School of Computing, Queen's University, Kingston, Ontario.
11. Byun, E-K.,Kee, Y-S., Deelman, E., Vahi, K., Mehta, G. and Kim, J-S. 'Estimating resource needs for time-constrained workflows', 2008 *Proceedings of the 4th IEEE International Conference on e-Science*.
12. Sudarsanam, A., Srinivasan, M., and Panchanathan, S. 'Resource estimation and Task Scheduling for Multithreaded Reconfigurable Architecture', *Proceedings of the 10<sup>th</sup> International Conference on Parallel and Distributed Systems* (2004).
13. Wieczorek, M., Podlipnig, S., Prodan, R. and Fahringer, T. 'Bi-Criteria Scheduling of scientific workflows for the grid', *Proceedings of the 8th ACM/IEEE International Symposium on Cluster Computing and the Grid -2008*.
14. Huang, R., Casanova, H. and Chien, A.A. 'Automatic resource specification generation for resource selection', *Proceedings of the 20th ACM/IEEE International Conference on High-Performance Computing and Communication, 2007*.
15. Hou, E.S.H., Ansari, N. and Ren, H. 'A genetic algorithm for multiprocessor scheduling', *IEEE Trans. Parallel and Distributed Systems*, Vol. 5, No. 2, pp. 113–120, February-1994.
16. Qu, Y., Soininen, J. and Nurmi, J. 'Static scheduling techniques for dependent tasks on dynamically reconfigurable devices', *Journal of Systems Architecture*, Vol. 53, No. 11, pp. 861–876, 2007.
17. Calheiros, R.N., Ranjan, R., De Rose, C.A.F. and Buyya, R., 'CloudSim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms', *Software: Practice and Experience*, Vol. 41, No. 1, pp. 23–50. 2011.

# A Novel Statistical Feature Selection Measure for Decision Tree Models on Microarray Cancer Detection

Janardhan Reddy Ummadi, B. Venkata Ramana Reddy  
and B. Eswara Reddy

**Abstract** Recently, machine learning techniques have become popular and widely accepted for cancer detection and classification. Prediction of cancer disease focuses on three main objectives: susceptibility prediction, recurrence prediction, and survivability prediction. Most of the conventional classification techniques deal with limited attributes and small datasets. Random forest classifier is one of the ensemble learning models, which is capable to handle datasets with a large number of attributes. Machine learning algorithms used for cancer prediction are supervised learning with high prediction rate. In this paper, a novel statistical attribute selection measure was implemented for cancer disease prediction. In this work, we have used different decision tree models such as random tree, random forest, Hoeffding tree to evaluate the performance of cancer disease prediction using proposed attribute selection measure. Experimental results are evaluated on different types of microarray cancer datasets including lung cancer, ovarian, lung cancer, and DLBCL-Stanford. The performance of each model is compared in order to find the most efficient and optimized algorithm. Experimental results show that proposed model has high computational efficiency in terms of accuracy and true positive rate.

**Keywords** Microarray cancer detection · Ensemble classification model  
Feature selection measures

---

J.R. Ummadi (✉)  
Vignan's University, Guntur, India  
e-mail: ummadi.janardan@gmail.com

B. Venkata Ramana Reddy  
Department of CSE, NEC, Nellore, India  
e-mail: busireddy100@gmail.com

B. Eswara Reddy  
JNTUACE, Kalikiri, India  
e-mail: eswarcejntua@gmail.com

## 1 Introduction

During the past two decades, feature selection methods and classification models have been used for prediction and diagnosis of different cancer diseases. Later, machine learning techniques became popular and widely accepted for cancer detection and classification. Prediction of cancer disease focuses on three important objectives, those are susceptibility prediction, recurrence prediction, and survivability prediction. Prediction susceptibility can be defined as the likelihood of generating cancer before the disease start. Recurrence prediction is the likelihood of regeneration of the disease after successful resolution. Survivability is the prediction of disease status. Conventional methods were capable of predicting the cancer diseases with few parameters and limited dimensions. Today, with the exponential growth of technology, it is merely impossible to use these conventional methods for cancer prediction due to the high dimensional features and parameters. Machine learning algorithms were proposed to satisfy these needs. It takes microarray data, clinical data, proteomic data, or a combination of these as input. The other traditional approaches consider the variables as independent and linear. But in case of a nonlinear and dependent variable, these traditional methods fail and machine learning over powers. Most of the biological systems are nonlinear and its parameters are interdependent, thus machine learning has become the better choice. Curse of dimensionality is another problem scenario where there exist more variables and fewer examples. Both machine learning and conventional methods suffer from this problem. This problem can be resolved by either decreasing the number of variables or increasing the number of training datasets. The sample-feature ratio must be more than 5:1 every time. Machine learning algorithms can be categorized into three broad types, they are: Supervised machine learning, Unsupervised machine learning and Reinforcement machine learning. Supervised learning consists of a prescient provider which provides the labeled training dataset as input to the algorithm and produces output after mapping. But on the contrary, for unsupervised machine learning, only training datasets are given as input without labels. Some of the examples of unsupervised learning are:- self-organizing feature maps, hierarchical clustering, k-means clustering, and so on. All machine learning algorithms used for cancer prediction come under supervised learning. Most widely applicable algorithms are: Artificial Neural Networks, Decision Trees, Genetic algorithms, Linear Discriminant Analysis,  $k$ -Nearest Neighbor, etc.

Different cancer datasets are used by the researchers for testing and prediction of cancer disease in many biomedical researches. Some commonly used cancer datasets are given below:-

*Blood Cancer* Blood cancer is categorized into different forms as leukemia, lymphomas, and myeloma. These diseases are divided on the basis of diagnosis, treatment, and their result. Cancer registries contain all the datasets of disease morphology which is recorded by International Classification of Diseases for Oncology.

*Breast Cancer:* It is heterogeneous cancer that is subdivided based on Estrogen Receptor, Progesterone Receptor, and Human Epidermal Growth Factor Receptor 2. Biomarkers are used to distinguish between cancerous and normal tissues. Along with these conventional classification based on biomarkers, nowadays genome DNA microarrays are considered for classification. Large transcriptomics datasets are available for the biomarkers to detect different breast cancers. The results are greatly affected by the divergence of datasets and various microarray platforms.

*Skin Cancer:* Lack of proper diagnosis and treatment of skin infections may lead to dangerous skin cancer. Each year numbers of skin cancer patients are increasing vastly as compared to all other cancer patients. Diagnosis and prediction of skin cancer are more difficult. Lesional morphology, body site distribution and scaling of lesions are mostly used methods of diagnosis which makes the process more complex. Many researchers integrated feature extraction with a classification for prediction of skin cancer like melanoma. Among these, some methods failed to evaluate large datasets of skin cancer for diagnosis.

*Oral Cancer:* Oral cancer refers to as any cancerous growth in the oral cavity. It is a subset of head and neck cancer. Primary lesion on oral tissue, metastasis, and growth of tissue from nasal cavity are some common examples for the growth of oral cancer. Most common form of oral cancer found in 90% of cases is squamous cell carcinomas. The oral cancer diagnosis and evaluation of different prediction strategy are performed by using publicly available oral cancer datasets.

There are many data mining models proposed for prediction of cancer disease, classification model is one of them. Two significant advantages make the model widely accepted are feature learning and parameter optimization. Previously, statistical methods were used for prediction of diseases. To overcome the problems of statistical approaches, machine learning techniques were proposed. Parameter optimization is another vital advantage of machine learning.

Besides these advantages, there exist some limitations which are described below.

- This approach is not suitable for high-dimensional features and sparse datasets.
- Feature selection is one of the major issues in the traditional classification models. As the size of the attributes increases, it is difficult to classify the test samples with at least 10% of attributes.
- A large amount of data are essential part of machine learning schemes like deep learning. It is very difficult and hectic to analyze, evaluate and manage such a vast amount of data.
- Above are some major issues found in machine learning models on cancer datasets found from previous researchers.

## 2 Related Work

Yusof et al. worked on classification of medical data and tried to improve the classification result by developing a feature selection-based machine learning approach for cancer detection [1]. They integrated the feature selection with pre-processing phase. They have implemented a novel filtering method on uncertain data in order to provide better classification results. They analyzed and compared feature selection approach with three cancer datasets and four machine learning algorithms. With respect to ROC and F-measure, they observed the performance ratio of feature selection-based classification and classification without feature selection. It can be concluded from this work through feature selection cannot satisfy all datasets (<https://archive.ics.uci.edu/ml/datasets/Lung+Cancer>), it definitely improves the performance of classification.

Pérez et al. introduced a feature selection classification using machine learning for prediction of breast cancer [2]. They merged various feature selection methods to improve the detection performance. Subsets of features are taken as input to the machine learning classifiers. The researchers performed validations of six cancer datasets by collecting data from public breast cancer databases. This model decreases features in every dataset by merging some feature selections with various evaluation functions. The algorithm shows significant results in both balanced and unbalanced datasets. They proved with series of experiment that, FFBP-neural network and R-Mean techniques are most efficient with significant AUC values.

Arafi et al. combined the concepts of support vector machines and particle swarm optimization and presented a new technique for breast cancer detection [3]. To decrease generalization errors, SVM was used. In order to find the optimized value of parameters, PSO was considered. They evaluated the performance of their algorithm on the basis of real benchmark datasets from VCI database (<http://tunedit.org/repo/UCI/breast-cancer.arff>). They split the whole dataset into three subparts with variable sizes; those are used in training, validation, and testing. The authors achieved increased performance as compared to four other classification methods of machine learning. Further work can be done on this model by including extended kernel functions for parameter optimization.

Begum et al. combined feature selection and kNN techniques to propose a new solution for classification problem on leukemia datasets [4]. They used Consistency-Based Feature Selection, Fuzzy Preference-Based Rough Set, Kernelized Fuzzy Rough Set and kNN classification algorithms on datasets. They performed experiments and found that CBFS algorithm outperforms the other two in terms of performance. Feature selection algorithms are responsible for filtering nonrelevant and duplicate features. The authors argued that their algorithm can be used for data classification of many cancer diseases (<http://eps.upo.es/biggs/datasets.html>). After combining feature selection with k-Nearest Neighbor algorithms in the previous work, a new approach to supervised learning based on feature selection was proposed by the same authors Begum et al. [5]. They concentrated on choosing biomarkers out of microarray dataset of genes, which can help in detection of



leukemia. The authors experimented their theory on three different machine learning algorithms with the same cancer datasets, that are SVM, kNN and NB. The performance of SVM is recorded far higher than that of the other two algorithms. This approach can be extended in future by applying other machine learning techniques on these leukemia datasets to analyze which method shows optimized performance above all other methods.

Bharathi and Natarajan presented Extreme Learning Machine approach for prediction of cancer disease [6]. They have proposed two phases for cancer prediction. The first phase consists of gene ranking algorithm and the second phase involves choosing minimum gene subset from the first phase.

They evaluated their method on Lymphoma and SRBCT datasets (<http://eps.upo.es/big5/datasets.html>). After applying fivefold cross validations on 62 samples by both SVM and ELM algorithms, 100% accuracy with less training time was recorded as compared to all other methods. They concluded that their proposed approach was fast and most efficient than that of other previously proposed models for detection and prediction of cancer diseases from datasets.

Chen et al. presented a new method to detect colon cancer by integrating neural network with supervised learning in their research [7]. They tried to reduce the risk of over learning by a cost function with Monte Carlo algorithm. A single change in the algorithmic parameter will affect the output and modify it. This algorithm decreases computation time greatly. The authors stated that their approach involving supervised model along with a voting scheme will perform much better than that of the only supervised model. Further future research can be done to include more datasets, time series prediction, and function fitness.

Dass et al. studied the major cause of cancer, i.e., gene mutation and introduced a new classification method for prediction of lung cancer [8]. Their theory concentrates on analyzing gene mutation and expression of datasets. The authors considered biomarkers datasets of Non-Small Cell Lung Cancer, and its categorizations that are Squamous Cell Cancer and Adenocarcinoma for their approach. The biomarkers consist of microRNAs (<http://eps.upo.es/big5/datasets.html>). They applied decision tree based classification technique on biomarkers. High accuracy (99.7%) was achieved because of integration of J48 algorithm with traditional decision tree classification by the researchers in the cross validation process of their theory.

Helmy et al. tried to reduce the computation time in classification and merged Extreme Learning Machine approach with Single-hidden Layer Feed-forward Neural-networks [9]. Their new approach is not only responsible for the random detection of hidden nodes, but also the analytical computation of output weights. ELM has the advantages over the problems of local minima and improper learning rate. They evaluated their approach on five datasets, those are Breast Cancer Wisconsin dataset, Pima Diabetes dataset, Heart-Statlog dataset, Hepatitis dataset, and Hypothyroid dataset. Out of all other activation function, the sigmoid function performs far ahead. They achieved good classification accuracy, decreased training time, and complexity. The authors concluded that their method can be implemented in high dimensional bioinformatics classification in future.

Klassen integrated SVM with random forest classification and applied on microarray cancer datasets [10]. Microarray cancer datasets contain gene expression which is used for detection of diseases. In order to find good classification rate, classifiers were evaluated with various numbers of genes. They validated their approach by experiments and found that their proposed scheme can efficiently work with microarray cancer datasets. Out of a large pool of genes, a small sample consisting of all genes resulted with great training and testing speed in Intel dual core 2 GHz processor. The approach is very fast than that of a neural network which was proved by the researchers.

For binary classification of microarray datasets, Arunkumar and Ramakrishnan proposed a new algorithm based on ELM [11]. They considered five datasets, that are-ALL/AML, CNS, Lung Cancer, Ovarian Cancer and Prostate Cancer. The performance of binary classification was analyzed (<http://eps.upo.es/big5/datasets.html>). Initially, feature extraction with correlation coefficient has been carried out. They experimented and showed that ELM shows significant performance as compared to many other traditional classification algorithms that make ELM a better option above all these prior algorithms. Classification accuracy of first and last datasets are slightly less, while other three datasets show better classification accuracy. The reason behind the lesser accuracy of the two said datasets is a poor correlation which is measured as less than 0.5.

Moulos et al. found that there have been very few works done since years on the stability of gene signatures, so they analyzed microarray cancer datasets by applying five feature selection approaches in high-dimensional throughput genomics [12]. The authors used this approach as a solution to encounter classification dimension problems. As gene signature decides between diseased state and normal state, the stability of gene signature must be a major concern. Further work is needed on inclusion of other datasets with more metrics of feature selection. Accuracy of each approach must be analyzed and compared to find the optimized method out of a number of conventional methods.

Nematzadeh et al. compared various works previously done on breast cancer classification techniques based on machine learning [13]. They analyzed  $k$ -fold (selected  $k = 10$ ) cross-validation to find the most significant and efficient approach among others resulting high accuracy rate. At first, they split their whole dataset into two subsets as training and validation sets. The numbers of fold are directly proportional to the computational cost because as the number of folds increases, the computational cost also increases. They stated that by increased value of  $k$  in  $k$ -fold cross-validation the accuracy rate is not good always. Thus it is proved that the values of  $k$  and accuracy rate are independent of each other. Later work can be carried out to find most accurate value of  $k$  by using genetic algorithm and PSO.

Ozcift studied various researches on feature extraction schemes and stated that there is need of an efficient universal variable selection scheme [14]. They introduced a BFS Random Forest wrapper technique to identify the optimal characteristics of the following biomedical datasets-colon cancer, leukemia, breast cancer, and lung cancer datasets. The authors tested their datasets and analyzed the accuracy rate of fifteen commonly applied different classification schemes. They also

tested and evaluated their proposed approach and proved that their suggested scheme resulted with better accuracy rate as compared to the other traditional approaches. The algorithm has the ability to reduce feature size of high dimensional datasets.

Machine learning plays a vital role in decision process based on the past and learned experiences. This enables the system to make an informed and experienced decision in similar situations. AI techniques for cancer detection and prognosis are in use for last three decades [15]. For detecting a specific type of cancer namely Multiple Myeloma, Waddell et al. [16], conducted a case study where single-nucleotide polymorphism (SNP), single positions of variation in DNA, profiles were used. Linear Support Vector Machine (SVM) was used to deal with the links to multiple features. They valued two classes as “pre-disposed” for the patients diagnosed before the age of 40 while “not pre-disposed” for the patients diagnosed after the age 70. The study reported above 70% of accuracy on the trained SVM.

Kim et al. [17] also used SVM for classification of patients who are at high risk of cancer recurrence. The authors also compared the result with artificial neural network (ANN), and traditional Cox-proportional hazard regression model (Cox regression). For prediction model, a total number of 7 different variables were used in the experiments including number of tumors, invasion and its size, number of metastatic lymph, and ER status. The prediction results in term of accuracy shows that SVM has outperformed ANN and Cox regression.

Similarly, Chih-Jen Tseng et al. [18], have used SVM for machine learning to predict the risk factors in recurrence of cervical cancer by considering four factors; pathologic\_S, pathologic\_T, cell type and RT target-summary. For the oral reoccurrence, Konstantinos P. Exarchos et al. [19] used three classifiers for clinical, imaging and genomic features and combined them for prediction and analysis of oral cancer. Authors employed Bayesian Networks, SVM, Decision Tree (DT), Random Forests and ANN for classifications considering the main oral cancer’s features like Smoker, p53 stain, extra-tumor spreading, TCAM and SOD2. They reported that BN has proven the best in feature selection in terms of accuracy, sensitivity, and specificity.

A predictive model was proposed by Stojadinovic A, Nissan et al. [20], to enrich the surgical decision for a patient with colon carcinomatosis. For this purpose, main features including Primary tumor histology, nodal staging, and extent of peritoneal cancer were used. The authors have reported the best performance of Bayesian classification in decision support. ANN, SVM, and SSL were used to classify the breast cancer survivability. The authors have compared the algorithm in terms of accuracy and reported that SSL performed well compared to other two. The authors used 6 features including Sex, age, T\_stage, N\_stage, LCK, and ERBB2 genes to model non-small cell lung cancer (NSCLC). The likelihood of risk was classified based on median survival time which was 36 months and considered at low risk while the remaining patients were assigned to high-risk group. After extensive training of the model, the authors reported more than 80% of the accuracy based on survival time in the validation set. SVM classifier was also used in another study to

classify the features including TNM\_stage and number of recurrences in oral cancer. After finding the most relevant and important feature, they reported over 97 and 100% rates for alive and dead patients respectively.

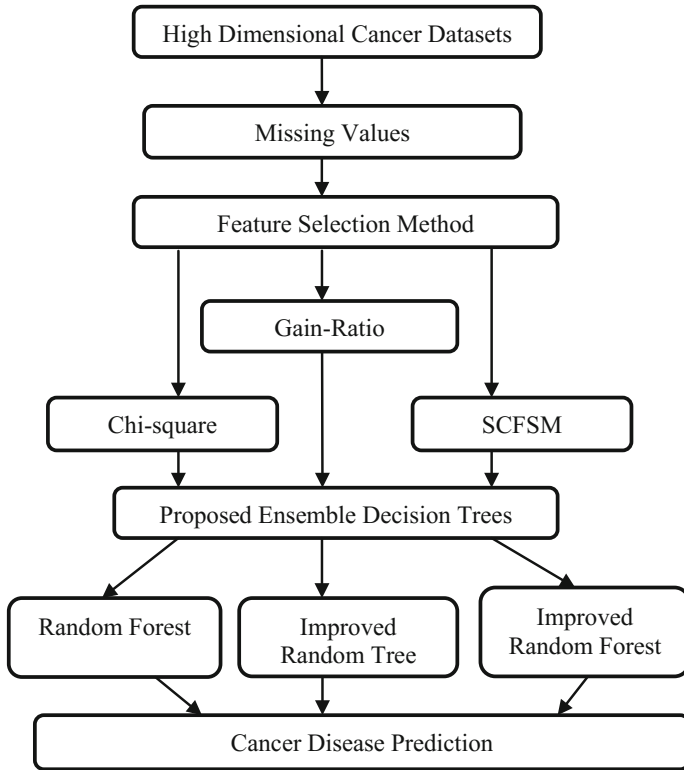
Bayesian networks were used in a study where authors developed a Bayesian information-value to present the disease history and other variables and their impact on the treatment. This study was specifically used to interpret the imperfect information values. Three models were evaluated which were termed as failure models. The study shows that power law and exponential failure models are more sensitive than linear failure model. The study also established the fact that exponential failure model may occur as the less realistic model. In cancer diagnosis and prognosis, there are some research studies where diverse classification models have been used and compared with the same data sets. They used ANN and DT along with a statistical technique called logistic regression and evaluated the results in terms of Accuracy, sensitivity, and specificity. SEER dataset of 433,000 cases was used after data cleansing. A binary output approach was used to identify the survivals where 1 was considered for survivals and 0 for non-survivals. The data set was divided into 10 mutually exclusive groups using stratified sampling methods. For training the models, they used 9 groups of data and repeated the process 10 times. The accuracy evaluation was computed by taking the average values. The study shows that DT method performed the best with the accuracy of 93%. Above are some major issues found in machine learning models on cancer datasets found from previous researchers [1–10].

### 3 Proposed Solution

In this paper, a novel ensemble classifier was designed and implemented to address the problem of attribute selection and high dimensionality issues. The main objective of our supervised ensemble classifier is to classify each high dimensional data for cancer prediction. Proposed ensemble model is usually designed and implemented to improve the cancer prediction rate on high dimensional data. Proposed ensemble model improves the traditional Random forest; Random tree and C4.5 classifiers in terms of accuracy, true positive rate and attribute selection measure.

Figure 1, represents the proposed ensemble model that describes the use of ensemble approach for cancer detection process. Proposed model takes high dimensional cancer datasets as input for data preprocessing. It has two main phases, the first phase is feature selection model and the second phase is ensemble learning model for cancer prediction process. Generally, ensemble learning model is generated from a group of base classifiers to predict the cancer detection process. In this paper, a novel ensemble model was designed and implemented for cancer detection process.

In this proposed model, we have used novel feature selection measures for ensemble classification model. Proposed three feature selection measures are used



**Fig. 1** Proposed ensemble learning framework

in traditional random forest and random tree classification models to improve classification rate on high dimensional datasets such as micro array cancer datasets.

In this framework, multiple cancer disease datasets are analyzed using the proposed ensemble model with high-dimensional feature set. In our model, different base classifiers such as C4.5, Naïve Bayes, Random forest, Random Tree and improved random forest are used to test the efficiency of the proposed model to the traditional models.

**Filtering Data:**

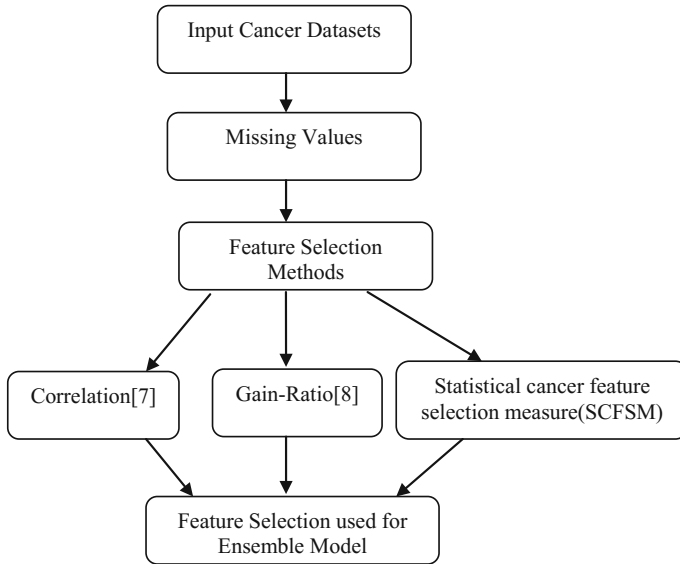
**Input:** Cancer datasets  $D(1), D(2) \dots D(n)$

**Output:** Filtered Data

**Procedure:**

Read dataset  $D(1), D(2) \dots D(n)$

For each instance  $Ins(i)$  in the Dataset  $D(1), D(2) \dots D(n)$



**Fig. 2** Proposed feature selection model

```

do
For each attribute A in the instance Ins(i)
do
if(isNumeric(Ai) && Ai(I) ==null)
then
    
```

$$A(Ins(i)) = \sum_{j=1/i \neq j}^n \left( \left( \sum X_j^2 \right) - \mu_{A(Ins(i))} \right) / \left( \text{Max}_{A(Ins(i))} - \text{Min}_{A(Ins(i))} \right) \quad (1)$$

```

end if
if(isNominal(Ai) && Ai(I) ==null)
then
    
```

$$A(Ins(i)) = \text{Prob}(\left( (A(Ins(j)) / C(k)) / j \neq i, \text{kth class} \right) / \left( \text{Max}_{A(Ins(i))} - \text{Min}_{A(Ins(i))} \right)); \quad (2)$$

```

end if
End for
    
```

In this filtering method, each attribute is tested for missing values. Most of the traditional methods are capable of filling missing values using numeric attributes. In

our preprocessing method, three major steps were executed on the cancer disease datasets.

- If the attribute is continuous and the value is null, then it is filled with computed value of Eq. (1).
- Similarly, if the attribute is nominal and the values are null, then it is filled with computed value of Eq. (2).
- Finally, if the class attribute is continuous, then it is converted to nominal and labeled with cancer disease types.

Preprocessing of the data is essential to improve the accuracy of the dynamic ensemble model as shown in Fig. 3.

### 3.1 Cancer Detection Using Optimized Random Forest Decision Tree Algorithm (CDORFDT)

Random forest is essentially an ensemble of decision trees where each individual tree acts as a base classifier and the classification is executed by choosing a vote based on the instance predictions made by each decision tree model. Random forest is one of the popular decision tree models for building decision trees for high dimensional medical datasets. But, for small to medium datasets the diversity falls drastically due to class imbalance problem. Traditional random forest model tends to use a simple randomization sampling and builds decision trees. Thus, we have several scopes of optimizing the traditional random forest classifier. In this paper, we have proposed a novel feature selection measure in the random forest and random tree algorithms for cancer detection.

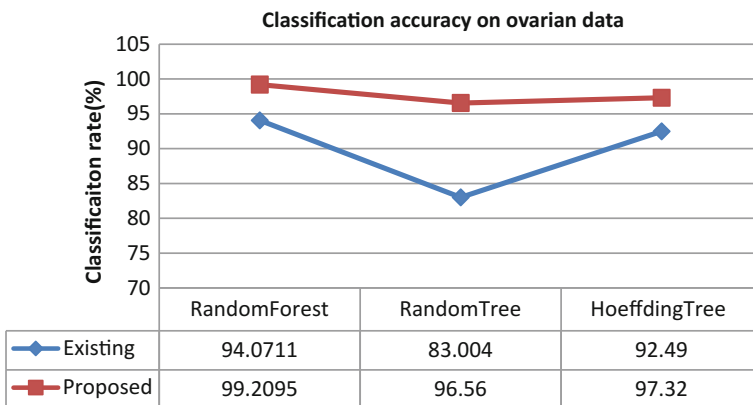


Fig. 3 Performance of ovarian cancer disease using classification models

### 3.2 Feature Selection Measures for Decision Tree Construction

Attribute selection measure is a heuristic factor for selecting the split functionality of decision tree construction. Entropy, Gain ratio, correlation, information gain are the well-known traditional attribute selection measures for decision tree construction.

In our research work, we used different split measures correlation [7], gain ratio [8], and proposed statistical cancer selecting measure to improve traditional random forest and random tree classifiers as shown in Fig. 2. In the Fig. 2, input cancer datasets are processed for missing values and it is replaced with the mean of the attribute. Since our proposed feature selection measure(SCFSM) is based on three measures such as Hellinger, Chi-square, and Conditional entropy measures, proposed feature selection measure select the optimal correlated features which are used to predict the cancer disease in ensemble classification model. Also, the number of features selected in our model highly depends on the correlated features and SCFSM computation values, this statistical measure is best applicable to high dimensional datasets and a large number of instances.

### 3.3 Proposed Statistical Cancer Feature Selection Measure (SCFSM)

Proposed attribute selection measure is a hybrid integrated measure of three computational measures: (1) Hellinger measure (2) Chi-square measure (3) Conditional entropy measure.

*Hellinger Measure:* Hellinger measure is used to quantify the similarity between the cancer disease class distributions. It is basically derived from F-divergence. In our feature selection measure, we have improved the traditional Hellinger measure using cubic polynomial. The computation formula used to measure the improved Hellinger measure is given in Eq. (3) as

$$\text{CubicHellinger} = \sqrt[3]{\left(\sum_{i=1}^m \sqrt[3]{D_i / |D_i|} - \sqrt[3]{D_j / |D_j|}\right)^2} \quad (3)$$

$D_i$  is the positive cancer data

$D_j$  is the negative cancer data

*Chi-Square Measure:* Chi-square can evaluate the cancer data by computing the chi-square statistic with respect to the disease class distribution. This is the



nonparametric statistical approach used to find the difference between the observed disease distribution to the actual non-disease distribution.

$$\begin{aligned} \text{Chisquare}(D_i, D_j) &= \sum (D_i - D_j)^2 / D_j \\ \text{YatesCorr}(\text{Chisquare}) &= \sum (|D_i - D_j - 0.5|)^2 / D_j \end{aligned} \quad (4)$$

*Conditional Entropy:* Let D be the cancer-related dataset with m-classes. Let  $D_i$  be the cancer-related data and  $D_j$  be the non-cancer instances data from the training dataset D. The expected conditional entropy of  $D_i$  given  $D_j$  is given by

$$\text{CEntropy}(D_i, D_j) = \sum \text{pro}(D_i) \log(D_i / D_j) \quad (5)$$

Proposed attribute selection measure is computed using (3), (4), and (5) as

$$\text{SCFSM} = - \frac{\sqrt[3]{\text{CEntropy}(D_i, D_j) * |D| * \text{CubicHellinger}(D_i, D_j)}}{\text{YatesCorr}(\text{Chisquare})} \quad (6)$$

### Ensemble Decision Tree Construction

**Input:** Ranked Features Data as FData;

**Output:** Disease prediction

**Procedure:**

Read cancer disease dataset as CData

For each Feature CData[i] in CData

Do

For each instance  $I(A_i)$  in  $A_i$  do

Do

For each attribute FData( $D_i$ ) do

Divide the data instances of FA( $D_i$ ) into 'k' independent sets.

Select classifier  $C_{i/i} = 1 \dots m$

Load training features and instances

- (a) Construct N subset of trained data and N subset of test data sampling with replacement.
- (b) In the tree growing phase, each and every node select k features at random from N, compute for best split computation using Eq. (6)
- (c) Sort the k individual trees according to cancer and non-cancer.
- (d) Select the majority voting available in each tree using ensemble learning.

End while

Calculate misclassified rate and statistical f-measure, accuracy and true positive rates;

Done

Done

### 4 Experimental Results

In this section, we have executed our proposed model on cancer microarray datasets and compared the results with traditional decision tree models. Dataset [21] used for experimental evaluation are summarized in Table 1. In the experimental results, 10% of the training data are used as testing data for performance evaluation.

Classification statistics of HoeffdingTree using proposed attribute selection measure on ovarian cancer dataset.

Correctly Classified	246	97.2332	%
Incorrectly Classified	7	2.7668	%

TP Rate	FP Rate	Recall	F – Measure	Class
1.000	0.967	1.000	0.786	Cancer
0.033	0.000	0.033	0.064	Normal
0.652	0.619	0.652	0.527	

==== Confusion Matrix ====

a	b	< – –	Classified as
162	0		a = Cancer
88	3		b = Normal

Classification statistics of HoeffdingTree without using proposed attribute selection measure on ovarian cancer dataset.

Correctly Classified Instances	234	92.4901	%
Incorrectly Classified Instances	19	7.5099	%

TP Rate	Precision	Class
0.889	0.993	Cancer
0.989	0.833	Normal
0.925	0.936	

Classification statistics of RandomForest decision tree using proposed attribute selection measure on ovarian cancer dataset.

Correctly Classified Instances	251	99.2095	%
Incorrectly Classified Instances	2	0.7905	%

**Table 1** Datasets and its properties

Dataset name	Features	Type
lungCancer_train	12,533	Numeric
DLBCL-Stanford	4027	Numeric
lung-Michigan	7130	Numeric

TP Rate	FP Rate	Precision	Class
1.000	0.022	0.988	Cancer
0.978	0.000	1.000	Normal
0.992	0.014	0.992	

Classification statistics of Random Forest decision tree without using proposed attribute selection measure on ovarian cancer dataset.

Correctly Classified Instances	238	94.0711 %
Incorrectly Classified Instances	15	5.9289 %

TP Rate	FP Rate	Class
0.981	0.132	Cancer
0.868	0.019	Normal
0.941	0.091	

Table 2, describes the improvement of individual weak classifiers by using our proposed ensemble classifier. Here the accuracy is computed in terms of true positive rate and true negative rate on the ovarian dataset. The classification rate was optimized on the ovarian dataset using our proposed feature selection measure (SCFSM) in the ensemble classifier.

The above graph describes the improvement of individual weak classifiers by using proposed statistical feature selection measure. Existing models used for evaluation are random forest, random tree, and Hoeffding tree. Different classifiers used in the proposed ensemble classification model are a random forest with (SCFSM), random tree with (SCFSM), and Hoeffding tree with (SCFSM) on ovarian cancer dataset. As shown in the figure, classification rate was optimized using proposed ensemble classifier with the traditional classifiers.

**Table 2** Performance analysis of classification accuracy over ovarian dataset

Ovarian dataset	Random forest	Random tree	Hoeffding tree
Existing Algorithm with default Attribute selection measures	94.0711	83.004	92.49
Proposed Ensemble Classifier with SCFSM	99.2095	96.56	97.32

## 5 Conclusion

In this paper, a new feature selection model is proposed for cancer microarray detection. Most of the conventional classification techniques deal with limited attributes and small datasets. Random forest classifier is one of the ensemble learning models, which is capable to handle datasets with a large number of attributes. In this paper, a novel statistical attribute selection measure was implemented for cancer disease prediction. In this work, we have used different decision tree models such as random tree, random forest, CART to evaluate the performance of cancer disease prediction using proposed attribute selection measure. Experimental results are evaluated on different types of microarray cancer datasets including lung cancer, ovarian, lung Cancer and DLBCL-Stanford. Performance of each model is compared in order to find the most efficient and optimized algorithm. Experimental results show that proposed model has high computational efficiency in terms of accuracy and true positive rate. In future, this work can be extended to a large number of instances with high-dimensional feature set.

## References

1. M. M. Yusof, R. Mohamed and N. Wahid, "Benchmark of Feature Selection Techniques with Machine Learning Algorithms for Cancer Datasets", "Proceedings of the International Conference on Artificial Intelligence and Robotics and the International Conference on Automation, Control and Robotics Engineering ACM", 2016.
2. N. Pérez, M. A. Guevara, A. Silva, I. Ramos and J. Loureiro, "Improving the performance of machine learning classifiers for Breast Cancer diagnosis based on feature selection", "Proceedings of Federated Conference on Computer Science and Information Systems, pp. 209–217", 2014.
3. A. Arafı, R. Fajr and A. Bouroumi, "Breast Cancer Data Analysis Using Support Vector Machines and Particle Swarm Optimization", "Complex Systems (WCCS), 2nd World Conference on IEEE", pp. 1–6, 2014.
4. S. Begum, D. Chakraborty and R. Sarkar, "Data Classification Using Feature Selection And kNN Machine Learning Approach", "International Conference on Computational Intelligence and Communication Networks", 2015.
5. S. Begum, D. Chakraborty and R. Sarkar, "Identifying cancer biomarkers from leukemia data using feature selection and supervised learning", "IEEE First International Conference on Control, Measurement and Instrumentation", 2016.
6. A. Bharathi and A. M. Natarajan, "Microarray Gene Expression Cancer Diagnosis Using Machine Learning Algorithms", "International Conference on Signal and Image Processing", 2010.
7. H. Chen, H. Zhao, J. Shen, R. Zhou and Q. Zhou, "Supervised Machine Learning Model for High Dimensional Gene Data in Colon Cancer Detection", "IEEE International Congress on Big Data", 2015.
8. M. V. Dass, M. A. Rasheed and M. M. Ali, "Classification of Lung cancer subtypes by Data Mining technique", "International Conference on Control, Instrumentation, Energy & Communication(CIEC)", 2014.

9. T. Helmy and Z. Rasheed, "Multi-Category Bioinformatics Dataset Classification using Extreme Learning Machine", "IEEE Congress on Evolutionary Computation, pp. 3234–3240, 2009.
10. M. Klassen, "Learning microarray cancer datasets by random forests and support vector machines", "5th International Conference on Future Information Technology IEEE", 2010.
11. C. Arunkumar and S. Ramakrishnan, "Binary Classification of Cancer Microarray Gene Expression Data using Extreme Learning Machines", "IEEE International Conference on Computational Intelligence and Computing Research", 2014.
12. P. Moulos, I. Kanaris, and G. Bontempi, "Stability of Feature Selection Algorithms for Classification in High-Throughput Genomics Datasets", "Bioinformatics and Bioengineering (BIBE) IEEE 13th International Conference, 2013.
13. Z. Nematzadeh, R. Ibrahim and A. Selamat, "Comparative Studies on Breast Cancer Classifications with K-Fold Cross Validations Using Machine Learning Techniques", "Control Conference (ASCC), 10th Asian IEEE, pp. 1–6, 2015.
14. A. Ozcift, "Enhanced Cancer Recognition System Based on Random Forests Feature Elimination Algorithm", "Journal of medical systems 36, no. 4", pp. 2577–2585, 2012.
15. Simes, R. John. "Treatment selection for cancer patients: application of statistical decision theory to the treatment of advanced ovarian cancer." *Journal of chronic diseases* 38.2 (1985): 171–186.
16. Waddell M, Page D, Shaughnessy Jr J. "Predicting cancer susceptibility from single-nucleotide polymorphism data: a case study in multiple myeloma". *ACM* 2005:21–8.
17. Kim W, Kim KS, Lee JE, Noh D-Y, Kim S-W, Jung YS, et al. "Development of novel breast cancer recurrence prediction model using support vector machine." *J Breast Cancer* 2012; 15: 230–8.
18. Tseng C-J, Lu C-J, Chang C-C, Chen G-D. "Application of machine learning to predict the recurrence-proneness for cervical cancer." *Neural Comput & Applic* 2014; 24: 1311–6.
19. Exarchos KP, Goletsis Y, Fotiadis DI. "Multiparametric decision support system for the prediction of oral cancer reoccurrence." *IEEE Trans Inf Technol Biomed* 2012; 16:1127–34.
20. Stojadinovic A, Nissan A, Eberhardt J, Chua TC, Pelz JOW, Esquivel J. "Development of a Bayesian belief network model for personalized prognostic risk assessment in colon carcinomatosis". *Am Surg* 2011; 77:221–30.
21. <http://datam.i2r.a-star.edu.sg/datasets/krbd/index.html>.

# A Novel Region Segmentation-Based Multi-focus Image Fusion Model

Garladinne Ravikanth, K.V.N. Sunitha and B. Eswara Reddy

**Abstract** Multi-focus image fusion scheme integrates multiple input images to obtain a composite fused image. Many research works have been carried out since years and various image fusion approaches were developed. The main idea behind the image fusion is to generate a fused image with enhanced quality and containing more information than that of individual source images. Nowadays, these image fusion techniques are implemented in many applications to combine multi-focus image data into a single composite image. Image fusion models can be categorized into two ways, spatial based fusion and transform based fusion. Transform based fusion is performed in three steps, (1) In the first step, transform coefficients from the input images are turned into transform domain frequencies. (2) In the second step, by applying the fusion rule, these transform coefficients are combined. (3) Through the process of inverse transform on the combined correlated images, fused composite image is generated. In this paper, we have introduced a novel region segmentation based multi-focus image fusion model and implemented it. Proposed model was thoroughly studied, analyzed, and compared with different multi-focus fusion models. Experimental results prove that the proposed model has high computational accuracy in terms of image quality and less error rate compared to traditional models.

**Keywords** Image fusion · Image segmentation · Spatial domain Multi-focus

---

G. Ravikanth (✉)  
BVC College of Engineering, Rajamahendravaram, India  
e-mail: garladinne.ravikanth@gmail.com

K.V.N. Sunitha  
BVRIT Hyderabad College of Engineering for Women, Hyderabad, India  
e-mail: k.v.n.sunitha@gmail.com

B. Eswara Reddy  
JNTUACE, Kalikiri, India  
e-mail: eswar.cse@jntua.ac.in

# 1 Introduction

Image fusion is a key model that plays an essential role in a large number of applications such as medical, remote sensing, image processing etc. Currently, a large number of image fusion algorithms such as pixel based and block based are implemented on the image datasets. To improve the robustness of image processing systems, many images of the same scenario are considered. It is not easy and convenient to analyze such a huge number of images. The main objective of spatial domain fusion is to obtain clearer resultant images as an outcome. Both the transform and spatial domain fusions are divided into numbers of different multi-focus image fusion methods. To overcome this problem many research works have been carried out since years and various image fusion approaches were developed. The basic idea behind these image fusion schemes are, they use necessary information from different images and produce a fused image. Nowadays, these image fusion techniques are implemented in many applications to combine multi-focus image data into a single composite image. Image fusion models can be categorized into two ways, spatial-based fusion and transform-based fusion. Transform based fusion is performed in three steps, (1) In the first step, transform coefficients from the input images are turned into transform domain frequencies. (2) In the second step, by applying the fusion rule, these transform coefficients are combined. (3) Through the process of inverse transform on the combined correlated images, fused composite image is generated. Spatial domain fusion is again classified into two: pixel-based fusion and region-based fusion. The main objective of spatial domain fusion is to obtain clearer resultant images as an outcome.

## 1.1 Transform Domain Fusion.

Transform domain image fusion consists of three phases. In this image fusion technique, the input source images are divided according to their transform coefficients. Then the process of fusion is carried out to merge these transform coefficients and a map is generated known as decision map. In the final phase, through the process of inverse transform on the combined coefficients, a fused composite image comes out as an outcome. The following approaches come under the category of transform domain fusion

**Discrete Cosine Transform.** In Discrete Cosine Transform, the input images are decomposed into blocks of size  $N \times N$  which do not overlap with each other. For each block DCT coefficients are evaluated and fusion rules are applied to get these coefficients.

**Discrete Wavelet Transform.** In this method of decomposition, filters are used at the successive layers to hold the details which are not available in the preceding layers.

**Redundant DWT.** DWT suffers from shift variance problem. To resolve this shift variance problem of DWT, redundant DWT was developed. The inputs are divided into three levels of RDWT. Daubechies filters are used for this process of decomposition. This method generates approximate wavelet bands.

**Stationary Wavelet Transform.** Stationary wavelet transform works almost same as DWT. But the only difference is that it does not support down sampling like DWT. Thus, this method is translation invariant.

**Pyramid Method.** This approach contains set of low-pass or a band-pass copy of image. Pattern information of each copy varies from each other. In this method, each level is represented as a factor of two smaller than that of its predecessor. The higher levels are dependent on the lower level's partial frequencies. The main problem of this method is that the pyramid does not have all information for reconstruction of the original image.

## ***1.2 Spatial Domain Fusion.***

When spatial domain fusion takes images as input, it reduces the signal-to-noise-ratio (SNR) of the composite image with a median filter. The main objective of spatial domain fusion is to obtain clearer resultant images as an outcome. The following are some of the methods that come under spatial domain fusion.

**Principal Component Analysis.** PCA is one of the statistical approaches that converts multivariate data with correlated variables to multivariate data with uncorrelated variables. Linear combination of original variables generates new variables.

**Brovey Method.** This method is also called as color normalization transform method because it contains RGB transform. The major objective of this method is to prevent the demerits of multiplicative method. This collects data from various sensors.

**Intensity Hue Saturation Fusion Method.** Due to the limitation of traditional RGB color space, IHS method uses perceptual color space. Intensity can be defined as total amount of light that reaches the eye. The predominant wavelength of a color is known as hue. Saturation is the purity or the amount of white light out of total amount of light.

## ***1.3 Other Popular Image Fusion Techniques***

Besides transform domain fusion and spatial domain fusion, some other image fusion techniques that have become popular are described below.



**Novel Cross-Scale Fusion.** To fuse large volume of medical images a cross-scale image fusion is developed. It encounters multi-scale decomposition. Fusion system communicates information within the decomposition levels. The information flows from lower to higher decomposition level. This information is essential in the process of evaluation of fusion coefficient.

**Unsupervised Change Detection.** Change detection algorithm in real time is very essential for the applications in the field of motion detection, environmental monitoring, remote sensing, biomedical analysis, and so on. The proposed unsupervised change detection algorithm is a combination of multi-focus image fusion along with kernel k-means clustering. It is highly recommended for better result of synthetic aperture radar images. Image fusion involves some other techniques like mean-ratio, log-ratio approaches, and DWT. As it is a nonlinear clustering, though it decreases false alarm rate, more accuracy can be achieved.

**Un-Decimated Wavelet Transform (UWT).** UWT divides the decomposition process and gives rise to two filtering operations by using spectral factorization of analysis filters. This method converts the one-dimensional signal to high-pass coefficients with the help of filter bank. It resolves an important issue of traditional fusion. The output image reduces unnecessary spreading of coefficients in overlapping images. This creates a problem in the selection and encounters reconstruction errors in the fused image. This technique is also not dependent on basic fusion rules.

**Wavelet-Based Fusion.** While diagnosing a tumor, detection of actual size, and place of the tumor is very important. To solve this wavelet-based fusion was developed. This detects tumor from the complementary and redundant medical images. It collects the input images and produces a fused composite image as an outcome. The approach plays an essential role in the area of tumor detection and diagnosis. Here the maximum values of coefficients with respect to the sharper brightness changes are taken. Images are decomposed like low-high, low-low, high-low, and high-high to support multi-scale transform and multi-resolution analysis. It also prevents overlapping of neighbors' problem of lower band signals.

In this paper, we have thoroughly studied, analyzed, and compared various multi-focus image fusion techniques with our proposed model. We have also identified the need of multi-focus image fusion, merits, and demerits of each and every approach and their performances with respect to the other pre-existing image fusion schemes.

## 2 Related Work

Zhang et al. together introduced a new method known as maximum local energy (MLE) for multi-focus image fusion in the field of mirror extended curvelet transform [1]. They used this to compute low-frequency coefficients of images and

the outputs are analyzed with mirror extended curvelet transform. Thus, edge features and details of images are improved significantly.

The authors described their approach in three major steps, which are

- By applying mirror extended curvelet transform, coefficients of two images are calculated.
- Maximum Local Energy technique is used to evaluate low-frequency coefficients, whereas Absolute Maximum Value technique is applied to find high-frequency coefficients.
- Inverse mirror extended curvelet transform method is used to produce the fused image. The researchers validated their theory with experiments and found that by using MLE approach the performance of image fusion is enhanced remarkably.

Zhang et al. presented a new image fusion algorithm by merging the concepts of variational decomposition and structure tensor analysis [2]. They implemented Rudin–Osher–Fatemi (ROF) model and Chambolle’s projection model in order to divide images into geometric and texture components. Both the components are fused independently. For the fusion of geometrical components, the weighted average approach is used. A vector reconstruction algorithm is taken into account to fuse texture components. Both the results are summed up at the end. The authors carried out a number of experiments and proved that their method outperforms other pixel level approaches. Future work can be done to merge acceleration algorithms with variational decomposition.

Zhang et al. thoroughly analyzed various image registration techniques and suggested a hybrid approach which merges feature-based as well as intensity-based methods [3]. Here, an extended edge-based image registration algorithm was introduced, which was combined with optical flow estimation to perform intensity-based registration smoothly. The authors also went through coarse-to-fine multiscale refinement. Empirical validation was performed by them and the proposed approach was evaluated through experiments. It can be concluded that this method is better than the other conventional image fusion techniques in terms of robustness and performance, which makes the approach more efficient one.

Zaveri et al. formed a new multi-focus image fusion technique and termed it as region-based image fusion scheme [4]. As compared to pixel-based image fusion, the proposed scheme shows better performance. They considered many numbers of source images to implement their work and the fusion outcomes are compared by means of standard reference and non-reference parameters. The researchers simulated their technique and compared it with three other techniques. The resultant outcomes proved that the proposed scheme provides enhanced results in terms of performance, not only than that of the pixel-based method but also than the regular region-based method.

Some of the merits of this approach are

- The unfocused parts of image produce perfect segmentation, enabling the user to retrieve extra details from the different focus parts.
- The algorithm uses two distinct fusion rules for data preservation of the output image, thus enhancing the robustness of the algorithm.
- The algorithm has minimal sensitivity to noise.

It can achieve the optimal result by implementing other different fusion rules and parameters. The major demerit of this approach is: Computation time is quite high as segmentation is implemented twice.

Yang, et al. identified the features of contourlet and presented a new multi-focus image fusion algorithm using contourlet transform [5]. According to their theory, the input images are split into the domain of contourlet transform. The process of fusion is then carried out in the form of sub-bands having distinct scale and direction. For the high-frequency sub-bands, fusion rules must have maximum absolute value coefficient. This rule has a dependency on spatial frequencies. In case of low-frequency sub-bands, the fusion rule has a dependency on spatial frequencies and contrast. Finally, the inverse transform is performed to get the fused image. After performing simulation, the authors found that their proposed scheme has achieved enhanced visual effect along with feasibility and stability than the conventional Laplacian pyramid approach and wavelet-based approach.

Xu et al. developed a new region-based image fusion scheme [6]. Some basic features of the above scheme are mentioned below.

- In-focus regions produce a clear visual, whereas the out of focus regions are blurred.
- Initially an image is split into homogeneous regions and those regions are merged by Local Perceived Sharpness (LPS) to enhance the ability of anti-noise.

By considering different sensitivity level of the human vision system, the researchers introduced this new region-based fusion scheme in the spatial domain. The authors performed a number of experimental evaluations to show that their proposed scheme out performs all other previously existing schemes in terms of objective and visual evaluations.

Yajie et al. suggested a new multi decision-based method for the process of multi-focus image fusion [7]. This method utilizes wavelet transform method. Evaluation function of low frequency can be represented as the sum of Laplacian operator in eight directions. Spatial frequency is used to calculate the evaluation function of high-frequency component. It is evaluated in only the direction of high frequency and also decreases workload significantly. In this research work, the researchers introduced some new universal measures along with the pre-existing measures (entropy and cross entropy). The universal measures are responsible for the integration of objective and subjective factors with spatial frequency. The universal measures are developed by three (horizontal, vertical, and diagonal) directions. No reference image is needed for this process of evaluation. Experiments on their approach produced better results (in case of both subjective and objective analysis) than the pre-existing methods.

Wan et al. tried to merge multiple images having different focus point to result in an all-in-focus type image [8]. To resolve the issue of multi-focus image fusion, they applied Robust Principal Component Analysis decomposition protocol to generate a sparse matrix. RPCA is used for the formation of composite feature space. The authors also mentioned that the features of sharp regions can be merged together. They decomposed the sparse matrix into smaller blocks and evaluated the standard deviation in every individual block. In order to smoothen the inter-block transitions, the researchers implemented a sliding window method. Eventually, they demonstrated that the quality of fusion is quite good with respect to visual and quantitative evaluations. It is applicable for both color and grayscale images. As the algorithm does not perform well in case of noisy source images, future work can be carried out in the direction of applying this approach to noisy images.

Suryavanshi suggested a new method of pixel significant-based multi-focus image fusion by using biorthogonal wavelets [9]. It works in 2 steps-analysis (decomposition) and synthesis (reconstruction). Here a weighted average of source pixels is considered to evaluate the fused pixel value. The weights associated with each pixel are calculated by parent-child relationship between the pixels in multi-resolution decomposition. The researchers simulated their method by testing the performance of this method with respect to five other methods by using some parameters (including Petrovic parameters). The above approach enhances the quality of the fused image.

Niu et al. introduced a new technique to find an optimal number of decomposition levels in wavelet-based multi-focus fusion [10]. The authors validated their technique and concluded that, the optimal number of decomposition level is not always constant. It varies according to the characteristics of original images and the number of optimal levels of decomposition can be monitored.

Ma et al. developed a new multi-focus image fusion algorithm for non-subsampled contourlet transform (NSCT) [11]. The authors considered the main component of NSCT band-pass sub-band coefficients as the target. They also constructed extreme problem for energy functional in order to get the closest target and termed it as fused coefficient. They implemented gradient-descent to reduce the functional and generate the numerical strategy.

Li used focused pixel detection method to present their new multi-focus image fusion technique [12]. The whole process is carried out in the following steps

- The conventional Multi-Scale Top-Hat transform is extended to extract focus information (pixels of the focused region).
- The decision map is updated according to the extended MTH values for each pixel. (3) The isolated regions are discarded from the decision map and the map is updated.
- A double sliding window method is used to increase the quality of resultant fused image and prevent interruption in the transition zone. This is responsible for fusion of transition zones.
- Both the decision maps are used in fusion to generate the ultimate fused image.

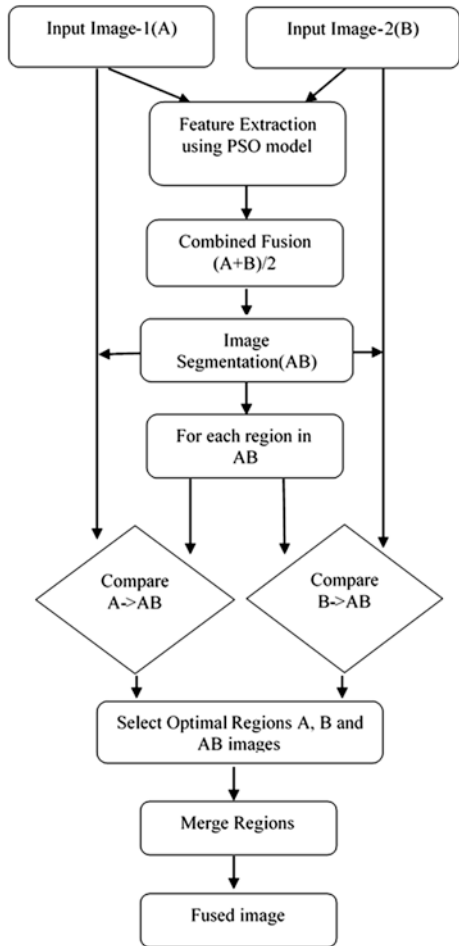
Hariharan et al. suggested a new approach to multi-focus image fusion by utilizing focal connectivity [13]. This approach can be implemented to a diversity of source images. They decomposed the input images into smaller blocks on the basis of their focal connectivity. Focal connectivity can be defined as the separating regions of an input image for that particular focal plane. The authors considered focal connectivity as the prime measure of this method, whereas this method discards the physical properties from its consideration. With the help of sharpness maps, the researchers were able to isolate and attribute image partitions in case of input images. The decomposed smaller blocks are combined together to produce the fused image. The authors tested their proposed technique and provided a comparison with respect to the other approaches and also proved that their approach is better than that of other pre-existing ones.

Gabarda et al. proposed a multi-focus image fusion scheme on the basis of pseudo-Wigner's distribution [14]. Here the source images having different spatial degradation patterns are taken as input images. Their fusion process emphasizes defocusing pixel-level measure. Pixel-level measure is a single dimensional pseudo-Wigner distribution (PWD) and is implemented on  $N$ -pixel window slices without overlapping. They applied the same procedure repeatedly till it covers the complete image. The authors experimented on a low-resolution image that is obtained by blurring and averaging two source images. Similarly, the above technique is applied to other images of the same scene (partly focused and partly defocused). They also argued that their scheme can be applied to any source image. It also eliminates mean square errors. Thus, the rate of correctness of this framework is high as compared to all existing methods and it also results in minimal computational cost

### 3 Proposed Model

In this section, a novel multi-focus image fusion with optimization model using region segmentation and local spatial frequency is presented. The basic proposed model is shown in Fig. 1. As shown in Fig. 1, two input source images  $I_A$ ,  $I_B$  are taken as input for feature extraction process. PSO based Feature extraction is used to find the spatial frequency of image A and image B to form a new image AB. After that, fusion based segmentation approach was applied on each image A, B, and AB. After segmentation, each region of A is compared to AB to find the optimal segmented fused regions for region merging process. Similarly, each region of B is compared to AB to find the optimal segmented fused regions for region merging. The multi-focus image fusion model can be accomplished using the following steps.

**Fig. 1** Flowchart of the proposed model



### 3.1 Image Feature Extraction

- Step 1 Divide the source image  $I_A, I_B$  into  $m, n$  blocks using PSO optimization model.
- Step 2  $BI_A(m, n)$  and  $BI_B(m, n)$  are the blocks of source images  $I_A$  and  $I_B$  using the PSO optimization model.
- Step 3 Compute row wise spatial frequency for a given  $BI_A(m, n)$  and  $BI_B(m, n)$  as

$$\phi_1 = \sqrt{\frac{1}{|m||n|} \sum_{i=1}^{|m|-1} \sum_{j=1}^{|n|-1} [v(i, j) - v(i, j - 1)]^2} \quad (1)$$

Step 4 Compute column wise spatial frequency for a given  $BI_{A(m, n)}$  and  $BI_{B(m, n)}$  as

$$\phi_2 = \sqrt{\frac{1}{|m||n|} \sum_{i=1}^{|m|-1} \sum_{j=1}^{|n|-1} [v(i, j) - v(i - 1, j)]^2} \quad (2)$$

Total spatial frequency

$$(\phi_1 + \phi_2) = \sqrt{\phi_1^2 + \phi_2^2} \quad (3)$$

Step 5 Construct the average image  $(A+B)/2$  using the steps 3 and 4 as a temporarily fused image.

Step 6 Compute F Segmentation( $I_A, k$ ), F Segmentation( $I_B, k$ ) and F Segmentation( $I_{AB}, k$ ). where  $k$  = number of classes, Default:  $k = 2$

F Segmentation( $I_A, k$ )

$$\lambda_0 = \frac{\sum I_{A(m,n)}(i, j)}{|I_A|} \quad (4)$$

Partition the image  $I_A$  into  $k$  classes as

$$\begin{aligned} C_1 &= \{0, 1, 2, \dots, \lambda\} \\ C_2 &= \{\lambda + 1, \lambda + 2, \dots, N - 1\}, N : \text{total number of grey levels of an image} \end{aligned} \quad (5)$$

Calculate the lower edge threshold level. Let  $p_{i,j}^1$  is the probability of occurrence of  $(i, j)$ th pixel in foreground classes of  $I_A$ , then we have,

$$p_{i,j}^1 = \frac{\text{Number of pixels of } I_A^1(i, j) \text{ in each block region of } I_A^1}{\text{Total number of pixels in } I_A^1} \quad (6)$$

$$P_{c_1}^1 = \sum_{i=0}^{\lambda} p_{i,j}^1 \quad (7)$$

Let  $p_{i,j}^2$  is the probability of occurrence of  $(i, j)$ th pixel in background classes of  $I_A$ , then we have,

$$p_{i,j}^1 = \frac{\text{Number of pixels of } I_A^1(i, j) \text{ in each block region of } I_A^1}{\text{Total number of pixels in } I_A^1} \quad (8)$$

$$p_{c_1}^1 = \sum_{i=0}^{\lambda} p_{i,j}^1 \quad (9)$$

Let  $p_{i,j}^2$  is the probability of occurrence of  $(i, j)$ th pixel in background classes of  $I_A$ , then we have,

$$\begin{aligned} p_{i,j}^2 &= 1 - p_{i,j}^1 \\ p_{c_2}^2 &= 1 - p_{c_1}^1 \end{aligned} \quad (10)$$

The average of two classes  $C_1$  and  $C_2$  can be computed as

$$\lambda_1 = \bar{X}_{c_1} = \sum_{i=0}^{\lambda} i * \frac{p_{i,j}^1}{p_{c_1}^1} \quad (11)$$

$$\lambda_2 = \bar{X}_{c_2} = \sum_{i=\lambda+1}^{N-1} i * \frac{p_{i,j}^2}{p_{c_2}^1} \quad (12)$$

Computing correlated inter- and correlated intra-class variance to the given image A as:

$$\text{var}_{c_1}(\lambda_1) = \frac{\sum_{i=0}^{\lambda} (i - \lambda_1)^2 \cdot p_{i,j}^1}{p_{c_1}^1} \cdot \text{corr}(c_1, c_2) \quad (13)$$

$$\text{var}_{c_2}(\lambda_2) = \frac{\sum_{i=0}^{\lambda} (i - \lambda_2)^2 \cdot p_{i,j}^2}{p_{c_2}^1} \cdot \text{corr}(c_1, c_2) \quad (14)$$

$$\text{var}_{\text{inter}}^{\lambda} = \text{var}_{c_1}(\lambda) + \text{var}_{c_2}(\lambda) \quad (15)$$

$$\text{var}_{\text{intra}}(\lambda_1, \lambda_2) = \frac{p_{c_1}^1 \cdot p_{c_2}^1 |\lambda_1 - \lambda_2|}{\text{Corr}(c_1, c_2)} \quad (16)$$

$\text{corr}(c_1, c_2)$  is the correlation between two block regions of  $c_1, c_2$



Step 7 Find all the pixel block regions of less than  $\lambda$  and mark them as '0', as  $SI_A[0]$

Step 8 Find all the pixel block regions with  $\geq \lambda$  to 255 values, as  $SI_A[1]$ .

#### F Segmentation( $I_B, k$ )

Step 9 Repeat Eqs. (4)–(16) for F Segmentation( $I_B, k$ ). Find all the pixel block regions of less than  $\lambda$  and mark them as '0', as  $SI_B[0]$

Step 10 Find all the pixel block regions with  $\geq \lambda$  to 255 values, as  $SI_B[1]$ .

#### F Segmentation( $I_{AB}, k$ )

Step 11 Repeat Eqs. (4)–(16) for F Segmentation( $I_{AB}, k$ ). Find all the pixel block regions of less than  $\lambda$  and mark them as '0', as  $SI_{AB}[0]$

Step 12 Find all the pixel block regions with  $\geq \lambda$  to 255 values, as  $SI_{AB}[1]$ .

Step 13 Compare segmented regions of  $SI_A, SI_B$  to  $SI_{AB}$  and merge all high-frequency pixels as the fused image.

## 4 Experimental Results

Experimental results are performed on the standard multi focus fusion image data sets taken from <http://dsp.etfbl.net/mif/>. In order to analyze the efficiency of our model, image fusion experiments are performed on several pairs of noisy multi-focus images. The average values of structural similarity index (SSIM), correlated measure and mean squared error (MSE) for the proposed model with the traditional models are performed on the datasets. Further, correlation coefficient is used as performance metric to provide the comparison between the segmented regions.

Figure 2 shows the input source image for multi-focus image fusion process. Here paper image with 50% noisy information was taken as image A.

Figure 3, shows the input source image for multi-focus image fusion process. Here paper image with 50% noisy information was taken as image B.

Figure 4, shows the fused image of two input images A and B using proposed model. The noisy information of the proposed image is reduced up to 15% in the fused image.

Similarly Figs. 5 and 6 show the input source images of a flower with noise and Fig. 7 shows the fused image with reduced noise.

SSIM: Structural similarity index is used to compute the image distortion level between the fused image (FImg) and the two source images (A & B). The measure is given by

Fig. 2 Input source A

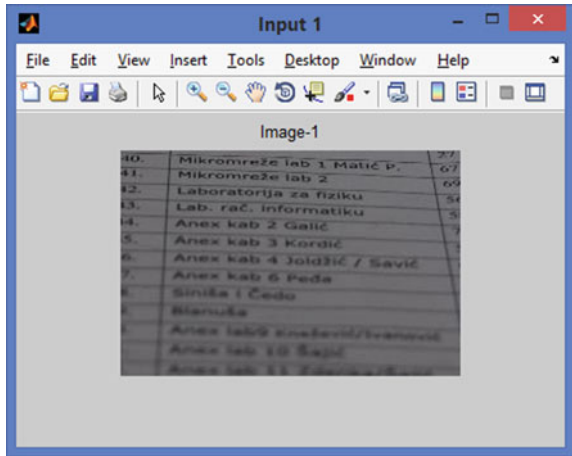
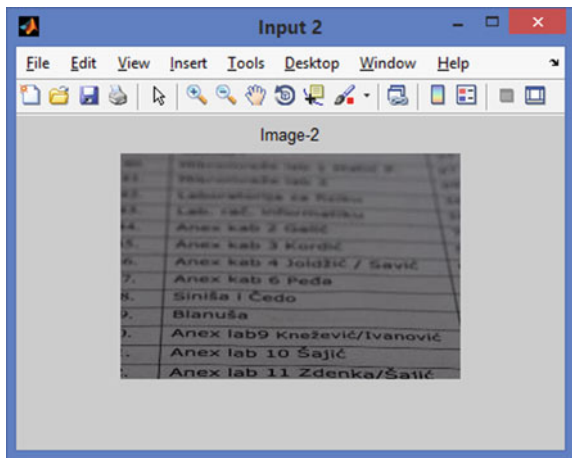


Fig. 3 Input source B



$$SSIM_{FIm_g}^{A,B} = \frac{1}{N} \sum_{i=1}^N [\varphi(i) \cdot \text{sim}(A_i, FIm_g) + (1 - \varphi(i)) \cdot \text{sim}(B_i, FIm_g)] \quad (17)$$

where ‘sim’ is the similarity index between two images and  $\varphi(i)$  is the covariance between the two images. Mutual information is used to evaluate the quantitative analysis of the fusion method. The MI and MSE are represented in the Tables 1 and 2.

From the Fig. 8, it is clearly observed that proposed model has high computational MI compared to traditional models on the standard multi focus image fusion test dataset.

Fig. 4 Fused image

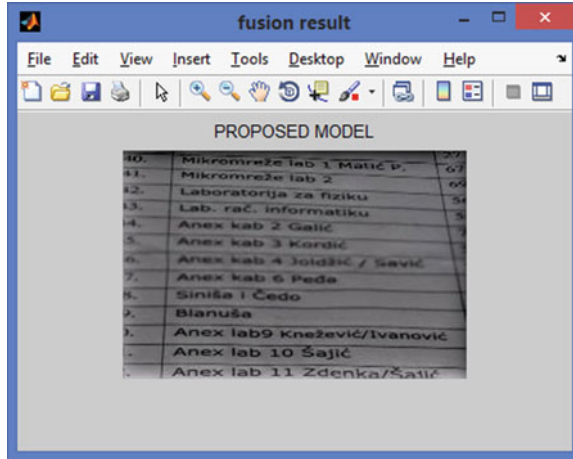
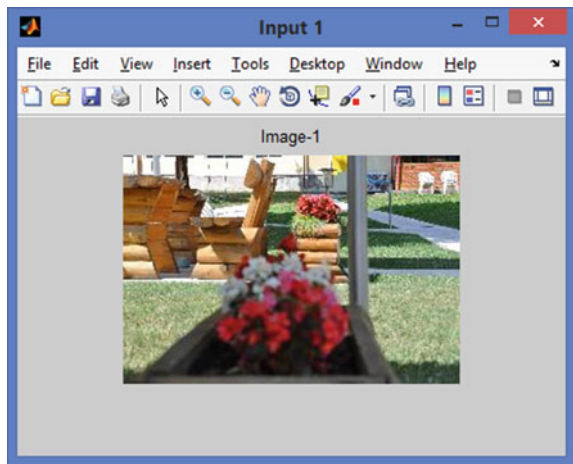


Fig. 5 Input source A



From the Fig. 9, it is clearly observed that proposed model has high computational SSIM compared to traditional models on the standard multi focus image fusion test dataset.

Fig. 6 Input source B

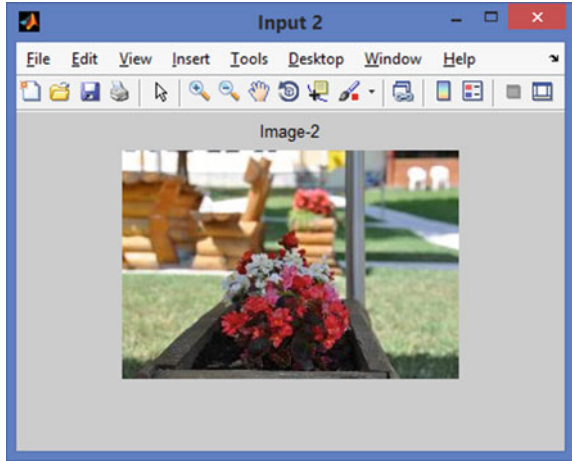


Fig. 7 Fused image

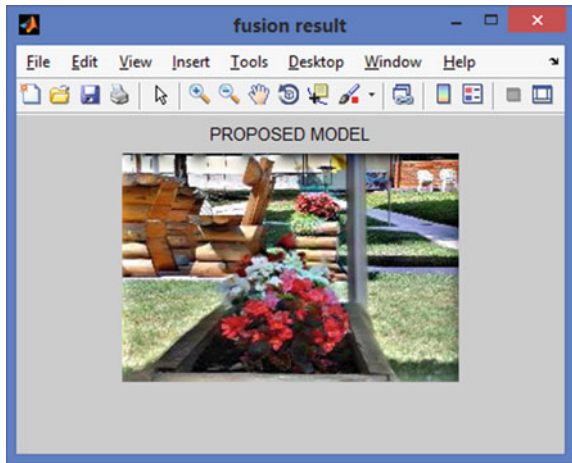
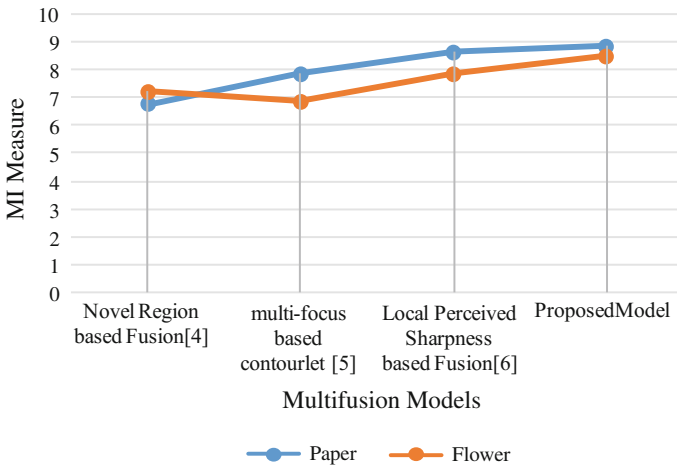


Table 1 MI performance analysis

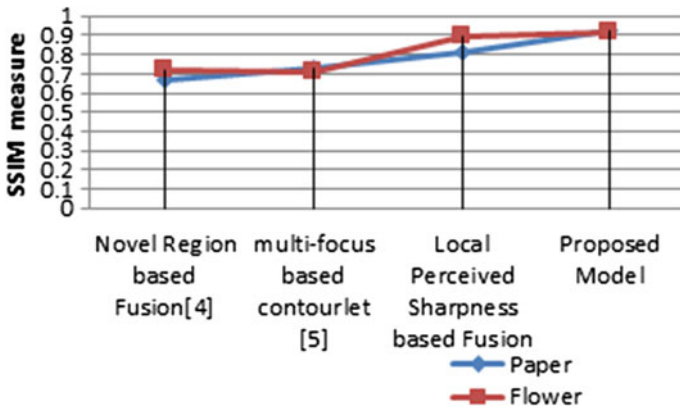
Mutual information (MI)	Novel region-based fusion [4]	Multi-focus based contourlet [5]	Local-perceived Sharpness-based Fusion [6]	Proposed model
Paper	6.743	7.845	8.625	8.857
Flower	7.246	6.864	7.854	8.481

**Table 2** SSIM performance analysis

SSIM	Novel region-based fusion [4]	Multi-focus-based contourlet [5]	Local-perceived sharpness-based fusion	Proposed model
Paper	0.67	0.73	0.815	0.93
Flower	0.72	0.704	0.897	0.916



**Fig. 8** MI performance of proposed model with the traditional models



**Fig. 9** SSIM performance of proposed model with the existing models

## 5 Conclusion

Noadays, image fusion algorithms are implemented in many applications to combine multi-focus image data into a single composite image. In this paper, we have proposed a novel region segmentation based multi-focus image fusion model and implemented it. In this method, the undistorted edge of image is enhanced to improve the quality of the fused image. Proposed model was thoroughly studied, analyzed and compared with various multi-focus image fusion techniques. This method uses the correlation coefficient value to measure the distortion similarity using the segmentation method. Experimental results show that the proposed model has high computational accuracy in terms of image quality and less error rate compared to traditional models. In future, this work can be extended to remote sensing images with noise filter techniques.

## References

1. Kim, Hyung-Tae et al. "Optical Distance Control for A Multi Focus Image In Camera Phone Module Assembly". *International Journal of Precision Engineering and Manufacturing* 12.5 (2011): 805–811. Web. 19 Mar. 2017.
2. Lee, Seung-Hyun et al. "Multi-Focus Image Fusion By Using A Pixel-Based SML Comparison Map". *Computer Science and its Applications* (2015): 615–621. Web. 19 Mar. 2017.
3. Shreyamsha Kumar, B. K. "Multifocus And Multispectral Image Fusion Based On Pixel Significance Using Discrete Cosine Harmonic Wavelet Transform". *Signal, Image and Video Processing* 7.6 (2012): 1125–1143. Web. 19 Mar. 2017.
4. T. Zaveri, M. Zaveri, V. Shah and N. Patel, "A Novel Region Based Multi-focus Image Fusion Method", "Journal of Digital Image Processing", pp. 50–54, 2009.
5. Xuejun, Li, and Wang Minghui. "Research Of Multi-Focus Image Fusion Algorithm Based On Sparse Representation And Orthogonal Matching Pursuit". *Communications in Computer and Information Science* (2014): 57–66. Web. 19 Mar. 2017.
6. L. Xu, J. Du, J. M. Lee, Q. Hu, Z. Zhang, M. Fang and Q. Wang. "Multi-focus Image Fusion Using Local Perceived Sharpness", "25th Chinese Control and Decision Conference (CCDC)", pp. 3223–3227, 2013.
7. "Multifocus Image Fusion Based On NSCT and Focused Area Detection - IEEE Xplore Document". *Ieeexplore.ieee.org*. N.p., 2017. Web. 19 Mar. 2017.
8. T. Wan, Z. Qin, C. Zhu and R. Liao, "A Robust Scheme for Multi-focus Images using Sparse Features", "Pattern Recognition Letters 34.9", pp. 1957–1961, 2013.
9. Zhong, Fuping, Yaqi Ma, and Huafeng Li. "Multifocus Image Fusion Using Focus Measure Of Fractional Differential And NSCT". *Pattern Recognition and Image Analysis* 24.2 (2014): 234–242. Web. 19 Mar. 2017.
10. Yong Yang, "A Novel DWT Based Multi-Focus Image Fusion Method – Science direct".
11. N. Ma, L. Luo, Z. Zhou and M. Liang, "A multifocus image fusion in non sub sampled contourlet domain with variational fusion strategy", "Seventh International Symposium on Multispectral Image Processing and Pattern Recognition (MIPPR 2011)", 2011.
12. H. Li, Y. Chai and Z. Li, "A new fusion scheme for multifocus images based on focused pixels detection", "Machine vision and applications 24.6", pp. 1167–1181, 2013.

13. H. Hariharan, A. Koschan and M. Abidi, "Multi-focus Image Fusion By establishing Focal Connectivity", "IEEE International Conference on Image Processing. Vol. 3", 2007.
14. S. Gabarda and G. Cristóbal, "Multifocus image fusion n through pseudo-Wigner Distribution", "Optical Engineering-44.4", 2005.

# OntoYield: A Semantic Approach for Context-Based Ontology Recommendation Based on Structure Preservation

G. Leena Giri, Gerard Deepak, S.H. Manjula and K.R. Venugopal

**Abstract** With the introduction of the Web 3.0 standards on the World Wide Web, there is a need to include semantic techniques and ontologies in the Web based Recommendation Systems. In order to build query relevant domains and make information retrieval more efficient, it required recommending ontologies based on the query. Most ontology recommendation systems do not preserve the associations and axioms between them rather ontology matching and clustering algorithms tend to deduce logics dynamically. In this paper, a semantic algorithm for ontology recommendation has been proposed, where query-relevant ontologies are recommended by preserving the relationships between the ontological entities. The semantic similarity is computed using the query and the concepts initially and further between the query and description logics which makes it a context-based ontology recommendation system. A strategic approach called as SemantoSim is proposed to compute the semantic similarity.

**Keywords** Ontologies · Ontology recommendation · Recommender systems  
Semantic similarity · Web 3.0

## 1 Introduction

The World Wide Web is the World's largest storehouse of information. The Web is constantly expanding and significantly growing beyond immeasurable capacity. There is urgency in organizing the information on the World Wide Web. The heterogeneity of information in the Web is the main reason for the need of its organization. Due to the enormous amount of Web data, retrieving the useful information is a tedious task. The task of enhancing the relevance of Web Search can

---

G. Leena Giri (✉) · G. Deepak · S.H. Manjula · K.R. Venugopal  
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, Bangalore University, Bangalore, Karnataka, India  
e-mail: leenagiri.g@dr-ait.org



be simplified by the incorporation of ontologies. Ontologies are explicit specification of conceptualizations [1]. They are the most basic entities of Semantic Web.

The problem of Web-based recommendation can be quite easily solved by the organization of the ontological entities. Ontologies are recommended by several types of evaluations. Although a few systems use semantics for recommending ontologies, the original structures of the ontological relations are never preserved. The preservation of ontological structures is highly important. If a specific ontology has more than one context, then such ontology is said to be ontologically committed. Most ontologies are associated with ontological commitments. If an ontology that is committed is “bat” which has two explicit meanings which is “a bird” and “sports bat” that can further be categorized as a “cricket bat”, “tennis bat” or even a “baseball bat”. “Bat” is as well associated with a “British Pub Game” named “Bat-and-Trap”. Likewise, “bat” can be used in various contexts like “to bat” or even “batting”. These are only a few contexts for the ontology “bat”. However, there are many more contexts for the term “bat” and the list goes on. Most ontology matching and recommendation systems only have a basic context check for the ontologies where inference is drawn by comparing the concepts and individuals. The axioms and structures are broken and made as per the context of the query using axiomating agents which contain previously defined axioms. However, the context is just assumed by these agents based on the previous. There is a need for preserving the original structure between the ontological entities but it is a challenge to preserve them during clustering of ontologies which is solved in the proposed system. An algorithm for recommending ontologies by preserving the original axiomatization between the concepts and individuals of Domain Ontologies is proposed.

**Motivation:** The traditional web searches that are based on ontologies mainly treat ontologies as independent terms with a certain degree of ontological commitments. The evident characteristic of these recommendations is the break and make the relationships between ontologies. It is quite essential and necessary to preserve the axiomatic relationships between several ontologies with multiple contexts. With preservation of axioms between ontology terms, the overall relevance of recommendation can be increased.

**Contribution:** An algorithm for cross-domain ontology clustering for terms with more than ontological context belonging to heterogeneous domains is proposed. The Interdomain relationships between the ontological entities are preserved. The clustering takes place based on the search term of a specific ontology. The query-relevant ontology recommendation is based on ontology matching using a semantic strategy. A SemantoSim measure for computing the semantic relatedness for matching of the ontologies is proposed. The semantic similarity is computed between the query words and the concepts initially. The description logics between the matching concepts and its individuals are preserved. Furthermore, the semantic heterogeneity between the query words and the description logics is computed to evaluate the correctness of the context of ontologies. Finally, the final concepts and its individuals along with their original relationships are clustered together and are yielded to the user. The precision, recall, and accuracy of the proposed system is enhanced.

Organization: The remaining of this paper is organized as follows. Section 2 provides a brief overview of Related Work. Section 3 depicts the Proposed Architecture. Section 4 discusses the Implementation in detail. The Performance Analysis & Results are discussed in Sect. 5. Section 6 concludes the paper.

## 2 Related Work

Ivan et al. [2] have improved the CORE [3] framework for recommending ontologies and its reuse. The proposed system uses informal description for domain specific ontologies and uses WordNet for refining them. The evaluations are manual and needs human collaborative assessment for determining if the ontologies were domain specific. Marcos et al. [4, 5] have proposed a methodology for Ontology Recommendation using Collaborative Knowledge. The ontologies are recommended with the motive of information organization for a set of initial terms by considering the semantic richness, coverage and the popularity of ontology in the existing Web. Marcos et al. [6] have proposed another Ontology Recommender, which evaluates biomedical ontology based on coverage, acceptance of the ontology, details of ontological classes, and specialization of the ontology with respect to the input domain data and have conducted evaluations.

Małgorzata et al. [7] have proposed a semantic methodology for validating the classes in UML models. The UML diagrams are validated based on their semantic correctness without the involvement of Domain Experts. Doan et al. [8] have proposed a system named GLUE for ontology matching based on one-on-one mapping of the ontological entities. The proposed system GLUE is based on machine learning that semi-automatically creates semantic mappings. Todd et al. [9] have discussed the fundamental problems that are involved in Ontology Alignment along with detailed discussions of the various complexities and anomalies faced by systems for aligning ontologies.

Sergio et al. [10] have proposed a technique for ontology matching by visualizing the problem of ontology matching as a binary classification problem. A pattern classification model by aligning the instances of heterogeneous ontologies is proposed. The ontology structure preservation and context based ontological matching or recommendation is never targeted. Ujwala et al. [11] have proposed an ontology matching technique by deriving the degree interoperability between informational sources. The experimentation is done for the geospatial domain thus constituting Geo-ontologies. An ontology matching framework by incorporating interoperability measurement is proposed.

Anam et al. [12] have proposed a Knowledge-Based Schema Matching technique for mapping ontologies. A machine learning approach is incorporated into the system which is further for classification and further incremental knowledge is installed into a graph schema. This approach is an integrated approach that combines machine learning and knowledge acquisition as a Hybrid Ripple Down Rules approach.

Ranjini et al. [13] have proposed an approach for identifying the semantically rich concepts by incorporating concept and relationship level classification. The approach is a weight-based iterative approach, where weights are assigned to clusters of ontologies and ranking is done based on these weights.

### 3 Proposed Architecture

The architecture of the proposed system is depicted in Fig. 1, to which a specific query for which ontologies should be recommended is entered. The query can be specified by a user or by ontology experts like the Knowledge Engineers or Domain Experts. The query can also be entered by bots or third-party systems seeking domain ontologies. As the system is queried, the query undergoes preprocessing. The query preprocessing is done by tokenization and stemming. A customized blank space Tokenizer tokenizes the query into individual query words. The query generally is to search and retrieve a class or a group of ontologies from an OWL ontology domain without breaking the axioms and the relationships between the ontological entities. The principal reason for stemming is to eliminate the unwanted and redundant stop words from the query. The query preprocessing yields a set of unique query terms for which the ontologies that are context relevant must be yielded.

OWL ontologies have a characteristic property of description logics. The concepts and individuals of owl ontologies are described using semantic description

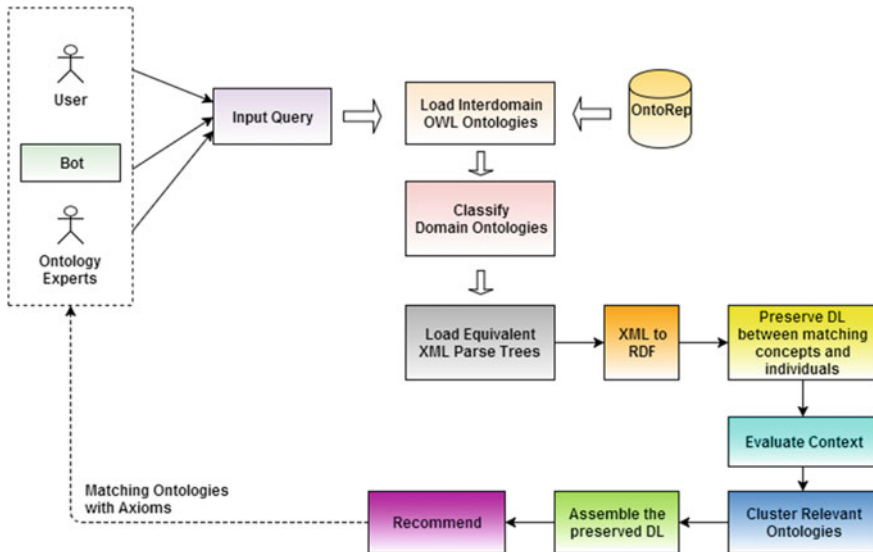


Fig. 1 Proposed architecture of OntoYield

logics. As only Ontological Experts and Knowledge Engineers are capable of interpreting a .owl file and to overcome this anomaly, the proposed OntoYield system is made capable of loading and classifying the .owl ontologies. It allows the users' to add Domain Level Ontologies as .owl files. OntoYield is able to convert the .owl file structure into its equivalent XML parse tree structure.

The ontological elements are both concepts and individuals that need to be organized. Once the equivalent XML schema for the OWL Ontological Domains is obtained, they are expressed into their triadic RDF format. The RDF or the Resource Description Framework is a metadata model that is used to describe the ontologies and directly relate them to each other based on their description logics. The description logic analysis of the concepts and individuals of ontologies will give a discernment of the context of the ontologies. The OWL ontologies are stored in the repository called as OntoRep. OntoRep is the storehouse of .owl formats of several Domain Level ontologies. The Semantic Similarity is computed initially between the query words and the conceptual nodes of the XML parse tree formulated from the Domain-Level OWL Ontologies. The RDF equivalence of the matching nodes is retrieved and the semantic similarity is computed between the description logic of the matching concepts and individuals.

The semantically relevant concepts and individuals with relevant description logics are clustered together. The description logics of the Ontologies which describe their hierarchy are also extracted when the clustering on ontologies takes place and are used for axiomitization by the Semantic Agents. The ontologies are arranged in the increasing order of their semantic similarity. The traditional graph-based approach for processing the hierarchical ontologies is overcome by incorporating a HashMap, HashTable Approach. The usage of Graphs for processing the ontologies is implemented as the relationships between the ontological entities can be expressed quite distinctively. The usage of graphs can be tedious and they tend to increase the overall complexity of the system. Moreover, the axioms retention between the concepts and individuals or between two concepts of the same domain is highly cumbersome. This anomaly can be solved using a HashMap and a HashTable coherently to store and map the concepts and individuals along with their axioms. The hierarchy of the concepts is maintained as it is in the same order and form the Key of the HashMap. The individuals for specific concepts are concatenated with a unique HashValue generated. The axioms that are description logics to the individuals are stored in the HashTable as values where the HashValue forms the Key. This is done in with the main objective of preserving the relationships, hierarchies and axioms between the concepts, individuals and axioms of a specific domain. There is a need for the generation of a unique HashValue to ensure that the description logics for the concepts and individuals are kept in order without losing its track.

The Semantic Similarity is computed between the query words and the concepts at first. Further, the description logics in the HashTable of matching concepts are loaded to compute its semantic similarity with the query words in order to derive the matching scenario and context. Only the concepts and description logics whose semantic similarities are matching are clustered together and further arranged in

their increasing order of their semantic similarity. Finally, the axioms in the form of description logics are induced between the concepts and the individuals based on the unique hash value. This definitely solves the problem of preserving the axioms between the concepts and individuals without any further problem. The Semantic Similarity is computed using the SemantoSim measure which is a semantic measure inspired from the Point wise Mutual Information [14] measure proposed by Church and Hanks.

$$\text{SemantoSim}(x, y) = \frac{\text{pmi}(x, y) + p(x, y)\log[p(x, y)]}{[p(x) \cdot p(y)] + \log(p(y, x))} \quad (1)$$

The SemantoSim measure is a semantic similarity measure that is derived from the Point wise Mutual Information measure and is a normalized semantic measure. The query terms are paired as  $(x, y)$  if there are two terms in the query. If there are three query terms then permutations of the three pairs are considered. If it's a single-term query, then semantic similarity for the term  $x$  and its most closely related semantically relevant term is considered. The query terms are the query words that are tokenized and stemmed. The SemantoSim yields the semantic relatedness between two terms  $(x, y)$ . The  $\text{pmi}(x, y)$  is computed using Eq. (1). The expression  $p(x, y)$  is the probability of the term  $x$  in its co-occurrence with  $y$ .  $p(x)$  is the probability of occurrence of the term  $y$  with  $x$ .  $p(x)$  and  $p(y)$  are the probabilities of the presence of the terms  $x$  and  $y$  respectively.

## 4 Implementation

The proposed system OntoYield is implemented in JAVA with MYSQL lite as the database. Netbeans was used as the preferred IDE for its ease of use. The HashMap and the HashTable are included from the Collections framework in JAVA. The front end is designed using JAVA SWINGS framework and HTML.

The ontologies used in the OWL format were domain specific and were collected from different sources. Since the Domain-Level OWL ontologies along with their axioms were collected from heterogeneous online ontology sources. Moreover, in order to suit the domains and increase the number of ontology entities and create an overall environment, ontologies were modeled using Protégé.

OntoCollab [15] was incorporated to add the more individuals and concepts with necessary axioms and description logics into the datasets considered for experimentation. The axiomitization between the concepts and individuals was made complex with extensively related description logics. The mapping between the concepts and individuals were many to many. The concepts behaved as individuals to a different set of concepts making the hierarchy more interesting and tough. Such datasets were chosen to achieve the recommendation results with the highest degree of relevance.

**Table 1** Ontologies and their Contexts used

OWL Ontologies	Contexts	No. of concepts	No. of individuals
Products Ontology	Electronics, Baking, Products, Food, Crockery	27	118
Books Ontology	Fiction, Work Books, School Text Books, Author Based and Computer Science Books	85	292
Cake Ontology	Flavor, Brand, Price	18	54
Furniture Ontology	Wooden Metal, Modern, Contemporary, Type of Wood	27	86
Cars Ontology	Brand Price, Model, Based on Sedan or Hatch Back, Engine	22	56

The Domain Ontologies considered for experimentation were stored in the OntoRep. One of the ontologies chosen was the “Book” Ontologies that had several contexts. Each context for the “Book” ontology was incorporated in separate .owl files. Though the term “Book” can be expressed as a specific descriptive ontology, it can have several contexts with respect to authors, subjects, type of books, target people of the book, etc. Similarly, another prospective ontology can be a simple “Product” ontology which can have its context in “Supermarket” ontology, “ECommerce”, etc. If the specific product is an electronic item, it can have its context in “Student Project” ontology or even in the “ElectronicItems” ontology. The details of ontologies, their different contexts and the number of concepts/individuals associated with the Ontology Datasets is depicted in Table 1.

A sequential algorithm for the proposed OntoYield system is depicted in Table 2. The query input by the system is subject to preprocessing which further yields the individual query words. The Ontology Repository is Looked Up for the keyword matches based on which the Domain Level OWL Ontologies are loaded. Further, the equivalent XML parse tree for the Domain Ontologies is loaded. It can be observed that the algorithm involves the computation of semantic similarity twice.

## 5 Performance Analysis and Results

The system was queried with distinct queries mentioned in Table 3 and the performance was evaluated for the proposed system for Ontology Recommendation. The ontologies that were specific to the query were clustered and recommended based on their SemantoSim values preserving the axioms between the concepts and individuals. Precision, Recall, and Accuracy are the metrics that were chosen for evaluating the Performance as they are the most preferred measures for any class of recommendation or information retrieval systems. Since OntoYield is a system that recommends and ontologies by preserving the ontological structures, the chosen metrics are comprehensible. Precision is defined as the ratio of the retrieved and

**Table 2** Proposed algorithm for recommending Contextual Ontologies in Heterogeneous Domains by preserving original axioms

Input: Query Q for recommending Ontologies either from a user or a Bot, Heterogeneous and Contextually related Domain Ontologies
Output: An artifact of Highly Relevant, Clustered Ontologies with their original axioms intact
<i>Begin</i>
Step 1: The query entered is Tokenized and Stemmed to remove all the stop words and yield a set of query words Q'
Step 2: for each Q'. tokens Lookup OntoRep and Load the initial Matches based on keyword matching
Step 3: Load the matching Domain-Level OWL ontologies and interpret as the XML parse trees based on the hierarchies of the concepts and individuals
Step 4: for each concept node c in XML tree ss = SemantoSim(q', c) if(ss < 0.25) HashMap H <sub>c</sub> ← c, ss
Step 5: for each c in H <sub>c</sub> for each individual I → C Φ <sub>i</sub> = Generate Unique HashValue C <sub>i</sub> = Concatenate < Φ <sub>i</sub> , I > Extract Description logics DL <sub>i</sub> of I → C HashMap H <sub>i</sub> ← c, C <sub>i</sub> HashTable H <sub>m</sub> ← I, DL <sub>i</sub>
Step 6: for each DL in H <sub>m</sub> ss1 = SemantoSim(q', DL <sub>i</sub> ) if(ss1 < 0.25) L = Tokenize and Remove Numeric Content from C <sub>i</sub>
Step 7: Using L Lookup H <sub>m</sub> , Retrieve current DL Axiomatize L and current concepts with DL
Step 8: Prepare an artifact comprising of current concepts and Matching individuals with context based axiomitization in the increasing order of the ss values of the concepts
<i>end</i>

**Table 3** Performance evaluation of OntoYield

Query	Precision %	Recall %	Accuracy %
Computer Science Books	85.19	88.46	86.83
Books	86.55	89.96	88.26
Petrol Cars below 25 lakhs	84.21	88.89	86.55
Modern Wooden Furniture	84.72	88.41	86.57
Cakes with Mixed Flavors	83.33	88.99	86.16
Average	84.8	88.94	86.87

relevant ontologies to the total number of retrieved ontologies. Recall is the ratio of the ontologies retrieved and relevant to the total number of relevant ontologies. Accuracy is defined as the average of the Precision and Recall Measures. Equations (2), (3) and (4) depicts the Precision, Recall and Accuracy of the system.

$$\text{Precision} = \frac{\text{No. of retrieved and relevant Ontologies}}{\text{Total No. of Ontologies retrieved}} \quad (2)$$

$$\text{Recall} = \frac{\text{No. of retrieved and relevant Ontologies}}{\text{Total No. of Ontologies that are relevant}} \quad (3)$$

$$\text{Accuracy} = \frac{\text{Precision} + \text{Recall}}{2} \quad (4)$$

OntoYield produces an average Precision of 84.8% and an average recall of 88.4%. An overall accuracy percentage of 86.87 is achieved by OntoYield. The reason for higher performance when the ontological structures are preserved is that, there is a query-based evaluation as well as a context based evaluation. This increases the overall relevance of the recommended ontologies and makes the recommendations highly specific to the query as well as the domain. Moreover, the SemantoSim measure proposed is an adaptation of the Pointwise Mutual Information algorithm with a semantic flavor added to it. With the intelligent integration of a semantic methodology for semantic similarity computation, retaining the original axiomitization between ontological entities along with estimating the semantic heterogeneity between the query words and the description logics increases the relevance of results.

Since the proposed OntoYield system is one of its kind and is the first system to recommend ontologies by preserving the ontological relationships and semantics between its concepts and individuals, its comparison is done by eliminating the Ontology Structure Preservation. The Ontologies are processed without preserving their structures by only recommending the ontologies by computing SemantoSim between the query words and the concepts/individuals. The overall precision, recall and accuracy measures dropped. The reason for a low relevance rate is mainly due to the ignorance of the context of the ontologies. An overall Precision of 81.78%, Recall of 84.99 and 83.39% were achieved when the structure preservation of ontologies was ignored. This clearly justifies the fact that OntoYield that preserves the original Description Logics between ontological entities performs way.

## 6 Conclusion

An approach for recommending ontologies by preserving the original structure and the semantics between the concepts and individuals is proposed. The proposed OntoYield system recommends ontologies that are relevant to the search query



along with the context-based axiomitization. OntoYield is one such system which not only recommends query relevant ontologies but also checks the context of ontologies and its ontological commitment. A SemantoSim measure for computing the semantic similarity between the query and concepts as well as the Description Logics is proposed. The OWL ontologies are converted into their equivalent XML trees and then into their RDF for processing. A HashMap and a HashTable approach is incorporated for processing the ontologies by preserving their original structure. OntoYield achieves an average precision of 84.8%, Recall of 88.94% and an accuracy of 86.87%. OntoYield is one of the first systems to recommend ontologies by preserving and analyzing the original semantics in Domain-Level Ontologies.

## References

1. Gruber, Thomas R. "Toward Principles for the Design of Ontologies used for Knowledge Sharing," *International Journal of Human-Computer Studies* 43, no. 5 (1995): 907–928.
2. Cantador, Iván, Miriam Fernández, and Pablo Castells. "Improving Ontology Recommendation and Reuse in WebCORE by Collaborative Assessments." (2007).
3. Fernández, Miriam, Iván Cantador, and Pablo Castells. "CORE: A Tool for Collaborative Ontology Reuse and Evaluation." (2006).
4. Romero, Marcos Martínez, José M. Vázquez-Naya, Cristian R. Munteanu, Javier Pereira, and Alejandro Pazos. "An Approach for the Automatic Recommendation of Ontologies using Collaborative Knowledge," *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 74–81, Springer, Berlin, Heidelberg, 2010.
5. Martínez-Romero, Marcos, José M. Vázquez-Naya, Javier Pereira, and Alejandro Pazos, "A Multi-Criteria Approach for Automatic Ontology Recommendation using Collective Knowledge," *Recommender Systems for the Social Web*, pp. 89–103. Springer, Berlin, Heidelberg, 2012.
6. Martínez-Romero, Marcos, Clement Jonquet, Martin J. O'Connor, John Graybeal, Alejandro Pazos, and Mark A. Musen. "NCBO Ontology Recommender 2.0: An Enhanced Approach for Biomedical Ontology Recommendation," *arXiv preprint arXiv:1611.05973* (2016).
7. Sadowska, Małgorzata, and Zbigniew Huzar. "Semantic Validation of UML Class Diagrams with the Use of Domain Ontologies Expressed in OWL 2." In *Software Engineering: Challenges and Solutions*, pp. 47–59. Springer International Publishing, 2017.
8. Doan, AnHai, Jayant Madhavan, Pedro Domingos, and Alon Halevy. "Ontology Matching: A Machine Learning Approach." In *Handbook on ontologies*, pp. 385–403. Springer Berlin Heidelberg, 2004.
9. Hughes, Todd C., and Benjamin C. Ashpole, "The Semantics of Ontology Alignment," Lockheed Martin Advanced Technology Labs, Cherry Hill NJ, 2004.
10. Cerón-Figueroa, Sergio, Itzamá López-Yáñez, Wade Alhalabi, Oscar Camacho-Nieto, Yenny Villuendas-Rey, Mario Aldape-Pérez, and Cornelio Yáñez-Márquez. "Instance-Based Ontology Matching for e-Learning Material using an Associative Pattern Classifier." *Computers in Human Behavior* 69 (2017): 218–225.
11. Bharambe, Ujwala, S. S. Durbha, Roger L. King, Nicolas H. Younan, and Kuldeep Kurte. "Use of Geo-Ontology Matching to Measure the Degree of Interoperability," *Geoscience and Remote Sensing Symposium (IGARSS), 2016 IEEE International*, pp. 7601–7604. IEEE, 2016.

12. Anam, Sarawat, Yang Sok Kim, Byeong Ho Kang, and Qing Liu. "Adapting a knowledge-based schema matching system for ontology mapping." In Proceedings of the Australasian Computer Science Week Multiconference, p. 27. ACM, 2016.
13. Ranjini, S., and K. Saruladha. "Concept Type and Relationship Type Classification based Approach for Identifying and Prioritizing Potentially Interesting Concepts in Ontology Matching," Emerging Trends in Engineering, Technology and Science (ICETETS), International Conference on, pp. 1–5. IEEE, 2016.
14. Church, Kenneth Ward, and Patrick Hanks, "Word Association Norms, Mutual Information, and Lexicography." Computational linguistics 16, no. (1990):22–29.
15. C. N. Pushpa, G. Deepak, J. Thriveni and K. R. Venugopal, "Onto Collab: Strategic Review Oriented Collaborative Knowledge Modeling using Ontologies," 2015 Seventh International Conference on Advanced Computing (ICoAC), Chennai, 2015, pp. 1–7.
16. Xiang, Chuncheng, Baobao Chang, and Zhifang Sui. "An Ontology Matching Approach Based on Affinity-Preserving Random Walks," Proceedings of the 24th International Conference on Artificial Intelligence, pp. 1471–1477, AAAI Press, 2015.
17. Xue, Xingsi, and Jeng-Shyang Pan, "A Segment-Based Approach for Large-Scale Ontology Matching." Knowledge and Information Systems (2017): 1–18.

# A Deep Autoencoder-Based Knowledge Transfer Approach

Sreenivas Sremath Tirumala

**Abstract** Deep Transfer Learning or DTS has proven successful with deep neural networks and deep belief networks. However, there has been limited research on to using deep autoencoder (DAE)-based network to implement DTS. This paper for the first time attempts to identify transferable features in the form of learning and transfer them to another network implementing a simple DTS mechanism. In this paper, a transfer of knowledge process is proposed where in knowledge is transferred from one Deep autoencoder network to another. This knowledge transfer has helped to improve the classification accuracy of the receiving autoencoder, particularly when experimented using corrupted dataset. The experiments are carried out on a text based hierarchical dataset. Firstly, a DAE is trained with regular undamaged dataset to achieve maximum accuracy. Then, a distorted dataset was used to train second DAEN for classification with which only 56.7% of the data is correctly classified. Then a set of weights are transferred from first DAEN to the second DAEN which resulted in an improvement of classification accuracy by about 22%. The key contribution of this paper is highlighting importance of knowledge transfer between two deep autoencoder networks which is proposed for the first time.

**Index terms** Deep autoencoders • Knowledge transfer • Hierarchical dataset  
Corrupted dataset

## 1 Introduction

Traditional problem solving approaches utilize previous knowledge to solve new unknown problems. For instance, software engineering uses existing reusable code to solve new problems to avoid starting from the scratch. However, Artificial Intelligence approaches like Neural Networks (Artificial Neural Networks or

---

S.S. Tirumala (✉)

Auckland University of Technology, Auckland, New Zealand  
e-mail: ssremath@aut.ac.nz

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_23](https://doi.org/10.1007/978-981-10-6319-0_23)

277

ANNs) lack this ability transferring knowledge from one ANN to another ANN. In other words, every ANN has to be trained from the scratch in spite of existing trained ANNs with state-of-the-art results. For instance, an ANN designed to solve a computer vision problem has to start with random initialization in spite of existing successful ANN models that have produced state-of-the-art results for image recognition. The generalization of an ANN has been an existing problem which has been widely discussed in the literature since 1990 till recently [1–3].

Deep Neural Networks (DNN) were success due to unique layer-wise training mechanism which was introduced in 2006 [4]. This enabled to extract features at various levels as well as to closely observe the representations in the weights. The learning mechanism of DNNs termed as Deep Learning had produced state-of-the-art results for various real-world applications [5–9] and been successful in its implementations with evolutionary computation, reinforcement learning, etc., [10]. However, there has been not much study on knowledge transfer with DNNs especially WITH autoencoder networks (DAEN). Unlike software engineering, DNNs lacks the ability of reusable components. In spite of existing successful DNN models, each DNN training mechanism has to start from the scratch other than very few recent attempts which may not be generalized. This paper presents an attempt to explore the possibility of transfer of knowledge between two DAEN. For the experiments a hierarchical synthetic dataset has been used from earlier work [11].

The paper is organized as follows. Section 2 introduces to DAEN along with recent works related to knowledge transfer in DNNs. Section 3 presents experimental results followed by discussion in Sect. 4. Conclusion and Future Directions in Sect. 5.

## 2 Related Work

There have been numerous attempts for improving the performance of Deep Neural Networks (DNNs). Transferring knowledge in on such approach where features learn by one DNN are transferred to another DNN to improve performance and accuracy. The importance of knowledge transfer between ANNs was identified in the early 1990s [12]. The process of knowledge transfer involves identification, extraction, and transfer of knowledge which is also referred as ‘Transfer Learning’. Deep Transfer Learning (DTL) is hypothesized by Yoshio Bengio in 2013 [13]. DTL attempts to identify transferable features in DNN and copy them to another DNN to improve performance and accuracy. The earlier attempt toward feature transfer between DNNs is attempted by Yosinski, a student on Benio [14]. This approach investigates on identifying layer(s) where generalization is occurring. In this approach, two Constitutional Neural Network ( $CNN_1$ ,  $CNN_2$ ) is trained on two equally divided parts of the ImageNet dataset. Then, the weight vectors of  $CNN_1$  and  $CNN_2$  are copied to new networks  $CNN_3$  and  $CNN_4$  three layers at a time while randomly selecting the weights of the other layers. After several experiments, Yosinski concluded that the generalization is occurring in the first two layers of

CNN. The transferable features exist only in the first two layers of a DNN for the same dataset. However, these results were not repeated when experiments are conducted on similar datasets, (dataset with a similar structure of ImageNet) which raise questions on the existence of transferable features only in the first three layers.

One of the initial attempts for knowledge transfer in DNNs using deep convolutional neural networks achieved limited success with small datasets [15]. The second notable implementation is to classify upper case Latin characters using an ANN that is trained on Chinese characters [16]. In another approach presented in 2014, ImageNet dataset is used for classification of images which concludes that first three layers consists of more generic features that can be transferred to another ANN for image classification problems [14]. The most recent work by Terekhov uses an alternate approach in which block of weights are introduced between a trained ANN to obtain a set of weights that are optimized with the values between the layers [3]. The new ANN is trained after introducing these set of blocks between the layers with which the training time is reduced. The approaches mentioned above depends on transferring a set of layers (weights) from one ANN to another. Further, there was no light on what exactly (knowledge) is being transferred. The first two approaches uses transfer of a set of layers. These layers are identified by comparing classification accuracy with freezing weights of 2–3 layers at a time on trial and error basis. However, there has been no known similar attempts with DAEN which forms the basis of this paper.

A new transduction transference approach was proposed to solve this issue [17]. Transductive learning approaches examine and learn from a specific training to a specific task drawn from the same distribution. So, in the case of source and target having different distributions, classification results are improved by transferring exploited labeled instances from trained network to new network with solving a similar problem. Experimental results on using Arabic digits to identify Latin digits (Character recognition) proved improved results both in performance and accuracy. However, the questions like where exactly the generalization is occurring are still unanswered. Deep Adaptation Network (DAN) architecture, presented in ICML 2015 is the most recent attempt toward understanding the learning process [18] toward generalization of deep CNNs. DAN generalizes the CNN toward domain adaptation scenario where task-specific features are identified and transferred. Further, with DAN it is confirmed that general features can moreover transferable and task-specific features are to be tailored to solve a different task. Autoencoders reconstruct the input by performing encoding and decoding mechanisms [4]. When more than one autoencoders are stacked together, it constitutes Deep Autoencoder Network or DAEN. Traditionally autoencoders are trained with unsupervised learning. This is same with DAEN. However, the fine-tuning and error propagation is carried out using a classifier layer at the end known as Softmax layer. This layer is trained with the known targets. So, in a typical DAEN, individual layers (autoencoders) are trained unsupervised whereas over all training is supervised making it a typical deep learning implementation.

### 3 Experiment Design

For the experiments, a biological taxon-based synthetic data set is used with feature hierarchies [11]. There are 90 organisms in the dataset with 6 different categories of species. Each species has a different set of features as shown in Figs. 1 and 2.

Each organism is represented in 20 bits categorized into Rank (4 bits), Group (4 bits), Sub-Group (4 bits) and features (8 bits) as shown in Fig. 3. The taxonomic Rank is determined by the shared features, Group and Sub-Group making this a hierarchical representation.

Firstly, DAEN ( $DAEN_1$ ) is trained with uncorrupted dataset until 100% classification accuracy is achieved. Then the dataset is corrupted by replacing values randomly as well as removing some values (changing to NaN). Classification was again performed on this dataset using a second DAEN ( $DAEN_2$ ). Then finally, the weights of  $DAEN_2$  are replaced with the weights of first  $DAEN_1$  making it  $DAEN_R$  and performed classification experiment without training the second DAEN. In other words, the weights are transferred from first DAEN to second DAEN for all autoencoders.

### 4 Experimental Results and Discussion

A 3-layered Deep autoencoder network (DAEN) is used to perform the experiments. Scaled Conjugate Gradient (SGD) algorithm is used for training. A symmetric node count of 50 each is chosen for all the autoencoders reason being

Fig. 1 Localist representation [11]

Features	Representation
C1 (Backbone)	00000001
C2 (Hair)	00000010
C3 (Hands and Feet)	00000100
C4 (hair on hands)	00001000

Fig. 2 Representation Multiple features [11]

Organism	Features	Representation
O1	C1 and C2	00000011
O2	C1, C3, C4	00001101

Fig. 3 Binary representation of organism

Organism	Rank	Group	Sub-Group	Features
1	0001	1101	1101	11011101
2	0010	1010	1010	10101010
3	0011	0101	0101	01010101
4	0100	0100	0100	01000100
5	0101	1001	1001	10011001
6	0110	1011	1101	10111001

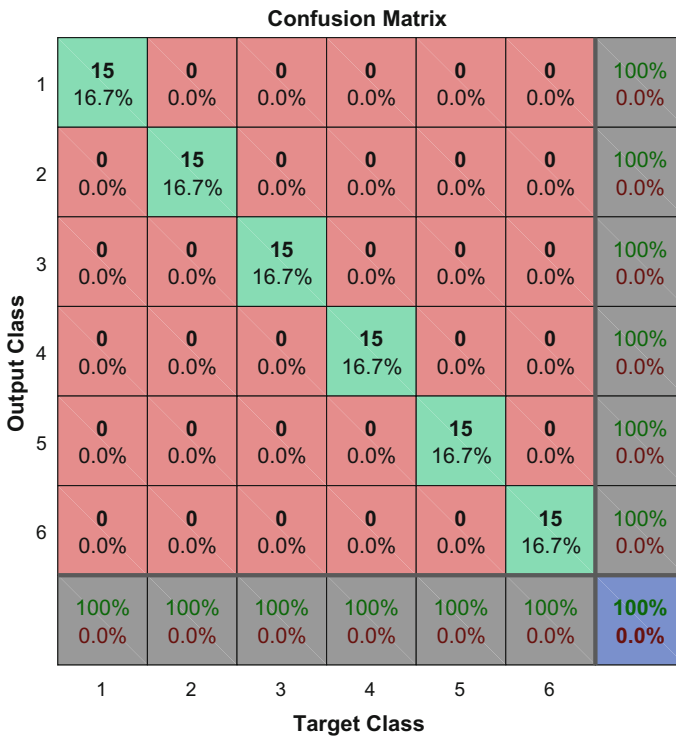
its efficiency in hierarchical data classification compared to asymmetry node count [11].

Support Vector Machines (SVM) is used for softmax layer. Each autoencoder is trained for 400,200,100 epochs and overall supervised training for softmax layers is performed for 100 epochs. The hierarchical dataset used for the experiments consists of 90 samples with 6 different species. Each experiment is performed 25 times. The main reason for selecting hierarchical dataset is that, it consist of known features. Further, the dataset is constructed using distributed representation which makes it easy to disturb the features and hierarchies.

The classification results with various DAENs and datasets is presented in Table 1. The confusion matrix for the experiment results with pure dataset, corrupted dataset are presented as Figs. 4 and 5 respectively. For pure dataset the

**Table 1** Classification results—NC: Non corrupted, C:corrupted

Deep autoencoder	Dataset	Accuracy (%)	Train rmse	Test rmse
$DAEN_1$	NC	100	0.003	0.0034
$DAEN_2$	C	56.7	0.663	0.5113
$DAEN_R$	C	78.9	–	0.252



**Fig. 4** Binary representation of organism

**Confusion Matrix**

<b>Output Class</b>	1	15 16.7%	6 6.7%	10 11.1%	5 5.6%	3 3.3%	11 12.2%	30.0% 70.0%
	2	0 0.0%	9 10.0%	1 1.1%	0 0.0%	0 0.0%	1 1.1%	81.8% 18.2%
	3	0 0.0%	0 0.0%	4 4.4%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
	4	0 0.0%	0 0.0%	0 0.0%	10 11.1%	1 1.1%	1 1.1%	83.3% 16.7%
	5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	11 12.2%	0 0.0%	100% 0.0%
	6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	2 2.2%	100% 0.0%
			100% 0.0%	60.0% 40.0%	26.7% 73.3%	66.7% 33.3%	73.3% 26.7%	13.3% 86.7%
		1	2	3	4	5	6	
		<b>Target Class</b>						

**Fig. 5** Binary representation of organism

classification accuracy is 100% where as when the dataset is damaged it fell to 56.7% due to corrupted data and misplaced hierarchies.

However, when the classification experiment is performed with the same corrupted dataset after transfer of weights (from  $DAEN_1$  are transferred to  $DAEN_2$ ) with new  $DAEN_R$ , there was a huge raise of 22.2% in the classification accuracy to 78.9% as shown in Fig. 6. One reason for this rise might be that ‘some knowledge’ is transferred unknowingly when weights are transferred. It is a fact that the principle components of any neural network is weights. However, weight is just numeric values and might not be significant by itself. However, collective weights might have some hidden representations that are responsible for this knowledge transfer. There hidden representations might constitute some form of knowledge which is being transferred and is responsible for improvement of accuracy.

When autoencoder is able to reconstruct the input, the weights might be storing the structure or some form in the weights. This has been utilized to replace the corrupted values such that the samples are classified correctly. However, it is still an open question that how representations can be extracted from weights.



**Confusion Matrix**

Output Class	1	11 12.2%	0 0.0%	1 1.1%	1 1.1%	2 2.2%	0 0.0%	73.3% 26.7%
	2	0 0.0%	15 16.7%	0 0.0%	0 0.0%	0 0.0%	2 2.2%	88.2% 11.8%
	3	4 4.4%	0 0.0%	14 15.6%	0 0.0%	0 0.0%	0 0.0%	77.8% 22.2%
	4	0 0.0%	0 0.0%	0 0.0%	14 15.6%	2 2.2%	7 7.8%	60.9% 39.1%
	5	0 0.0%	0 0.0%	0 0.0%	0 0.0%	11 12.2%	0 0.0%	100% 0.0%
	6	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	6 6.7%	100% 0.0%
			73.3% 26.7%	100% 0.0%	93.3% 6.7%	93.3% 6.7%	73.3% 26.7%	40.0% 60.0%
		1	2	3	4	5	6	
		<b>Target Class</b>						

Fig. 6 Binary representation of organism

## 5 Conclusion and Future Work

In this paper, a Deep autoencoder-based knowledge transfer approach was presented. Artificial Neural Networks (ANNs) lack the reusability in spite of existing models for problem solving. There were very few attempts to investigate on knowledge transfer in deep neural networks and none in the case of autoencoders.

Experiments are carried out using synthetic data set with known feature hierarchies on a three layer Deep autoencoder Network (DAEN). When the proposed transfer of weights is applied, the classification accuracy for a damaged dataset has improved by over 22% from 56.7%. Then the weights of second DAEN are replaced with to 78.9%. It is noteworthy to observe that, some knowledge is been transferred unknowingly which is responsible for this jump. The source and composition of knowledge is still need to be investigated.

One of the future direction is to expose the weights of each layer of ANNs and try to map with the input features to get a better understanding of representations hidden in weights of ANNs. The further systematic investigation is needed by applying this approach on larger data sets especially typical benchmark data sets like MNIST and non-hierarchical and complex data sets like gene expression.

## References

1. D. K. Milligan and M. J. D. Wilson, *Fundamental Structure/Behaviour Relationships in Synchronous Boolean Neural Networks*, 1990, pp. 997–1000.
2. Z. Waszczyszyn, *Fundamentals of artificial neural networks*. Springer, 1999, pp. 1–51.
3. A. V. Terekhov, G. Montone, and J. K. O’Regan, *Knowledge Transfer in Deep Block-Modular Neural Networks*, 2015, pp. 268–279.
4. Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, “Greedy layer-wise training of deep networks,” in *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, pp. 153–160.
5. Y. Bengio and Y. LeCun, “Scaling learning algorithms towards AI,” in *Large Scale Kernel Machines*. MIT Press, 2007.
6. K. He, X. Zhang, S. Ren, and J. Sun, “Spatial pyramid pooling in deep convolutional networks for visual recognition,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
7. C. Xiong, L. Liu, X. Zhao, S. Yan, and T. Kim, “Convolutional fusion network for face verification in the wild,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2015.
8. D. Hingu, D. Shah, and S. S. Udmale, “Automatic text summarization of wikipedia articles,” in *Communication, Information Computing Technology (ICCICT), 2015 International Conference on*, Jan 2015, pp. 1–4.
9. A. Graves and J. Schmidhuber, “Offline handwriting recognition with multidimensional recurrent neural networks,” pp. 545–552, 2009.
10. S. S. Tirumala, “Implementation of evolutionary algorithms for deep architectures,” in *Proceedings of the 2nd International Workshop on Artificial Intelligence and Cognition (AIC), Torino, Italy, November, 2014*, pp. 164–171.
11. S. S. Tirumala and A. Narayanan, “Hierarchical data classification using deep neural networks,” in *Neural Information Processing*. Springer International Publishing, 2015, pp. 492–500.
12. E. Y. Li, “Artificial neural networks and their business applications,” *Information & Management*, vol. 27, no. 5, pp. 303–313, 1994.
13. Y. Bengio, A. Courville, and P. Vincent, “Representation learning: A review and new perspectives,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.
14. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, “How transferable are features in deep neural networks?” in *Advances in Neural Information Processing Systems*, 2014, pp. 3320–3328.
15. S. Gutstein, O. Fuentes, and E. Freudenthal, “Knowledge transfer in deep convolutional neural nets,” *International Journal on Artificial Intelligence Tools*, vol. 17, no. 03, pp. 555–567, 2008.
16. D. C. Cireřan, U. Meier, and J. Schmidhuber, “Transfer learning for latin and chinese characters with deep neural networks,” in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–6.
17. C. Kandaswamy, L. M. Silva, L. A. Alexandre, J. M. Santos, and J. M. Sá, *Artificial Neural Networks and Machine Learning - ICANN 2014: 24th International Conference on Artificial Neural Networks, Hamburg, Germany, September 15–19, 2014. Proceedings*. Cham: Springer International Publishing, 2014, ch. Improving Deep Neural Network Performance by Reusing Features Trained with Transductive Transference, pp. 265–272.
18. M. Long, Y. Cao, J. Wang, and M. Jordan, “Learning transferable features with deep adaptation networks,” in *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, D. Blei and F. Bach, Eds. JMLR Workshop and Conference Proceedings, 2015, pp. 97–105. [Online]. Available: <http://jmlr.org/proceedings/papers/v37/long15.pdf>.

# Performance Preview on Image Super Resolution Using Wavelets Transform Based on Samples

Vicharapu Balaji, Ch. Anuradha, P.S.R. Chandra Murty  
and Grandhe Padmaja

**Abstract** In Image analysis with Wavelet change, super resolution is amazingly critical. In our proposed work to acquire Super resolution of the input image two prominent wavelet transforms are used namely Discrete and Stationary Wavelet Transform. In general Single frame resolution can be refined by different augmentation procedures like interpolation also leads to Blur and obscure edges. Thus, this paper input is taken as any sample image from the set and then applying basic wavelet filters as specified, i.e., DWT and SWT to get a super resolved image. Then on the super resolved image wavelet filters are applied.

**Keywords** Super resolution · DWT and SWT · Interpolation · High frequency (HF) · Low frequency (LF)

## 1 Introduction

The method to reconstruct a high resolution images assumes an essential part in various medical imaging and electronic applications as high resolution images are craved and often required. Pixel density is more in high resolution images as it gives more subtle elements of data which is required in the basic application, for example, medical diagnosis, satellite perception and mammography images. One essential methodology for single edge super determination in interjection in which

---

V. Balaji (✉) · P.S.R. Chandra Murty (✉)  
Department of CSE, ANUCET, Guntur Dt, AP, India  
e-mail: v12.balaji@gmail.com

P.S.R. Chandra Murty  
e-mail: Chandra\_psr@rediffmail.com

Ch. Anuradha  
Department of CSE, VRSEC, Krishna Dt, AP, India

G. Padmaja  
Department of CSE, PSCMR CET, Vijayawada, AP, India  
e-mail: padmajagrandhe@gmail.com

high recurrence data is expelled from low determination picture and predication is done for complete information in the high resolution image. They are some implicit existing methodologies for the super resolution which depends upon the standard representation techniques like pixel replication, linear interjection, bilinear, and bicubic that constructs the pixel number excluding the details [1, 2]. However, addition-based super determination systems presented the obscure impact in boundaries of image, so diverse super resolution systems in context of different images to maintain a strategic distance from obscure in images. For the most part, super determination system can be confined to three types: spatial zone reconstruction, Frequency, and probability-based strategies, Tasi and Huang [3] are the essential who made considered super determination using frequency Domain. Keren et al. developed a spatial domain procedure using rotation and global translation models to perform image registration. Irani and Peleg [4] proposed dynamic images of an object, and more mind boggling motions than unadulterated translational motion in the image plan to produce high resolution images. The proposed work depends on creating an arrangement of reproduced images with low resolution. Contrasts between this arrangement of pictures and the genuine watched images with minimal resolution are projected back using a back-projected kernel, onto an initial estimation of the high-determination image. Cohen, Arvin, and Dinstein proposed an algorithm for producing a high resolution image by augmenting Irani and Peleg's work, in which images with high resolution is produced by use of projection of each and every pixel. In any case, their strategy is limited to input pictures that are a basic interpretation of the first picture to acquired super resolved picture. The issue with many of super resolution techniques is with edges of the image. Edge boundaries may associated with noise and blur due to moving objects and errors caused by motion. Ji and Fermuller [5] used a standard bi-orthogonal wavelet channel bank(cdf-9/7) algorithm to produced desired target image with better resolution. Jiji et al. [2] proposed single edge picture super resolution Methodology in which high frequency segments are retrieved by applying wavelet coefficient and then pixel values in various frequency sub-bands [6]. Using Fourier transform, this method focuses on frequency analysis in local filtering than global filtering.

Enhancement provided by same researchers by utilization of DWT and SWT to transfer the given image into various high frequency sub-bands. These sub-bands and info image are consolidated subsequent to applying interpolation on both. Here the attempt has utilized same filter bank for this super resolution scheme. Chappali and Bose had the idea of edge level on reconstructed image quality in lifting scheme based wavelet super resolution. In their algorithm, they attempt to evacuate, however much of the corrupted noise as could reasonably be expected without influencing the reconstructed image quality because of obscure presented in the super resolution process.

## 2 Previous Work

There are different algorithms and approaches using wavelets and spatial domains for obtaining super resolution image. Gholamreza Anbarjafari and Hassan Demirel evolved interpolation based a super resolution technique using Discrete Wavelet Transform [7]. In their paper, low resolution image is transformed to different sub-bands frequencies using discrete wavelet transform. These high frequency sub-bands are interpolated using various interpolation. Super-resolved image is obtained by inverse transformed of combination of interpolated high frequency sub-bands and input image they have proved that quality of image is enhanced using this wavelet-based technique as compared to super resolved image obtained by different interpolation method. The method proposed by Demirel and Anbarjafari in Fig. 1.

Another super resolution technique is introduced by Gholamrez Anbarjafari Hassan Demirel which is the extension of the existing works. The simulation technique of super resolution with a new concept as interpolation utilizing with wavelet transformation techniques. In this technique, the decimated values of DWT HF sub-bands are predicted by SWT HF sub-bands through interpolation technique. Here high frequency sub-bands acquired from DWT are introduced and added to sub-bands got structure SWT [1]. Long-lasting super resolution image is acquired by combination of subgroups and interpolated input image. This technique proposed by Demirel et al. Give improvement in quality measure of the image [8, 9].

## 3 Proposed Work

Resolution has been every now and again alluded as an essential part of an image. Images are being prepared with a specific end goal to acquire more enhanced resolution [9]. There are three surely understood interpolation techniques, to be

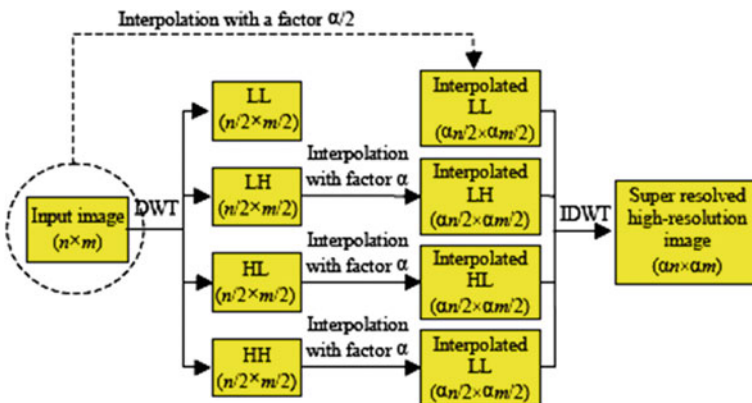
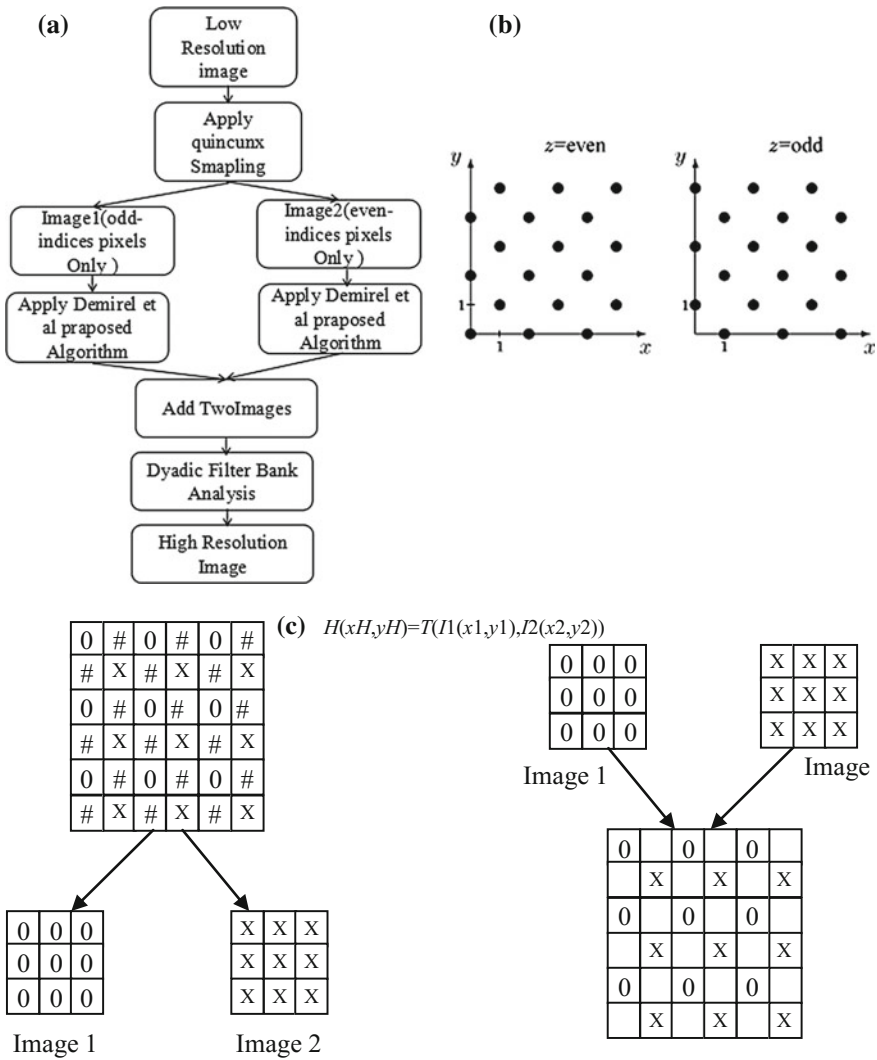


Fig. 1 Block diagram for method proposed by Gholamreza Anbarjafari



**Fig. 2** a Block diagram for proposed method. b Representation of quincunx sampling theorem. c Example for quincunx sampling

specific nearest neighbor interpolation, bilinear interpolation, and bicubic interpolation. Our proposed algorithm identified with the super resolution system implemented in view of modified demirel et al. along with sampling theorem [7]. This algorithm used to enhance the resolution for the specific low resolution image, by taking the samples from the low resolution image then build the high resolution image-based sampling pixels as shown in Fig. 2.

**Algorithm:**

- Step 1: In this first step read the low resolution image, this images related to the general images (lenna image), medical images, satellite images, real-time objects.
- Step 2: In second step, apply quincunx sampling on the original image. Quincunx sampling obtained only even and odd indices pixel values only
- Step 3: after that apply super resolution algorithm which is proposed by Demirel et al. using this algorithm obtained high resolved image.
- Step 4: after obtained high resolution image applies dyadic filter analysis, for removing noise and blur.

**Modified demirel et al. algorithm:**

This modified demirel et al. algorithm actualized same as proposed algorithm of demirel, the fundamental adjustment done at wavelet transform, here in this altered algorithm we are expanding the level of DWT and SWT. Really in this proposed calculation, we take 3-level DWT and SWT. DWT works under decimation process implies factor 2, we are applying the DWT on any image, we get four sub-bands decimated by factor 2. It implies the size of sub-bands lessened by half of the original image. In this decimation procedure we misfortune some data at sub-bands. You need to anticipate those data by utilizing different interpolation techniques and Stationary Wavelet transform. The remaining procedure is done same as demirel et al. calculation.

**Quincunx sampling Theorem:**

There exist several sampling theorems in digital image processing. In our proposed method, we are using quincunx sampling theorem for taking the samples from the image. Actually, quincunx sampling takes the alternative pixels from the image, this sampling technique divides the original image into two images which are having the only even indices and odd indices.

$$H(xH, yH) = T(I1(x1, y1), I2(x2, y2))$$

which is based on half-pixel shift done in both directions (rows and columns). Through this process, the image is decomposed into two images as  $I_1$  and  $I_2$  shows in the following figure. The pixel image of  $6 \times 6$  into two  $3 \times 3$  images, here zeros indicate odd indices pixel and the cross indicates the even indices pixel.

After completion of the sampling process, then applying the Demirel et al. algorithm on both the images individually. So we got two high resolution images. After getting this two resolution images, add both the images into single image. After getting the single high resolution image applying the Dyadic Filter Analysis, this filter is used to remove the noise and blur at the edges [11, 12].

For, quality assessment between the original image and reconstructed image find the mean, variance, mode, median, and find the noise ratio between two images find MSE and PSNR values.

**Mean:** The mean is the average of all pixels intensities in the image. It's simply add up all the intensities, and then divide by the size of the matrix. In other words, it is the sum divided by the count.

$$\text{Mean}(\mu) = \frac{\sum_{i,j}^{m,n} f(i,j)}{m * n} \quad (1)$$

**Variance:** Variance is used to find out how the pixels spread over the image plan. Low variance means the pixel values in the image clustered close together. Higher variance means the pixel values are more spread out.

$$\text{Variance}(\sigma^2) = \frac{\sum_{i,j}^{m,n} (\mu - f(i,j))^2}{m * n} \quad (2)$$

**Standard deviation:** is a measure that is used to quantify the amount of variation or dispersion of pixel values in image.

$$\text{Standard deviation} = \sqrt{\sigma^2} \quad (3)$$

**Median:** The *median* of a finite list of pixels can be found by arranging all the observations from lowest value to highest value and picking the middle one.

If  $n(m*n)$  is **odd** then Median ( $M$ ) = value of  $((n + 1)/2)$ th item term.

If  $n(m*n)$  is **even** then Median ( $M$ ) = value of  $[(n)/2]$ th item term +  $((n)/2 + 1)$ th item term]/2

**Mean Square Error:**

$$\text{MSE} = \frac{1}{mn} \left( \sum_{i=1}^m \sum_{j=1}^n (t'(i,j) - t(i,j))^2 \right) \quad (4)$$

**Peak Signal to Noise Ratio:**

$$\text{PSNR} = -10 \log_{10} \left( \frac{\max^2}{\text{MSE}} \right) \quad (5)$$

## 4 Results, Outcomes, and Discussion

The Proposed work is implemented based on take the samples from the original image and then reconstruct the original image with maximizing resolution of the image. If the input of size of the image is  $256 \times 256$ , then it will increase the resolution nearly  $2048 \times 2048$ . Proposed algorithm gives at most possible reconstruction of the increased the resolution of image. For the performance analysis for the finding the level of reconstruction finds the mean, variance, mode, median. And



**Fig. 3** **a** Input image.  
**b** Image which contains only even indexed intensity value.  
**c** Image which contains only odd indexed intensity values.  
**d** Super resolved image



finding the error ration between these two images find the MSE and PSNR value between original and reconstructed image as shown in Fig. 3.

### 4.1 Performance Analysis for Reconstruction

Result For Reconstructed SR_Image		
	Original_image	SR_Image
Mean	128.232	128.172
Variance	9.81028e+06	9.58407e+06
Standard_deviation	3132.14	3095.82
Median	119	119
mode	93	94

## 5 Conclusion

This proposed algorithm reaches almost perfect reconstruction of the low resolution image to high resolution image. In future work, plan to work for implement high resolution with smooth edges, to enhance the quality of the image and also to reduce the time Complexity.

## References

1. Gajjar, Prakash P., and Manjunath V. Joshi. "New learning based super-resolution: use of DWT and IGMRF prior." *IEEE Transactions on Image Processing* 19.5 (2010): 1201–1213.
2. Jiji, C. V., Manjunath V. Joshi, and Subhasis Chaudhuri. "Single frame image super resolution using learned wavelet coefficients." *International journal of Imaging systems and Technology* 14.3 (2004): 105–112.
3. Tsai, R. Y., and Thomas S. Huang. "Multiframe image restoration and registration." *Advances in computer vision and Image Processing* 1.2 (1984): 317–339.
4. Irani, Michal, and Shmuel Peleg. "Motion analysis for image enhancement: Resolution, occlusion, and transparency." *Journal of Visual Communication and Image Representation* 4.4 (1993): 324–335.
5. Ji, Hui, and Cornelia Fermuller. "Robust wavelet-based super-resolution reconstruction: theory and algorithm." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31.4 (2009): 649–660.
6. Nguyen, Nhat, and Peyman Milanfar. "A wavelet-based interpolation-restoration method for superresolution (wavelet superresolution)." *Circuits, Systems and Signal Processing* 19.4 (2000): 321–338.
7. Demirel, Hasan, and Gholamreza Anbarjafari. "Image resolution enhancement by using discrete and stationary wavelet decomposition." *IEEE transactions on image processing* 20.5 (2011): 1458–1460.
8. Iqbal, Muhammad Zafar, Abdul Ghafoor, and Adil Masood Siddiqui. "Satellite image resolution enhancement using dual-tree complex wavelet transform and nonlocal means." *IEEE geoscience and remote sensing letters* 10.3 (2013): 451–455.
9. Bagawade Ramdas, P., S. Bhagawat Keshav, and M. Patil Pradeep. "Wavelet transform techniques for image resolution enhancement: A study." *International Journal of Emerging Technology and Advanced Engineering* 2.4 (2012): 167–172.
10. Naik, Sapan, and Nikunj Patel. "Single image super resolution in spatial and wavelet domain." *arXiv preprint arXiv:1309.2057* (2013).
11. Anbarjafari, Gholamreza, and Hasan Demirel. "Image super resolution based on interpolation of wavelet domain high frequency subbands and the spatial domain input image." *ETRI journal* 32.3 (2010): 390–394.

# Handwritten Symbol Recognition Using Hierarchical Shape Representation Model Based on Shape Signature

M. Raja Babu, T. Gokaramaiah and A. Vishnuvardhan Reddy

**Abstract** The Signature represents visual object shape 2D contour in 1D to recognition shape of the objectQuery. This 1D shape representation translated into Centroid Distance Histogram (CDH) Gokaramaiah et al. (Comput Graph Image Process 25:357–370, 1974 [16]) to achieve invariant transformations such as translation, scale, rotation, flip. The CDH representation performs well in content-based image retrieval system with low computational complexity and this representation method insensitive to noise of boundary. The CDH fails to represent concave shape object because the signature function maps some of the angle to more than one length from the centroid to contour. This problem solved by modifying the shape signature function which returns the average centroid length when the angle difference between two contour points approximately equals to 0.873 by traversing contour points in a clockwise direction. The starting point for clock traversing is minimum distance point from the centroid to contour. The Average Centroid Lengths (ACL) converted into histogram which makes shape representation independent of transformations. To improve recognition, more information of contour obtained by first-order and second-order difference histogram of the modified signature. This first-order and second-order difference Gokaramaiah et al. (IEEE Comput Soc, 2010 [1]) shape signature represented as hierarchical ACL. This ACL representation suitable for the Handwritten symbol recognition because small changes in the contour of shape adopted in Hierarchical ACL representation.

---

M. Raja Babu (✉)

Department of Information Technology, Aditya Engineering College,  
Surampalem, East Godavari 533437, India  
e-mail: raaja\_525@yahoo.com

T. Gokaramaiah

Department of Computer Science and Engineering, Hyderabad Institute  
of Technology and Management, Gowdavalley, Hyderabad 501401, India  
e-mail: tgokari@gmail.com

A. Vishnuvardhan Reddy

Department of Computer Science and Engineering, G.Pulla Reddy Engineering  
College (Autonomous), Kurnool 518007, India  
e-mail: avijavishnu@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

N. Chaki et al. (eds.), *Proceedings of International Conference on Computational Intelligence and Data Engineering*, Lecture Notes on Data Engineering and Communications Technologies 9, [https://doi.org/10.1007/978-981-10-6319-0\\_25](https://doi.org/10.1007/978-981-10-6319-0_25)

The Handwritten symbol recognized based on k-nearest neighbor classifier (k-NNC) on sample database symbols.

**Keywords** Handwritten symbol representation · Pattern recognition  
Centroid distance histogram

## 1 Introduction

Hand-drawn symbols are natural for human beings to work with computing machines. Nowadays, most of the mobile phones and PDAs are equipped with a touch screen that can be used for pen inputs. The gestures are an easy way to communicate with machines, various hand-drawn symbols can be used to as command gestures (like symbol @ on the touch screen open accounting software), to launch applications, for authentication purpose, etc. The hand-drawn symbol composed be writing various strokes which are lines and or curves on touch screen. Shape is an important visual feature of hand-drawn symbol which distinguishes it from others and is done easily by living beings. A shape is invariant to translation, scale, rotation, mirror-reflection and for small distortions. Shape recognition has several applications, like content-based image retrieval [2], object-classification where input is image of an object [3], character recognition [4], online hand-drawn text or shape recognition [5], writer recognition [6], etc.

The online handwritten text or shape recognition is based on low-level characteristics of shape [5, 7] are used. These methods are high computational time requirements, does not provide satisfactory results. The object shape representation methods are categorized as (i) *contour-based* methods, and (ii) *region-based* approaches. In contour-based approaches, solitary the edge of object shape is considered while in region-based approaches the whole region is considered. Every classification is additionally separated into *structural* and *global* approaches. Structural approach isolates the object as comprising of a few sections alongside their links [8]. In global approach, the shape is articulated to overall. These approaches can be additionally partitioned into a changing area and space area strategies, based on whether the image is changed into another space (e.g., by applying medial axis transformation [9], Fourier transformation [10], and so forth) or not. Different shape rendering or depiction policies proposed in the past syndicates *shape signature*, *signature histogram*, *shape context*, *shape invariants*, *curvature*, *moments*, *shape matrix*, *spectral features*, etc. A detailed review of several methods is given by Zhang et al. [2].

In an image recognizing, a 3D object is complicated because its two-dimensional projection which is captured by the camera depends on the camera angle. The Same 3D object can produce entirely different 2D images. This paper does not consider this issue but rather expect that 2D pictures are given from where the contour of the protest can be effectively extricated by applying standard division and edge location techniques [10].

This paper proposes a shape representation scheme which represents to both curved and noncurved shapes. This is a speculation of our prior shape representation method called signature histogram and  $k^{\text{th}}$  order augmented histogram [1] which are appropriate just for curved shapes.

The remaining of the paper is organized as follows. Section 2 inspects contour-based representation schemes. Section 3 is about proposed shape representation method. Section 4 describes the representation method  $k^{\text{th}}$  order augmented. Section 5 gives the experimental information that were conducted to find the validation of proposed scheme against a similar method. Finally, Sect. 6 finishes up about representation method.

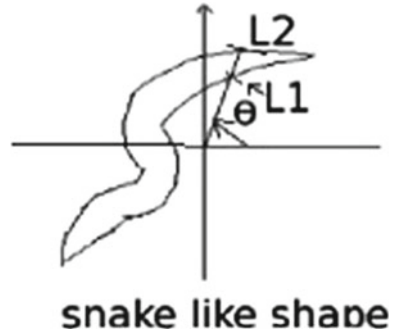
## 2 Contour-Based Representation Schemes

The boundary information processes the contour-based methods, whereas the region of the object is considered the region-based methods. Contour-based methods classified *structural* approaches and *global* approaches. The structural approaches break the boundary as several segments/sections. The structural approaches are *chain code representation* [11, 12], *polygon decomposition* [13, 14], *smooth curve decomposition* [15], etc.

*Circularity* (proportion of squared perimeter and area), *area*, *eccentricity* (proportion of major axis' length and minor axis' length), *bending energy*, *major axis orientation* are some of simple global methods [16]. These global methods are used as coarse representation. *Shape signatures* [17], *boundary-moments* [18], etc, are other relevant global approaches.

Shape signature is a one-dimensional function which maps 2D contour points of the visual object and there are several shape signatures like *centroidal profile*, *complex coordinates*, *tangent angle*, *centroid distance*, *curvature*, *two segmented angle function* [2, 17, 19], *turning function*, exist in the Literature. The centroid distance signature [2] or centroid contour distance (CCD) curve [3] is a function of predefined angle to distance from the centroid to contour points. This function representation is invariant to translation but not to scale, rotation and flip. The scale invariance achieved by CCD has normalized it and is called as *normalized centroid distance signature*. This representation is not invariant to rotation and flip the shift match, reverse match overcome the rotation and flip invariant problems, respectively. These are computationally expensive for shift matching and reverse matching. The Signature was quantized into signature histogram [2] to decrease computational cost. This representation fails for concave shape because the Signature function some predefined angle maps more than one distance. Figure 1 shows centroid contour signature function maps  $\theta$  to two centroid contour distances for the handwritten symbol.

Fig. 1 Handwritten symbol



### 3 Proposed Representation Scheme

The Shape signature fails to represent the concave object. The centroid to contour Euclidean distance and the angle between centroid to contour calculated by traversing clockwise direction. The centroid to contour angle difference between two points approximately equals 0.873 then it returns average centroid length of these two points. Let us consider object shape of contour as  $SP = \{P_1, P_2, \dots, P_n\}$ . The Euclidean distance from centroid to contour given by

$$L(P_i) = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$$

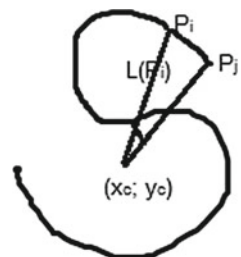
where  $(x_c, y_c)$  is centroid point as shown in Fig. 2. The  $\dot{P}$  is a contour point which has minimum the Euclidean distance from centroid to contour.

The angle  $\theta$  is calculate from centroid to contour point  $P_i$  defined as

$$\theta(P_i) = \text{atan}\left(\frac{y_i - y_c}{x_i - x_c}\right)$$

The average distance length Avg function defined between two pair of points when angle difference  $\theta(P_i, P_j) > T$  where T is threshold and its value is 0.873.

Fig. 2 Handwritten symbol 'S' with centroid contour distance



$$\text{Avg}(P_i, P_j) = \frac{\sum_{k=i}^j l(P_k)}{|i - j|}$$

$$\text{ACL} = \{L_i, |\text{Avg}(P_i, P_j)| \forall i, j \text{ such that } \theta(P_i, P_j) \equiv 0.873\}$$

the  $\dot{P}$  is a start point and ACL is set of all average centroid lengths calculated by traversing all contour points in clockwise direction. These ACL are normalized so that representation invariant to scale. ACL values are used to construct histogram.

### 4 $K^{\text{th}}$ Order Histogram Representation from Average Contour Lengths

The histogram constructed from the average lengths  $A$ . These values distributed over equally divided bins. The bins formed by dividing the  $[0 - 1]$  interval into 10 parts with equal range, each part are called as a bin. The bin is  $[0.0 - 0.10, 0.11 - 0.20, 0.21 - 0.30, 0.31 - 0.40, 0.41 - 0.50, 0.51 - 0.60, 0.61 - 0.70, 0.71 - 0.80, 0.81 - 0.90, 0.91 - 1.00]$  The histogram X-axis represent bin intervals and Y-axis represents frequency. These Histogram representations robust to contour noise because a small change in the position of contour point will reflect in the Centroid to contour distance. The reflected change in the length may fall into the same bin. The AL used to construct the histogram. To reduce false positives in recognition  $K^{\text{th}}$  order representation [1] of contour derived from average distance lengths ACL which store information of contour.

The Average distance lengths  $\text{ACL} = \{L_1, L_2, L_3 \dots L_n\}$ .  $k^{\text{th}}$  order difference distances is defined as

$$\text{ACL}^k(i) = \begin{cases} |L_i - L_{i+k}| & \text{for } i = 1 \text{ to } M - k \\ |L_i - L_{i+k-M}| & \text{for } i = M - k + 1 \text{ to } M \end{cases}$$

$\text{ACL}^k(i)$  is a set different distances in  $k^{\text{th}}$  order, These values used to derive first-order histogram, second-order histogram. Three histograms (ACL, First-order histogram, second-order histogram) are combined to represent the shape. This representation is invariant to scale, translation, rotation, and flip. Figure 3 shows ACL, first-order histogram of the hand written symbol ‘S’.

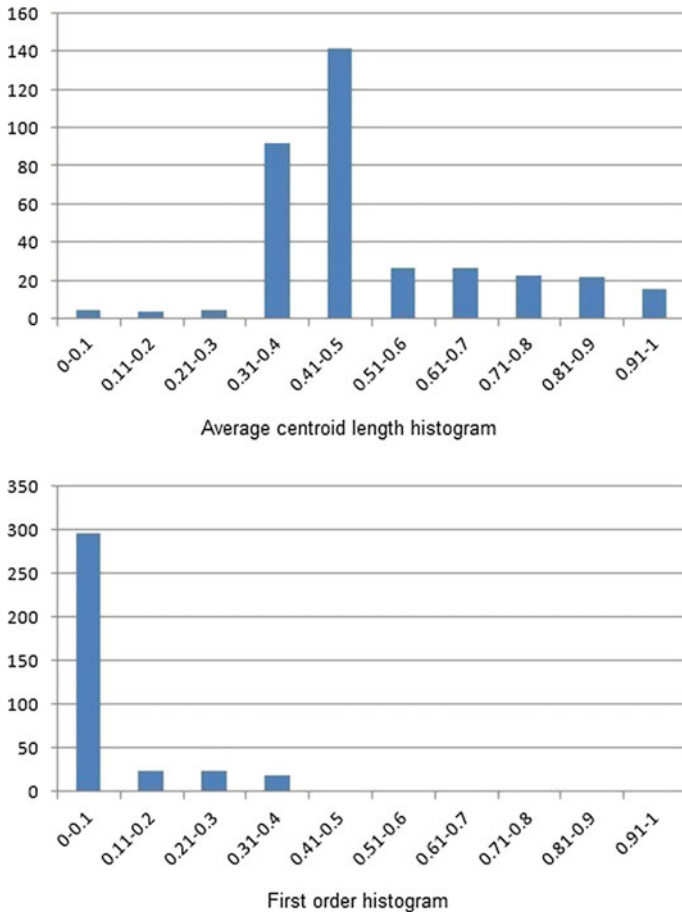


Fig. 3 ACL, First-order histogram for Handwritten Symbol ‘S’

### 5 Experimental Results

Twenty different persons are asked to draw four symbols, each symbol for ten times. So, a total of 800 hand-drawn symbols database is created. Each symbol drawn on  $512 \times 512$  writing pad to collect the coordinates of symbol. 100 randomly chosen symbols (out of 800 symbols) are separated by doing sampling without replacement and is used as the test set. Remaining 700 symbols constitutes the training set. The writer independent hand written symbol recognition problem addressed in the paper. Some sample hand-drawn symbols are shown in the Fig. 4.

The classifier used is the k-nearest neighbor classifier (k-NNC). Two representation schemes, viz., Zernike moments based one and the proposed probability



**Fig. 4** Sample hand-drawn symbols



distributions are compared. The proposed method is compared against the similar system and is found to be a better one as far as the test database is considered.

## 6 Conclusions and Future Works

Centroid contour distance signature method best among various signature methods but it fails to represent concave shapes. This draw back solved in this paper with centroid to contour angle difference calculated by traversing counter of shape in a clockwise direction. The signature converted into histogram which makes shape independent of the transformations and it decreases computational cost. The proposed method is compared against the similar system and is found to be a better one as far as the test database is considered. The angle difference is fixed at 0.873, we can work on decided factor of angle difference which one best.

**Author Declaration** We the undersigned declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere. We confirm that all ethical approvals have been granted to the authors regarding data managed in this research. We further confirm that any aspect of the work covered in this manuscript has been conducted with the ethical approval of all relevant bodies and that such approvals are acknowledged within the manuscript.

## References

1. T. Gokaramaiah, P. Viswanath, and B. E. Reddy. A novel shape based hierarchical retrieval system for 2d images. *IEEE Computer Society*, pages 10–14, 2010.
2. D. Zhang and G. Lu. Review of shape representation and description techniques. *Pattern Recognition*, 37:1–19, 2004.
3. Z. Wang, Z. Chi, and D. Feng. Shape based leaf image retrieval. In *IEE Proc.-Vis Image Signal Process.*, volume 150, Feb 2003.
4. V. S. Chakravarthy and B. Kompella. The shape of handwritten characters. *Pattern Recognition Letters*, 24:1901–1913, 2003.
5. T. Artieres, S. Marukatat, and P. Gallinari. Online handwritten shape recognition using segmental hidden markov models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):205–217, 2007.
6. S. N. Srihari, S.-H. cha, H. Arora, and S. Lee. Individuality of handwriting. *Journal of Forensic Sciences*, 47(4):1–17, 2002.
7. T. Gevers and A.W. Smeulders. Combining color and shape invariant features for retrieval. *IEEE Transactions on image processing*, 9(1):102–119, 2000.
8. L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt. Shape representation and classification using the Poisson equation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1991–2005, 2006.
9. H. Blum. *A transformation for extracting new descriptors of shape*, in: W. Whaten-Dunn (Ed.), *Models for the Perception of Speech and Visual Forms*. MIT Press, Cambridge, MA, 362–380, 1967.
10. R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Second Edition, Pearson Education, 2002.
11. H. Freeman. On the encoding of arbitrary geometric configurations. *IRE Trans. Electron. comput. EC-10*, pages 260–268, 1961.
12. H. Freeman and A. Saghri. Generalized chain codes for planar curves. *Proceedings of the Fourth International Joint Conference on Pattern Recognition*, Kyoto, Japan:701–703, 7–10 November 1978.
13. W. Groskey, P. Neo, and R. Mehrotra. Index-based object recognition in pictorial data management. *Computer Vision Graphics Image Processing*, 52:416–436, 1990.
14. W. Groskey, P. Neo, and R. Mehrotra. A pictorial index mechanism for model-based matching. *Data Knowledge Engineering*, 8:309–327, 1992.
15. S. Berretti, A. Bimbo, and P. Pala. Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Transactions on Multimedia*, 2(4):225–239, 2000.
16. I. Yong, J. Walker, and J. Bowie. An analysis technique for biological shape. *Computational Graphics and Image Processing*, 25:357–370, 1974.
17. P.J. Van Otterloo. *A Contour-Oriented Approach to Shape Analysis*. Prentice-Hall International(UK) Ltd, Englewood Cliffs, NJ, 2 edition, 1991.
18. M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Analysis and Machine Vision*. Chapman & Hall, London, UK, NJ, 2 edition, 1993.
19. O. Starostenko, C. K. Cruz, A. Chavenz-Aragon, and R. Contreras. A novel shape indexing method for automatic classification of lepidoptera. *IEEE Computer Society*, 2007.

# Determination of Optimal Clusters for a Non-hierarchical Clustering Paradigm *K*-Means Algorithm

T.V. Sai Krishna, A. Yesu Babu and R. Kiran Kumar

**Abstract** Non-hierarchical procedures usually require the user to specify the number of clusters before any clustering. The problem of deciding on the number of clusters which suitably fit a dataset as well as the evaluation of the clustering results is subjected to rigorous research. Therefore, we propose three methods of testing significance for determining the optimal number of clusters for a given dataset such as elbow, silhouette and gap statistic methods. A total of 52 drugs (known to act against 5-HT receptor) with their properties such as Molecular Weight, logP, Heavy Atoms, H-bond Donors (HBD), H-bond Acceptors (HBA), polar surface area (PSA), number of freely rotatable bonds (RB) and half-life period of the drug created in the form of a table was subsequently used for analysis. Before performing optimal number of clusters, the dataset is tested for clusterability using Hopkins statistic. For 5-HT receptor drug compounds dataset, the Hopkins statistic was found to be 0.2357, which indicates that the data is highly clusterable. Different methods for determining the optimal number of clusters include elbow and silhouette methods as well as gap statistic. It is evidenced that none of the methods is able to reach a consensus and estimate the number of optimal clusters. Therefore, NbClust package with 30 indices showed consensus toward the identification of the optimal number of clusters,  $k$  for the 5-HT receptor dataset where it resulted in 3 cluster solutions by maximum indices.

**Keywords** Clustering · *K*-means · Clusterability · Elbow · Silhouette Gap statistic

---

T.V. Sai Krishna  
JNTUK, Kakinada, A.P, India  
e-mail: tvsai.kris@gmail.com

A. Yesu Babu (✉)  
Sir C.R.R. College of Engineering, Eluru, A.P, India  
e-mail: adimulam9@gmail.com

R. Kiran Kumar  
Krishna University, Machilipatnam, A.P, India  
e-mail: kirankreddi@gmail.com

## 1 Introduction

Dividing data objects into clusters is a non-trivial task in data mining. Clustering data depends on the inherent properties of observations within a dataset. It should be noted that objects in a cluster should be similar to each other. Different clusters should have different distributions of associated properties of data and hence separate clusters shall be obtained. The goal in clustering is that the objects within a group share a degree of similarity to one another and different from the objects in other groups. The greater the similarity within a group and greater the difference between groups, the better or more distinct is the clustering. This would result in enhancing a balanced, compact, and separated partitioning of data. Many algorithms have been proposed to endure the clustering task which tries to find clusters which are compact, separated, balanced and parsimonious. A well-separated cluster depends on good measure of distance for the data. Different clustering algorithms will give different results on the same data. In few cases, the same clustering algorithm may give different results, for example,  $k$ -means, which involves some arbitrary initial condition.

The  $k$ -means algorithm results in a simple or flat partition, because it provides a single set of clusters, with no particular organization or structure within them. However, some clusters might be closely related to others and few are more distantly related. Non-hierarchical procedures do not involve the tree-like construction outputs. As an alternative, these methods assign objects into clusters once the number of clusters to be formed is specified prior run. The number of clusters may either be specified in advance or determined as part of the clustering procedure. Non-hierarchical methods start either from an initial partition of items into groups or an initial set of seed points, which will form the nuclei of clusters.

To perform clustering, a number of algorithms have been put forward that include  $k$ -means, hierarchical clustering, self-organizing maps, support vector machines, fuzzy clustering and others. Among them,  $k$ -means and hierarchical clustering algorithms are the most commonly used [1].

Non-hierarchical procedures usually require the user to specify the number of clusters before any clustering is accomplished and hierarchical methods routinely produce a series of solutions ranging from  $n$  clusters to a solution with only one cluster present. As such, the problem of deciding on the number of clusters which suitably fit a data set, as well as the evaluation of the clustering results, have been subject to several research efforts [2]. Therefore, we propose three methods of testing significance for determining the optimal number of clusters for a given dataset such as elbow, silhouette and gap statistic methods.

## 2 Materials and Methods

### 2.1 *R Programming*

Recent advances in molecular biology extended our chances to discover and comprehend the complexity of biological systems. Data integration has been made possible with the advent of computational tools, statistical methods and allied soft wares [3]. R program is a free software environment for statistical computing and graphics. Packages are being developed to address the specific needs of systems biology approaches, such as network modeling, simulation formal verification, and graph visualization [4]. Castelo and Roverato [5] developed an R package for molecular network discovery using microarray data. Le Meur and Gentleman [6] identified multiprotein complexes and pairs of multiprotein complexes that share an unusually high number of genetic interactions. The structure of the R software is a base program, providing basic program functionality, which can be added onto with smaller specialized program modules called packages. R is an integrated suite of software facilities for data manipulation, calculation, and graphical display. There is a difference in philosophy between R and some other statistical software, since in R a statistical analysis is normally done as a series of steps, with intermediate results being stored as objects.

### 2.2 *Dataset*

A dataset of clinically approved drugs which are known to act against Serotonin [5-hydroxytryptamine (5-HT)] receptors which are associated with many disease conditions including depression, anxiety, social phobia, schizophrenia, obsessive-compulsive and panic disorders; migraine, hypertension, pulmonary hypertension, eating disorders, vomiting, and irritable bowel syndrome were selected [7]. The serotonergic system seems to be important in bulimia nervosa (BN). Modifications in brain serotonin function contribute to different aspects of eating disorders [8]. The data set used in this research contains physicochemical properties [9] of 5-HT receptor drugs extracted from malacards database [10]. A total of 52 drugs with their properties such as Molecular Weight, logP, Heavy Atoms, H-bond Donors (HBD), H-bond Acceptors (HBA), polar surface area (PSA), number of freely rotatable bonds (RB), and half-life period of the drug created in the form of a table was subsequently used for analysis.

### 3 Assessing the Clusterability

#### 3.1 Hopkins Statistic

Before performing optimal number of clusters, the dataset is tested for clusterability using *Hopskin statistic*. It is used to assess the “clustering tendency” of a dataset by measuring the probability that a given dataset is generated by a uniform data distribution. In other words, it tests the “spatial randomness” of the data.

*Hopkins statistic* ( $H$ ) is calculated as the mean nearest neighbor distance in the random dataset divided by the sum of the mean nearest neighbor distances in the real and across the simulated dataset, given by the formula:

$$H = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n x_i + \sum_{i=1}^n y_i} \quad (1)$$

A value of  $H$  about 0.5 means that  $\sum_{i=1}^n y_i$  and  $\sum_{i=1}^n x_i$  are close to each other, and thus the data  $D$  is uniformly distributed.

If the value of *Hopkins statistic* is close to zero, then we can reject the null hypothesis and conclude that the dataset is significantly a clusterable data.

### 4 Determination of Optimal Clusters

The optimal clustering is somehow subjective and depends on the method used for measuring similarities and the parameters used for partitioning. Different methods for determining the optimal number of clusters include *elbow* and *silhouette* methods as well as *gap statistic*.

#### 4.1 Elbow Method

The basic idea behind partitioning methods, such as  $k$ -means clustering, is to define clusters such that the total intracluster variation (known as total within-cluster variation or total within-cluster sum of square) is minimized.

#### 4.2 Average Silhouette Method

Silhouette analysis measures how well an observation is clustered and it estimates the average distance between clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters.

### 4.3 Gap Statistic

The *gap statistic* can be applied to any clustering method proposed by R Tibshirani et al. [11]. The gap statistic compares the total within intracluster variation for different values of  $k$  with their expected values under null reference distribution of the data, i.e., a distribution with no obvious clustering.

## 5 Results and Discussion

The dataset used for analysis is given in Table 1. Data was normalized to improve prediction accuracy and thereby not to allow a particular feature impact the prediction due to variations in numeric range values.

```
>res$hopkins_stat
```

```
Hopskin statistic: 0.2357666
```

The value of *Hopkins statistic* is significantly  $<0.5$ , indicating that the data is highly clusterable. Additionally, it is observed that the ordered dissimilarity image (Fig. 1) contains patterns (i.e., clusters). The ordering of dissimilarity matrix is done using hierarchical clustering. For 5-HT receptor drug compounds dataset, the *Hopkins statistic* was found to be 0.2357, which indicates that the data is highly clusterable.

Three clustering techniques such as hierarchical,  $k$ -means and Partitioning around medoids (PAM) clustering were employed to assess the predictability of optimal clusters using elbow, silhouette, and gap statistic methods.

### 5.1 The Elbow method

From Figs. 2, 3, and 4, it is evidenced that elbow method for all the three algorithms was able to predict 3 clusters as optimal. It should be noted that the elbow method is sometimes ambiguous. An alternative is the average silhouette method [12] which can be also used with any clustering approach.

The program resulted in silhouette width of each observation, is given below followed by silhouette plots for all methods.

Cluster	Neighbor	Sil_width
[1,]	4	5 0.06978294
[2,]	4	5 0.36988344
[3,]	4	5 0.28755813
[4,]	4	5 0.39741189

(continued)

(continued)

Cluster	Neighbor	Sil_width
[5,]	4	5 0.29955088
[6,]	4	5 0.31411832
[7,]	4	4 -0.08133679
[8,]	4	4 0.37825473
[9,]	4	5 0.18734503
[10,]	4	5 0.38303800
[11,]	4	5 0.38247396
[12,]	4	5 0.52707498
[13,]	4	1 0.23067480
[14,]	4	1 0.17384197
[15,]	4	5 0.33233555
[16,]	4	5 0.53063366
[17,]	4	4 0.05986465
[18,]	4	3 -0.04939680
[19,]	4	4 0.18841109
[20,]	4	5 0.22710683
[21,]	4	4 0.10024237
[22,]	4	5 0.19819243
[23,]	4	5 0.25937503
[24,]	4	3 0.25228619
[25,]	4	4 0.41789549
[26,]	4	5 0.36944551
[27,]	4	5 0.35041498
[28,]	4	4 0.19632132
[29,]	4	5 0.05650501
[30,]	4	5 0.39019241
[31,]	4	5 0.36704796
[32,]	4	4 0.21678005
[33,]	4	4 0.34615794
[34,]	4	5 0.54400712
[35,]	4	5 0.17384129
[36,]	4	1 0.23397622
[37,]	4	4 0.04782804
[38,]	4	4 0.14901861
[39,]	4	5 0.57563602
[40,]	4	5 0.42291865
[41,]	4	1 0.35126930
[42,]	4	2 0.33882966
[43,]	4	4 0.12717226
[44,]	4	4 0.29148951

(continued)



(continued)

Cluster	Neighbor	Sil_width
[45,]	4	4 0.22114918
[46,]	4	5 0.12937591
[47,]	4	3 0.15725411
[48,]	4	5 0.56816451
[49,]	4	5 0.25353968
[50,]	4	5 0.01601285
[51,]	4	5 0.16736097
[52,]	4	4 -0.06993628

From the above silhouettes, it can be inferred that variables 1, 17, 37, and 50 lies between two clusters as the values are near zero. Observations with a negative  $S_i$  are observed with variables 4, 18, and 52 which suggest that they might have been placed in the wrong cluster. Moreover, an average silhouette width 0.26 suggests that the data is well clustered.

From Figs. 5, 6, and 7, it was observed that each method reported different optimal clusters and coordination in observations were not identified. *K*-means reported 10 clusters as optimal whereas PAM and hierarchical resulted in 7 and 2 clusters. Therefore, gap statistic method is studied.

The *clusGap* function from the *cluster* package calculates a *goodness of clustering* measure, called the *gap statistic*. Once all the gaps are calculated (potentially adding confidence intervals), select the number of clusters as optimal which has maximum gap.

Finally from gap statistic, only one cluster solution is suggested. It should be noted that each method displayed different clusters. Table 2 given below compares the methods employed.

From Table 2, it is evidenced that none of the methods is able to reach a consensus and estimate the number of optimal clusters. Though elbow method predicted 3 cluster solutions in all clustering algorithms, the elbow method is sometimes ambiguous. The average silhouette method resulted in varying degrees of cluster solutions whereas a single cluster solution is suggested by gap statistic method (Figs. 8, 9 and 10).

Hence, an R package *NbClust*, was used, which aims to gather all indices available in SAS or R packages together in only one package to generate possible optimal number of clusters. The output resulted in consensus from nearly 30 indices and the majority rule is followed to select optimal clusters. From Fig. 11, it can be concluded that the optimal number of clusters,  $k$  for the 5-HT receptor dataset was found to have 3 cluster solutions.

**Table 1** Dataset of drugs against 5-HT receptor extracted from Malacards database

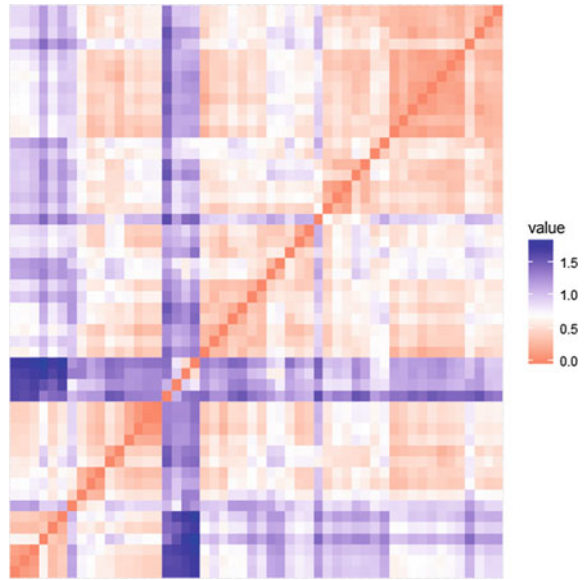
Row names	Mwt	logP	Heavy atoms	HBD	HBA	TPSA	RB	Half-life
Paroxetine	329.371	3.327	24	1	3	44	4	21
Serrraline	306.236	5.18	20	1	0	16	2	24
Citalopram	324.399	3.813	24	1	2	37	5	35
Clomipramine	314.86	4.528	22	1	1	7	4	32
Escitalopram	324.399	3.813	24	1	2	37	5	27
Fluoxetine	309.331	4.435	22	1	1	25	6	1
Fluvoxamine	318.339	3.202	22	1	3	58	9	15.6
Cocaine	303.358	1.868	22	1	4	57	3	0.5
Desipramine	266.388	3.533	20	1	1	19	4	7
Duloxetine	297.423	4.631	21	1	2	25	6	12
Imipramine	280.415	3.875	21	1	1	7	4	16
Methamphetamine	149.237	1.837	11	1	0	16	3	4
Methylphenidate	233.311	2.085	17	1	2	42	3	1
Milnacipran	246.354	1.771	18	1	1	47	5	6
Nortriptyline	263.384	3.826	20	1	0	16	3	16
Phentermine	149.237	1.966	11	1	0	27	2	7
Venlafaxine	277.408	3.036	20	2	2	33	5	5
Vilazodone	441.535	4.03	33	3	4	103	7	25.4
Amoxapine	313.788	3.429	22	1	3	41	0	8
Atomoxetine	255.361	3.725	19	1	1	25	6	5
Desvenlafaxine	263.381	2.733	19	3	2	44	4	10
Dexfenfluramine	231.261	3.246	16	1	0	16	4	32
Doxepin	279.383	3.962	21	1	1	13	3	6
Minaprine	298.39	2.196	22	1	5	50	5	2
Nefazodone	470.017	3.552	33	0	7	55	10	2
Protriptyline	263.384	4.302	20	1	0	16	4	6

(continued)

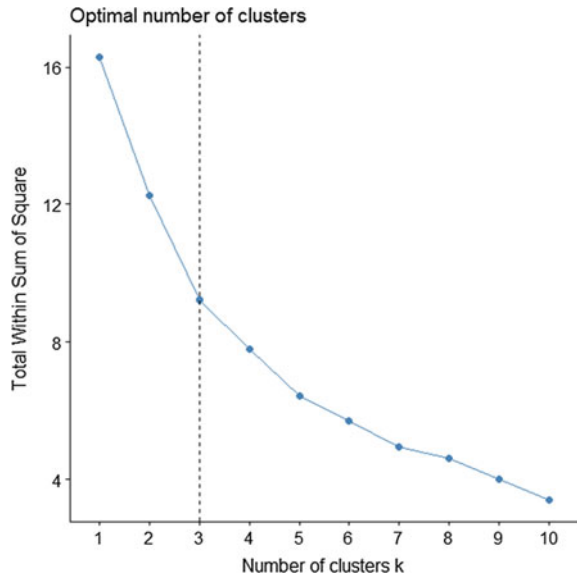
Table 1 (continued)

Row names	Mwt	logP	Heavy atoms	HBD	HBA	TPSA	RB	Half-life
Sibutramine	279.855	4.738	19	1	0	4	5	1.1
Tramadol	263.381	2.635	19	2	2	33	4	6.3
Trazodone	371.872	2.362	26	0	6	45	5	3
Trimipramine	294.442	4.121	22	1	1	7	4	11
Amiripryline	277.411	4.169	21	1	0	4	3	10
Mirtazapine	265.36	2.479	20	0	3	19	0	20
Mazindol	284.746	2.609	20	1	3	35	1	10
Pseudoephedrine	165.236	1.328	12	2	1	36	3	9
Vortioxetine	298.455	3.864	21	1	2	19	3	66
Dexmethyphenidate	233.311	2.085	17	1	2	42	3	2
Dextromethorphan	271.404	3.383	20	1	1	13	1	3
Mianserin	264.372	3.084	20	0	2	6	0	10
Amphetamine	135.21	1.576	10	1	0	27	2	10
Dopamine	153.181	0.599	11	3	2	68	2	0.02
Meperidine	247.338	2.213	18	1	2	30	3	3
Verapamil	454.611	5.093	33	1	5	65	13	2.8
Loxapine	327.815	3.771	23	0	4	28	0	4
Olanzapine	312.442	1.746	22	1	5	30	0	21
Ondansetron	293.37	3.129	22	0	4	39	2	5.7
Quetiapine	383.517	2.856	27	1	6	48	5	6
Ribavirin	324.186	2.894	21	3	11	195	5	9.5
Phenelzine	136.198	0.692	10	2	2	38	3	1.2
Alitretinoin	300.442	5.603	22	0	2	40	5	2
Tegaserod	301.394	2.815	22	4	2	87	7	11
Fenfluramine	231.261	3.246	16	1	0	16	4	20
Amineptine	337.463	4.499	25	1	2	56	8	0.48

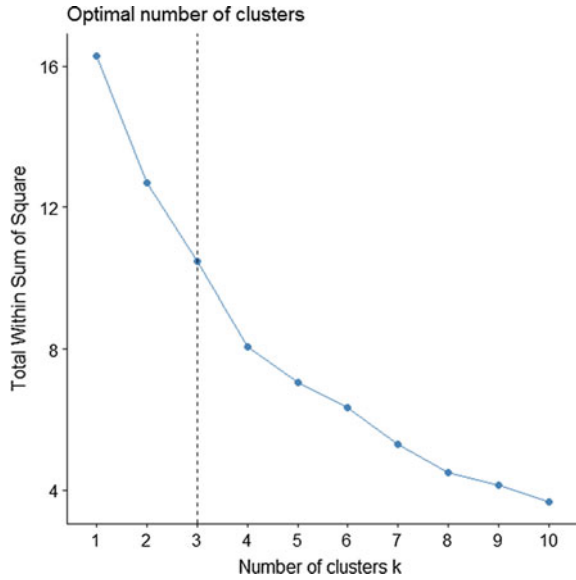
**Fig. 1** Dissimilarity matrix of the dataset



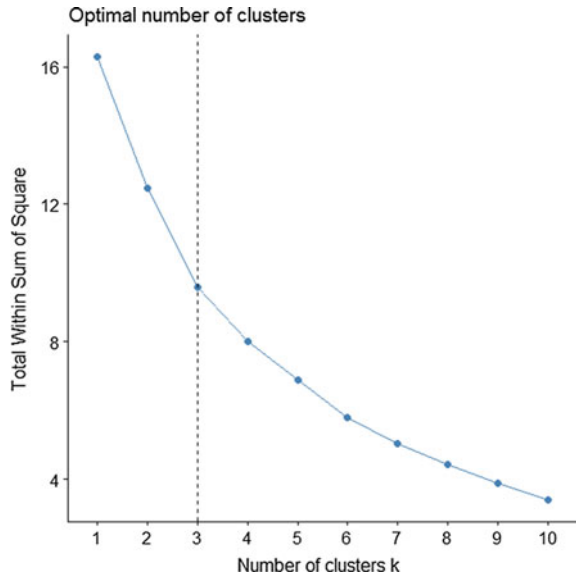
**Fig. 2** Elbow method for *k*-means clustering showing 3 cluster solutions as optimal number of clusters



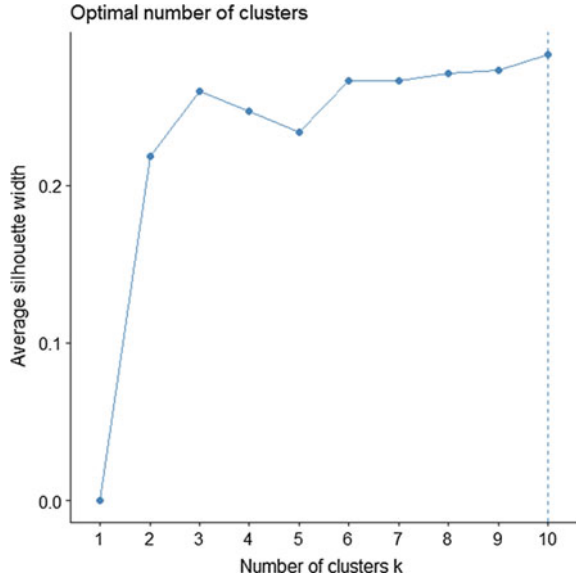
**Fig. 3** Elbow method for PAM clustering showing 3 cluster solutions as optimal number of clusters



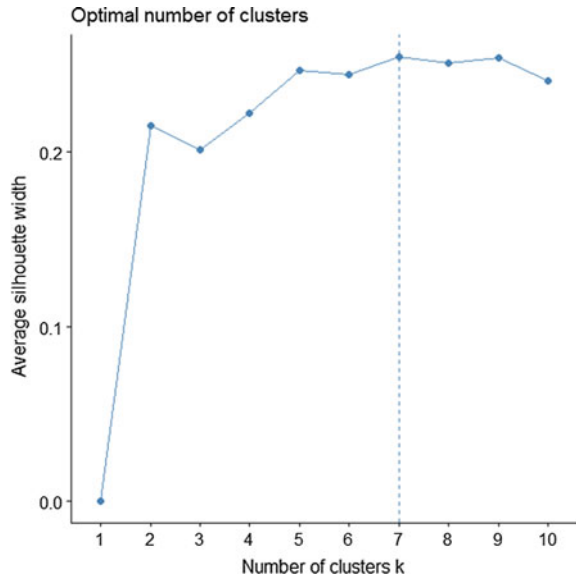
**Fig. 4** Elbow method for hierarchical clustering showing 3 cluster solutions as optimal number of clusters



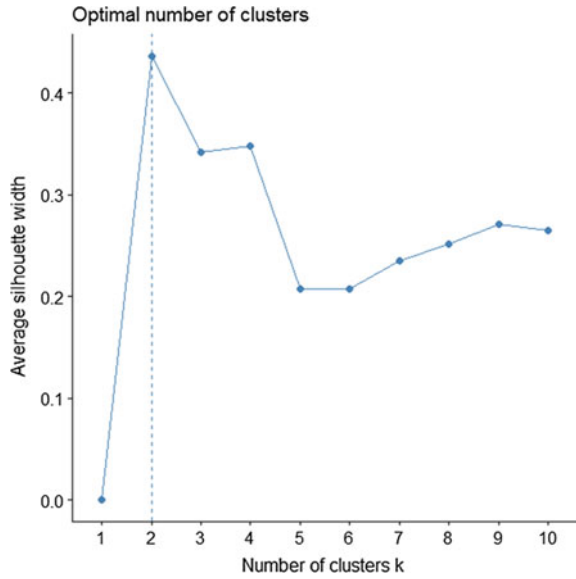
**Fig. 5** Average silhouette method for *k*-means clustering showing 10 clusters being optimal



**Fig. 6** Average silhouette method for PAM clustering showing 7 clusters as optimal



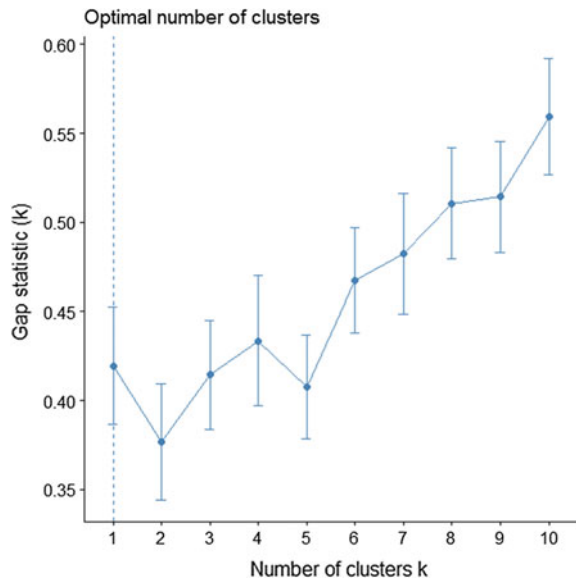
**Fig. 7** Average silhouette method for hierarchical clustering showing 2 clusters as optimal



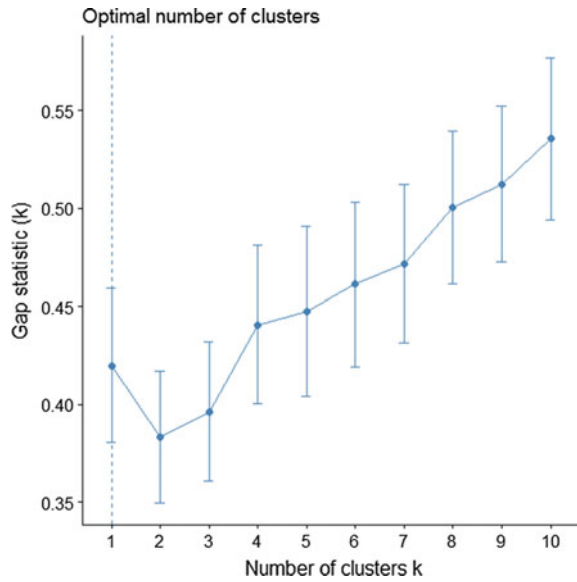
**Table 2** Comparative data on optimal clusters generated by elbow, silhouette and gap statistic methods

Method	Number of optimal clusters		
	<i>k</i> -means	PAM	Hierarchical
Elbow	3	3	3
Silhouette	10	7	2
Gap statistic	1	1	1

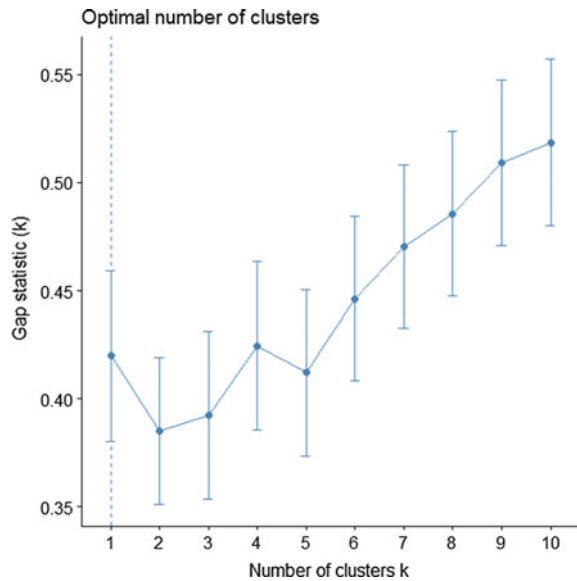
**Fig. 8** Gap statistic method for *k*-means clustering suggesting 1 optimal cluster



**Fig. 9** Gap statistic method for PAM clustering suggesting 1 optimal cluster

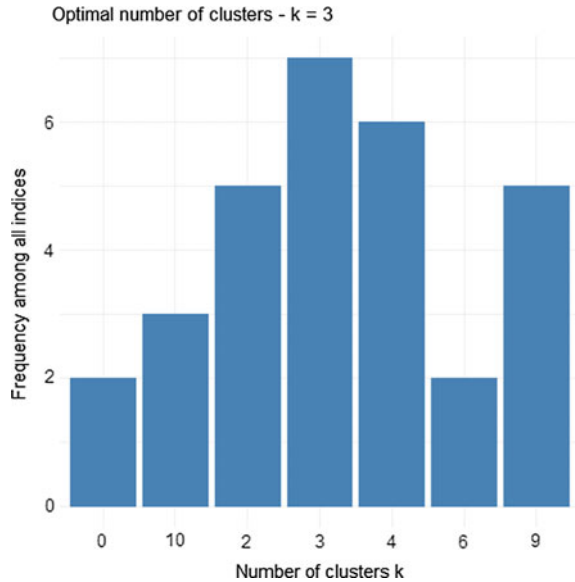


**Fig. 10** Gap statistic method for hierarchical clustering suggesting 1 optimal cluster





**Fig. 11** Consensus on optimal number of clusters obtained from NbClust package



## 6 Conclusion

In this work, a study has been made possible to evaluate the number of optimal clusters for any non-hierarchical clustering technique, such as  $k$ -means where initial clusters should be mentioned prior analysis. The value of *Hopkins statistic* is significantly  $<0.5$ , indicating that the data is highly clusterable. Three clustering techniques such as hierarchical,  $k$ -means and Partitioning around medoids (PAM) clustering employed to assess the predictability of optimal clusters using elbow, silhouette, and gap statistic methods resulted in ambiguous data. Hence, an R package NbClust was implemented to evaluate the optimal number of clusters where the consensus approach resulted in 3 cluster solutions as the best option.

## References

1. Mohammad Shabbir Hasan and Zhong-Hui Duan. Hierarchical  $k$ -Means: A Hybrid Clustering Algorithm and Its Application to Study Gene Expression in Lung Adenocarcinoma. Chapter 4 In: "Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology". 2015, Pages 51–67 DOI:[10.1016/B978-0-12-802508-6.00004-1](https://doi.org/10.1016/B978-0-12-802508-6.00004-1).
2. Charrad, Malika, et al. Nb Clust: An R Package for Determining the Relevant Number of Clusters in a Data Set. *J. Stat. Soft* 61 (2014): 1–36.
3. Nolwenn Le Meur and Robert Gentleman. Analyzing Biological Data Using R: Methods for Graphs and Networks. Chapter 19.
4. Huber W, Carey VJ, Long L, Falcon S, Gentleman R. (2007) Graphs in molecular biology. *BMC Bioinformatics*, 8(6):S8.

5. Castelo R, Roverato A. (2009) Reverse engineering molecular regulatory networks from microarray data with qp-graphs. *J Comput Biol*, 16(2):213–227.
6. Le Meur N, Gentleman R. (2008) Modeling synthetic lethality. *Genome Biol*, 9(9):R135.
7. <https://pdfs.semanticscholar.org/bedf/1761ec0ab9d54634c353618447079aceb1f3>.
8. Howard Steiger. Eating disorders and the serotonin connection: state, trait and developmental effects. *J Psychiatry Neurosci*. 2004 Jan; 29(1): 20–29.
9. <https://pubchem.ncbi.nlm.nih.gov/>.
10. <http://www.malacards.org>.
11. Tibshirani, R., Walther, G. and Hastie, T. (2001). Estimating the number of data clusters via the Gap statistic. *Journal of the Royal Statistical Society B*, 63, 411–423.
12. Kaufman, L. and Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York.

# Nonlinear Tensor Diffusion Filter Based Marker-Controlled Watershed Segmentation for CT/MR Images

S.N. Kumar, A. Lenin Fred, H. Ajay Kumar  
and P. Sebastian Varghese

**Abstract** The segmentation is the process of extraction of the desired region of interest and it plays a vital role in computer vision and image processing. The conventional watershed segmentation is prone to over segmentation due to the presence of noise. The nonlinear tensor diffusion filtering is used for preprocessing the CT/MR images prior to segmentation and its performance is superior to conventional spatial domain filters like median, Gaussian, and bilateral filter. The preprocessed image was subjected to marker-controlled watershed segmentation; satisfactory results were produced when compared with the morphological gradient and area open morphological gradient watershed approaches. The algorithms were developed in Matlab 2010a and tested on abdomen CT and knee MR images.

**Keywords** Nonlinear tensor diffusion · Segmentation · Marker-controlled watershed · Morphology gradient approach

## 1 Introduction

The role of segmentation is inevitable in medical image processing in terms of applications like localization of tumor and anomalies, study of anatomical structure, diagnosis, measurement of tissue volumes, computer-guided surgery, and treatment planning. The thresholding, region growing, and edge detection are classical segmentation algorithms and other semiautomated and automated algorithms for medical images are described in [1].

---

S.N. Kumar (✉)  
Sathyabama University, Chennai, India  
e-mail: appu123kumar@gmail.com

A. Lenin Fred (✉) · H. Ajay Kumar  
Mar Ephraem College of Engineering and Technology, Elavuvilai, Marthandam, India  
e-mail: leninfred.a@gmail.com

P. Sebastian Varghese  
Metro Scans and Laboratory, Trivandrum, India  
e-mail: sebastin464@gmail.com

The morphological gradient, area open morphological gradient, and marker-controlled watershed segmentation approaches were analyzed in this paper. The watershed segmentation algorithms were analyzed under three conditions; without filtering, with median filtering, and with nonlinear tensor diffusion (NLTD) filtering. The marker-controlled watershed approach with NLTD filtering produces efficient results than morphological gradient, area open morphological gradient approaches.

The hybrid segmentation algorithm based on watershed and fast region algorithm with edge preserving noise reduction was robust for synthetic and MR brain images [2]. The conventional watershed algorithm issues like over segmentation, noise sensitivity, and detection of thin structures were solved by the incorporation of prior information for the segmentation in MR images of knee cartilage and brain [3]. The morphological opening and closing play a vital role in gray scale image filtering and segmentation [4, 5]. Similarly, the morphological reconstruction also plays a dominant role in various image analysis applications [4]. The marker-controlled watershed segmentation algorithm based on gray scale morphology was found to be efficient for color, gray scale MR medical images, and aerial images [6]. The morphological operations are employed for the segmentation of hyper spectral images and the algorithms were implemented in multiprocessor system [7, 8]. A hybrid segmentation algorithm comprising of watershed and  $K$ -means clustering with prior shape information was employed for the segmentation of MR images of the corpus callosum [9]. The thresholding and watershed algorithm was also coupled for the segmentation of MR images of the brain, the preprocessing was done by the median filter and Gaussian high-pass filter [10]. The anisotropy diffusion filter was employed for the preprocessing of images prior to watershed segmentation for the segmentation of heart and brain MR images [11]. The marker-based watershed algorithm with non-sampled contourlet transform was used for the automatic multi organ segmentation of prostate MR images [12].

## 2 Method and Materials

In watershed segmentation, the gradient of the image is processed rather than taking the raw image itself [13]. The image is viewed as a topographical surface in which the  $(x, y)$  coordinates correspond to the position and  $z$  coordinate corresponds to gray scale Intensity. The catchment basin corresponds to points where the drops of water certainly fall to single minima. The watershed lines correspond to points where drops of water will likely to fall in more than one minimum. In watershed segmentation, the image is clustered based on gray scale intensity similar to the dam construction preventing water flow into different locations. The principle of watershed segmentation is depicted below in Fig. 1.

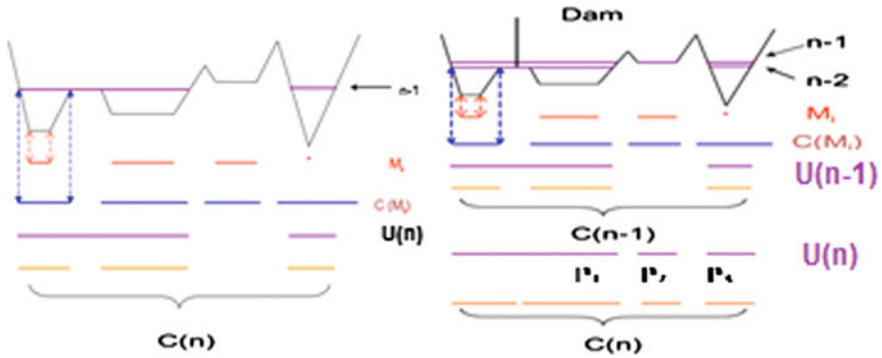


Fig. 1 Principle of watershed algorithm

- Let  $M_1, M_2, \dots, M_R$  represent the coordinates of the points in the regional minima of the gradient of the image  $I(x, y)$ .
- The  $C(M_i)$  are the coordinates of the points in the catchment basin related with regional minimum  $M_i$  and  $U[n]$  are the set of coordinates  $(s, t)$  for which  $I(s, t) < n$ .
- During each flooding, the topography is viewed as a binary image and it is done with integer flood increments from  $n = \min + 1$  to  $n = \max + 1$ , where  $\min$  and  $\max$  represent the minimum and maximum gray value of the image  $I(x, y)$ .
- Let  $C_n(M_i)$  depict the set of coordinates of points in the catchment basin associated with minimum  $M_i$  at flooding stage  $n$ .

$$C_n(M_i) = C(M_i)U[n] \tag{1}$$

$$C_n(M_i) = \subseteq U[n] \tag{2}$$

- Let  $C_n(M_i)$  depict the set of coordinates of points in the catchment basin associated with minimum  $M_i$  at flooding stage  $n$ .
- The  $C[n]$  represents the union of the flooded catchment basin. For each step in  $n$ ,  $C[n - 1]$  has been constructed and the goal is to obtain  $C[n]$  from  $C[n - 1]$ .
- The  $Q[n]$  represents the set of connected components in  $U[n]$  and for each  $p \in Q[n]$ , there are three possibilities
- Case 1:  $p \cap C[n - 1]$  is empty ( $p_1$ ), a new minimum is encountered and  $p$  is incorporated into  $C[n - 1]$  to form  $C[n]$ .
- Case 2:  $p \cap C[n - 1]$  comprises one connected component ( $p_2$ ) of  $C[n - 1]$ ,  $p$  is incorporated into  $C[n - 1]$  to form  $C[n]$ .
- Case 3:  $p \cap C[n - 1]$  contains more than one connected components ( $p_3$ ) of  $C[n - 1]$ , the ridge separating two or more catchment basins has been encountered and a dam is built within  $p$  to prevent overflow between the catchment basins.
- Repeat the procedure until  $n = \max + 1$ .

In this paper, the improvements in classical watershed algorithm were analyzed and marker-controlled watershed with NLTD filtering produces efficient results

### 3 Preprocessing

The preprocessing plays a vital role in watershed segmentation algorithm since the classical watershed algorithm is prone to noise and results in over segmentation, failure in the detection of thin objects. The conventional median filter approach is widely used in many applications; however, the non-noisy pixels are also affected and the edge preservation is also poor. The NLTD filtering approach is a partial differential equation-based approach and it produces better results than conventional spatial domain filters, Perona Malika (PM) model and nonlinear scalar diffusion (NLSD) filtering approach.

The NLTD restoration mathematical model is depicted below:

$$\partial_t I = \nabla \cdot (T \nabla I) \quad (3)$$

where  $T$  is a positive semi definite symmetric diffusion sensor.

The preprocessing of 2D images are considered in this paper, hence  $T$  is a  $2 \times 2$  matrix as follows:

$$T = \begin{bmatrix} P & Q \\ Q & R \end{bmatrix} \quad (4)$$

where

$$P = \frac{c_1 g_x^2 + c_2 g_y^2}{I_G^2 + \varepsilon} \quad (5)$$

$$Q = \frac{(c_2 - c_1) g_x g_y}{I_G^2 + \varepsilon} \quad (6)$$

$$R = \frac{c_1 g_y^2 + c_2 g_x^2}{I_G^2 + \varepsilon} \quad (7)$$

The term  $I_G$  represents the Gaussian smoothed version of the input image and  $g_x$ ,  $g_y$  are the components of  $I_G$ . The terms  $c_1$  and  $c_2$  are the diffusion constants.

$$c_1 = \exp\left(-\left(\frac{I_G}{K}\right)^2\right) \quad (8)$$

$$c_2 = \frac{1}{5} * c_1 \tag{9}$$

The element of tensors are functions of the image characteristics. The NLTD in terms of Cartesian coordinates.

$$\partial_t I = [\partial_x \quad \partial_y] \begin{bmatrix} P & Q \\ Q & R \end{bmatrix} \begin{bmatrix} \partial_x I \\ \partial_y I \end{bmatrix} \tag{10}$$

$$\partial_t I = [\partial_x \quad \partial_y] \begin{bmatrix} P\partial_x I + Q\partial_y I \\ Q\partial_x I + R\partial_y I \end{bmatrix} \tag{11}$$

$$\partial_t I = \partial_x(P\partial_x I + Q\partial_y I) + \partial_y(Q\partial_x I + R\partial_y I) \tag{12}$$

$$\partial_t I = \partial_x(P\partial_x I) + \partial_x(Q\partial_y I) + \partial_y(Q\partial_x I) + \partial_y(R\partial_y I) \tag{13}$$

While comparing this partial differential equation with the NLSD model, two new terms arise  $\partial_x(Q\partial_y I)$  and  $\partial_y(Q\partial_x I)$ .

The NLTD restoration model in discrete form is as follows:

$$I_{ij}^{t+\lambda} = I_{ij}^t + \lambda \left[ Q_1 I_{i-1,j+1} + R_1 I_{i,j+1} + Q_2 I_{i+1,j+1} + P_1 I_{i-1,j} - \left( \frac{P_{i-1,j} + 2P_{ij} + P_{i+1,j} + R_{i-1,j} + 2R_{ij} + R_{i+1,j}}{2} \right) I_{ij} + P_2 I_{i+1,j} + Q_3 I_{i-1,j-1} \right] \tag{14}$$

where

$$P_1 = \frac{P_{i-1,j} + Q_{ij}}{4} \tag{15}$$

$$P_2 = \frac{P_{i+1,j} + P_{ij}}{2} \tag{16}$$

$$Q_1 = \frac{-Q_{i-1,j} + Q_{ij+1}}{4} \tag{17}$$

$$Q_2 = \frac{Q_{i+1,j} + Q_{ij+1}}{4} \tag{18}$$

$$Q_3 = \frac{Q_{i-1,j} + Q_{ij+1}}{4} \tag{19}$$

$$R_1 = \frac{R_{i,j+1} + R_{ij}}{2} \tag{20}$$

## 4 Approaches in Watershed Segmentation

### 4.1 Morphological Gradient Approach

The morphological operations are a nonlinear transformation that applies a structural element to an input image and creates an output image of same size [14]. The erosion and dilation are preliminary mathematical morphology operations and it depends upon the structuring element shape and size.

The dilation allows objects to expand and hence effectively fill small holes and connect disjoint sets. The gray scale dilation of  $I$  by  $S$ , denoted by  $I \oplus S$  is defined as follows:

$$(I \oplus S)(s, t) = \max \left\{ I(s-x, t-y) - \frac{S(x, y)}{(s-x)}, (t-y) \in D_I; (s, y) \in D_S \right\} \quad (21)$$

where  $D_I$  and  $D_S$  are the domains of  $I$  and  $S$ , respectively.

The characteristics of dilation are as follows:

- i. When the origin of structuring element coinciding with the image pixel is “white”, there is no change and position changes to next pixel.
- ii. When the origin of structuring element coinciding with the image pixel is “black”, all the pixels covered by the structuring element are changed to black.

The erosion shrinks objects there by removing the boundaries. The gray scale erosion of  $I$  by  $S$ , denoted by  $I \ominus S$  is defined as follows:

$$(I \ominus S)(s, t) = \max \left\{ I(s-x, t+y) - \frac{S(x, y)}{(s-x)^{prime}}, (t+y) \in D_I; (s, y) \in D_S \right\} \quad (22)$$

where  $D_I$  and  $D_S$  are the domains of  $I$  and  $S$ , respectively.

The characteristics of erosion are as follows:

- i. When the origin of structuring element coinciding with the image pixel is “white”, there is no change and position change to next pixel.
- ii. When the origin of structuring element coinciding with the image pixel is “black”, if at least one of the pixels in the image under the structuring element is in “white” pixel range then the pixel value of image beneath the center of structuring element is changed from ‘black’ to “white.”



The morphological gradient function comprises of dilation followed by erosion is employed in this approach prior to the classical watershed algorithm. The classical watershed algorithm is described above.

### 4.2 Area Open Morphological Gradient Approach

The connected opening  $C_i(I)$  of a set  $I \subseteq M$  at point  $i \in M$  is the connected component of  $I$  containing  $i$ , if  $i \in I$  and  $\phi$  otherwise. The area opening  $\Gamma_\lambda^a$  is defined on subsets of  $M$  which is represented as follows:

Consider  $I \subset M$  and  $\lambda \geq 0$ . The area opening operation of the parameter  $\lambda$  of  $I$  is given by

$$\Gamma_\lambda^a(I) = \{i \in I \mid \text{Area}[C_i(I) \geq \lambda]\} \tag{23}$$

After the area open operation, the morphological gradient function is applied prior to the classical watershed segmentation approach.

### 4.3 Marker-Controlled Morphological Gradient Approach

The marker-controlled watershed segmentation was found to be efficient for the segmentation of objects with closed contours. The internal marker limits the number of regions specifying the region of interest. Similar to the seed point generation in region growing algorithm, markers can be manually or automatically selected. The regions without internal markers are merged and no dam is built. The external marker represents the pixels that belong to the background. The watershed lines are termed as external markers as they belong to the similar background region. The automatic generation of markers are employed in this work and it is based on dilation-based image reconstruction and erosion-based image reconstruction. For the extraction of foreground marker, the dilation-based image reconstruction and erosion-based image reconstruction are sequentially applied the resultant image is subjected to regional maximum function for the extraction of foreground marker. The reconstructed image after dilation and erosion is subjected to canny edge detector for the extraction of background marker. The gradient image in this approach is determined from the area open morphological gradient image, fore ground marker, and back ground marker.

Let  $I, J$  be the two images defined on the same domain and  $J \leq I$ . The dilation-based image reconstruction is described as follows. The reconstruction of  $I$  from  $J$  denoted as  $I_{\text{rec}}(J)$  is obtained by iterating elementary geodesic dilation of  $J$  under  $I$  until stability is reached.

$$I_{\text{rec}}(J) = \bigcup_{n \geq 1} \psi^{(n)}(J) \quad (24)$$

where  $\delta^{(n)}(J)$  is obtained by iterating an elementary geodesic dilation and the geodesic dilation is defined as

$$\psi^{(n)}(J) = (J \oplus S) \cap I \quad (25)$$

( $S$  is the disk-structuring element of size 5 and  $\cap$  stands for point wise minimum).

Similarly the erosion-based reconstruction  $I_{\text{rec}}(J)$  is obtained by iterating elementary geodesic erosion of  $J$  above  $I$  until stability is reached

$$I_{\text{rec}}(J) = \bigcap_{n \geq 1} \alpha^{(n)}(J) \quad (26)$$

where  $\alpha^{(n)}(J)$  can be obtained by iterating  $n$  elementary geodesic dilation and the geodesic dilation is defined as

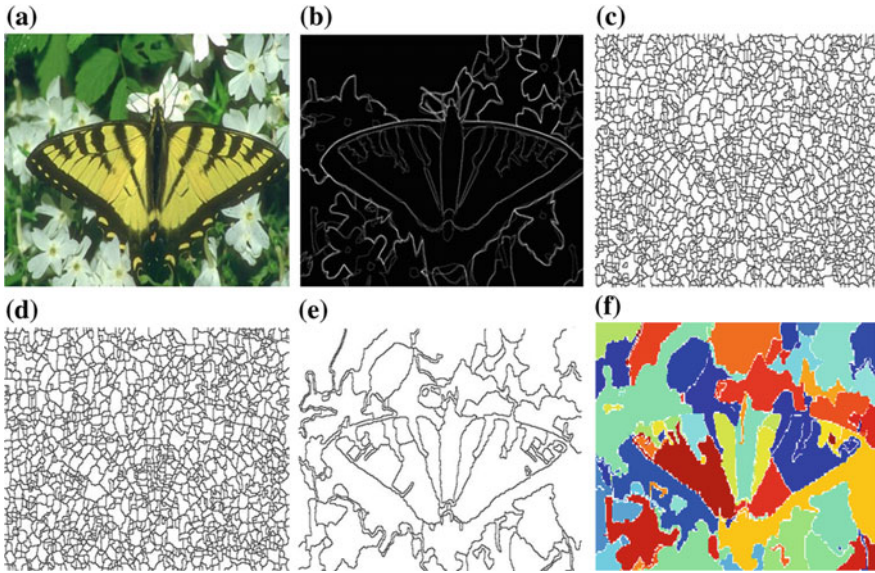
$$\alpha^{(n)}(J) = (J \ominus S) \cup I \quad (27)$$

( $S$  is disk-structuring element of size 5 and  $\cup$  stands for point wise maximum).

## 5 Results and Discussion

This paper proposes different approaches of watershed algorithm for the segmentation of medical images. The algorithms were tested on the desktop computer with specifications—Intel i3 processor; 4 GB RAM, 64 bit operating system. Since real-time medical images are used here, the ground truth generation is difficult; hence for the evaluation of watershed segmentation algorithm approaches, images from Berkley database are used. The ground truth images are available in the Berkley database [15] that can be used to compare the machine generated segmentation result. Figure 2 below depicts the input image from the database, ground truth image and the result of various watershed segmentation approaches. The visual inspection itself is showing that the marker-controlled morphological gradient watershed algorithm generates efficient results than the morphological gradient and area open morphological gradient approach.

The performance analysis was done by error metrics like over detection error (ODE), under detection error (UDE), and localization error (LE). Let  $S_{\text{ref}}$  be the ground truth image and  $S_c$  be the extracted contour obtained from the segmentation



**Fig. 2** **a** Input image **b** Ground truth image **c** Morphological watershed segmentation **d** Area open morphological watershed segmentation **e, f** Marker-controlled morphological watershed segmentation

result of image  $I$ . The ODE represents the contours that do not coincide with the  $S_{ref}$ . The UDE represents pixels of  $S_{ref}$  that have not been detected. The LE determines the distance between the misclassified pixels and the nearest pixels of  $S_{ref}$ . The segmentation algorithm is said to be good if the three errors are minimum. In the expressions for ODE, UDE, and LE, the term  $card(I)$  represents the number of contour pixels of the image which is subjected to segmentation.

Under detection error

$$UDE = \frac{card(S_{ref/C})}{card(S_{ref})} \tag{28}$$

Over detection error

$$ODE = \frac{card(S_C) - card(S_C \cap S_{ref})}{card(S) - card(S_{ref})} \tag{29}$$

Localization error

$$LE = \frac{card(S_{ref/C} \cup S_{C/ref})}{card(I)} \tag{30}$$

**Table 1** Performance analysis of watershed segmentation approaches for Berkley database images

Image ID	Error measure	Morphological watershed algorithm	Area open morphological watershed algorithm	Marker-controlled morphological watershed algorithm
35010	Over	0.9486	0.8036	0.7486
35010	Under	0.0516	0.0416	0.0316
35010	Loc	0.2396	0.1986	0.1012
42049	Over	0.9365	0.8759	0.7935
42049	Under	0.0419	0.0400	0.0395
42049	Loc	0.2245	0.2189	0.1923
118035	Over	0.8946	0.7989	0.7123
118035	Under	0.0339	0.0224	0.0198
118035	Loc	0.2145	0.2233	0.1145
227092	Over	0.8225	0.7659	0.6985
227092	Under	0.0331	0.0212	0.0112
227092	Loc	0.1989	0.1812	0.1655

From Table 1, it is clear that the marker-controlled morphological watershed algorithm produces efficient results than the morphological watershed and area open morphological watershed algorithm. For the real-time medical images, marker-controlled watershed algorithm was tested under three cases; without filtering, with median filtering and with NLTD filter (Fig. 3).

The algorithms were tested on real-time abdomen CT and knee MR images. The images were obtained from Metro scans and Research Laboratory, Trivandrum. The segmentation results depict the marker-controlled morphological watershed without filter, with median filter and with NLTD filter. The marker-controlled morphological watershed segmentation with NLTD filter was found to produce efficient results than the other two cases (without filter and with median filter).

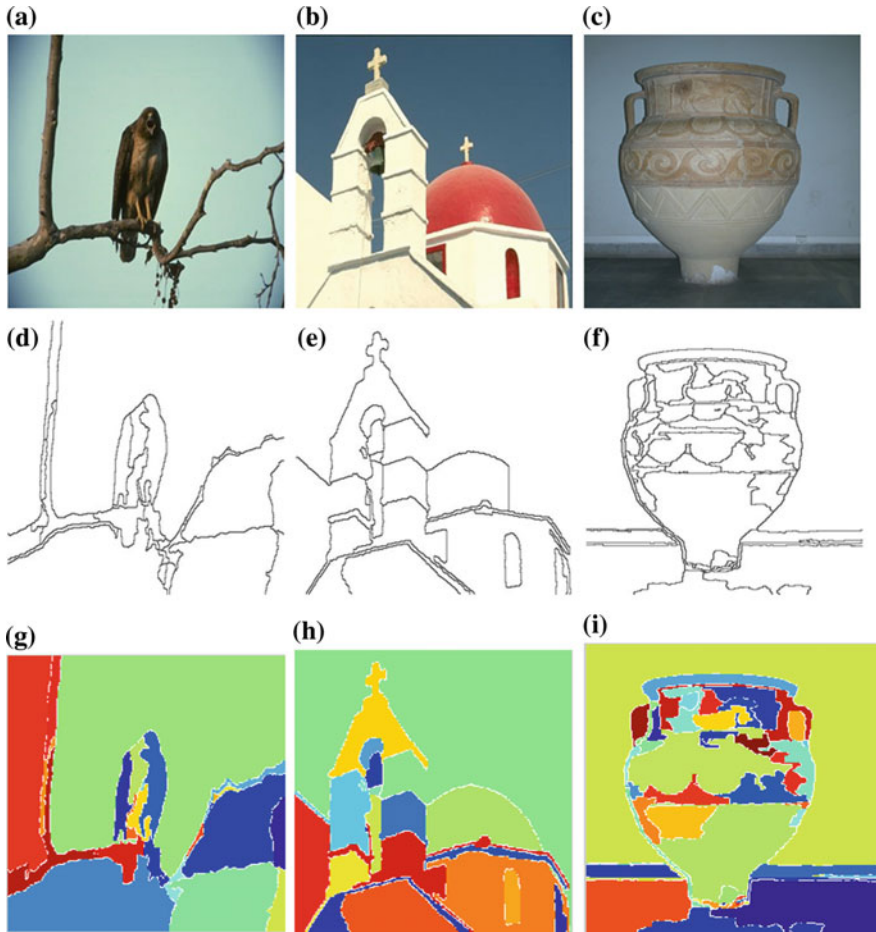
In each figure depicted below, the first column represents the input images, the second column represents the marker-controlled watershed segmentation without filter, the third column represents the marker-controlled watershed segmentation with median filter, and the fourth column represents the marker-controlled watershed segmentation with NLTD filter. Figure 4 depicts the segmentation in healthy abdomen CT images.

Figure 5 depicts the segmentation in abdomen CT images with liver tumor.

Figure 6 depicts the segmentation in abdomen CT images with liver cirrhosis.

Figure 7 depicts the segmentation in abdomen CT images with lesions in kidney.

Figure 8 depicts the segmentation in MR knee images.



**Fig. 3** First row represents the input images from Berkley database, the second row represents the contour of marker-controlled watershed segmentation and third row represents marker-controlled watershed segmentation results

## 6 Conclusion

This paper proposes three approaches in watershed segmentation. The marker-controlled watershed was found to be better than morphological watershed and area open morphological watershed approaches. The result was validated by performance metrics on Berkley database images. The marker-controlled watershed segmentation was tested on abdomen CT and knee MR images under three conditions; without filter, with median filter and with NLTD filter. The marker-controlled watershed algorithm with NLTD filter was found to produce efficient results.

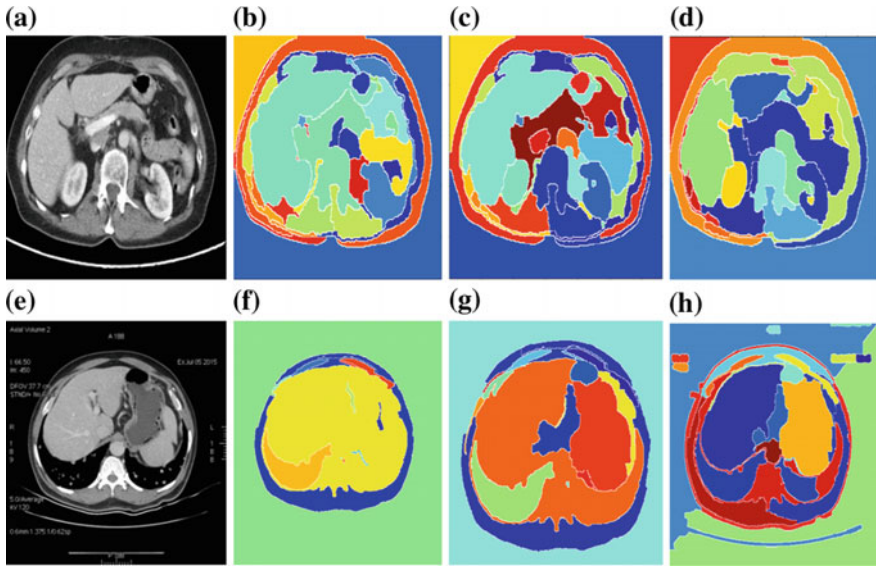


Fig. 4 Healthy abdomen CT liver segmentation results

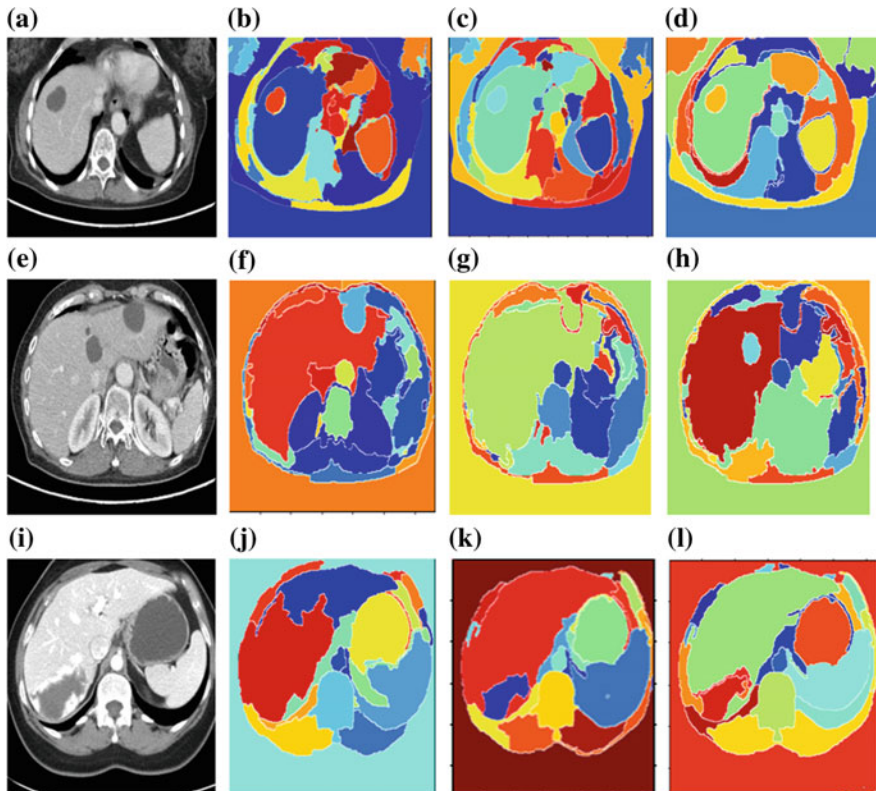


Fig. 5 Abdomen CT liver with lesions segmentation results

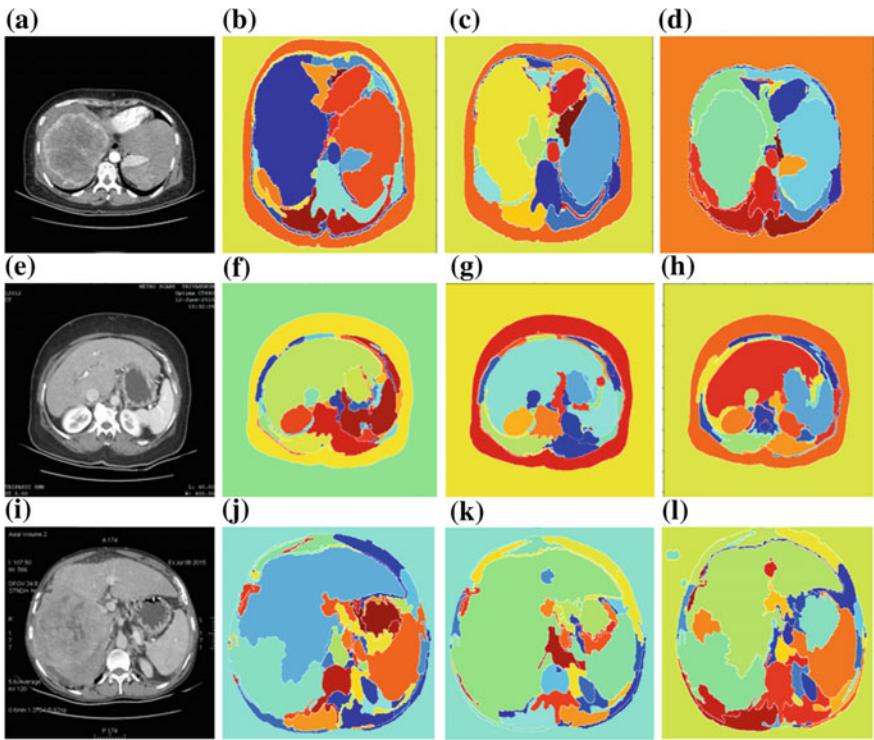


Fig. 6 Abdomen CT liver cirrhosis segmentation results

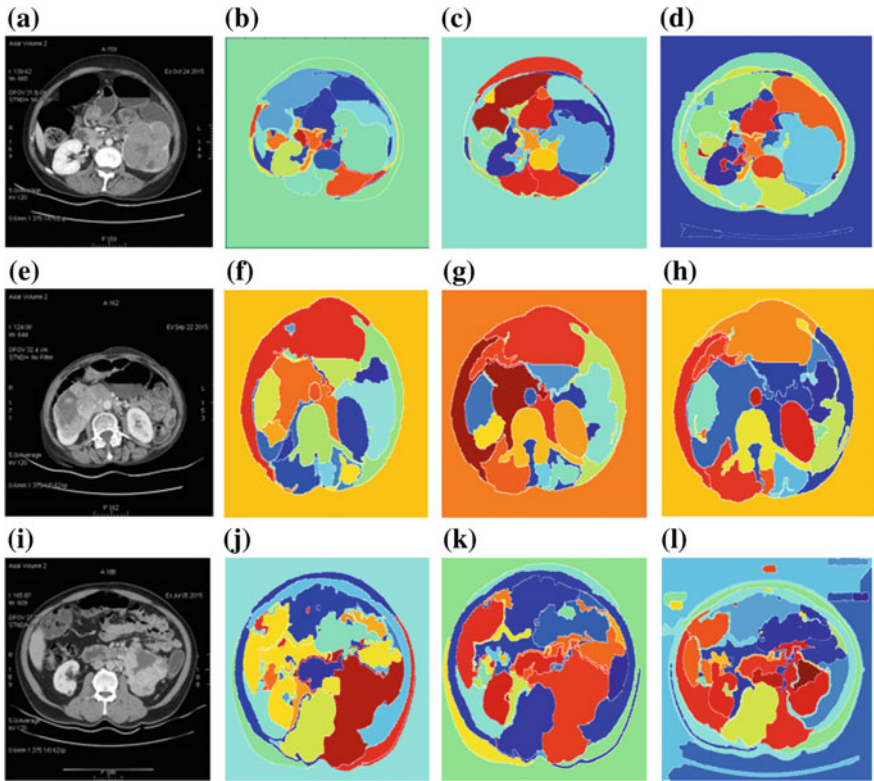


Fig. 7 Abdomen CT kidney with lesions segmentation results

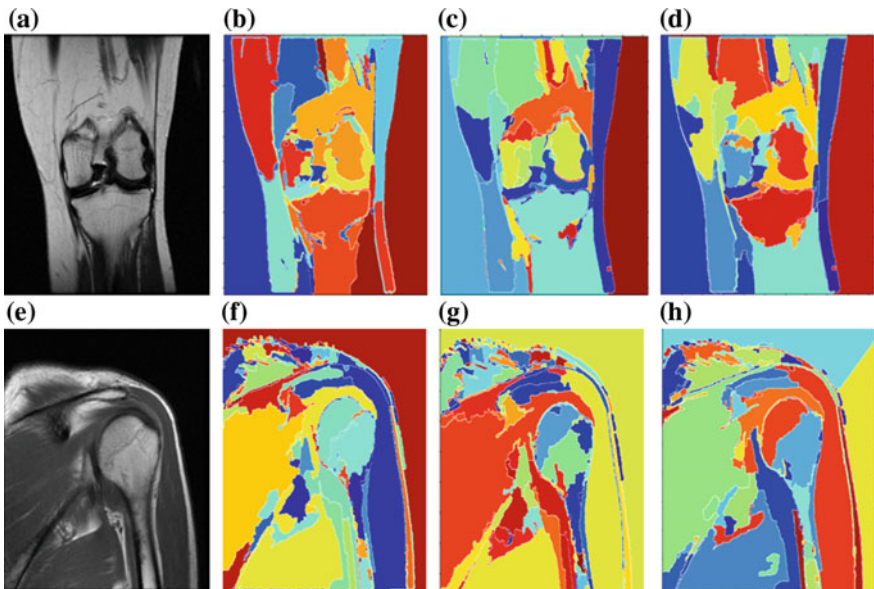


Fig. 8 Knee MR segmentation results



**Acknowledgements** The authors would like to acknowledge the support provided by DST under IDP scheme (No: IDP/MED/03/2015).

## References

1. Dzung L. Phamy, Chenyang Xu, Jerry L. Prince, A Survey of Current Methods in Medical Image Segmentation, Annual Review of Biomedical Engineering, January 19, 1998.
2. Kostas Haris, Serafim N. Efstratiadis, Nicos Maglaveras, and Aggelos K. Katsaggelos, Hybrid Image Segmentation Using Watersheds and Fast Region Merging, IEEE Trans. on Image Processing, 7(12), 1684–1699, (1998). doi:[10.1109/83.730380](https://doi.org/10.1109/83.730380).
3. V. Grau, A. U. J. Mewes, M. Alcañiz, Member, R. Kikinis, and S. K. Warfield, Improved Watershed Transform for Medical Image Segmentation Using Prior Information, IEEE Trans. on Medical Imaging, 23(4), 447–458, (2004), doi:[10.1109/TMI.2004.824224](https://doi.org/10.1109/TMI.2004.824224).
4. Luc Vincent, Grayscale area openings and closings, their efficient implementation and applications, Proc. EURASIP Workshop on Mathematical Morphology and its Applications to Signal Processing, Barcelona, Spain, pp. 22–27, May 1993.
5. Luc Vincent, Morphological Area Openings and Closings for Grayscale Images, Proc. NATO Shape in Picture Workshop, Driebergen, The Netherlands, Springer-Verlag, pp. 197–208, September 1992.
6. K. Parvati, B. S. Prakasa Rao, and M. Mariya Das, Image Segmentation Using Gray-Scale Morphology and Marker-Controlled Watershed Transformation, Hindawi Publishing Corporation, Discrete Dynamics in Nature and Society, Volume 2008, Article ID 384346, 8 pages, doi:[10.1155/2008/384346](https://doi.org/10.1155/2008/384346).
7. David Valencia and Antonio Plaza, Efficient Implementation of Morphological Opening and Closing by Reconstruction on Multi-Core Parallel Systems, First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing, 2009. WHISPERS '09. doi:[10.1109/WHISPERS.2009.5289002](https://doi.org/10.1109/WHISPERS.2009.5289002).
8. Martino Pesaresi and Jon Atli Benediktsson, A New Approach for the Morphological Segmentation of High-Resolution Satellite Imagery, IEEE Trans. on Geoscience and Remote Sensing, 39(2), 309–320, February 2001, doi:[10.1109/36.905239](https://doi.org/10.1109/36.905239).
9. Ghassan Hamarneh, Xiaoxing Li, Watershed segmentation using prior shape and appearance knowledge, Image and Vision Computing, 27(1–2) 59–68, 1 January 2009, doi: <http://dx.doi.org/10.1016/j.imavis.2006.10.009>.
10. Anam Mustaqeem, Ali Javed, Tehseen Fatima, An Efficient Brain Tumor Detection Algorithm Using Watershed & Thresholding Based Segmentation, I.J. Image, Graphics and Signal Processing, 10, 34–39, 2012, DOI:[10.5815/ijigsp.2012.10.05](https://doi.org/10.5815/ijigsp.2012.10.05).
11. Mithun Kumar PK, Md. Gauhar Arefin, Mohammad Motiur Rahman and A. S. M. Delowar Hossain, Automatically Gradient Threshold Estimation of Anisotropic Diffusion for Meyer's Watershed Algorithm Based Optimal Segmentation, I.J. Image, Graphics and Signal Processing, 12, 26–31, 2014, DOI:[10.5815/ijigsp.2014.12.04](https://doi.org/10.5815/ijigsp.2014.12.04).
12. Luc Vincent, Morphological Grayscale Reconstruction in Image Analysis: Applications and Efficient Algorithms, IEEE Tran. on Image Processing, 2(2), 176–201, 1993, DOI:[10.1109/83.217222](https://doi.org/10.1109/83.217222).
13. R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd ed., Upper Saddle River, NJ, Prentice Hall, 2002, pp. 541–553, 639–647.
14. Tulsani, Hemant, Saransh Saxena, and Mamta Bharadwaj. “Comparative study of techniques for brain tumor segmentation.” In Multimedia, Signal Processing and Communication Technologies (IMPACT), 2013 International Conference on, IEEE., 117–120, 2013, DOI:[10.1109/MSPCT.2013.6782100](https://doi.org/10.1109/MSPCT.2013.6782100).
15. D. Martin and C. Fowlkes and D. Tal and J. Malik, A Database of Human Segmented Natural Images and its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics, Proc. 8th Int'l Conf. Computer Vision, 2, 416–423, 2001, website: <https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/>.

# Improved Ensemble Methods to Solve Multi-class Imbalance Problem Using Adaptive Weights

K. Vasantha Kokilam and D. Ponmary Pushpa Latha

**Abstract** Nowadays, usage of the internet-of-things and oftenest of mobile phones, the technology, and networks of sensors has led to a huge and sudden increase in the amount of data commonly available in a streaming fashion. Online learning plays a vital role in classification which paves way for adaptive learning algorithm. Online multi-class imbalanced learning is a virgin problem that coalesces to the disputes of both online learning and multi-class imbalance learning. Sometimes the data streams for many applications like disease diagnosis, fraud detection, etc., may result in skewed class distribution. Many works in the literature are focused on online class imbalance learning with two-class problem. In this paper, a novel approach of EMOOB and EMUOB algorithms was proposed that overcome multi-class imbalance problem, where data comes in an online fashion. A comprehensive analysis was made in terms of class imbalance status, data distributions, imbalance rate, and in the performance of classifier. Based on the comprehension gained, two new ensemble methods EMOOB and EMUOB with adaptive weights are proposed in the name of WEOB.

**Keywords** Multi-class imbalance · Online learning · Oversampling  
Under sampling · Class distribution

## 1 Introduction

Smart home, healthcare technologies usages are increasing each and every day. Online multi-class imbalance learning [1, 2] is an evolving technique that acquires the attention by so many researchers. It intends to take on the combined issue of online learning and multi-class imbalance learning. Multi-class imbalanced dataset

---

K. Vasantha Kokilam (✉) · D. Ponmary Pushpa Latha  
Karunya University, Coimbatore, India  
e-mail: t.kokilam@gmail.com

D. Ponmary Pushpa Latha  
e-mail: ponmarymca@gmail.com

significantly represents unequal distribution among more than two classes ( $N > 2$ ). Where  $N$  represents a number of class labels in the dataset. Learning the incoming data can be done in two ways. One is batch learning where data are stored and processed chunk-by-chunk. Another way is online learning where data are arrived in online fashion and are processed one by one by applying incremental learning method [3] without storing and reprocessing the observed samples.

Many real-world problems [4] contain multi-class labels like spam filtering in email communication, fault diagnosis in computer monitoring system, Healthcare applications using body worn sensor. In all these application, data often arrive over time in a stream like fashion.

- Each tuple/sample in the Dataset is assumed to belong to a predefined class which is determined by the class label attribute. Let us consider Dataset =  $[x(n), c(n)]$ , where  $x(n)$  represents attribute values from  $(1 \dots n)$  and  $c$  represents a corresponding class label for which the particular sample or record belongs.  $c$  takes the values from  $(1 \dots n)$ . Class Labels can be categorized as single ( $c = 1$ ) or one class, binary ( $c \leq 2$ ) or multi-class ( $c > 2$ ). In most of the databasesm,  $c \leq 2$ . For example, class labels in breast cancer UCI Dataset either belongs to Benign (Positive Sample) or Malignant (Negative Sample) called as as binary class problem. In case of spam filtering dataset various spams like content spam ( $c = 1$ ), link spam ( $c = 2$ ), cloaking spam ( $c = 3$ ), and combined spam ( $c = 4$ ) has to be detected and the number of records in each category also vary this is classic example of multi-class problem.
- Class imbalance learning is one type of classification problems, where some classes are highly underrepresented or sometimes overrepresented compared to other classes. Few skewed distribution makes many conventional machine-learning algorithms as less effective, especially in predicting minority class examples. Imbalance refers drastic difference in the number of records between major and minor classes. For example, various techniques have been proposed to solve the problems associated with class imbalance. The modern years brought raised attention in applying machine-learning methods to the complicated real-world problems, most of which are classified through imbalanced data. Thus, the class imbalance problem is an important topic and investigated nowadays through data mining approaches.
- In case of multi-class, two primitive types of multi-class occur: one minority and multiple majority class, one majority and multiple minority class. This scenario leads to even more complications in classifying the incoming sample with good performance.

## 2 Background Study

Various methods have been proposed to hostage the binary class imbalance issue. In the existence of multi-class Imbalance, more number of problem arises even critical than the binary problem. A brief study of various algorithms for multi-class imbalance is listed.

## 2.1 *Class Decomposition of Multi-class*

Most existing solutions for multi-class imbalance problems use class decomposition schemes to handle multi-class and work with two-class imbalance techniques to handle each imbalanced binary sub-task. Normally multi-class imbalance problem can be resolved using three major ways they are OVO (One vs. One), OVA (One vs. Rest), ECOC (Error Correcting Output Codes).

### 2.1.1 One-versus-One

One-versus-one (OVO), splitting the multi-class problem into many numbers of two-class problem strategy is used. In this strategy, the multi-class problem is simplified into sets of binary class problem and the traditional algorithms are applied on the dataset, in order to perform Classification.

### 2.1.2 One-Versus-Rest

The one-versus-rest (OvR) or one-versus-all, (OvA) strategy involves training a single classifier per class, with the samples of that class as positive samples and all other samples will be considered as negatives. Decisions are taken based on the base classifiers confidence score rather than just a class label.

For example, Let us consider 4 classes namely  $C_1$ ,  $C_2$ ,  $C_3$ , and  $C_4$ . If we have 4 classes,  $C_1$  will be considered as majority and  $C_2$ ,  $C_3$ ,  $C_4$  will be a minority and based on this recall value is calculated. Similarly in the second iteration,  $C_2$  will be considered as majority and  $C_3$ ,  $C_4$  will be minority for which recall value is derived. In the third iteration  $C_3$  is majority class and  $C_4$  will be minority for which recall value is calculated.

Suppose  $X_t$  is the sample arrived during the time  $t$ . All the models perform their iteration and results are calculated in terms of recall values. The perfect iteration which gives positive recall value will be considered. The number of combinations will be calculated using the formula  $N(N-1)/2$ . In case of 4 class problem, the number of iterations will be 6 like  $C_1:C_2$ ,  $C_1:C_3$ ,  $C_1:C_4$ ,  $C_2:C_3$ ,  $C_2:C_4$ ,  $C_3:C_4$ .

### 2.1.3 ECOC

ECOC (Error Correcting Output Codes) works in matrix format. Class labels are stored in the form of Matrix  $M$ . Multi-class prediction is made by finding the codeword closest in Hamming distance to this sequence of binary predictions on the test example.

## 2.2 *Sampling Methods*

Sampling methods can be classified as undersampling and oversampling. In undersampling, if the number of majority class is huge when compared to that of minority class. The process of removing samples from majority class is called as under sampling. Over sampling generates artificial data to increase the minority samples. Advanced sampling is another technique in which boosting effectively alters the distribution of the training data.

Random undersampling was proposed to deal with imbalanced data streams. The majority class examples have a lower probability to be selected for training. It is based on the assumption that before learning, a training set is initialized with any required classification model and the imbalance rate does not undergo change over time and the majority/minority class information is known.

Undersampling and oversampling were proposed to deal with class imbalance in online learning. Based on the results of parameter on Online Bagging's [5] sampling rate it leads to either oversampling or undersampling. However, the sampling parameters cannot be adjusted to change imbalance rates, as they are pre-set prior to learning.

## 2.3 *Cost-Sensitive Algorithms*

Cost-sensitive ensemble algorithm [6] was proposed which covers multi-class imbalance directly without using class decomposition. The advantages of this algorithm are used to find an appropriate cost matrix with multiple classes and it also provides the appropriate cost for an algorithm. Under this algorithm, optimal setup cost setup, classwise, was arrived by the application of a Genetic Algorithm (GA). With reference to the training objective of the given problem, the fitness is tested on G-mean and F-measure. Integration of the cost vector so obtained was done into M1 algorithm which is a cost-sensitive version of the Adboost [7], which has the capability to process multi-class data sets and is denoted by AdaC2 [8]. Since the process of cost vector search is very time consuming, as is the feature of the Genetic Algorithm, there arises a need for a new model to deal with efficiency and effectiveness, on multi-class imbalance problems.

As an advancement, MuSeRA [9] and REA [10] algorithms were proposed for data processing in chunks/batches for imbalanced data streams (Batch learning). The SERA (Selectively Recursive Approach) [11] adopted a balancing approach by absorption of minority samples, on a selective basis, from the previous chunk, into the current training. The selection of minority examples was done based on a similarity measure. In this framework, predictions are done by single hypothesis based on current training chunk.

A state-of-the-art algorithm called Learn++NIE [12] was proposed. Earlier, a batch-based learning algorithm namely Learn++CDS (Concept Drift with SMOTE)

[13, 14] was proposed for NSE (Non-Stationary Environment). Under Learn++, along with the ensemble, a classifier is trained and added, with each incoming chunk of data and is an ensemble learning framework. Generally, the minority class samples are rare in imbalance datasets and incremental learning methods have to discard once samples are learnt, which is known as meeting of single-pass requirement but in the case of Learn++NIE, this process is violated, as it does not remove old classifiers from the ensemble and hence in NSE [15], it works well.

As a recent development, perceptron-based approach has been explored, different misclassification costs to classes are assigned for this approach, which helps in the adjustment of the weights between perceptrons. Two models RLSACP [16] and WOS-ELM [17] were proposed by different researchers. If any error committed in the minority class would result in huge impact and the problem results in higher cost.

Both the above-mentioned models were tested in static scenarios and a fixed imbalance rate was maintained and both were effective; RLSACP adopts a window-based strategy to update misclassification costs at a predefined speed. It is based on the number of examples in each class. WOS-ELM requires a validation set to adjust misclassification costs and it is based on classification performance.

## ***2.4 Multi-class Imbalance Problem in Online Learning***

VWOS-ELM algorithm [18] was proposed to solve class imbalance problems in multi-class data streams. Weighted online sequential extreme learning is a perceptron-based algorithm which supports both chunk-by-chunk and one-by-one method. VWOS-ELM is formed by multiple WOS-ELM base classifiers. Before initiating the sequential learning, it requires a dataset initialization. Based on the incoming data set, the class weights are maintained to handle class imbalance.

CBCE (Class-Based Ensemble Method) [19] works on One-against-all class decomposition technique which is used for handling multiple classes. Under sampling is applied to overcome class imbalance induced by the class evolution. Though CBCE resolves multi-class problem with the best accuracy, usage of class decomposition looks to be drawback.

## **3 Data Description and Experiment**

This paper identifies the enhancement of multi-class [20, 21] imbalance problem. Basically, in very rare cases, datasets of multi-class are balanced. If it is balanced there is no need of under sampling and over sampling. It can directly pass the classification.

The following concepts are discussed in this paper.

First, resampling setting strategies in MOOB and MUOB is enhanced. Then the focus is on the performance of EMOOB (Enhanced Multi-class Oversampling based online bagging) and EMUOB (Enhanced Multi-class Undersampling based online bagging) on dynamic data streams, based on current imbalance status.

Second, Imbalance ratio makes a gap in fulfilling the expected recall value. In dynamic cases, any type of imbalance ratio is possible. In this paper, 5, 10, 20, 30 imbalance ratio results alone is discussed.

Third, Based on the achieved results, for better accuracy and robustness under dynamic scenarios, two ensemble strategies are proposed that maintain both EMOOB and EMUOB with adaptive weight adjustment, called WMOB (Weighted Multi-class Online Bagging).

### 3.1 *EMUOB and EMOOB*

#### 3.1.1 Data

Two multi-class datasets like primary lymphography datasets, heart disease Cleveland datasets are taken with different imbalance rate simulated as data streams. Multi-minority and multi-majority class play a vital role in classification.

In IR (imbalance rate), the probability of occurrence of the minority class, is a lead factor that affects any classification accuracy performance in data streams. Throughout the online processing, it is impossible to get the altogether picture of data it arrives in a streaming fashion.

The number of occurrence of minority class and data sequence becomes a key factor to a greater extent in online learning than in offline learning. Class data distributions may be a major component in causing the abasement of classification performance.

Four major types of data distributions in the minority class include safe, outliers, borderline, and rare examples [21, 22] (Table 1).

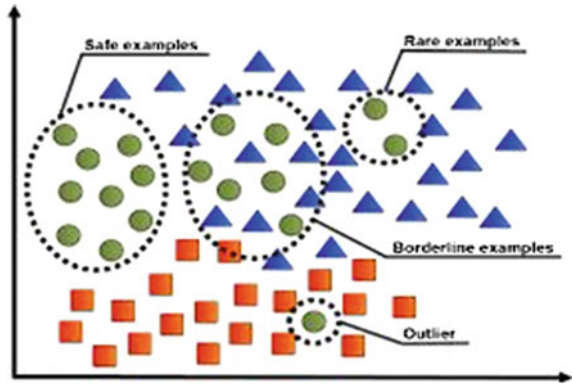
Borderline, rare and outlier data sets were found to be the real source of difficulties in real-world data sets, which are the major research issues in online applications which will be done in future (Fig. 1).

Binary Class have only one minority class and one majority class whereas multi-class have more than one minority or majority class it is denoted as

**Table 1** Data distributions in dataset

Safe	Data that are located in the homogenous regions populated by the example from one class only
Borderline	Data which scattered in the boundary regions between classes, where the examples from both classes overlap
Rare and outliers	Rare and outliers are singular examples located deeper in the region dominated by the majority class

**Fig. 1** Different types of class distribution



multi-majority or multi-minority classes. Suppose let us consider the scenario where even though there are multi-classes, if the data arrived at each class is balanced. Then, there is room for any imbalanced concepts like oversampling and under sampling. But if the number of class in multi-class is imbalanced in course of multi-majority and multi-minority then sampling strategy has to be adopted. EMOOB and EMUOB are running the resampling method even the data is balanced. So there is a need to find out the current Imbalance ratio of multi-class.  $\lambda = 1$  denotes the data stream becomes balanced and if it is not be equal to 1 the class imbalance detection method has to be applied. In order to avoid the event of routinely running the resampling even if the data is balanced  $\lambda$  has been introduced. An ensemble consist of  $M$  base learners, current training example  $(x_t, y_t)$  and the current class size  $w_1^{(t)}, w_2^{(t)} \dots w_n^{(t)}$ . The class size ratio is determined.

### 3.1.2 EMOOB and EMUOB algorithm for multi-class

**Input:** an ensemble with  $M$  base learners, current training example  $(x_t, y_t)$  and current class size  $w_1^{(t)}, w_2^{(t)} \dots w_n^{(t)}$

For each base learner  $f_m (m = 1, 2, \dots, M)$  do

If  $y_t = i$  and  $\begin{cases} w_i^{(t)} < w_{i+n}^{(t)} & \text{for EMOOB} \\ w_i^{(t)} > w_{i+n}^{(t)} & \text{for EMUOB} \end{cases}$

Set  $K \sim \text{Poisson} ((w_{\downarrow}(i+n)^{\uparrow}((t)))/(w_{\downarrow}i^{\uparrow}((t))))$

Else if  $y_t = i + n$  and

$$\begin{cases} w_{i+n}^{(t)} < w_i^{(t)} & \text{for EMOOB} \\ w_{i+n}^{(t)} > w_i^{(t)} & \text{for EMUOB} \end{cases}$$



```

Set  $K \sim \text{Poisson} ((w_{\downarrow}(i))^{\uparrow}((t)))/(w_{\downarrow}(i+n))^{\uparrow}((t)))$ 
Else
Set  $K \sim \text{Poisson} (1)$ 
Else if
Update  $f_m$   $K$  times
End for

```

### 3.2 Imbalance Ratio

Imbalance ratio within classes also impacts the performance of classification algorithm. In non-stationary environment, imbalance ratio cannot be predicted. The impact of Imbalance ratio is discussed with the consideration of both smaller IR value.

### 3.3 WMOB—Weighted Multi-class Online Bagging

By calculating recall of all classes, we can obtain current G-mean. It is used to determine the weights of EMOOB and EMUOB. Their weighted ensemble, denoted as WMOB, is expected to be both accurate and robust in dynamic environments, as it adopts the better strategy (OOB or UOB) for different situations. Two weight-adjusting strategies are proposed and compared here. Suppose OOB has G-mean value  $g_o$  and UOB has G-mean value  $g_u$  at the current moment. Let  $\alpha_u$  and  $\alpha_o$  denote the weights of OOB and UOB, respectively.

#### 3.3.1 Weighted Ensemble of OOB and UOB

OOB G-mean =  $g_o$   
 UOB G-mean =  $g_u$

$\alpha_u$  and  $\alpha_o$  denotes weights of UOB and OOB  
 Normalized G-mean values are used to update the weight

$$\alpha_{oi} = \frac{g_{oi}}{g_{oi} + g_{ui+n}} \quad (1)$$

$$\alpha_{ui} = \frac{g_{ui+n}}{g_{oi} + g_{ui+n}} \quad (2)$$

$$\alpha_o = \frac{1}{n} \sum \alpha_{oi} \tag{3}$$

$$\alpha_u = \frac{1}{n} \sum \alpha_{ui} \tag{4}$$

In the case of multiple class, the weights can be taken from any values (1 ... n). In other words, the final prediction will solely depend on the online model with the higher G-mean. Let’s denote this method as WMOB. From the statistical point of view, combining the outputs of several classifiers by averaging can reduce the risk of an unfortunate selection of poorly performing classifiers and thus provide stable and accurate performance, although the averaging may or may not beat the performance of the best classifiers in the ensemble.

### 4 Results and Discussion

Our measures of WMOB and EMOOB and EMUOB were analyzed and both gave improved results in terms of Accuracy, Precision, F-measure, and Recall. Among them, WMOB provides superior results than EMOOB and EMUOB, which is presented in Tables 2 and 3.

Further extension of work can be focused on data distribution and concept drift in case of multi-class problem.

**Table 2** Comparison of Accuracy and precision with EMOOB and EMUOB with WMOB

Dataset	Imbalance Ratio	Accuracy	Accuracy	Precision	Precision
		EMUOB and EMOOB	WMOB	EMOOB and EMUOB	WMOB
Heart disease cleveland datasets	0	85.9320	86.9822	0.8345	0.8531
	5	78.8000	86.5306	0.8060	0.8657
	10	77.1053	82.5858	0.7238	0.7830
	20	76.6667	83.1703	0.6741	0.7337
	30	77.0671	83.5165	0.6287	0.6823
Lymphography dataset	0	75.7650	81.9632	0.4597	0.5020
	5	76.4059	81.6626	0.4690	0.5386
	10	76.1614	81.5403	0.4525	0.4969
	20	76.1322	81.0513	0.4364	0.4848
	30	76.4706	80.3178	0.4434	0.5027

**Table 3** Comparison of Recall and F-measure with EMOOB and EMUOB with WMOB

Dataset	Imbalance	Recall	Recall	F-measure	F-measure
	Ratio	EMUOB and EMOOB	WMOB	EMUOB and EMOOB	WMOB
Heart disease cleveland datasets	0	0.8399	0.8575	0.8372	0.8531
	5	0.8053	0.9114	0.8056	0.8709
	10	0.7836	0.8628	0.7525	0.8210
	20	0.7912	0.8652	0.7279	0.7941
	30	0.7759	0.8741	0.6946	0.7664
Lymphography dataset	0	0.7994	0.8575	0.5837	0.6333
	5	0.8137	0.9114	0.5950	0.6771
	10	0.7855	0.8067	0.5742	0.6150
	20	0.6766	0.7694	0.5306	0.5948
	30	0.7266	0.8414	0.5507	0.6294

## 5 Conclusion

In this paper, we developed a method called EMOOB and EMUOB to solve multi-class classification problems by constructing a multi-class classifier using the online learning with classification algorithms. In this new method, we have developed WMOB which is a weighted ensemble of OOB and UOB, using G-mean value, resulting in better accuracy. The calculations also factor in different class imbalance rates for different class distribution. Hence, the results are much effective and assure higher level of accuracy, which is important in critical decision making. Since online data streaming is used for inputs, it provides real-time solutions.

**Acknowledgements** I hereby thank and acknowledge the support and guidance received from my guide Dr. D. Ponmary Pushpa Latha., Associate Professor, Karunya University, Coimbatore.

## References

1. Zhang, Z., Krawczyk, B., Garcia, S., Rosales-Pérez, A., Herrera, F.: Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. *Knowledge-Based Systems*. 106, 251–263 (2016).
2. Wang, S., Minku, L., Yao, X.: online class imbalance learning and its applications in fault detection. *International Journal of Computational Intelligence and Applications*. 12, 1340001 (2013).
3. Wang, S., Minku, L.L, Yao, X.: A learning framework for online class imbalance learning. in *Proc. IEEE Symp. Comput.Intell. Ensemble Learn.*36–45 (2013).
4. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. *Intell. Data Anal.* pp. 429–449. (2002).
5. Breiman, L.: Bagging predictors. *Machine Learning*. 24. 123–140, (1996).

6. Cao, P., Li, B., Zhao, D., Zaiane, O.: A novel cost sensitive neural network ensemble for multiclass imbalance data learning. In *The 2013 International Joint Conference on Neural Networks*. pp. 1–8 (2013).
7. Sun, Y., Wong A.K., and Wang, Y.: Parameter inference of cost-sensitive boosting algorithms. in *Proc. 4th Int. Conf. Mach. Learn. Data Mining Pattern Recognit.*, pp. 21–30 (2005).
8. Sun, Y., Kamel, K.S., Wong, A.K., and Wang, Y.: Cost-sensitive boosting for classification of imbalanced data. *Pattern Recognit.* pp. 3358–3378 (2007).
9. Chen, S., He, H., Li, K. and Desai.: Musera: Multiple selectively recursive approach towards imbalanced stream data mining. In *Neural Networks (IJCNN)*, 1–8 (2010).
10. Chen, S., He, H.: Sera: selectively recursive approach towards nonstationary imbalanced stream data mining. In *Neural Networks. IJCNN. International Joint Conference*, 522–529 (2009).
11. Ryan Hoens, T., Chawla.: Learning in non-stationary environments with class imbalance. In *18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp 168–176 (2012).
12. Hoens, T., Polikar, R., Chawla, N.V.,: Learning from streaming data with concept drift and imbalance: an overview. *Progress in Artificial Intelligence*, 1, pp. 89–101 (2012).
13. Ditzler, G. and Polikar, R.: An ensemble based incremental learning framework for concept drift and class imbalance. In *Neural Networks (IJCNN), The 2010 International Joint Conference*. pp. 1–8 (2010).
14. Ditzler, G., Polikar, R.: Incremental learning of concept drift from streaming imbalanced data. *IEEE transactions on knowledge and data engineering*, pp. 2283–2301 (2013).
15. Ditzler, G., Roveri, M., Alippi, C. and Polikar, R.: Learning in nonstationary environments. A survey. *IEEE Computational Intelligence Magazine*, pp. 12–25 (2013).
16. Ghazikhani, A., Monsefi, R. and Yazdi, H.S.: Recursive least square perceptron model for non-stationary and imbalanced data stream classification. *Evolving Systems*, 4(2), pp. 119–131 (2013).
17. Mirza, B., Lin, Z., Toh, K.: Weighted Online Sequential Extreme Learning Machine for Class Imbalance Learning. *Neural Processing Letters*. 38, 465–486 (2013).
18. Mirza, B., Lin, Z., Cao, J., Lai, X.: Voting based weighted online sequential extreme learning machine for imbalance multi-class classification. *IEEE International Symposium on Circuits and Systems (ISCAS)*. pp. 565–568 (2015).
19. Sun, Y., Tang, K., Wang, S., Yao, X., Minku, L.: Online ensemble learning of data streams with gradually evolved classes. *IEEE Transaction on Knowledge and Data Engineering* (2016).
20. Hulse, J.V., Khoshgoftaar, T.M., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In *Proceedings of the 24th international conference on Machine learning*. pp. 935–942 (2007).
21. Wang, S., Yao, X.: Multi-class imbalance problems: Analysis and potential solutions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*. 42, 1119–1130 (2012).
22. Wang, S., Minku, L., Yao, X.: Resampling-Based Ensemble Methods for Online Class Imbalance Learning. *IEEE Transactions on Knowledge and Data Engineering*. 27, 1356–1368 (2015).

# Opinion Content Extraction from Web Pages Using Embedded Semantic Term Tree Kernels

Veerappa B. Pagi and Ramesh S. Wadawadagi

**Abstract** Rapid proliferation of user-generated content (UGC) published over the Web in the form of natural language has made the task of automatic Information Extraction (IE) a challenging issue. Despite numerous models proposed in the literature to address Web IE issues, still there is a growing demand for researchers to develop novel techniques to cope up with new challenges. In this paper, an approach to extract opinion content from Web pages using Embedded Semantic Term Tree Kernels (ESTTK) is addressed. In traditional tree kernels, the similarity of any two given production rules is determined based on exact string comparison between the peer nodes in the rules. However, semantically identical tree fragments are forbidden, even they can contribute to the similarity of two trees. A mechanism needs to be addressed, which accounts for the similarity of nodes with different vocabulary and phrases holding knowledge that are relatively analogous. Hence, the primitive tree kernel function is reconstructed to obtain the similarity of nodes by searching keywords in opinion lexicon embedded as vectors. Experimental results reveal that ESTTK results in better prediction performance compared to the conventional tree kernels.

**Keywords** Opinion extraction · DOM tree · Tree kernels · Classification  
Sentiment analysis

---

V.B. Pagi · R.S. Wadawadagi (✉)  
Department of Computer Science and Engineering,  
Basaveshwar Engineering College, Bagalkot, India  
e-mail: rswlib@yahoo.co.in

V.B. Pagi  
e-mail: veereshpagi@gmail.com

## 1 Introduction

Evolution of Web 2.0 and related technologies have democratized the process of Web content creation allowing Web users to become not only publishers but also distributors of content [1–3]. Using social media, people can relate to each other by their interests, demands, and preferences to form the social networks [4]. Enormous amount of Web pages are being generated in the form of structured, unstructured, and semi-structured format. Furthermore, the Web page contents are extremely heterogeneous: they have different content types (text, images, audio, video, etc.), different representation formats (html, json, rdf, xml, etc.), different authors, different authoritativeness, and different degrees of reliability [5]. Besides these properties, the temporal patterns of Web content changes constantly [6]. The heterogeneous, voluminous, and dynamic nature of the Web always triggers a demand for construction of new extraction systems. In this context, many IE systems use application specific APIs to extract information from Web pages. However, these approaches are not practicable when APIs usage is limited only for specific sources of information need to be examined [7]. On the other hand, techniques based on Web page segmentation that divides a Web page into multiple semantically coherent parts are also addressed. But, these systems fail to explore the diversity of the datasets examined, as they are evaluated using Web pages from specific Websites. Similarly, systems based on Natural Language Processing (NLP) techniques require writing specialized rules for each of the Web domain. Though hand tailored content extractors are assumed to be more precise, generalized automated extractors are often found sufficient and less laborious.

In this paper, a hybrid model that combines features of Web page segmentation and NLP is proposed. The problem associated with utilization of Web page segmentation can be effectively addressed by analyzing the structure of Document Object Model (DOM) of an HTML page by learning the semantics and density of tag structures [8, 9]. However, most Web pages are embedded with additional information such as banners, advertisements, duplicate pages, copyright, etc., which are not related to the actual content and are treated as noise. To measure the usefulness of the DOM tree node contents, the readability score associated with the different nodes of DOM tree is determined. Readability score is a normalized value that determines how easy a reader can comprehend a written text. In natural language context, the readability of text depends on the complexity of its vocabulary and syntax involved in representing its content. Some statistical features, such as the total number of words, sentences, and syllables in text and hyper text present at different nodes in the DOM tree are analyzed to determine the significance of the node in presenting the text content. However, the statistical features may vary from one node to another node, the quantities need to be normalized with respect to each node in order to achieve an optimal performance on different styled Web pages. Flesch–Kincaid grade level test [10] is used efficiently in this approach to determine the normalized readability score for each node. Once the readability score is obtained, depending upon the appropriate threshold value, the nodes values are

classified as significant or not significant. Then, the node values labeled as significant are further subjected to classification for determining the node content is opinion content (subjective) or not. The proposed model ESTTK can be used as a binary classifier to predict the presence of opinion material on the given node. Finally, the node values determined to be subjective content are extracted from the nodes for further analysis.

Learning techniques based on kernels present a flexible framework for applying apriori knowledge about the problem domain by making combinations and parameterizing in different ways of the employed kernel function. Especially, Tree Kernels (TK) have been utilized as a fundamental tool for encoding the syntactic structure of the text content in the form of parse trees and have performed better in many natural language applications. A variation of TK known as Subset Tree Kernels (STK), based on counting matching subset trees, has gained greater importance in learning semantically rich information [11]. However, STK strictly considers tree fragments containing complete grammar rules for matching. Yet another class of TK, called Syntax and Semantic Tree Kernels (SSTK) which exploits linguistic structure and previous knowledge of the semantic dependencies of various terms is also proposed [12]. SSTK adopts a semantic smoothing approach to increase similarity matching of tree fragments containing terminal nodes. An improved SSTK, specifically designed to work on text categorization tasks, named Semantic Syntactic Tree Kernels are also developed [13]. These kernels employ a specially designed trees known as an Embedded Semantic Term Tree Kernels (ESTTK) and a leaf weighting component for content matching. These kernels allow partial matches between tree fragments, where a partial match between two subtrees occurs when they differ only by their terminal symbols. To improve the performance of such kernels, in the proposed model the tree kernel function is reconstructed, so that the similarity of two tree fragments is computed based on the synonymy of terms present in the terminal nodes using opinion lexicon embedding as terms vector.

The remainder of this paper is structured as follows. Section 2 provides a literature review on contemporary approaches of Web IE systems. A detailed discussion on the proposed methodology of opinion content extraction from Web pages is presented in Sect. 3. Section 4 gives an account on the performance evaluation of the proposed models conducted as a series of experiments. Finally, Sect. 5 concludes with a summary of the contributions, final remarks and a discussion on future work.

## 2 Literature Review

Web IE has been studied extensively during the past decade, and many research works come about, as well as systems are being produced. A significant part of the work detailed in the literature is primarily contemplated to identify and semantically annotate the segments of the Web pages and fewer studies deals with the problem

of removing non-informative segments. Other systems in the literature are devoted to extract subjective information from the Web pages based on NLP and Machine Learning (ML) concepts.

Zou et al. [14] present a hybrid approach for segmentation of online medical journal articles. The approach follows geometric layout of the Web page to represent Web content as a zone tree structure. For a given journal article, a zone tree is constructed by combining DOM tree analysis and recursive X-Y cut algorithm. Additionally, the approach combines other visual cues such as background color, font size, font color, to segment the page into homogeneous regions. A new approach to Web page segmentation using quantitative linguistics, derived from the area of Computer Vision (CV) is also discussed [15]. In this approach, authors utilize the notion of text-density as a measure to identify the individual text segments of a Web page. This reduces the problem of segmentation into a 1D-partitioning task. Further, in [16] Liu et al. present were: a system that automatically extracts user reviews from Web pages. They employ a different technique called level-weighted tree matching algorithm to compute the similarity between two subtrees. Using this algorithm, noise can be eliminated and the boundaries of the review records are clearly identified. Then, the consistency of each node of review records in the DOM tree is measured. Based on the node consistencies the minimum subtree that contains the pure review content is extracted. A fast, accurate and generic model for extracting informative content from diverse Web pages called content extraction via text-density (CETD) using DOM tree node text-density is presented [17]. The weight of the each nodes is measured using two statistical text-density techniques, namely simple text-density and composite text-density. Further, to extract content intact, a technique called DensitySum is proposed to replace Data Smoothing. A similar kind of technique that operates on DOM tree to evaluate each tree node and associated statistical features such as link-density and text distribution across the node to predict node's weight towards overall content provided by the document is discussed [9]. Many formatting features like fonts, styles and the position of the nodes in the Web page are evaluated to identify the nodes with similar formatting as compared to the significant nodes. This hybrid model is basically derived from two different models, i.e., one is based on statistical features and other based on formatting characteristics and have achieved the best accuracy. An extension to the previous works presented, Lin et al. [18] develop BlockExtractor, a tool that identifies informative content in three steps. First, it looks for blocks containing Block-Level Elements (BLE) and Inline Elements (InE) which are designed to roughly segment pages into blocks. Second, the densities of each BLE and InE blocks is computed to eliminate noise. Finally, all the redundant BLE and InE blocks that have emerged in other pages from the same site are removed. In contrast with above techniques, Lopez et al. [19] present a novel technique for content extraction that uses the DOM tree to analyze the hierarchical relations exists among various elements in the Web page. The technique follows the notion of Chars-Nodes Ratio (CNR), which shows the relation between text content and tags content of each node in the DOM tree. Further, the approach is formalized to compute the CNR for each node in the DOM tree and those nodes with a higher



CNR are selected for processing. Yet another method for content extraction similar to [19], a model based on the Words/Leafs Ratio in the DOM Tree is also presented [20]. To improve the mining results, cleaning the Web pages before mining algorithms are applied becomes a critical step. Hence, [21] focuses on identification and removal of local noise present in the Web pages. They propose a novel and simple tree structure called featured DOM Tree for the detection and removal of local noise content.

A sort of techniques based on NLP and ML is also studied in the context of literature review. Ashraf et al. [22] employed a multi-objective genetic algorithm based clustering technique to automatically extract information from Web pages comprising semi-structured data. The system parses and tokenizes the data from an HTML document using domain-specific information provided by the user. Further, the data is partitioned into clusters containing similar elements and estimates an extraction rule based on the pattern of occurrence of data tokens. A framework based on Bayesian learning for solving the wrapper adaptation with new feature selection is presented [23]. The approach is designed to adapt automatically the IE knowledge previously learned from a source Website to a unknown site, on the other hand, discovering previously unseen attributes that are not specified in the learned or adapted wrapper. It is also possible to discover semantic labels for the new attribute sets discovered. In traditional Web IE, the main focus is on processing static documents and are difficult to reflect the dynamic content on the Web. To address this challenge, Peng et al. [24] propose a new methodology based on shallow parsing combined with a set of rules defined. The rules are generated according to the syntactical features of English grammar, such as the tense of verbs, the usage of modal verbs and so forth. A novel method for Web IE from a set of knowledge base to answer user consultations using natural language is addressed [25]. The system uses a Fuzzy Logic engine, which takes advantage of its flexibility for managing sets of accumulated knowledge. Another important contribution of this paper deals with automatic term weighting for Vector Space Model (VSM) using Fuzzy Logic-based term weighting. This method substitutes the classical TF-IDF term weighting scheme with its flexible model. Yang and Cardie [26] formulate opinion expression extraction as a segmentation problem. To capture user interactions, they present joint learning approach which combines opinion segment detection and attribute labeling into a single probabilistic model, and estimate parameters for this joint model to extract opinion expression. An unsupervised learning framework for extracting features of popular products from product description pages originated from different E-commerce Web sites is addressed [27]. They develop a discriminative graphical model based on Hidden Conditional Random Fields. Unlike existing IE methods that do not consider the popularity of product attributes, this model is able to not only identify popular product features from a collection of customer reviews but also map these popular features to the related product attributes. This model can also bridge the vocabulary gap between the text in product description pages and the text in customer reviews. It is evident from the above literature study that reasonable amount of research is reported on Web IE systems both form Web page segmentation and ML domains.

### 3 Proposed Methodology

Figure 1 illustrates the experimental setup of the proposed methodology for Web information extraction. The process involves three stages: (i) The algorithm exploits informative and non-informative characteristics of a given Web page encapsulated in DOM tree nodes by recursive traversal and analysis for readability test. (ii) Those tree nodes that are qualified to be informative are further examined for subjectivity detection using ESTTK. (iii) Finally, the content returned to be subjective material comprising opinions are directly extracted from the nodes for further analysis.

#### 3.1 DOM Tree as an HTML Document Model

Document Object Model [28] is a standardized, platform and language independent interface for accessing and updating content, structure, and style of HTML documents. Each HTML page is represented as a DOM tree with tags corresponds to internal nodes and the content as leaf nodes. Further, every node is described by various characteristics such as parent node, child nodes, tag, style, cardinality, class, id, and so forth. Example 1 demonstrates a segment of HTML code and Fig. 2 is its corresponding DOM tree.

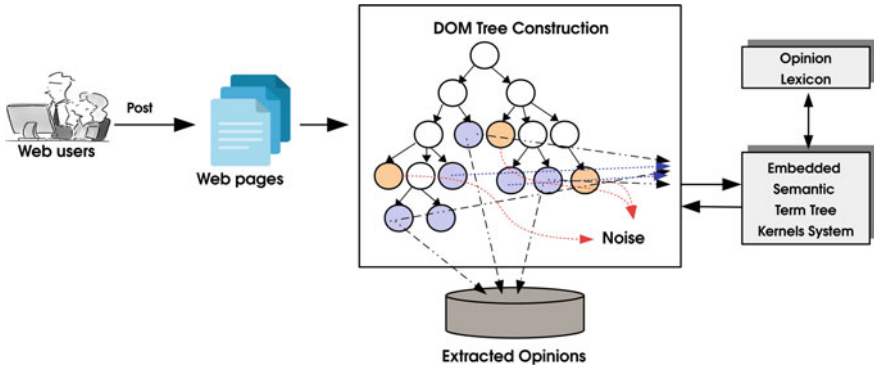
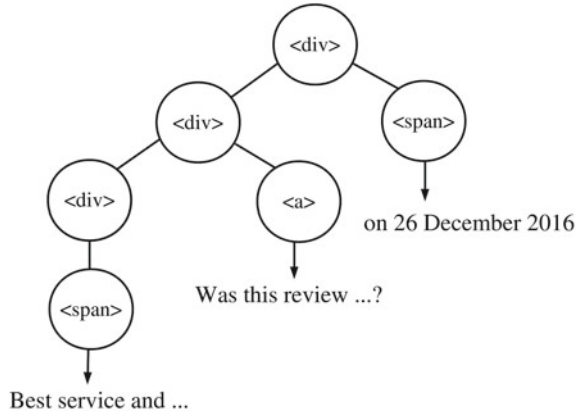


Fig. 1 Proposed methodology for web information extraction system

**Fig. 2** The DOM tree of Example 1



*Example 1* A brief segment of HTML code from review sites

```

1. <div id="R2IC8VIK0J2DWS" data-hook="review" class="a-section review">
2.   <div class="top-small review-comments">
3.     <div class="a-row review-data">
4.       <span class="review-text">
5.         Best service and timely delivery by amazon, good packaging
6.         and excellent device
7.       </span>
8.     </div>
9.     <a> Was this review helpful to you?</a>
10.  </div>
11. <span> on 26 December 2016 </span>
12. </div>

```

### 3.2 Node Density by Readability Score

As mentioned earlier, the method processes Web pages and segments it into informative blocks while it ignores noisy content. We propose a new technique called readability test to discard the non-informative content. Readability score is to determine the ease with which a reader can understand a written text. Flesch–Kincaid grade level test [10] is used to determine the normalized readability score. The grade level is calculated with the following formula:

$$S = 0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.59. \quad (1)$$

The result gives a score ‘ $S$ ’ as a U.S grade level, making it easier for teachers, parents, librarians, and others to judge the readability level of various books and texts. Lower the score, better is the readability of the text. For example, the readability score for the product review quoted in Example 1 is obtained as 12.3. However, the example sentence is too short and the score seems to be bit higher than the expected. The objective is to capture more trusted reviews from the sites, the threshold value ‘ $t$ ’ is set to 10.0. Meaning, the content possessing readability score ‘ $S \leq 10.0$ ’ are considered as informative content otherwise non-informative. The qualified informative contents are further analyzed by ESTTK system for subjectivity detection.

### 3.3 *Embedded Semantic Term Tree Kernels*

The key idea of using tree kernels is to determine the number of common sub-structures between two trees  $t_1$  and  $t_2$  without explicitly considering the whole fragment space. A general formula for such a tree kernel between two trees  $t_1$  and  $t_2$  can be determined using Haussler’s formulation [30]

$$k(t_1, t_2) = \sum_{f \in F} w(f) c_1(f) c_2(f), \quad (2)$$

where  $F$  is the set of all tree fragments and different concepts of tree fragments define different tree kernels,  $c_1(f)$  and  $c_2(f)$  return the counts for fragment  $f$  in trees  $t_1$  and  $t_2$ , respectively, and  $w(f)$  assigns a weight to fragment  $f$ . More technically, the kernel is a weighted dot product over vectors of fragment counts.

In this paper, we focus only on Semantic Syntactic Tree Kernels (SSTK), first introduced by Moschitti et al. [13]. These kernels are extended tree kernels with embedded semantic terms and a leaf-density component. They grant partial matching between tree fragments, where a partial match between two subtrees occurs only when they differ by their terminal symbols. The partial match between terminal nodes is performed according to a predefined kernel  $k_s$ . The tree fragment kernel is defined as:

**Algorithm 1** OpinionExtraction( $T$ )**Data:** A DOM Tree  $T$  of a Web page  $W$ **Result:** Extracts Opinion Content from a given Web page  $W$ 


---

```

1: Flag = 0
2: if  $T == Null$  then
3:   return -1
4: end if
5: if  $T.LeafNode() == True$  then
6:    $S = FleschKincaidGradeScore(T.Value)$ 
7:   if  $S \leq 10.0$  then
8:      $Flag = EmbeddedSemanticTermTreeKernel(T.Value)$ 
9:   end if
10:  if  $Flag == 1$  then
11:    return  $ExtractContent(T)$ 
12:  end if
13: else
14:   $OpinionExtraction(L)$ 
15:   $OpinionExtraction(R)$ 
16: end if

```

---

$$k(t_1, t_2) = \sum_{n_1 \in N_1} \sum_{n_2 \in N_2} k_s \Delta(n_1, n_2), \quad (3)$$

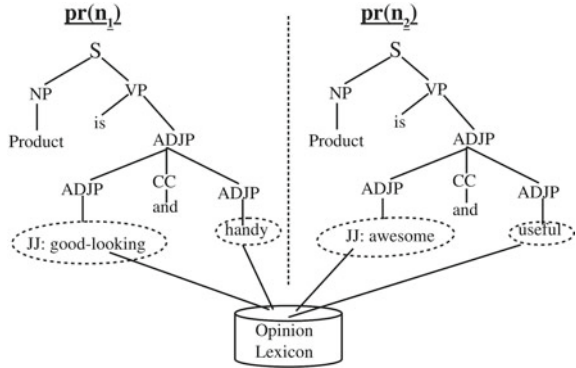
where  $\Delta$  calculates the similarity between every two nodes in the tree as follows:

$$\Delta(n_1, n_2) = \begin{cases} 0 & \text{if } pr(n_1) \neq pr(n_2) \\ 1 & \text{if } pr(n_1) = pr(n_2), n_1 \text{ and } n_2 \text{ are preterminals} \\ \prod_{i=1}^{nT(pr(n_1))} k_s(C^i(n_1), C^i(n_2)) & \text{otherwise} \end{cases} \quad (4)$$

where  $pr(n_1)$  and  $pr(n_2)$  are the grammar production rules at node  $n_1$  and  $n_2$ , respectively,  $C^i(n)$  is the  $i$ -th child of node  $n$ . In Eq. 4, the  $\Delta$  computes the similarity of two production rules based on exact string matching between the peer nodes in the rules. However, semantically identical but not exactly similar tree fragments are ignored even though they can contribute to the similarity of two trees. Thus, the traditional tree kernel is reconstructed  $\Delta$  to determine the similarity of two production rules by identifying the synonymy of terms present at peer nodes in opinion lexicon embedded as a vector. Formally, it is defined as follows:

$$\Delta(n_1, n_2) = \begin{cases} 0 & \text{if } pr(n_1) \neq pr(n_2) \text{ or } \prod_{i=1}^{nT(pr(n_1))} t(pr(n_1)_i), pr(n_2)_i) = 0 \\ 1 & \text{if } n_1 \text{ and } n_2 \text{ are preterminals and } \prod_{i=1}^{nT(pr(n_1))} t(pr(n_1)_i), pr(n_2)_i) = 1 \\ \prod_{i=1}^{nT(pr(n_1))} (C^i(n_1), C^i(n_2)) & \text{otherwise} \end{cases} \quad (5)$$

**Fig. 3** A ESTTK tree with its similarity matching



where  $pr(n_1)$  and  $pr(n_2)$  are the grammar production rules at node  $n_1$  and  $n_2$ , respectively,  $pr_i(n_1)$  and  $pr_i(n_2)$  are the  $i$ -th peer nodes of the two production rules and  $t$  is a threshold function defined as follows:

$$t(n_1, n_2) = \begin{cases} 0 & \text{if } \text{Synonymity}(V_{n_1}, V_{n_2}) < \theta \\ 1 & \text{if } \text{Synonymity}(V_{n_1}, V_{n_2}) \geq \theta \end{cases} \quad (6)$$

where  $V_{n_1}$  and  $V_{n_2}$  are the word embedding vectors of two input nodes and  $\theta$  is the threshold above which the two nodes are considered similar for the kernel computation, if the percentage portion of the term vectors contained in opinion lexicon. Figure 3 depicts the process of how similarity of two production rules are determined using ESTTK with its tree fragments.

## 4 Experimental Setup

A series of experiments are conducted to demonstrate the effectiveness of the proposed model. For this purpose, three opinion datasets from different domains (i) MPQA Opinion Corpus (ii) Movie Reviews Data Set (MRDS), and (iii) Opinosis Review Dataset (ORD) have been used as the benchmark datasets. The MPQA [29] opinion annotated corpus provides an infrastructure for sentiment annotation that is not provided by other sentiment NLP corpora, and is much more varied in topic, genre, and publication source. However, Movie Reviews Data Set is a collection of movie reviews and Opinosis Review Dataset [30] is a collection of topic related sentences extracted from user reviews. The reviews are obtained from multiple sources - Tripadvisor (hotels), Edmunds.com (cars), and Amazon.com (various electronics). The proposed model is effectively implemented using Python 3.3 and Python Implementation of Tree Kernels (PITK) [31], a set of classes implementing tree kernel functions. Currently PITK includes Subtree Kernel, the Subset Tree Kernel, the Partial Tree Kernel, and the Position Aware Kernels. For better

evaluation of the models, the performance of ESTTK algorithm and the proposed Web IE system are separately evaluated.

#### 4.1 Evaluation of ESTTK as a Binary Classifier

In this subsection, we conduct an experiment for subjectivity detection as a binary classification problem using ESTTK on each of the three datasets. The efficiency of ESTTK algorithm can be analyzed by comparing its performance against two traditional tree kernels STK and SSTK. Standard metrics were used to evaluate and compare the performance of these approaches. In particular, precision, recall, and F1-scores were calculated for comparative analysis of each method. For experimentation, the value of  $\alpha$  is set to 0.6, indicates 60% of the terms present in union vector are also present in opinion lexicon. The numbers of documents for training and testing from each dataset is selected in the ratio of 60:40. Table 1 gives the summary of classification results obtained. The experimental results reveal that the proposed ESTTK outperforms the traditional tree kernels.

#### 4.2 Evaluation of the Overall System for Opinion Extraction

Finally, the performance of the overall system proposed for opinion content extraction from Web pages is evaluated. For evaluation purpose, Web pages from [www.trustedreviews.com](http://www.trustedreviews.com) containing user reviews about budget smartphones are collected. Total count of most trusted reviews contained in Web page is manually annotated for experimentation. The recognition rate of the proposed model is then determined by computing the average classification accuracy of ten non-overlapping trials. The classification results are presented in Table 2.

**Table 1** Performance evaluation of ESTTK compared with other Tree Kernels

Dataset	Tree Kernels	Precision	Recall	F1-Measure
3*MPQAOC	STK	0.7432	0.7621	0.7525
	SSTK	0.7286	0.8139	0.7689
	ESTTK	0.8354	0.8263	0.8308
3*MRDS	STK	0.7136	0.7913	0.7504
	SSTK	0.7348	0.8338	0.7812
	ESTTK	0.8921	0.9090	0.9004
3*ORD	STK	0.7015	0.7116	0.7065
	SSTK	0.8013	0.7968	0.7990
	ESTTK	0.9121	0.8890	0.9004

**Table 2** Classification results of overall opinion extraction system

Trial	Total reviews in a web page	Reviews correctly extracted	Accuracy (%)
1	23	19	82.61
2	25	21	84.00
3	32	26	81.25
4	20	18	90.00
5	26	23	88.46
6	24	20	83.33
7	23	18	78.26
8	21	17	80.95
9	23	21	91.30
10	24	21	87.50
Average	23.5	20	85.05

## 5 Conclusion

In this chapter, a hybrid model that combines DOM tree features and Tree Kernels is presented. It evaluates each DOM tree node and associated readability score to predict the significance of the content in a node. Further, we investigate how syntactic structures of natural language text can be exploited using tree kernels. The power of tree kernels can be extended by reconstructing the primitive kernel functions with a new Embedded Semantic Term Tree Kernels. Then, the ESTTK model can be applied for subjectivity detection as a binary classification problem. The performance of proposed model is evaluated against several opinion datasets. The results show that the proposed model outperforms the conventional tree kernels. In future, it would be interesting to study different models of semantic tree kernels based on embedded lexicons and semantic structures.

## References

1. Carsten Ullrich et al., Why web 2.0 is good for learning and for research: principles and prototypes, Proceedings of the 17th International Conference on World Wide Web, Beijing, China. 705–714 (2008).
2. Bower, M, Deriving a typology of Web 2.0 learning technologies, British Journal of Educational Technology. 47 (4) 763–777 (2016).
3. Jakub Piskorski and Roman Yangarber, Information Extraction: Past, Present and Future, Multi-source, Multilingual Information Extraction and Summarization Part of the series Theory and Applications of Natural Language Processing, Springer, 23–49 (2012).
4. Sebastian A. R., Felipe A., Web Intelligence on the Social Web, Advanced techniques in Web intelligence-1, SCI311, Springer, (2010) 225–249.
5. Ricardo B. Y., Paolo B., Web Structure Mining, Advanced techniques in Web intelligence-1, SCI311, Springer, 113–142 (2010).



6. Maria C. Calzarossa, Daniele T., Modeling and predicting temporal patterns of web content changes, *Journal of Network and Computer Applications*, 56 (3) (2015) 115–123.
7. Pappas, Nikolaos, Katsimpras, Georgios and Stamatatos, Efstathios, Extracting Informative Textual Parts from Web Pages Containing User-Generated Content, *Proceedings of 12th International Conference on Knowledge Management and Knowledge Technologies*, (4) 1–8 (2012).
8. G. Vineel. Web page DOM node characterization and its application to page segmentation. In *Proceedings of the 3rd IEEE International conference on Internet multimedia services architecture and applications, IMSAA'09, NJ, USA*, (2009).
9. Pir Abdul Rasool Qureshi, Nasrullah Memon, Hybrid model of content extraction. *J. Comput. Syst. Sci.* 78(4) 1248–1257 (2012).
10. Kincaid, J.P., Fishburne, R.P., Rogers, R.L., and Chissom, B.S., Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. *Research Branch Report*, 8–75 (1975).
11. M. Collins and N. Duffy, New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *ACL02*, (2002).
12. Stephan Bloehdorn and Alessandro Moschitti, Combined Syntactic and Semantic Kernels for Text Classification, *Advances in Information Retrieval-Proceedings of the 29th European Conference on Information Retrieval, Rome, Italy, Springer LNCS*, (4425) 2–5 (2007).
13. S. Bloehdorn and A. Moschitti, Structure and semantics for expressive text kernels. In *CIKM'07: Proceedings of the 16th ACM Conference on information and knowledge management, New York, NY, USA*, 861–864 (2007).
14. J. Zou, D. Le, and G.R. Thoma, Combining DOM tree and Geometric Layout Analysis for Online Medical Journal Article Segmentation, *Proceeding of 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 119–128 (2006).
15. C. Kohlschutter and W. Nejdl. A Densitometric, Approach to Web Page Segmentation. In *ACM 17th Conference on Information and Knowledge Management*, 1173–1182 (2008).
16. Wei Liu, Hualiang Yan and Jianguo Xiao, Automatically extracting user reviews from forum sites. *Computers and Mathematics with Applications*, 62(7) 2779–2792 (2011).
17. Fei Sun, Dandan Song, and Lejian Liao, DOM based content extraction via text density, *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, (2011).
18. Shuang Lin, Jie Chen, and Zhendong Niu, Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction, *Tsinghua Science and Technology*, 17(3) 256–264 (2012).
19. Sergio LÃ³pez, Josep Silva, and David Insa, Using the DOM Tree for Content Extraction, *WWV* 46–59 (2012).
20. David Insa, Josep Silva, and Salvador Tamarit, Using the words/leafs ratio in the DOM tree for content extraction. *J. Log. Algebr. Program*, 82(8) 311–325 (2013).
21. Shine N. Das, Pramod K. Vijayaraghavan, and Midhun Mathew, Eliminating Noisy Information in Web Pages using featured DOM tree, *International Journal of Applied Information Systems (IJ AIS)*, Foundation of Computer Science FCS, New York, USA 2(2) 2249–0868 (2012).
22. Fatima Ashraf, Tansel Ozyer, Reda Alhaji, Employing Clustering Techniques for Automatic Information Extraction from HTML Documents, *IEEE Transactions on Systems, Man, Cybernetics*, 38(5) 2008.
23. Tak-Lam Wong and Wai Lam, Learning to Adapt Web Information Extraction Knowledge and Discovering New Attributes via a Bayesian Approach, *IEEE Transactions on Knowledge and Data Engineering*, 22(4) 2010.
24. Min Peng, Xiaoxiao Ma, Ye Tian, Ming Yang, Hua Long, Quanchen Lin, Xiaojun Xia, The Web Information Extraction for Update Summarization Based on Shallow Parsing. *3PGCIC* 109–114 (2011).

25. Jorge Ropero, Ariel Gómez, Alejandro Carrasco, and Carlos Lóen, A Fuzzy Logic intelligent agent for Information Extraction: Introducing a new Fuzzy Logic-based term weighting scheme, *Expert Systems with Applications*, 39, 4567–4581 (2012).
26. Bishan Yang, Joint Modeling of Opinion Expression Extraction and Attribute Classification, *Transactions of the Association for Computational Linguistics* 2, 505–516 (2014).
27. Lidong Bing, Tak-Lam Wong, and Wai Lam., Unsupervised Extraction of Popular Product Attributes from Ecommerce Web Sites by Considering Customer Reviews, *ACM Transactions on Internet Technology (TOIT)*. 16(2) 1–17 (2016).
28. W3C document object model. Website, 2009. <http://www.w3.org/DOM>.
29. Lingjia Deng and Janyce Wiebe, MPQA 3.0: Entity/Event-Level Sentiment Corpus. 2015 Conference of the North American Chapter of the Association for Computational Linguistics “Human Language Technologies, Denver, Colorado, USA (2015).
30. [http://text-analytics101.rxnlp.com/2011/07/user-review-datasets\\_20.html](http://text-analytics101.rxnlp.com/2011/07/user-review-datasets_20.html).
31. <http://joedsm.altervista.org/pythontreekernels.htm#svmlight>.

# Combining Fuzzy C-Means and KNN Algorithms in Performance Improvement of Intrusion Detection System

B. Sujata and P. Ravi Kiran Varma

**Abstract** One of the major issues in Intrusion Detection System (IDS) is misclassifications that leads to either false positives or false negatives. From the literature, it was found that among the various categories of IDS datasets, User-to-Root (U2R) attacks and Remote-to-Local (R2L) attacks are the most misclassified categories. Abnormal samples are identified with high accuracy by anomaly detection and normal samples are identified better by misuse detection methods. To reduce the false positives and false negatives, a hybrid two-phase mixture of anomaly and misuse detection are proposed with the assistance of various machine-learning techniques. In the first phase, unsupervised fuzzy C-means clustering (FCM) is used to cluster normal and anomalous data samples. In the second phase, two elements of K nearest neighbor (KNN) are used. One for checking normal and another one for checking abnormal samples. The proposed systems are evaluated using KDD 1999 IDS dataset and also compared with similar works and found to be beneficial.

**Keywords** Intrusion detection system · IDS · Anomaly detection  
Misuse detection · Fuzzy C-means · KNN algorithm

## 1 Introduction

Firewall and Intrusion Detection System (IDS) plays a major role in protecting the network perimeter [1]. Intrusion detection systems act against malicious instances where the confidentiality, integrity, and availability of information resources get compromised. Unauthorized access to a system can be detected with the help of intrusion detection systems [2], whereas firewalls can only filter network traffic

---

B. Sujata · P. Ravi Kiran Varma (✉)  
MVGR College of Engineering (A), Vizianagaram, Andhra Pradesh, India  
e-mail: ravikiranvarmap@gmail.com

B. Sujata  
e-mail: sujata.mvgr@gmail.com

based on a policy [1, 3]. Intrusion detection systems are designed both for computer systems and networks. These particular systems act as network sniffers for monitoring a network in a promiscuous mode. All types of mischievous network traffic and computer utilization that includes network attacks contrary to susceptible services, data focussed attacks on applications, host centered attacks such as privilege escalation, unauthorized logins, and access to subtle files and malware can be proficiently found by these systems. Administrators get alerted by system alerts whenever a rule violation occurs in network packets upon pattern matching with the help of the suitable algorithm. In current approaches, there are two types of intrusion detection systems called misuse or signature-based detection techniques and anomaly-based detection techniques.

Misuse detection or signature-based techniques are proficient in discovering the known attacks by matching the signatures or the attack descriptions in contradiction of the audit data stream. Here the suspicious traffic is classified into four types: Dos attacks, Probe attacks, and U2R, R2L. Abnormal or anonymous attacks that deviate from normal behavior can be detected using the Anomaly-based detection techniques. Hybrid Intrusion detection is a combination of both misuse and anomaly detection which fetches better results than applying individually. Fuzzy c-means clustering algorithm is used in this paper for acquiring better results in identifying the misuse, anomaly, and hybrid detections by the use of various machine-learning techniques.

In this paper, anomaly detection component 1 applies Fuzzy C-means clustering in phase one where Fuzzy C-means (FCM) is a method of clustering that allows one piece of data to fit into two or more clusters. Pattern recognition plays a crucial role here, and minimization of the objective function is the main criteria. It works on the principle of applying the membership functions for identifying whether the instance belongs to which particular cluster. In turn, this procedure helps in the identification of the normal and abnormal instances in phase one. The unusual cases found in phase one are carry forwarded to the phase two for further detection of normal and attack instances with the help of KNN algorithm. In this paper with the support of FCM and KNN algorithms, high detection rate and low false positive rates can be achieved.

## 2 Related Work

Kumar and Spafford [4] proposed a mechanism of pattern matching to identify the known attacks misuse detection techniques with the help of Color Petri Nets and found beneficial in the context of generality, portability, and flexibility. However, its evaluation and implementation were not performed. Cannady [5] proposed misuse detection based on the analytical strength of artificial neural networks that provide a perspective to identify and classify network activity grounded on limited, incomplete, and nonlinear data sources. However, here the system is unable to receive the inputs directly, hence demands further research in this regard. Prakash

et al. [6], proposed misuse detection techniques in contrast to the two previous cases where pattern matching with known attack signatures is done with the help of various data-mining techniques corresponding to classification, clustering, outlier detection, and association rule mining to address the security issues in E-Commerce.

Patcha and Park [7], performed a survey on various anomaly detection techniques and found that known attacks are occurring with low false positive rate upon the usage of statistical anomaly detection, and data-mining techniques. However, the major drawback of the signature detection approach is that such systems typically require a signature to be defined for all possible attacks that an attacker may launch against a network. Teodoro et al. [8], discovered known attacks in an NIDS by applying mechanisms like knowledge-based technologies, machine-learning-based NIDS schemes, Bayesian networks, Markov models, genetic algorithms, clustering and outlier detection, neural networks, and fuzzy logic techniques on KDD 1999 datasets for improving security and protection of networks. However, low detection efficiency, low throughput, and high cost were observed to be the drawbacks. Wang et al. [9], proposed the mechanism of finding unknown attacks or the abnormal behaviors with the help of traffic classification, Anomaly detection, extreme learning machine, Support Vector Machines (SVM), L1-norm minimization and finally proposed that ELM was found advantageous. However, resource requirement was not adequate, and performance was comparatively low.

Depren et al. [10], proposed the combination of misuse and anomaly detection in the form of a hybrid approach for getting better results than applying individually. J.48 decision tree was applied for classifying various attacks in misuse detection component, and self-organizing map structure is proposed for identifying the normal behavior and anomaly detection components. However, the dataset can be further classified and supplied for the two components to obtain better results. Liao et al. [11], reviewed signature-based detection (SD), Anomaly-based detection (AD), and stateful protocol analysis (SPA) and proposed numerous rule-based approaches for detection of unfamiliar attacks. However, this approach could not identify unknown attacks completely. Kim et al. [12], proposed that anomaly detection can be improved when combined with known attacks in the form of hybrid detection with the help of one-class SVM and C 4.5 Decision tree to reduce the false positive and false negatives. However, for significant improvement in the performance of hybrid detection, the data used can be even more divided into subsets.

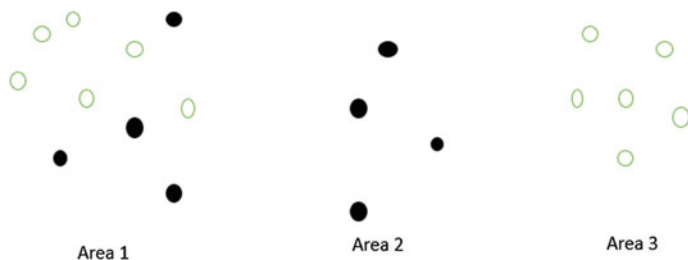
Guo et al. [13] exploited the strengths of misuse detection and anomaly detection, an intensive focus on intrusion detection combines the two. A hybrid approach was proposed towards achieving a high detection rate with a low false positive rate. The authors proposed an anomaly detection component 1 in level one with the help of change of clusters method using K-means clustering algorithm and the outcomes based on normal and attack based instances will be supplied to the both anomaly detection component 2 and misuse detection component in level 2 with the help of KNN algorithm to identify the exact normal behavior and attacks. However,

K-means only works for the instances that are outliers that are the instances which do not belong to any cluster. In the proposed system, it is described that if an instance is present in more than one cluster then there comes, the Fuzzy C-means clustering into the picture which becomes helpful in dealing with the fuzziness of the data.

### 3 Methodology

Proposed hybrid intrusion detection helps in improving the ability to identify known and unknown attacks where this hybrid detection combines both misuse and anomaly detection for achieving high detection rate and low false positive rates. Anomaly detections and misuse detections are performed phase wise to identify the known and unknown attacks much accurately, by applying anomaly detection component 1 in phase one on a dataset and this phase gives the output by dividing some instances as normal instances and some others as abnormal instances (attacks). Now, this output is in turn supplied to phase two where again anomaly detection and misuse detection are performed for further identifying the false positives, and false negatives and are stored in their respective lists. Figure 1 shows a set of sample instances with a combination of both normal and abnormal instances.

In Fig. 1, area 1 consists of a set of instances which consists of both normal and malicious instances. Normal instances are depicted with white color and attack instances are denoted with black color. The attack instances are separated from area 1 and are shown in area 2, and normal instances are separately shown in area 3. Figure 2 describes the hybrid detection mechanism adopted by the proposed hybrid approach. In phase one the anomaly detection component 1 is used to identify the normal and abnormal instances in the given data. These classified normal and abnormal instances are further applied with both anomaly detection component 2, and misuse detection component to further filter the exact normal and attack instances and store them separately.



**Fig. 1** Schema of relation of instances

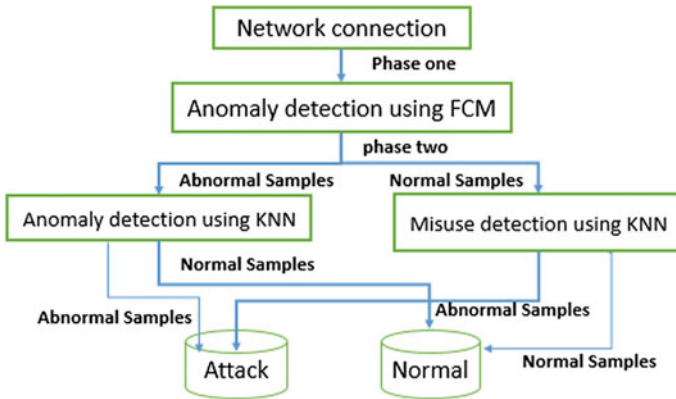


Fig. 2 Overview of detection procedure of proposed hybrid approach

### 3.1 Phase One: Anomaly Detection Component 1

Anomaly detection component 1 is applied on data in phase one by performing fuzzy C-means (FCM) [14] clustering on the given data to find whether the given instance belongs to more than one cluster. If it is discovered that the instance belongs to more than one cluster, then the cluster centers are calculated with the help of similarity measure, and then member function calculation is performed to know that the given instance is belonging to which cluster exactly. The following objective function can be minimized as shown in Eq. (1)

$$J_k = \sum_{i=1}^N \sum_{j=1}^M u_{ij}^k \|x_i - c_j\|^2 \tag{1}$$

$$1 \leq k \leq \infty$$

here k is some real number greater than 1, the degree of membership of  $x_i$  is  $u_{ij}$  in the cluster  $j$ ,  $i$  th measure of d-dimensional data is  $x_i$ , The d-dimensional center of the cluster is  $c_j$ , and to express the similarity between any measured data and the center  $\|*\|$  some normal entity is used. Through an iterative optimization of the objective function, fuzzy partitioning can be carried out as shown in Eq. (1) with the help of updating the membership of  $u_{ij}$  and the cluster centers  $c_j$  as shown in Eq. (2):

$$u_{ij} = \frac{1}{\sum_{t=1}^C \left( \frac{\|x_i - c_j\|}{\|x_i - c_t\|} \right)^{\frac{2}{k-1}}} \tag{2}$$

$$c_j = \frac{\sum_{i=1}^N u_{ij}^k \cdot x_i}{\sum_{i=1}^N u_{ij}^k} \quad (3)$$

$\max_{ij} \left\{ \left| u_{ij}^{(n+1)} - u_{ij}^{(n)} \right| \right\} < \Delta$ , where  $\Delta$  is a termination criterion between 0 and 1 and stops the iteration up to  $n$  steps. This process a local minimum point  $J_k$ .

**Algorithm 1.** Anomaly-based detection method based on similarity measure

*Input:* Normal training dataset  $X$ , value  $\alpha$  predefined for setting anomaly threshold.

*Output:* The class of  $x_p$  (normal or abnormal)

- 1: Collect clusters  $C_1, C_2, \dots, C_k$  and their centers  $t_1, t_2, \dots, t_k$  for  $X$  from Eq. (3) by Fuzzy C-means.
- 2: Obtain the reference instances  $t_1, t_2, \dots, t_m$  for  $t_1, t_2, \dots, t_m$ .
- 3: for each instance  $x_i \in X$  do
- 4: find the nearest cluster center  $t_i$  for  $x_i$  from  $t_1, t_2, \dots, t_m$
- 5: Calculate membership function  $J_k$  from Eq. (1) and Eq. (2)
- 6: If  $\max_{ij} \left\{ \left| u_{ij}^{(n+1)} - u_{ij}^{(n)} \right| \right\} < \Delta$
- 7: stop
- 8: else
- 9: repeat step 1 to 6.
- 10: end for
- 11: Local minimum point of  $J_k$  is considered.
- 12: Determine the anomaly threshold  $\partial$  based on  $\alpha$ .
- 13: Build the profile  $(\{t_1, t_2, \dots, t_m\}, \partial)$
- 14: Find nearest cluster center  $t_p$  for  $x_p$  from  $t_1, t_2, \dots, t_m$ .
- 15: Calculate  $J_k$  for  $x_p$  using Eq. (1)
- 16: If  $J_k > \partial$  then
- 17:  $x_p$  is abnormal.
- 18: else
- 19:  $x_p$  is normal.
- 20: end if

Figure 3 shows the convergence of centroids for a dataset with the help of membership function calculations.

However, because of less computational complexity anomaly detection component-1 cannot identify the false positives and false negatives perfectly. For this to achieve a hybrid approach is adopted to attain high detection rate and low false positive rate by continuing a similar procedure in phase two.



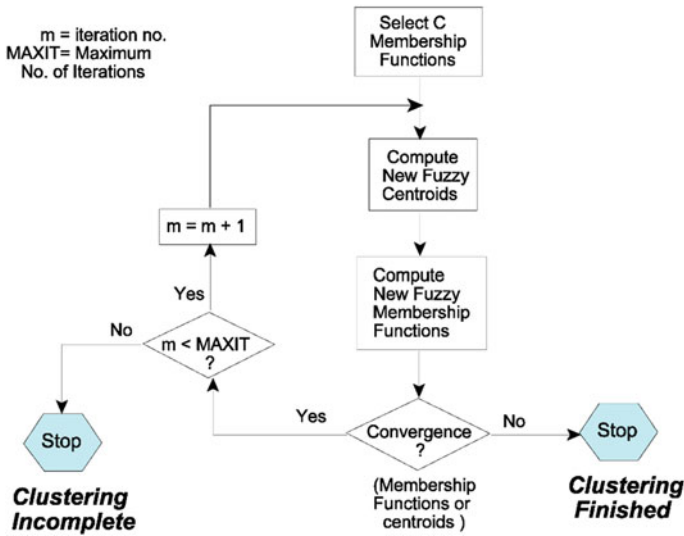


Fig. 3 Fuzzy C-means algorithm flow diagram

### 3.2 Phase Two: Applying Anomaly Detection Component 2 and Misuse Detection Component with KNN

The normal instances and the abnormal instances from phase one are again tested with the help of misuse and anomaly detections in phase two. Using KNN algorithm, both the detection techniques are performed by querying the similarity of k nearest neighbors.

#### Anomaly detection component 2

The anomaly threshold obtained from phase one is applied to cosine function for identifying the false positives formed in anomaly detection component 1.

$$\cos(x_1, x_2) = \frac{x_1 x_2^T}{\|x_1\| \|x_2\|} \tag{4}$$

where  $x_1, x_2$  are used for matching similarity of  $x_1$  with all the instances of  $x_2$ .

**Algorithm 2:** KNN model built in anomaly detection component 2.

*Input:* Normal training dataset  $X'$  with scores sequence  $(J_k = 1, \dots, k)$ , anomaly threshold  $\partial$ , value  $\alpha'$  predefined for setting the outlier threshold, number of nearest neighbors  $k$ , similarity threshold  $\partial'$ , instance  $x_p$  declared abnormal by anomaly detection component 1

*Output:* The class label of  $x_p$  (normal or attack)

Selecting instances for training dataset  $X'$

1.  $X' = \phi$
2. Obtain the outlier threshold  $\partial'$  from  $J_k$  sequence based on  $\alpha'$
3. for each instance  $x_p \in X'$  do
4. if  $\partial < J_k \leq \partial'$  then
5. Add  $x_p$  in  $X'$
6. end if
7. end for
8. for each instance  $x_j \in X'$
9. Calculate  $\cos(x_p, x_j)$
10. if ( $\cos(x_p, x_j) == 1$ ) then
11.  $x_p$  is normal; exit
12. end if
13. end for
14. find  $k$  largest scores of  $\cos(x_p, x_j)$
15. The average scores of  $k$  greatest scores of  $\cos(x_p, x_j)$  are obtained.
16. if the average score is greater than  $\partial'$  then
17.  $x_p$  is normal
18. else
19.  $x_p$  is an attack
20. end if

### Misuse detection component

New arrival instance  $x_p$ , declared normal by anomaly detection component 1, will be further checked by the misuse detection component, which verifies whether  $x_p$  is a normal instance or an attack. Therefore, the job of the misuse detection component in phase two is to identify the false negatives produced by anomaly detection component 1 because some attacks termed as U2R and R2L seem to have the behavior almost nearer to that of normal behavior. Hence there is a threat of getting misclassified as normal instances by anomaly detection component 1. For this objective, another KNN model is to be presented similar to the one defined in Algorithm 2. In this section, the built KNN continues to use a cosine metric to measure the similarity between  $x_p$  and each instance in the corresponding training dataset  $\hat{X}$ , and then checks whether  $x_p$  is an attack or a normal instance.

**Algorithm 3.** KNN model built in the misuse detection component.

*Input:* A dataset  $\hat{X}$ , composed of  $M$  U2R and R2L attacks, anomaly detection component 1 ( $\{t_1, t_2, \dots, t_m\}$ , anomaly threshold  $\partial$ ), number of nearest neighbors  $k$ , similarity threshold  $\partial'$ , instance  $x_p$  declared normal by anomaly detection component 1.

*Output:* The class label of  $x_p$  (normal or attack)

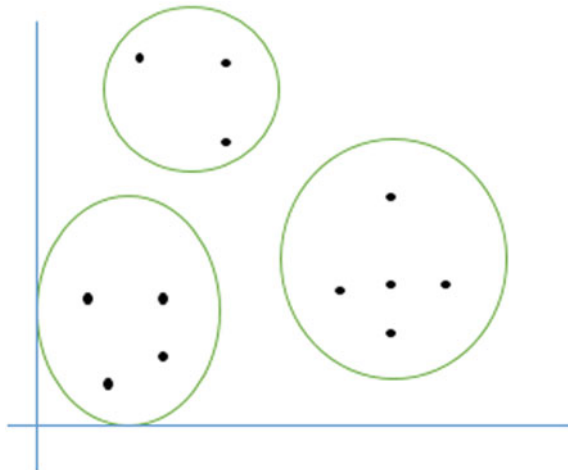
1.  $\hat{X} = \phi$
2. for each instance  $x_p \in \hat{X}$ , do
3. Calculate outlier score  $d_i$  for  $x_i$  by using Eq. (3)

4. if  $d_i \leq \partial$  then
5.  $\widehat{X} = \widehat{X} \cup \{x_p\}$
6. end if
7. end for
8. for each instance  $x_j \in \widehat{X}$  do
9. Calculate  $\cos(x_p, x_j)$
10. if (  $\cos(x_p, x_j) == 1$ ) then
11.  $x_p$  is an attack; exit
12. end if
13. end for
14. find  $k$  largest scores of  $\cos(x_p, x_j)$
15. Obtain the average score of the  $k$  greatest scores of  $\cos(x_p, x_j)$
16. if the average score is greater than  $\partial'$  then
17.  $x_p$  is an attack
18. else
19.  $x_p$  is normal
20. end if

### 4 Experimental Results

To verify the performance ability of the proposed hybrid approach, set of sample instances from KDD 1999 dataset are taken and experimentation is conducted as shown in Fig. 4 where the data is grouped into clusters and the clusters may consist of both normal and attack instances. Here the same instance can be a member of

**Fig. 4** Sample data divided into three clusters



more than one cluster. Each record in these datasets contains 41 features, and a label provides its type. All of the attack records in the KDD 1999 data are depicted into four basic attack classes namely DoS, Probe, U2R, and R2L. In the conducted experiments, attack instances are used to build anomaly detection components 1 and 2. M number of U2R and R2L attacks build misuse detection components.

When applied with the first algorithm on the sample data taken, with the membership function based on the membership order identifies that which instances is placed twice in different clusters. Then on the instances, the algorithms in both the phases are applied to identify whether the instance is a normal instance or attack.

#### 4.1 Performance Evaluation Metrics

Several widely used metrics are used to assess the performance of the proposed hybrid approach in order to achieve high detection rate and low false positive rate. They are detection rate (DR), true negative rate (TNR), false positive rate (FPR), and accuracy (ACC). These metrics can be calculated based on four primary metrics; they are true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). FP are those normal instances incorrectly predicted as attacks, whereas FN denotes the attacks being projected as normal instances. DR, TNR, FPR, and ACC are obtained by:

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

$$TNR = \frac{TN}{TN + FP} \quad (6)$$

$$FPR = \frac{FP}{FP + TN} \quad (7)$$

$$ACC = \frac{TN + TP}{TN + TP + FN + FP} \quad (8)$$

In the next step, a comparison on detection of actual normal data and the probable attacks namely Dos attack, Probe attack, U2R attack, and R2L attacks, that get deviated from normal behavior and that are vulnerable are detected. With the help of this approach, normal and abnormal behaviors can be identified easily. This approach has been applied with the help of many machine-learning techniques Table 1 compares the results obtained by applying various algorithms and explains that FCM proves to be more efficient when compared to the C4.5 decision tree, Hybrid artificial immune system and SOM and KNN. FCM has detection rate for normal instances as 99.65%, DoS as 99.32%, Probe as 99.14%, U2R as 38.76%, and R2L as 9.86%, these results have out performed when compared to applying the previous two which were used with KDD 1999 [15, 16] dataset.

**Table 1** Comparison of the results of misuse detection on applying various algorithms for detecting Normal, DoS, Probe, U2R, and R2L

Approach	Normal (%)	DoS (%)	Probe (%)	U2R (%)	R2L (%)
C 4.5 [17]	96.80	99.19	99.71	66.67	89.14
Hybrid artificial immune system and SOM [18]	99.4	96.8	64.7	34.6	5.2
KNN [13]	99.26	97.26	94.96	28.51	9.72
FCM	99.65	99.32	99.14	38.76	9.86

## 5 Conclusion

The combination of misuse detection and anomaly detection named as hybrid detection systems are proved to be more efficient in this paper. It has been observed that the hybrid detection is found more beneficiary when applied with FCM for finding out the anomalous data in anomaly detection component 1 and then supplied to next phase to in turn detect the normal and attack data with the help of KNN. By applying KNN equally on anomaly detection component 2 and misuse detection component the normal and abnormal data can be acquired. In this way, the data is filtered twice for getting better results. In the future work attribute reduction for IDS data set as proposed in [19, 20], hybrid classification algorithms shall be considered.

## References

1. Ravi Kiran Varma P, Valli Kumari V, Srinivas Kumar S: Ant colony optimization-based firewall anomaly mitigation engine. *Springerplus* 5(1032), 1–32 (2016).
2. Kemmerer, R., Vigna, G.: *Intrusion Detection: A Brief History and Overview*. *Computer* 35(4), 27–30 (2002).
3. Ravi Kiran Varma P, Valli Kumari V, Srinivas Kumar S: Packet Filter Firewall Rule Anomalies and Mitigation Techniques: A Technical Review. *CiiT International Journal of Networking and Communication Engineering* 9(4), 101–108 (2017).
4. Kumar, S., Spafford, E.: A Pattern Matching Model for Misuse Intrusion Detection. In: 11th National Computer Security Conference, pp. 11–21 (1994).
5. Cannady, J.: Artificial Neural Networks for Misuse Detection. In: National Information Systems Security Conference (1998).
6. Jay Prakash, Rakesh Kumar, Saini, J.: Survey on Misuse Detection Systems using Intelligent Agents. *International Journal of Emerging Technology and Advanced Engineering* 5(1) (2015).
7. Patcha, A., Park, J.: An Overview of Anomaly Detection Techniques: Existing Solutions and Latest Technological Trends. *Computer Networks* 51(12), 3448–3470 (2007).
8. P Garcia-Teodoro, J Diaz-Verdejo, G Maci-Fernandez, Vazquez, E.: Anomaly-based Network Intrusion Detection: Techniques, Systems and Challenges. *Computers and Security* 28(1–2), 18–28 (2009).

9. Yibing Wang, Dong Li, Yi Du, Pan, Z.: Anomaly detection in traffic using L1-norm minimization extreme learning machine. *Neurocomputing* 149, 415–425 (2015).
10. O Depren, M Topallar, E Anarim, Ciliz, M.: An intelligent intrusion detection system for anomaly and misuse detection in computer networks. *Expert Systems with Applications*. *Expert Systems with Applications* 29, 713–722 (2005).
11. Hung-Jen Liao, Chun-Hung Richard Lin, Ying-Chih Lin, Tung, K.-Y.: Intrusion detection system : A comprehensive review. *Journal of Network and Computer Applications* 36(1), 16–24 (2013).
12. G Kim, S Lee, Kim, S.: A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. *Expert Systems with Applications* 41(4), 1690–1700 (2014).
13. Chun Guo, Yuan Ping, Nian Liu, Luo, S.-S.: A two-level hybrid approach for intrusion detection. *Neurocomputing* 214, 391–400 (2016).
14. Yinghua Lu, Tinguai Ma, Chanhong Yin, Xiaoyu Xie, Tian, W.: Implementation of the Fuzzy C-Means Algorihm in Meteorological Data. *International Journal of database theory and applications* 6(6), 1–18 (2013).
15. Naahid, M.: Analysis of KDD CUP 99 Dataset using clustering based datamining. *International Journal of Database Theory and Application* 6(5), 23–34 (203).
16. Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, Ghorbani, A.: A Detailed Analysis of the KDD CUP 99 Data Set. In : *IEEE Symposium on Computational Intelligence in Security and Defense Applications (CISDA)* (2009).
17. Cheng Xiang, Png Chin Yong, Meng, L.: Design of multi level classifier for Intrusion detection system using Bayesian clustering and decision trees. *Pattern Recognition Letters* 29 (7), 918–924 (2008).
18. Simon T Powers, He, J.: A hybrid artificial immune system and Self Organizing Map for network intrusion detection. *Information Sciences* 178(15), 3024–3042 (2008).
19. Varma PRK, Kumari VV, Kumar SS: A novel rough set attribute reduction based on ant colony optimisation. *Int. J. Intelligent Systems Technologies and Applications* 14(3/4), 330–353 (2015).
20. Ravi Kiran Varma P, Valli Kumari V, Srinivas Kumar S: Feature selection using relative fuzzy entropy and ant colony optimization applied to real-time intrusion detection system. *Procedia Computer Science* 85(2016), 503–510 (2016).

# Author Index

## A

Ajay Kumar, H., 317  
Ali, Shahid, 181  
Anuradha, Ch., 285  
Anuradha, T., 105

## B

Balaji, Vicharapu, 285  
Bala Sindhuri, K., 133  
Bhanu Jyothi, Kella, 191

## C

Chandra Murty, P. S. R., 285  
Cherukuri, Hemanthi, 95

## D

Deepak, Gerard, 265  
Devarakonda, Nagaraju, 113  
Durga Teja, K., 133

## E

Erothi, Uma Shankar Rao, 85  
Eswara Reddy, B., 229, 247

## G

Garg, Naresh Kumar, 125  
Giri, M., 201  
Goel, Lavika, 65  
Gokaramaiah, T., 293  
Guttikonda, Prashanti, 95

## H

Hande, V. Shreyas, 29  
Hima Bindu, K., 191

## J

Jilo, Chala Tura, 1  
Jyothi, S., 201

## K

Kariyappa, B.S., 29  
Kiran Kumar, R., 301  
Kumar, C.S. Hemanth, 29  
Kumar, Munish, 125  
Kumar, Naman S., 29  
Kumar, S.N., 317

## L

Lenin Fred, A., 317  
Leena Giri, G., 265

## M

Manjula, S.H., 265  
Mantri, Raghav, 65  
Mundukur, Nirupama Bhat, 95

## N

Nagamani, Ch., 113  
Narayan, K. Gowri Raghavendra, 41  
Narendrababu Reddy, G., 215

## P

Padmaja, Grandhe, 285  
Pagi, Veerappa B., 345  
Phani Kumar, S., 215  
Pinna, Francesco, 11  
Ponmary Pushpa Latha, D., 333  
Ponnusamy, R., 145  
Prasad, A., 173

## R

Raja Babu, M., 293  
Raju, P. Pothu, 41  
Rakesh, P., 133  
Rao, Chinta Someswara, 1  
Rao, Kolakaluri Srinivasa, 159  
Rao, T. Srinivasa, 41

Ravikanth, Garladinne, [247](#)  
 Ravi Kiran Varma, P., [359](#)  
 Rodda, Sireesha, [85](#)  
 Rupa, Ch., [173](#)

**S**

Sai Krishna, T.V., [301](#)  
 Sajjan, Mallikarjun V., [29](#)  
 Sakshi, [125](#)  
 Sebastian Varghese, P., [317](#)  
 Shah, Bansari, [55](#)  
 Shaik, Subhani, [113](#)  
 Sirajuddin, MD., [173](#)  
 Srinivasu, Lingamallu Naga, [159](#)  
 Subhani, Shaik, [113](#)  
 Sudhakar, P., [41](#)  
 Sudhanva, N.G., [29](#)  
 Sujata, B., [359](#)  
 Sujatha, B., [1](#)  
 Sunitha, K.V.N., [247](#)  
 Sunitha Ram, C., [145](#)  
 Swamy, Mallikarjun, [65](#)  
 Swapna, B., [105](#)

**T**

Tirumala, Sreenivas Sremath, [277](#)  
 Tiwari, Shailendra, [55](#)

**U**

Udaya Kumar, N., [133](#)  
 Ummadi, Janardhan Reddy, [229](#)

**V**

Vasantha Kokilam, K., [333](#)  
 Venkata Ramana Reddy, B., [229](#)  
 Venugopal, K.R., [265](#)  
 Verma, Ananda Prakash, [55](#)  
 Vishnuvardhan Reddy, A., [293](#)

**W**

Wadawadagi, Ramesh S., [345](#)

**Y**

Yesu Babu, A., [301](#)

**Z**

Zedda, Mariangela, [11](#)