

Lecture Notes in Electrical Engineering 907

Sanjeev Sharma · Sheng-Lung Peng ·  
Jitendra Agrawal · Rajesh K. Shukla ·  
Dac-Nhuong Le *Editors*

# Data, Engineering and Applications

Select Proceedings of IDEA 2021

# Lecture Notes in Electrical Engineering

## Volume 907

### Series Editors

Leopoldo Angrisani, Department of Electrical and Information Technologies Engineering, University of Napoli Federico II, Naples, Italy

Marco Arteaga, Departament de Control y Robótica, Universidad Nacional Autónoma de México, Coyoacán, Mexico

Bijaya Ketan Panigrahi, Electrical Engineering, Indian Institute of Technology Delhi, New Delhi, Delhi, India

Samarjit Chakraborty, Fakultät für Elektrotechnik und Informationstechnik, TU München, Munich, Germany

Jiming Chen, Zhejiang University, Hangzhou, Zhejiang, China

Shanben Chen, Materials Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

Tan Kay Chen, Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore

Rüdiger Dillmann, Humanoids and Intelligent Systems Laboratory, Karlsruhe Institute for Technology, Karlsruhe, Germany

Haibin Duan, Beijing University of Aeronautics and Astronautics, Beijing, China

Gianluigi Ferrari, Università di Parma, Parma, Italy

Manuel Ferre, Centre for Automation and Robotics CAR (UPM-CSIC), Universidad Politécnica de Madrid, Madrid, Spain

Sandra Hirche, Department of Electrical Engineering and Information Science, Technische Universität München, Munich, Germany

Faryar Jabbari, Department of Mechanical and Aerospace Engineering, University of California, Irvine, CA, USA

Limin Jia, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

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

Alaa Khamis, German University in Egypt El Tagamoa El Khames, New Cairo City, Egypt

Torsten Kroeger, Stanford University, Stanford, CA, USA

Yong Li, Hunan University, Changsha, Hunan, China

Qilian Liang, Department of Electrical Engineering, University of Texas at Arlington, Arlington, TX, USA

Ferran Martín, Departament d'Enginyeria Electrònica, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, Spain

Tan Cher Ming, College of Engineering, Nanyang Technological University, Singapore, Singapore

Wolfgang Minker, Institute of Information Technology, University of Ulm, Ulm, Germany

Pradeep Misra, Department of Electrical Engineering, Wright State University, Dayton, OH, USA

Sebastian Möller, Quality and Usability Laboratory, TU Berlin, Berlin, Germany

Subhas Mukhopadhyay, School of Engineering & Advanced Technology, Massey University, Palmerston North, Manawatu-Wanganui, New Zealand

Cun-Zheng Ning, Electrical Engineering, Arizona State University, Tempe, AZ, USA

Toyoaki Nishida, Graduate School of Informatics, Kyoto University, Kyoto, Japan

Luca Oneto, Department of Informatics, Bioengineering, Robotics, University of Genova, Genova, Genova, Italy

Federica Pascucci, Dipartimento di Ingegneria, Università degli Studi "Roma Tre", Rome, Italy

Yong Qin, State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China

Gan Woon Seng, School of Electrical & Electronic Engineering, Nanyang Technological University, Singapore, Singapore

Joachim Speidel, Institute of Telecommunications, Universität Stuttgart, Stuttgart, Germany

Germano Veiga, Campus da FEUP, INESC Porto, Porto, Portugal

Haitao Wu, Academy of Opto-electronics, Chinese Academy of Sciences, Beijing, China

Walter Zamboni, DIEM - Università degli studi di Salerno, Fisciano, Salerno, Italy

Junjie James Zhang, Charlotte, NC, USA

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering—quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact [leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com).

To submit a proposal or request further information, please contact the Publishing Editor in your country:

#### **China**

Jasmine Dou, Editor ([jasmine.dou@springer.com](mailto:jasmine.dou@springer.com))

#### **India, Japan, Rest of Asia**

Swati Meherishi, Editorial Director ([Swati.Meherishi@springer.com](mailto:Swati.Meherishi@springer.com))

#### **Southeast Asia, Australia, New Zealand**

Ramesh Nath Premnath, Editor ([ramesh.premnath@springernature.com](mailto:ramesh.premnath@springernature.com))

#### **USA, Canada**

Michael Luby, Senior Editor ([michael.luby@springer.com](mailto:michael.luby@springer.com))

#### **All other Countries**

Leontina Di Cecco, Senior Editor ([leontina.dicecco@springer.com](mailto:leontina.dicecco@springer.com))

**\*\* This series is indexed by EI Compendex and Scopus databases. \*\***

Sanjeev Sharma · Sheng-Lung Peng ·  
Jitendra Agrawal · Rajesh K. Shukla ·  
Dac-Nhuong Le  
Editors

# Data, Engineering and Applications

Select Proceedings of IDEA 2021

 Springer



*Editors*

Sanjeev Sharma  
School of Information Technology  
Rajiv Gandhi Technical University  
Bhopal, Madhya Pradesh, India

Jitendra Agrawal  
Department of Computer Science  
and Engineering  
Rajiv Gandhi Technical University  
Bhopal, Madhya Pradesh, India

Dac-Nhuong Le  
Department of Information Technology  
Haiphong University  
Haiphong, Vietnam

Sheng-Lung Peng  
Department of Creative Technologies  
and Product Design  
National Taipei University of Business  
Taiwan, Taiwan

Rajesh K. Shukla  
Department of Computer Science  
and Engineering  
Oriental Institute of Science  
and Technology  
Bhopal, Madhya Pradesh, India

ISSN 1876-1100

ISSN 1876-1119 (electronic)

Lecture Notes in Electrical Engineering

ISBN 978-981-19-4686-8

ISBN 978-981-19-4687-5 (eBook)

<https://doi.org/10.1007/978-981-19-4687-5>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

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

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

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

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

# Contents

<b>Distortion Controlled Secure Reversible Data Hiding in H.264 Videos</b> .....	1
Jaladi Vivek, Baswaraj Gadgay, and D. C. Shubhangi	
<b>A Method for Improving Efficiency and Security of FANET Using Chaotic Black Hole Optimization-Based Routing (BHOR) Technique</b> .....	15
Mayank Namdev, Sachin Goyal, and Ratish Agarwal	
<b>Machine Learning Techniques for Intrusion Detection System: A Survey</b> .....	29
Saby Singhal and Pradeep Yadav	
<b>Software Fault Detection by Using Rider Optimization Algorithm (ROA)-Based Deep Neural Network (DNN)</b> .....	41
Shilpa Garg, Deepak Kumar, Suresh Chand Gupta, and Vijay Anant Athavale	
<b>An Approach for Predicting Admissions in Post-Graduate Programme by Using Machine Learning</b> .....	57
Shivam Sharma and Hemant Kumar Soni	
<b>A Survey on Various Representation Learning of Hypergraph for Unsupervised Feature Selection</b> .....	71
Rana Pratap Singh, Divyank Ojha, and Kuldeep Singh Jadon	
<b>Adoption of Blockchain Technology for Storage and Verification of Educational Documents</b> .....	83
Vijay Anant Athavale, Shakti Arora, and Anagha Athavale	
<b>Obstacle Collision Prediction Model for Path Planning Using Obstacle Trajectory Clustering</b> .....	99
Samir Ajani and Salim Y. Amdani	

<b>Human Saliency Based Object Detection from Natural Images</b> .....	113
Naveen Chandra and Himadri Vaidya	
<b>Dynamic Education Background: Procure the Maximum Initiation from PBL for Education Naïve Bayes Algorithm for Machine Learning</b> .....	129
Vishnu Kumar Mishra, Megha Mishra, Jitendra Sheetlani, Rahul Deo Sah, Anup Mishra, and Satyendra Kurariya	
<b>Hybrid Microarray Gene Data Classification Based on GA-ACO Optimization</b> .....	141
N. Karunya, V. A. Kanimozhi, N. Muthumani, M. P. Karthikeyan, and Dac-Nhuong Le	
<b>An In-Field Real-Time Automatic Weed Detection Using Deep Learning Techniques</b> .....	153
Siddharth Dutt Choubey and Rohit Singh Thakur	
<b>Sentiment Analysis of Indians Due to Conflict Between India and China at the Actual Line of Control in Galwan Valley, Ladakh</b> .....	167
Ravi Kumar and Deepak Kumar	
<b>Feature Selection Using Information Gain for Software Effort Prediction Using Neural Network Model</b> .....	177
Sushma Khatri and Pratosh Bansal	
<b>A Comprehensive Survey of Web and Mobile Apps for Fishermen</b> .....	199
M. R. Jivthesh, M. R. Gaushik, N. B. Sai Shibu, Dhanesh Raj, and Sethuraman N. Rao	
<b>Study of Machine Learning Classifiers for Intrusion Detection System</b> .....	213
Akshita Mishra and Archana Thakur	
<b>The Analysis of Time Series Data</b> .....	225
Charu Kathuria, Deepti Mehrotra, Shalini Bhartiya, and Navnit Kumar Misra	
<b>Computed Tomography Image Processing Methods for Lung Nodule Detection and Classification: A Review</b> .....	237
Ebtasam Ahmad Siddiqui, Vijayshri Chourasia, Madhu Shandilya, and Vivek Patel	
<b>Implementation and Analyzing SURF Feature Detection and Extraction on WANG Images Using Custom Bag of Features Model</b> .....	255
Roohi Ali and Manish Maheshwari	

**Machine Learning Approaches for Image-Based Screening of Cervical Cancer** ..... 269  
 Priyanka Rastogi, Kavita Khanna, and Vijendra Singh

**A Comparative Analysis of Image Steganography Based on Discrete Wavelet Transform (DWT) and Exploiting Modification Direction (EMD)** ..... 283  
 Abhishek choubey and Shruti Bhargava Choubey

**Web Attack Detection Using Machine Learning** ..... 291  
 Raturaj Malavade, Harshali Upadhye, Heena Jamadar, Deepak Kshirsagar, and Jagannath Aghav

**Pragmatic Analysis of Web Service Discovery Models and Architectures from a Qualitative Perspective** ..... 301  
 Moumita Majumder Sarkar and Manish Dhananjay Sawale

**An Integrated Semi-Supervised Learning Framework for Image Compression Using DCT, Huffman Encoding, and LZW Coding** ..... 317  
 Jamandlamudi Sravya Sruthi, Rekh Ram Janghel, and Lokesh Singh

**Analyzing the Need for Video Summarization for Online Classes Conducted During Covid-19 Lockdown** ..... 333  
 Shikha Sharma and Madan Lal Saini

**Local Roughness Binary Pattern for Texture Classification** ..... 343  
 Sumit Kumar Gupta, Susheel Yadav, Dharendra Pratap Singh, and Jaytrilok Choudhary

**Speech Enhancement Using VAD for Noise Estimation in Compressive Sensing** ..... 357  
 Vasundhara Shukla and Preety D. Swami

**A Comparative Study on Single Image De-Raining Using Convolutional Neural Network** ..... 371  
 Poornima Shrivastava, Roopam Gupta, and Asmita A. Moghe

**A Hybrid Translation Model for Pidgin English to English Language Translation** ..... 385  
 Saviour Oluwatomiya, Sanjay Misra, John Wejin, Akshat Agrawal, and Jonathan Oluranti

**An Incremental Load Balancing Algorithm in Federated Cloud Environment** ..... 395  
 Nzanzu Vingi Patrick, Sanjay Misra, Emmanuel Adetiba, and Akshat Agrawal

**An Intelligent Hydroponic Farm Monitoring System Using IoT** ..... 409  
 Jalani H. Naphtali, Sanjay Misra, John Wejin, Akshat Agrawal, and Jonathan Oluranti

<b>IoT and Machine Learning Based Anomaly Detection in WSN for a Smart Greenhouse</b> .....	421
Molo Mbaso Joaquim, Abednego Wamuhindo Kamble, Sanjay Misra, Joke Badejo, and Akshat Agrawal	
<b>Content Based Deep Factorization Framework for Scientific Article Recommender System</b> .....	433
Akhil M. Nair, Oshin Anto, Anchana Shaji, and Jossy George	
<b>On Discrimination Power of Image Feature Vector</b> .....	443
Sushila Palwe	
<b>Design a Mechanism for Opinion Mining</b> .....	453
Samir N. Ajani and Parul Bhanarkar	
<b>Event Detection in Live Twitter Streams Using Tf-Idf and Clustering Algorithms</b> .....	469
Tavishi Jain, Bhavya Singh, and Rupesh Kumar Dewang	
<b>Event Detection and Summarisation of Live Tweets Using SCAN Algorithms</b> .....	481
Tavishi Jain, Bhavya Singh, and Rupesh Kumar Dewang	
<b>Spam Review Detection Using Okapi Relevance Method for Negative Reviews</b> .....	493
Saloni Juneja, Shubham Goyallal, Sonali Agarwal, Saransh Agrawal, Rohit Kumar, Rupesh Dewang, and Arvind Mewada	
<b>Predicting Time-Series Data Using Linear and Deep Learning Models—An Experimental Study</b> .....	505
Ahmad Alsharef, Sonia, Monika Arora, and Karan Aggarwal	
<b>A Review on Community Detection Methods and Algorithms in Social Networks: Open Trends and Challenges</b> .....	517
Ranjana Sikarwar, Shashank Sheshar Singh, and Harish Kumar Shakya	
<b>Three-Stage Heterogeneous Data Clustering Using Unsupervised Multiple Kernel and Extreme Learning Machine</b> .....	531
Ankit R. Mune and Soheli A. Bhura	
<b>Artificial Intelligences-Based Approaches for Generating Image Caption</b> .....	541
S. P. Ingale and G. R. Bamnote	
<b>Composite Reversible Data Hiding Scheme for Secure Image Reconstruction</b> .....	553
Nandni Tandon and Abhishek Sharma	
<b>Semantic Relation-Based Modularity-Optimized Community Detection Algorithm for Heterogeneous Networks</b> .....	565
Rishank Rathore and Ravi Kumar Singh Pippal	

**Sentimental Analysis with Emojis by Using Machine Learning** ..... 583  
Balajee Maram, B. Srinivas Kumar, and P. Swaroopni

**Security Risk Analysis and Design Reengineering for Smart Healthcare** ..... 599  
Madhu Sharma Gaur, Navneet K. Gaur, Sanjeev Kumar, and Prem Sagar Sharma

**Prediction of the Risk of Heart Attack Using Machine Learning Techniques** ..... 613  
Pinaki Ghosh, Umesh Kumar Lilhore, Sarita Simaiya, Atul Garg, Devendra Prasad, and Ajay Kumar

**An Integrated CBIR Approach for Medical Image Retrieval System** .... 623  
Anubhav Sharma and Shiv Shakti Shrivastava

**Precise Forecasting of Stock Market Pricing Using Weighted Ensemble Machine Learning Method** ..... 637  
Umesh Kumar Lilhore, Sarita Simaiya, Advin Manhar, Shilpi Harnal, Pinaki Ghosh, and Atul Garg

**Graph-Based Mechanism to Prevent Structural Attack Over Social Media** ..... 649  
Jitendra Patel and Ravi Kumar Singh Pippal

**Image Forgery Detection Using Supervised Learning Algorithm** ..... 663  
R. Cristin, T. Daniya, and S. Divyatej

**A Machine Learning Method for Customer Sentiment Analysis on Social Media** ..... 683  
Chetan Agrawal, Anjana Pandey, and Sachin Goyal

**A Novel Hybrid Approach for the Designing and Implementation of Dogri Spell Checker** ..... 695  
Shubhnandan S. Jamwal and Parul Gupta

# Distortion Controlled Secure Reversible Data Hiding in H.264 Videos



Jaladi Vivek, Baswaraj Gadgay, and D. C. Shubhangi

**Abstract** Reversible data hiding in H.264 videos aims to keep the quality of recovered video resulting after extracting the information to be of the same quality as that of the original cover video. Most reversible data embedding algorithms in H.264 video streams are based on the Intra-prediction mode. Embedding decision is made only based on the loss occurring to IPM blocks. IPM-based embedding decision results in distortion getting drifted across multiple blocks reducing the quality of the reconstructed video. This work exploits the IPCM mode in intra-prediction to embed secret information with the constraints of minimizing distortion and maximizing embedded capacity. This work proposes a IPCM frame selection for embedding secret information. The selection is done in a way to minimize the distortion cost. The proposed solution increases the embedding capacity by adaptive control of the IPCM block generation rate. Security is provided for the embedded content by encrypting them before embedding. Comparison is done in terms of distortion introduced in video and the embedding capacity achieved. The proposed solution is able to achieve at least 6.3% higher embedding capacity, 8.9% higher Peak Signal-to-Noise Ratio (PSNR) and 4.66% lower MSE compared to existing works.

**Keywords** H.264/AVC · Video data hiding · IPCM · Encryption

## 1 Introduction

Data hiding is a technique to hide data in an imperceptible way into a multimedia cover [1]. The wireless sensor network is a network of the sensor. Watermarking [2], Steganography [3] and reversible data hiding (RDH) [4] are some of the well-known data hiding techniques. While all these three methods are for hiding a secret content in a cover signal, the objectives are different. The goal of watermarking is to

---

J. Vivek (✉)

Lingaraj Appa Engineering College, Bidar, India

e-mail: [vivec53@gmail.com](mailto:vivec53@gmail.com)

B. Gadgay · D. C. Shubhangi

Department of PG Studies, VTU, Kalaburagi, India

hide content in cover such that it is robust against any deformation attacks. The goal of steganography is to hide data in such a way that it becomes secure against any statistical detection. In RDH, the goal is to recover cover and the embedded information without any loss. Both embedding capacity and signal quality are important parameters in RDH and they are positively correlated with rate distortion. Among the various cover sources used for RDH like images, audios and videos, videos are better preferred due to their higher embedding capacity at a comparatively lower distortion. The existing methods for hiding data in H.264 videos can be categorized into four types: intra-prediction mode (IPM) [5], motion vector [6], discrete cosine transform coefficients (DCT) [7] and entropy coding coefficients [8]. Due to higher video quality distortion during recovery, DCT and entropy coding coefficients are hardly preferred for data embedding in H.264 videos. Intra-prediction mode is the most adopted method for data hiding in H.264 videos due to its low computational complexity and higher redundancy volume for hiding secret information. The current IPM-based data hiding methods calculate the distortion at the local IPM block level and embed content if distortion is less than the threshold. This local decision can introduce intra-frame distortion drift into subsequent frames as distortion gets accumulated. Due to intra-frame distortion drift, the reconstructed video is not the same as that of the original video.

In this work, a distortion controlling strategy is proposed to select the IPCM blocks considering the distortion at a global level instead of a local decision. The IPCM blocks are selected in such a way to minimize both distortion cost and distortion drift cascading to multiple IPCM blocks. In addition to selection, the proposed method also adaptively controls the rate of IPCM block generation based on the volume of information available to embed and the distortion cost of the IPCM blocks. To provide additional security against steganalysis attacks, the secret information to be hidden is scrambled using chaos encryption. The encrypted secret information is then hidden in the selected IPCM blocks. Following are the contributions of the proposed solution

1. Selection criteria for IPCM block selection considering both distortion cost and distortion drift.
2. Adaptive rate control of IPCM blocks based on the residual volume of information to be hidden and distortion cost of IPCM blocks.
3. Security against steganalysis attack by chaos encryption of secret information before embedding.

## 2 Related Works

Zhang et al. in [5] used linear code over  $Z_4$  to increase the embedding capacity of H.264 intra-prediction steganography. By using the properties of abelian groups and multiplicative commutative semi-groups, the secret information to be hidden in the video is transformed. This transformed information is at a comparatively lower size compared to original secret information. The property of linear code is that it



can withstand about 50% distortion and still be able to provide the secret information. But the method is not sensitive to the distortion and distortion drift introduced into the intra-prediction blocks. Zhao et al. in [9] assigned cost to the candidate blocks based on characteristics of the complex texture. Adaptive embedding perturbation is done over the candidate blocks representing complex textures. Though the method was secure against steganalytic methods, it reduced the embedded capacity and was not sensitive to distortion drift. Nie et al. in [10] proposed an effective intra-prediction mode-based video steganography. A distortion function based on the sum of absolute difference (SAD) prediction deviation is defined. The inter prediction frames are scored based on the sum of SAD and embedding is done on blocks whose SAD value is below a tolerance threshold. Syndrome-trellis code (STC) is used for data embedding. The method is robust against steganalysis attacks. But it cannot achieve higher data embedding capacity. MaungMaung et al. in [11] used audio-video synchronization information in the MP4 container to hide secret information. The synchronization information stored in ssts box of the MP4 container is manipulated at least significant bits to embed hidden data. This method is not secure against steganalysis attacks. Wang et al. in [12] proposed a tunable data hiding method for H.264/AVC streams. A group of coefficients are selected for hiding secret information. The method is flexible to user requirements on embedding capacity and distortion. Modifying the coefficients distorts the video quality and this method is not suitable for RDH. Lakshmi et al. in [13] proposed a data hiding method with two objectives of reversible data hiding and improving the contrast of the frames. The frames with poor illumination are selected and contrast improved. By improving the contrast, the least significant bits are modified to hide secret information. The method is not suitable for all videos and it is not secure against steganalysis attacks. Chen et al. in [14] proposed a data hiding method by modifying the QDCT coefficients of high-frequency area, thereby there is no effect on the visual quality of H.264 videos. The macroblocks within intra-frame 4\*4 prediction mode are selected. The last zero QDCT coefficients are paired and modified according to a mapping rule to hide secret information. By modifying the coefficient, there is not much impact on the visual quality of the H264 video. Though the embedding capacity is higher and distortion is low in this method, it is insecure against steganalysis attack. The same author in [15] used the 2n nonzero QDCT coefficients of luminance components of macroblock to hide secret information. But affecting the luminance creates a higher distortion drift in subsequent blocks. Chen et al. in [16] selected the blocks for data embedding based on three conditions preventing distortion propagation. QDCT coefficients are modified in the selected 4\*4 luminance blocks to hide secret information. The method is not sensitive to distortion drift. Fallahpour et al. in [17] proposed a high-capacity data embedding algorithm in H.264 videos. The secret data is embedded into QDCT coefficients of I frames. The embedding is done in such a way that intra-frame distortion drift is minimized. The coefficients are split into two groups embedding an averting group. Secret information is carried in the embedding group. Averting group prevents distortion drift in adjacent blocks. The distortion drift mitigation is limited to the reduced number of blocks. Lin et al. in [18] used the Shamir secret sharing algorithm to split the secret information into n

parts. Out of these  $n$  parts, only  $k$  parts ( $k < n$ ) are needed for the reconstruction of the secret. The  $n$  parts are embedded into the intra-frame of H.264 videos. Even if  $n-k$  parts are affected due to video quality distortion, it is still possible to reconstruct the data using the Shamir algorithm. But in case distortion affects more than  $n-k$  parts, it is not possible to reconstruct the secret information. Kim et al. in [19] proposed a data hiding scheme for H.264 scheme with complete reversibility for I and P frames. Data embedding is done by the clever pairing of coefficients in the I and P blocks. But the effect of distortion drifting to subsequent frames cannot be solved in this method. Xu et al. in [20] proposed a solution to solve the distortion of H.264 video introduced by transmission errors. The motion vector of a microblock is embedded into other microblocks within the same frame. In this way, corrupted microblocks can be reconstructed. A two-dimensional RDH scheme exploiting the distribution of motion vector data and histogram shifting is proposed. But the embedding capacity is limited in this approach. The same authors in [21] selected the most suitable embedding region which does not affect the visual quality and modified the coefficients in those embedding regions. Nguyen et al. in [22] avoided intra-frame distortion drift and guaranteed low distortion by selecting the suitable coefficient pair in the I frames. But the method is susceptible to transmission errors. Xu et al. in [23] embedded secret information in H.264 videos by modulating the prediction modes of  $4 \times 4$  luminance blocks. The secrecy is improved in two ways: chaotic encryption of secret information and random selection of  $4 \times 4$  luminance blocks. The video quality distortion is higher in this approach. Chen et al. in [24] proposed an efficient cost assignment method to measure the distortion cost and select the I blocks. Syndrome-trellis code (STC) is used for data embedding in the selected I blocks. Yao et al. in [25] proposed a model to estimate the embedding distortions caused by modifying different residual coefficients. The block with a distortion cost lower than the threshold is selected and data embedding is done using the histogram shifting technique. But the method is insecure against steganalysis attacks. Ma et al. in [26] used the direction of intra-frame prediction to avert the distortion drift while data embedding in I frames. Data embedding is done by using the paired coefficients of  $4 \times 4$  DCT blocks. But the method is not resilient against transmission errors. Wang et al. in [27] used adaptive hybrid coding methods to increase the embedding capacity in H.264 videos. Hybrid coding is applied on different trailing DCT coefficients in such a way to increase the embedding capacity with low bit rate growth. But the problem with this approach is it becomes difficult to reconstruct the data in case of distortion in blocks due to transmission errors.

### 3 Proposed Method

The proposed Distortion Controlled Secure RDH is built on the IPCM mode used by H.264 encoder during intra-prediction [30]. In IPCM mode, the values of samples are sent directly without prediction, transformation or quantization. IPCM mode is usually preferred in H.264 to represent regions of the image without any loss

of fidelity. Since IPCM blocks are picture values, data embedding can be done with higher capacity compared to coefficient-based embedding. The data embedding capacity ( $D_c$ ) in H.264 video with  $N_I$  IPCM blocks is given as

$$D_c = 256 * N_I * L_{bits} + 2 * 64 * N_I * C_{bits} \quad (1)$$

where  $L_{bits}$ ,  $C_{bits}$  is the number of low bits per IPCM luma and chroma samples to be used for data hiding. Even with a single IPCM macroblock, with  $L_{bits}$  and  $C_{bits}$  as 4, an embedding capacity of 1536 bits can be achieved. But since the IPCM blocks are generated to represent regions without loss of fidelity, not all IPCM blocks are suitable for data embedding. The distortion introduced by data embedding can cause a loss of fidelity and also affects the reconstructed video quality. Thus, there is a need to select the IPCM block for data embedding based on the distortion cost. In approaches based on IPCM data hiding, data embedding on the sequence of IPCM blocks is avoided to reduce the effect of occasional distortion drift. This reduces the embedding capacity. But evaluating the sequence of IPCM blocks for data embedding at a joint level of distortion can solve the problem of reduced embedding capacity and at the same time minimizes the overall distortion. In approaches based on IPCM data hiding, IPCM blocks are generated based on periodic intervals without consideration for the data volume to be hidden and characteristics of past IPCM blocks, output video volume expansion, etc. By adaptive control of IPCM block generation, maximal embedding can be achieved with lower volume expansion. The proposed FDACS-RDH achieves the following two objectives in IPCM data hiding. Firstly, minimize the overall distortion cost by modeling distortion on a joint sequence of IPCM blocks instead of individual blocks. Secondly, adaptive rate control of IPCM block generation is based on the volume of information to be hidden, characteristics of past IPCM blocks and the number of video frames pending so that embedding is achieved with lower video volume expansion. The architecture of the proposed FDACS-RDH solution is given in Fig. 1.

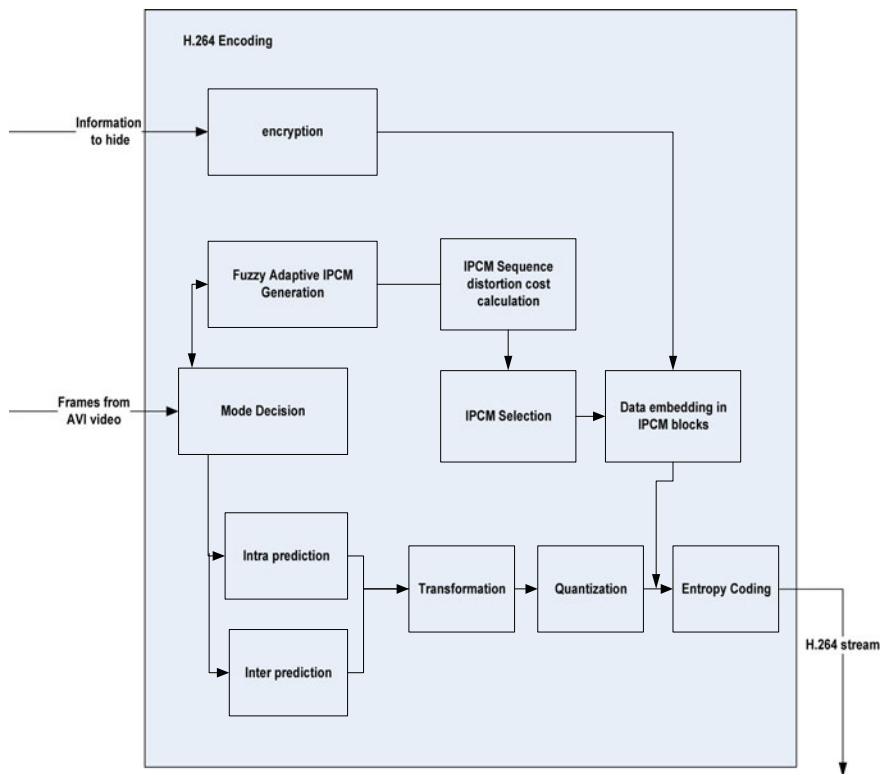
### 3.1 Distortion Cost Calculation and IPCM Selection

Every  $k$  IPCM blocks are jointly analyzed for distortion before embedding data into the IPCM blocks. Let  $V$  be the decision variable deciding if embedding needs to be done on the IPCM blocks. It is expressed as

$$V = \{v_1, v_2, v_3, \dots, v_k\}. \quad (2)$$

$V = 0$  if hiding cannot be done in block and  $V = 1$  if hiding is to be done in block.

Let  $D$  be the distortion cost for embedding data into the IPCM block. It is expressed as



**Fig. 1** Architecture of proposed distortion controlled reversible data hiding

$$D = \{d_1, d_2 \dots d_k\}. \quad (3)$$

The distortion cost is calculated using a fuzzy c mean clustering on histogram values on DCT coefficients after thresholding the DCT coefficients. The coefficient resulting from the application of DCT on the IPCM image signal  $x_i$  is given as

$$X_i = \text{DFT}(x_i) = \sum_{n=0}^{N-1} x_i(n) e^{-j \frac{2\pi}{N} kn}. \quad (4)$$

The threshold function  $T$  transforming  $X_i$  can be given as

$$T(X_i) = \begin{cases} 0, & \forall |X_{ni}| < T_1 \\ X_i(k) \omega(|X_{ni}(k)|), & |X_{ni}| \geq T_1. \end{cases} \quad (5)$$

$X_{ni}$  is calculated based on the estimation of the mean and standard deviation of coefficients as

$$X_{ni} = \frac{X_i(k) - \mu_m(k)}{\sigma_m(k)}. \quad (6)$$

The mean and standard deviation estimation is as follows:

$$\mu_m(k) = \frac{\sum_{i=0}^{I-1} X_i(k)}{I} \quad (7)$$

$$\sigma_m(k) = \sqrt{\frac{\sum_{i=0}^{I-1} X_i(k)^2}{I} - \mu_m(k)^2}$$

The resulting  $T(X_i)$  is split into 10 histograms by dividing the coefficient range into 10 equal side bins. This histogram values over 10 bins are used as features. A training dataset of  $N$  images is taken with different visual qualities. Each of the images are converted to a histogram feature vector. Fuzzy C Means clustering is done on the training data of  $N$  feature vectors with the number of clusters as  $P$ . The cluster center after the fuzzy C means clustering is defined as

$$D = \{D_{e,q}, e = 1, 2 \dots P \text{ and } q = 1, 2, 3, \dots 10\} \quad (8)$$

where  $D_{e,q}$  is the coordinating cluster  $e$ .

The distortion cost is calculated as the sum of the weighted distortion score of membership values of each cluster as follows:

$$D_{ck} = \sum_{e=1}^P \Phi_{r,e} \times DS_e \quad (9)$$

where  $DS_e$  is the distortion score of the cluster  $e$ .

Calculating distortion in the frequency domain is done in this work as the frequency domain captures the visual distortion information better than the spatial domain.

Let  $E$  be the video expansion cost given as

$$E = \{e_1, e_2 \dots e_k\}. \quad (10)$$

$e_i$  is the cost of expansion of video due to IPCM block  $i$ .

The expansion cost depends on the number of least significant bits used in the IPCM block. The choice can be from 1 to 4. The distortion also varies in proportion to the number of least significant bits used in IPCM blocks.

### 3.2 *Data Embedding in IPCM Block*

The secret information to be hidden in the video is encrypted to provide additional security against steganalysis attacks. For text secret information, the chaos encryption method proposed in [28] is used. For image secret information, the chaos encryption method proposed in [29] is used. Both these chaos encryptions provide encrypted content with the same size as that of the original. The encrypted content is converted to a bit stream. The bit stream is inserted into the last three least significant bits of selected IPCM blocks in order.

## 4 Results






The proposed algorithm is implemented in MATLAB and the performance is measured for different videos in terms of the following metrics: Peak Signal-to-Noise ratio (PSNR), Mean Square Error (MSE) and Embedding capacity (EC). Distortion is measured in terms of PSNR and MSE. The performance is compared against both IPM- and IPCM-based methods. For comparison against IPM methods, two recent works of embedding using linear block code [5] and Cost assignment-based method [24] are used. For comparison against IPCM, the IPCM macroblock approach [30] is used.

The test is conducted with Lena image ( $512 \times 512$ ) of size 786 KB as the secret. The image is chaos encrypted and converted to a binary stream before embedding. The performance is tested for the video sequences given in Table 1.

### 4.1 *PSNR Comparison*

The comparison of PSNR results between the original and the recovered video for different solutions is presented in Table 2. The PSNR has improved in the proposed solution due to data embedding in the least significant bit of IPCM blocks whose distortion cost is very low. Due to this, there is no visual difference in the reconstructed video. But in IPM solutions, the coefficients are modified and this creates a minor visual difference. Compared to IPCM macroblocks, the proposed method is able to achieve higher PSNR due to the evaluation of IPCM for distortion cost and selection of IPCM with lower distortion for data embedding in the proposed solution.

**Table 1** Test video sequences

Sequences	Frame size	No. of frames
Akyio 	176 × 144	300
Foreman 	176 × 144	300
Mobile 	176 × 144	300
Container 	176 × 144	300
City 	352 × 288	300

(continued)

**Table 1** (continued)

Sequences	Frame size	No. of frames
Crew 	352 × 288	300
Harbour 	352 × 288	300
Soccer 	352 × 288	300

**Table 2** Comparison of PSNR

Video sequences	PSNR			
	Linear block code [5]	IPCM macroblock [30]	Cost assignment method [24]	Proposed FDCS-RDH
Akyio	39.32	43.72	46.21	50.83
Foreman	36.81	41.61	43.12	48.21
Mobile	34.22	39.21	42.34	46.12
Container	36.90	38.23	41.11	45.12
City	35.77	38.12	41.23	44.89
Crew	37.90	40.11	43.21	47.89
Harbour	32.61	35.23	38.32	42.11
Soccer	36.63	38.21	40.56	43.79
Average	36.27	39.30	42.01	46.12



**Table 3** Comparison of MSE

Video sequences	MSE			
	Linear block code [5]	IPCM macroblock [30]	Cost assignment method [24]	Proposed FDGS-RDH
Akyio	22.06	19.86	20.30	18.74
Foreman	74.34	70.93	72.03	61.99
Mobile	185.62	183.88	182.79	160.12
Container	28.62	26.99	27.17	21.25
City	228.64	214.95	218.46	200.12
Crew	474.33	460.31	465.30	453.87
Harbour	732.14	700.65	724.73	680.12
Soccer	393.22	371.12	383.82	361.21
Average	267.37	256.08	261.82	244.67

## 4.2 MSE Comparison

The comparison of MSE results between the original and the recovered video for different solutions is presented in Table 3. The average MSE in the proposed solution is comparatively low and consistent for different videos. The MSE in the proposed solution is on average 9.27% lower compared to Linear block code, 4.66% lower compared to IPCM macroblock and 7% lower compared to the cost assignment method.

This MSE is low in the proposed solution compared to IPM-based methods due to data embedding in IPCM blocks without any effect on coefficients. The data capacity in one IPCM block is about 1500 bits but to achieve the same, at least 3000 coefficients have to be modified in the IPM method. Compared to the IPCM macroblock method, MSE is lower in the proposed solution due to the selection of better IPCM blocks for data embedding.

## 4.3 Embedding Capacity Comparison

The comparison of embedding capacity results between the original and the recovered video for different solutions is presented in Table 4. The average EC in the proposed solution is 2% higher compared to Linear block, 3.77% higher compared to the Cost assignment method and 6.73% lower compared to the IPCM macroblock method. The EC in the proposed solution is higher compared to the IPM method due to the large capacity for hiding in IPCM compared to IPM coefficients. The proposed solution has lower EC compared to the IPCM macroblock method as the number of IPCM generated in the proposed solution is lower compared to the IPCM macroblock. While

**Table 4** Comparison of embedding capacity (EC)

Video sequences	Embedding capacity (EC)			
	Linear block code [5]	IPCM macroblock [30]	Cost assignment method [24]	Proposed FDGS-RDH
Akyio	9579	12,750	6499	11,045
Foreman	15,057	18,457	17,604	17,611
Mobile	14,357	19,456	13,462	14,166
Container	11,792	15,654	12,453	13,430
City	51,823	55,706	54,161	53,601
Crew	51,822	56,120	50,091	54,115
Harbour	56,973	59,342	55,311	58,172
Soccer	52,956	53,256	52,517	50,258
Average	33,044	36,342	32,762	34,049

the IPCM macroblock method generates IPCM in regular intervals, the proposed work generates it adaptively.

## 5 Conclusion

In this work, a distortion controlled secure Reversible Data Hiding is proposed. The proposed method used the IPCM mode of H.264 videos. The distortion cost of IPCM blocks is measured and blocks with the least distortion cost are selected for data embedding. The proposed solution adaptively controls the rate of generation of the block using fuzzy logic. Due to adaptive control of IPCM generation rate and distortion controlled selection of IPCM block, the proposed solution is able to achieve lower distortion compared to both IPM and IPCM approaches. Distortion measured in terms of MSE is lowest in the proposed solution. Distortion measured in terms of PSNR is highest in the proposed solution. The proposed Solution has at least 8.9% higher PSNR and a constant 4.66% lower MSE. The embedding capacity is higher in the proposed solution compared to IPM methods and 6.73% lower compared to the IPCM macroblock method.

## References

1. Zhang X, Yin Z (2017) Data hiding in multimedia. *Chin J Nat* 39(2):87–95
2. Asikuzzaman M, Pickering MR (2017) An overview of digital video watermarking. *IEEE Trans Circuits Syst Video Technol* 28(9):2131–2153
3. Fridrich J, Pevny T, Kodovsk J (2007) Statistically undetectable jpeg ‘ steganography: dead ends challenges, and opportunities. In: *Proceedings of the 9th workshop on Multimedia & security*. ACM, pp 3–14

4. Shi Y-Q, Li X, Zhang X, Wu H-T, Ma B (2016) Reversible data hiding: advances in the past two decades. *IEEE Access* 4:3210–3237
5. Zhang L, Chen D (2020) The large capacity embedding algorithm for H.264/AVC intra-prediction mode video steganography based on linear block code over Z4. *Multimed Tools Appl* 79:12659–12677
6. Zhu B, Ni J (2018) Uniform embedding for efficient steganography of H.264 video. In: *Proceedings of the 2018 25th IEEE International conference on image processing (ICIP)*, Athens, Greece, 7–10 October 2018, pp 1678–1682
7. Lee W, Sun W (2019) Reversible steganography scheme based on position-recording in DCT coefficients. In: *Proceedings of the 2019 15th International conference on computational intelligence and security (CIS)*, Macao, Macao, 13–16 December 2019, pp 424–428
8. Kim CR, Lee SH, Lee JH, Park J (2018) Blind decoding of image steganography using entropy model. *Electron Lett* 54:626–628
9. Zhang L, Zhao X (2017) An adaptive video steganography based on intra-prediction mode and cost assignment. In: *Proceedings of the digital forensics and watermarking*, Magdeburg, Germany, 23–25 August 2017, pp 518–532
10. Nie Q, Xu B, Feng B, Zhang LY (2018) Defining embedding distortion for intra prediction mode-based video steganography. *Comput Mater Contin* 55–59
11. MaungMaung I, Wong K, Tanaka K (2016) Reversible data hiding methods based on audio and video synchronization in MP4 container. *Int Symp Intell Signal Process Commun Syst (ISPACS)* 2016:1–6. <https://doi.org/10.1109/ISPACS.2016.7824699>
12. Wang W, Lin Y (2015) A tunable data hiding scheme for CABAC in H.264/AVC video streams. In: *International symposium on next-generation electronics (ISNE)*, pp 1–4
13. Lakshmi M, K P, Arjun & N M, Sreenarayanan & Arya, K.A. (2016) Reversible data hiding in videos for better visibility and minimal transfer. *Proc Technol* 25:256–263. <https://doi.org/10.1016/j.protcy.2016.08.105>
14. Yi, Chen, Wang H, Wu H, Liu Y (2019) Reversible video data hiding using zero QDCT coefficient-pairs. *Multimedia Tools Appl* 78. <https://doi.org/10.1007/s11042-019-7635-z>
15. Chen Y, Wang H, Wu H, Liu Y (2018) An adaptive data hiding algorithm with low bitrate growth for H.264/AVC video stream. *Multimedia Tools Appl* 77. <https://doi.org/10.1007/s11042-017-5411-5>
16. Chen Y, Wang HX, Wu HZ, Chen YL, Liu Y (2018) A data hiding scheme with high quality for H.264/AVC video streams. In: *International conference on cloud computing and security*, Springer, Cham, pp 99–110
17. Fallahpour M, Shirmohammadi S, Bhanbari M (2015) A high capacity data hiding algorithm for H.264/AVC video. *Secur Commun Netw* 8(16):2947–2955
18. Liu Y, Chen L, Hu M, Jia Z, Jia S, Zhao H (2016) A reversible data hiding method for H.264 with shamir's (t, n)-threshold secret sharing. *Neurocomputing* 188:63–70
19. Kim H, Kang S (2018) Genuine reversible data hiding technology using compensation for H.264 bitstreams. *Multimedia Tools Appl* 77(7):8043–8060
20. Xu DW, Wang RD (2016) Two-dimensional reversible data hiding-based approach for intra-frame error concealment in H.264/AVC. *Signal Process Image Commun* 47:369–379
21. Xu DW, Wang RD, Shi YQ (2014) An improved reversible data hiding-based approach for intra-frame error concealment in H.264/AVC. *J Visual Commun Image Representation* 25(2):410–422
22. Nguyen D-C, Nguyen T-S, Chang C-C, Hsueh H-S, Hsu F-R (2018) High embedding capacity data hiding algorithm for H.264/AVC video sequences without intraframe distortion drift. *Secur Commun Netw* 2018(2029869):11
23. Xu DW, Wang RD, Wang JC (2012) Prediction mode modulated data-hiding algorithm for H.264/AVC. *J Real-Time Image Process* 7(4):205–214
24. Chen Y et al (2021) Adaptive video data hiding through cost assignment and STCs. *IEEE Trans Depend Secure Comput* 18(03):1320–1335
25. Yao Y, Zhang W, Yu N (2016) Inter-frame distortion drift analysis for reversible data hiding in encrypted H.264/AVC video bitstreams. *Signal Process* 128:531–545

26. Ma X, Li Z, Tu H, Zhang B (2010) A data hiding algorithm for H.264/AVC video streams without intra-frame distortion drift. *IEEE Trans Circuits Syst Video Technol* 20(10):1320–1330
27. Wang T, Wang H, Li Y (2016) Reversible data hiding with low bit-rate growth in H.264/AVC compressed video by adaptive hybrid coding. *Proc Int Conf Cloud Comput Secur* 48–62
28. Baptista MS (1998) Cryptography with chaos. *Phys Lett A* 240(1–2):50–54
29. Li Q, Qian G (2017) A new image encryption algorithm based on chaotic maps. In: *Proceedings of the 9th International conference on signal processing systems (ICSPS 2017)*. Association for Computing Machinery, New York, NY, USA, pp 65–69
30. Kapotas SK, Skodras AN (2009) Real time data hiding by exploiting the IPCM macroblocks in H.264/AVC streams. *J Real-Time Image Proc* 4:33–41

# A Method for Improving Efficiency and Security of FANET Using Chaotic Black Hole Optimization-Based Routing (BHOR) Technique



Mayank Namdev, Sachin Goyal, and Ratish Agarwal

**Abstract** The power and competence of drones are exploited in numerous areas of applications like surveillance, healthcare, and disaster management in Flying Ad-hoc Network (FANET). The incongruous routing and incompetent flight period are major issues in drones due to minimum stability and a small amount of battery capability. Then, the communication is unproductive and ineffectual between drones, which decreases the performance of FANET by increasing the cost of message delivery. Therefore, a Chaotic Black Hole Optimization-based Routing (Chaotic BHOR) technique is implemented to improve the communication efficiency and security among drones by reducing the energy cost and enhancing the key contribution over FANET. The MATLAB 2021a tool is used to implement the Chaotic BHOR, and outcomes illustrate the superior effectiveness of Chaotic BHOR on the basis of end-to-end delay, packet delivery ratio, and throughput and power expenses against previous techniques like OLSR, MP-OLSR, and ML-OLSR-PMS.

**Keywords** Black hole optimization · Chaotic · Drones · FANET · Packet delivery ratio · Throughput

## 1 Introduction

The assorted and escalating drones [1–3] are presented to offer pragmatic and sturdy communiqué in FANET. A miscellaneous organization symphony is accomplished in obtaining a drone-based packet communication utilizing for numerous persistent usages such as healthcare, surveillance, security, and disaster management reliant

---

M. Namdev (✉)

Department of CSE, University Institute of Technology, RGPV, Bhopal 462023, India  
e-mail: [mayank.namdev@gmail.com](mailto:mayank.namdev@gmail.com)

S. Goyal · R. Agarwal

Department of IT, University Institute of Technology, RGPV, Bhopal 462023, India  
e-mail: [sachingoyal@rgtu.net](mailto:sachingoyal@rgtu.net)

R. Agarwal

e-mail: [ratish@rgtu.net](mailto:ratish@rgtu.net)

on different architectural works [4]. The video and audio data packets of Internet of Things (IoT) devices are transmitted through Unmanned Aerial Vehicles (UAVs) easily and effectively by utilizing 5G techniques. The numerous amounts of IoT instruments created the issues like overhead, privacy, and sturdiness in on-demand path establishment strategy for several environments [5]. These types of issues of UAVs are diminished by using the 3D-SWAP method for efficient packet transmission with low cost and improving the packet delivery ratio in FANET [6]. The load balancing is another important concern in FANET due to high mobility with reduced battery power. It is reduced by enhancing the coverage area of the base station in on-demand routing and also improving the packet transmission capability of drones in minimum delivery time [7, 8]. The Mobile Ad-hoc Network (MANET) uses the clustering of drones to enhance the accessibility and throughput of the network by transmitting the packets with negligible delay and the least packet drop ratio [9, 10]. The cluster head is elected by some selection procedure in each cluster, and after that, the cluster head has the responsibility to receive the packets from cluster members and transmit the packets to another cluster head or base station [11, 12]. The selection of cluster head and members is performed by using the drone's properties (stability, velocity, and energy) and applying various optimization techniques like gray wolf optimization and particle swarm optimization [13].

The former key concerns of routing among drones such as explosive and arbitrary mobility are abridged by introducing the competent routing schemes in the course of optimization algorithms in FANET [14]. The various routing schemes like Optimized Link State Routing (OLSR) are used for finding the shortest paths from source to destination drones in FANET to minimize the packet slump in terms of stability and number of drones [15]. The shortest path is selected from multiple paths among drones in OLSR at every instant in Multi-Path OLSR (MP-OLSR), and neighbors' participation is evaluated to know about neighbors of each drone, which are further utilized for communication in Fog computing [16, 17]. The prediction of the shortest path having minimum energy consumption is also used in OLSR to reduce the packet transmitting time in several dimensional areas. The mobility-based routing is well organized for selecting the IoT instruments to transmit the video and audio data packets with the least cost consumption [18]. Another method Mobility and Load aware OLSR with First In First Out (ML-OLSR-FIFO) is introduced to combine the features of drones like mobility, first in first out concern, and load sharing, and this method evaluates the outcomes in terms of data precedence and scheduling schemes [19]. Bird Swarm Algorithm-based Routing (BSAR) is used to reduce the limitations of drones like improper routing and flight time by enhancing the performance in terms of energy [20].

In the previous discussion, various features of drones such as velocity, power, stability, and neighbors involvement are well utilized for efficient communication through multiple types of OLSR routing in FANET. However, entire drone features are not exploited to achieve the efficient and optimal communication in FANET. So, a Chaotic Black Hole Optimization-based Routing (Chaotic BHOR) technique is implemented to reduce the energy cost and enhance the key contribution for improving the communication efficiency and secrecy among drones in FANET. The

effectiveness of Chaotic BHOR is examined on the basis of end-to-end delay, packet delivery ratio, and throughput and power expenses against previous techniques like OLSR, MP-OLSR, and ML-OLSR-PMS.

## 2 Proposed Chaotic Black Hole Optimization-Based Routing (Chaotic BHOR) Technique

The proposed Chaotic BHOR is applied to generate the best routing scheme utilizing the optimal link selection of drones in FANET. Initially, the number of keys of each link is evaluated for security purposes, and the energy cost of each link is obtained using link involvement in FANET. Then, the Chaotic BHOR is applied to the Absolute Link Involvement (ALI) combining the Link Energy Cost (LEC) and Key Link Exploitation (KLE) to evaluate the optimal links of drones. At last, the optimal drones for the best routing are obtained with the help of optimal ALI values. These optimal drones are used to transmit the maximum packets for improving the performance of FANET.

### 2.1 Key Link Exploitation

One key is used to encrypt one data packet and a drone sends that encrypted data packet to another neighboring drone via a link. It means a drone sends numerous encrypted data packets to neighboring drone after performing encryption through numerous keys. These keys are obtained by evaluating Key Link Exploitation (KLE) through Link's Keys (LK) using Eq. (1). The highest KLE values of the highest links are used to select the best drones.

$$KLE_{e,f} = \begin{cases} \sum_{k=1}^{KEYS} LK_{e,f} & \text{if } k \in D_e, D_f \\ 0 & \text{Otherwise} \end{cases} \quad (1)$$

where

- $KLE_{e,f}$  Key Link Exploitation between drones  $D_e$  &  $D_f$ .
- $LK_{e,f}$  Link's Keys between drones  $D_e$  &  $D_f$ .
- KEYS Total Keys.

## 2.2 Link Energy Cost

Firstly, the contribution of each and every link is evaluated by exchanging the pairs of HELLO-Acknowledgment packets between neighboring drones under the Transmission Ranges ( $R_T$ ) of each drone. For this purpose, a Drone Contribution (DC) is obtained to evaluate the Total Neighbors (TN) of every drone by using Eq. (2).

$$DC_e = |TN(e)| = \sum_{e \neq f} \{dst(e, f) < R_T\} \quad (2)$$

where

$dst(e, f)$  Distance between drones  $D_e$  &  $D_f$ .  
 $|TN(e)|$  Total Neighbors of drone  $D_e$ .  
 $DC_e$  Drone contribution of drone  $D_e$ .

Further, the Link Contribution (LC) is evaluated for each and every link by dividing the  $DC_e$  value from 1 for obtaining similar distribution among links (Eq. 3).

$$LC_{e,f} = \frac{1}{DC_e} \quad (3)$$

where

$LC_{e,f}$  Link contribution of a link between drones  $D_e$  &  $D_f$ .

At last, Link Energy Cost (LEC) is obtained by calculating the power consumption of each and every link to exchange the data. A drone is well suitable for communication having the highest links with the lowest LEC value (Eq. 4).

$$LEC_{e,f} = \frac{LC_{e,f}}{AC_{e,f}} \quad (4)$$

where

$LEC_{e,f}$  Link energy cost of a link between drones  $D_e$  &  $D_f$ .  
 $AC_{e,f}$  Accessing Cost of a link between drones  $D_e$  &  $D_f$ .

## 2.3 Absolute Link Involvement

The Absolute Link Involvement (ALI) is generated for each link by exploiting two factors ( $LEC_{e,f}$  &  $KLE_{e,f}$ ) (Eq. 5).

$$Maximize ALI_{e,f} = (w_1 \times KLE_{e,f}) + \left( w_2 \times \frac{1}{LEC_{e,f}} \right) \quad (5)$$



where

$w_1$  &  $w_2$  Weight terms ( $w_1 + w_2 = 1$ ).

$ALI_{e,f}$  Absolute link involvement of a link between drones  $D_e$  &  $D_f$ .

### 2.4 Black Hole Optimization (BHO) Technique

After getting Eq. (5) for every link in FANET, the BHO is applied to generate the maximum ALI values for gaining the optimal links from existing links. A field of distance-time (x, y, t) through a robust and dominant gravitational area is notated as a black hole, where no one could go off of it. An incident perspective is an algebraically described area regarding a black hole to facilitate outcomes the spot of no go back. If a little gets close to the horizon, then it should be captivated into a black hole and perpetually fade away. The black hole has a few basic features like mass, charge, and momentum, in which the general black hole has no charge and momentum; yet, greater gravitational power is achieved by a greater mass black hole.

The BHO starts with principal candidate solutions population for a quandary of optimization, in which the population is selected by using some chaotic functions. The best candidate is allocated as a black hole, and all remaining candidates are normal stars in each generation. The black hole applies to cart the stars closer to him after the primary generation. The black hole soaks the extremely closer stars, and the latest candidate solution is illogically generated for initiating the next exploration subsequent to situate the latest star to exploration area.

#### Fitness Value Calculation

In the first stage, the candidate solutions (stars) population.

$P(S) = \{S_1^t, S_2^t, S_3^t, \dots, S_N^t\}$  is generated by using some chaotic function and positioned in the exploration area for optimization. The complete fitness of populations is generated by utilizing Eqs. (6) and (7).

$$Ftn_j = \sum_{j=1}^{L^P} F(P(t)) \tag{6}$$

$$Ftn_{BH} = \sum_{j=1}^{L^P} F(P(t)) \tag{7}$$

where

$Ftn_j$  Value of  $j$ th star fitness.

$Ftn_{BH}$  Value of black hole fitness.

$F(P(t))$  Fitness functions for population (time t).

$L^P$  Population length.

The calculation of the population is performed and the best candidate having the best fitness  $Ftn_j$  is allocated as a black hole and the remaining are normal stars. The black hole begins to enchant the closer stars and the remaining stars embark on going to closer the black hole subsequent to the primary stage.

**Black Hole Attraction Speed of Stars**

The black hole embarks on enchanting the stars in the area of it and the remaining stars. The black hole begins to fascinate the stars in the region of it and other stars embark on going closer to the black hole. The black hole attraction speed is generated by utilizing Eq. (8).

$$S_j(t + 1) = S_j(t) + random \times (S_{BH} - S_j(t)) \tag{8}$$

where

- $j$  1, 2, . . . . . ,  $N$ .
- $S_j(t)$  and  $S_j(t + 1)$   $j$ th star location at iteration  $t$  and  $(t + 1)$ .
- $S_{BH}$  Black hole location in the exploration area.
- $random$  uninformed value within  $(0, 1)$ .
- $N$  total candidate solution (Stars).

A star possibly will achieve, at least cost, a location comparable to a black hole nevertheless stirring close to the black hole. Consequently, the black hole changes to that latest star location and vice versa. The generation will be sustained by means of the latest black hole, and the remaining stars embark on enchanting closer to that next black hole location.

**Opportunity of the Incident Horizon Channel for Stirring Stars**

The opportunity of the incident horizon channel for stirring stars is developed to generate a higher optimal solution for optimization strategy. All candidate results (star) route the black hole’s incident horizon, it will be distributed by a black hole and a candidate destroys by shocking during the black hole’s complete instance of time. Consequently, the latest candidate is occupied and detached illogical in exploration area to offer next exploration in population by means of steady candidate results. The subsequent generation performs afterward the enchanting of entire stars. The horizon radius (Rad) is obtained by utilizing Eq. (9).

$$Rad = \frac{Ftn_{BH}}{\sum_{j=1}^N Ftn_j} \tag{9}$$

If distance between a candidate and black hole is getting lesser than horizon radius (Rad), then the candidate is emaciated and the latest candidate is obtained and disseminated illogically in the exploration area.

**2.4.1 Chaotic Function**

The chaotic function is exceedingly stooped on early circumstances and well exploited for casual number seed over the logistic system. The chaotic function is represented by using Eq. (10).

$$\varphi_{m+1} = \tau \times \varphi_m \times (1 - \varphi_m) \tag{10}$$

where

- $\varphi$  Variables ( $\varphi \in [0, 1]$ ,  $m = 0, 1, 2, 3, \dots$ ).
- $\tau$  Constants within  $[1, 4]$ .

The population of stars is subscribed by utilizing chaotic function (Eq. 10) for improving the efficiency of the BHO technique with total exploitation of the solution area.

**2.5 Absolute Drone Involvement**

Once applying Chaotic BHOR, the optimal ALI values are obtained for every link in FANET. After that, all ALI values of entire links of drone  $D_e$  are combined to obtain a matrix  $ALI_e$  for each drone in FANET (Eq. 11).

$$ALI_e = [ALI_{e,f}]_{1 \times DC_e} \tag{11}$$

where

$ALI_e$  Optimal ALI values of a drone  $D_e$ .

At last, the Absolute Drone Involvement (ADI) is calculated by utilizing Eq. (12).

$$ADI_e = \frac{\sum_{f=1}^{DC_e} ALI_{e,f}}{DC_e} \tag{12}$$

where

$ADI_e$  Optimal ADI value of a drone  $D_e$ .

Similarly, the ADI values for every drone are obtained and the best drone is selected based on optimal ADI values for optimal, secure, and cost-efficient routing in FANET.

## 2.6 Working and Operation of Chaotic BHOR Technique

The working steps of Chaotic BHOR are as follows:

- Step 1 A Key Link Exploitation (KLE) is evaluated for a link by utilizing Eq. (1).
- Step 2 A Link Energy Cost (LEC) is evaluated for a link by utilizing Eq. (2) to Eq. (4).
- Step 3 The Absolute Link Involvement (ALI) is generated for a link by combining KLE and LEC with the help of Eq. (5).
- Step 4 Step 1 to step 3 are continuously performed for all links in FANET and ALI equation for all links is obtained, which is again used by Chaotic BHOR as the objective function.
- Step 5 The chaotic function (Eq. 10) is used to generate the initial population of stars in the BHO technique.
- Step 6 The BHOR is initiated, where the BHO (Eqs. 6–9) is functionalized over objective function ALI (Eq. 5) to generate optimal links.
- Step 7 The Absolute Drone Involvement (ADI) values (Eq. 11 to 12) for every drone are evaluated by using ALI values of links and the best drone is elected based on optimal ADI values for optimal, secure, and cost-efficient routing for communiqué in FANET.

## 3 Result and Analysis

The proposed Chaotic BHOR is examined in MATLAB 2021a tool to evaluate the competence of Chaotic BHOR in FANET. The competence of Chaotic BHOR is calculated depending on some circumstances (Table 1).

The total drones (30, 60, 90, and 120) are scattered above 1000 m × 1000 m area, 100–500 m transmission range, and 0–40 m/s velocity range in FANET. The Chaotic BHOR processes above 200 generations on the basis of the 200 population length of Chaotic BHOR.

The proposed Chaotic BHOR is evaluated depending on parameters like end-to-end delay, packet delivery ratio, and throughput and power expenses against previous techniques like OLSR, MP-OLSR, and ML-OLSR-PMS.

**Table 1** Examination circumstances

Examination circumstances	Values
Total drones	30, 60, 90, 120
Network area	1000 m × 1000 m
Velocity range of drones	0–40 m/s (random)
Transmission range	100–500 m
Population length	200
Number of generations	200

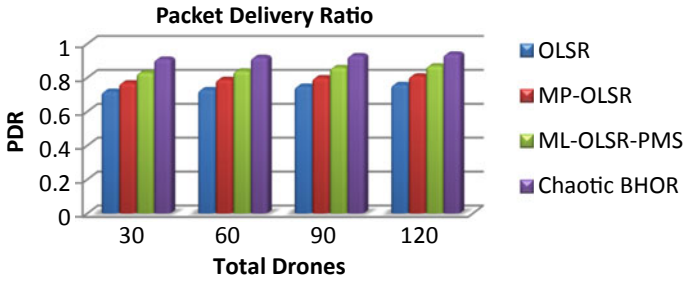


Fig. 1 Packet delivery ratio versus total drones (1000 m × 1000 m network area)

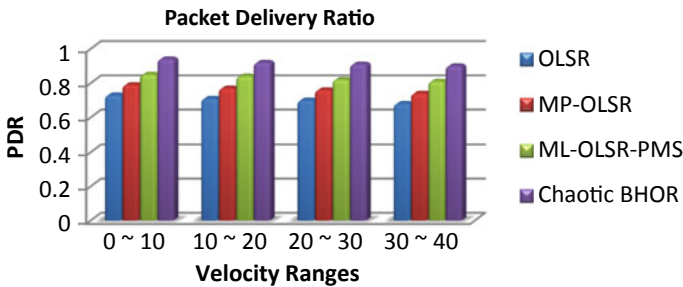


Fig. 2 Packet delivery ratio versus velocity ranges (1000 m X 1000 m network area)

### 3.1 Packet Delivery Ratio (PDR)

The PDR is obtained by the division of sending packets arriving at the base location in a packet sent by whole drones by means of the best drones in FANET.

It explains that Chaotic BHOR generates the utmost PDR of 94% and 94% dependent on the total drones and velocity ranges for 1000 m × 1000 m (Figs. 1 and 2) network area as examined in opposition to earlier methods OLSR (76% and 73%), MP-OLSR (81% and 79%), and ML-OLSR-PMS (87% and 85%). The PDR is improved by enhancing the total drones because of additional drones to packet sent. The PDR is reduced by enhancing the velocity ranges because of increasing the chance of packet drop with high mobility drones.

### 3.2 End-To-End Delay (EED)

The EED is obtained by combining the route recognition and routing period describing the association and potency of FANET. The optimal routing-based communication is achieved by minimizing the EED values for packet transmission in FANET.

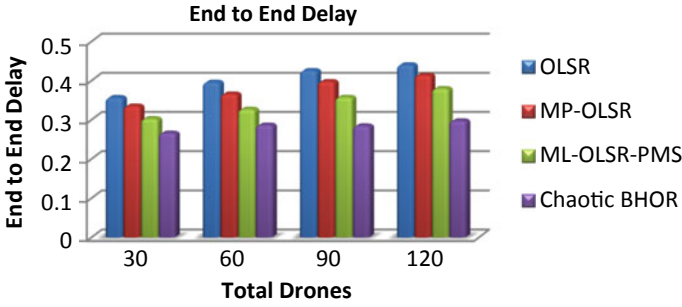


Fig. 3 End-to-end delay versus total drones (1000 m × 1000 m network area)

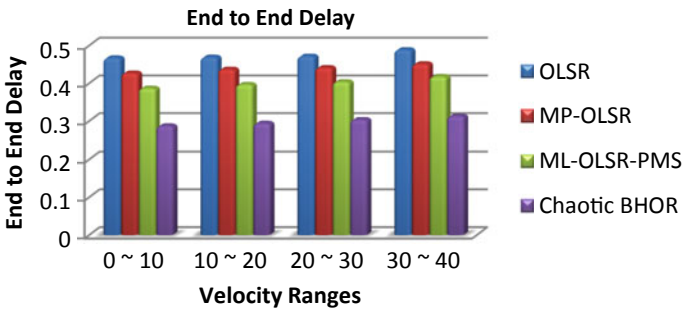


Fig. 4 End-to-end delay versus velocity ranges (1000 m × 1000 m network area)

It explains that Chaotic BHOR generates minimum EED (Seconds) of 0.2657 and 0.2867 depending on total drones and velocity ranges for 1000 m × 1000 m (Figs. 3 and 4) network area as examined in opposition to earlier methods OLSR (0.3567 and 0.4668), MP-OLSR (0.3348 and 0.4267), and ML-OLSR-PMS (0.3027 and 0.3867). The EED is improved by increasing the total drones because of additional drones to packet sent. The EED is improved by enhancing the velocity ranges because of the high mobility of drones increasing the distance between drones.

### 3.3 Power Expenses

The power expense is explained as the consumed energy by drones to send and exchange the packets of information through optimal routing in FANET. The survival period and constancy of drones are improved by decreasing the power expenses.

It explains that Chaotic BHOR generates the minimum power expenses (Joules) of 0.967 and 4.102 depending on total drones and velocity ranges for 1000 m × 1000 m (Figs. 5 and 6) network area as examined in opposition to earlier methods OLSR (6.3 and 14.36), MP-OLSR (4.6 and 12.68), and ML-OLSR-PMS (2.8 and 7.3). The

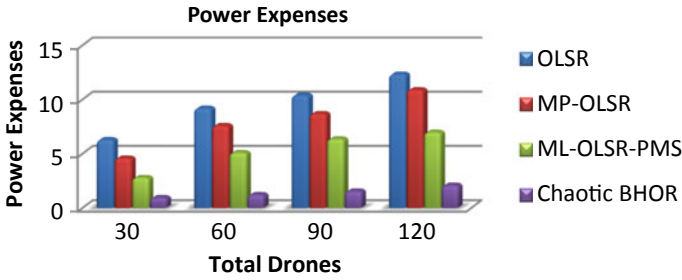


Fig. 5 Power expenses versus total drones (1000 m × 1000 m network area)

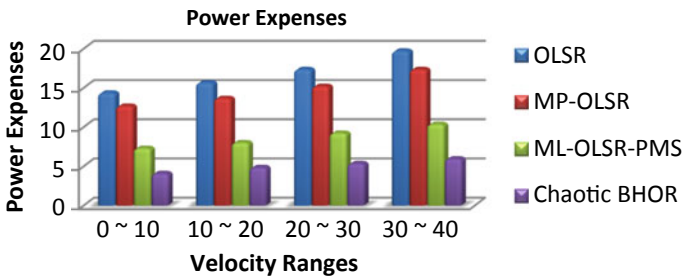


Fig. 6 Power expenses versus velocity ranges (1000 m × 1000 m network area)

power expense is improved by increasing the total drones because of additional drones to power consumption. The power expense is improved by enhancing the velocity ranges due to the minimum stability of drones increasing the distance between drones.

### 3.4 Throughput

The Throughput is obtained by the division of sending packets arriving at the base location, per unit time, where the packets are sent by whole drones by means of the best drones in FANET.

It explains that Chaotic BHOR generates the utmost throughput of 10,867 and 10,854 dependent on the total drones and velocity ranges for 1000 m × 1000 m (Figs. 7 and 8) network area as examined in opposition to earlier methods OLSR (6784 and 6896), MP-OLSR (7215 and 7459), and ML-OLSR-PMS (8036 and 8362). The throughput has improved by enhancing the total drones because of additional drones to packet sent. The throughput is reduced by enhancing the velocity ranges because of increasing the chance of packet drop with high mobility drones.

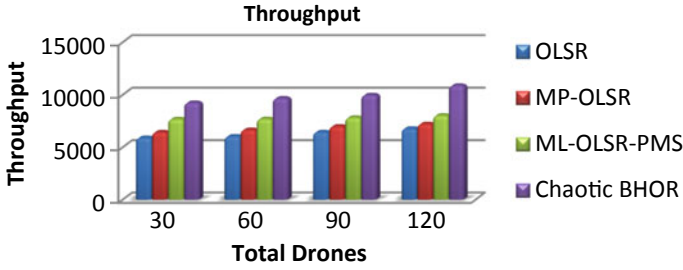


Fig. 7 Throughput versus total drones (1000 m × 1000 m network area)

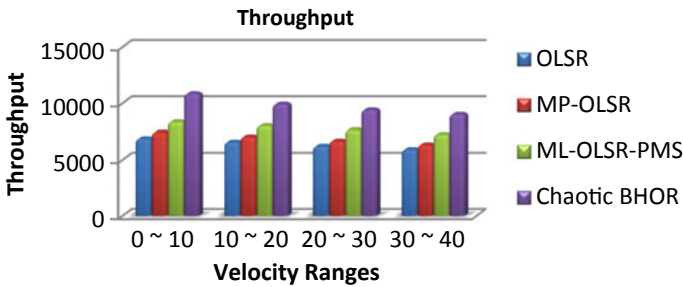


Fig. 8 Throughput versus velocity ranges (1000 m × 1000 m network area)

### 4 Conclusion

Several application fields like surveillance, healthcare, and catastrophe supervision are oppressed by controlling and proficiency of drones in FANET. The least amount of stability and a small amount of battery potential are key concerns of drones, which generates the incompatible routing and inept flight period. Then, the routing-based communication is infertile and unimpressive between drones, which diminishes the efficiency of FANET by enhancing message delivery cost. Here, a Chaotic BHOR technique is introduced to improve the communication competence and protection among drones by dropping the cost of energy and increasing the key contribution over FANET. The Chaotic BHOR is implemented to MATLAB 2021a tool, and outputs explain that the Chaotic BHOR generates 6%, 12% & 18% superior PDR value, 12%, 20% & 24% superior end-to-end delay value, 66%, 78% & 84% superior power expenses, and 35%, 50% & 60% superior throughput as compared to ML-OLSR-PMS, MP-OLSR, and OLSR, respectively.



## References

1. Sharma V, Sabatini R (2016) UAVs Assisted delay optimization in heterogeneous wireless networks. *Commun Lett IEEE*, pp 1–5
2. Khan NA, Jhanjhi NZ, Brohi SN, Nayyar A (2020) Emerging use of UAV's: secure communication protocol issues and challenges. Elsevier, pp. 37–55
3. Pandey A, Shukla PK, Agrawal R (2020) An adaptive flying ad-hoc network (FANET) for disaster response operations to improve quality of service (QoS). *Mod Phys Lett B* 34(10):1–28
4. Ahn T, Seok J, Lee I, Han J (2018) Reliable flying IoT networks for UAV disaster rescue operations. *Mob Inf Syst Hindawi* 1–13. <https://doi.org/10.1155/2018/2572460>
5. Erdelj M, Uk B, Konam D, Natalizio E (2018) From the eye of the storm: an IoT ecosystem made of sensors, smartphones and UAVs. *Sensors*, MDPI 18:1–20. <https://doi.org/10.3390/s18113814>
6. Ferrera E, Alcántara A, Capitán J, Castaño AR, Marrón PJ, Ollero A (2018) Decentralized 3D collision avoidance for multiple UAVs in outdoor environments. *Sensors*, MDPI 18:1–20. <https://doi.org/10.3390/s18124101>
7. Hu B, Wang C, Chen S, Wang L, Yang H (2018) Proactive coverage area decisions based on data field for drone base station deployment. *Sensors*, MDPI 18:1–14. <https://doi.org/10.3390/s18113917>
8. Namdev M, Goyal S, Agrawal R (2021) An optimized communication scheme for energy efficient and secure flying ad-hoc network (FANET). *Wirel Personal Commun* 1–16. <https://doi.org/10.1007/s11277-021-08515-y>
9. Huang J, Fan X, Xiang X, Wan M, Zhuo Z, Yang Y (2016) A clustering routing protocol for mobile ad hoc networks. *Math Prob Eng Hindawi* 1–10. <https://doi.org/10.1155/2016/5395894>
10. Chaturvedi A, Agrawal R, Goyal S (2019) Classifying various clustering techniques for MANET. *Int J Sci Technol Res* 8(10):2225–2230
11. Ganesan R, Raajini XM, Nayyar A, Sanjeev I kumar P, Hossain E, Ertas AH (2020) BOLD: bio-inspired optimized leader election for multiple drones. *Sensors MDPI* 20:1–20
12. Valentino R, Jung WS, Ko YB (2018) A design and simulation of the opportunistic computation offloading with learning-based prediction for unmanned aerial vehicle (UAV) clustering networks. *Sensors*, MDPI 18:1–14. <https://doi.org/10.3390/s18113751>
13. Fahad M, Aadil F, Rehman ZU, Khan S, Shah PA, Muhammad K, Lloret J, Wan H, Lee JW, Mehmood I (2018) Grey wolf optimization-based clustering algorithm for vehicular ad-hoc networks. *Comput Electr Eng Elsevier* 1–19
14. Zheng X, Qi Q, Wang Q, Li Y (2017) An adaptive density based routing protocol for flying Ad Hoc networks. In: 2nd International conference on materials science, resource and environment engineering (MSREE) AIP Conf Proc 1–8. <https://doi.org/10.1063/1.5005315>
15. Lenovo AV, Litvinov GA (2018) Simulation-based packet delivery performance evaluation with different parameters in flying ad-hoc network (FANET) using AODV and OLSR. In: International conference information technologies in business and industry, IOP Conf Series J Phys 1–16. <https://doi.org/10.1088/1742-6596/1015/3/032178>
16. Yi J, Adnane HA, David S, Parrein B (2018) Multipath optimized link state routing for mobile ad hoc network. *HAL* 1–17
17. Radu D, Cretu A, Parrein B, Yi J, Avram C, Astilean A (2018) Flying ad hoc network for emergency applications connected to a fog system. *HAL* 1–13. <https://www.hal.archives-ouvertes.fr/hal-01763827>
18. Wei Z, Liu X, Han C, Feng Z (2018) Neighbor discovery for unmanned aerial vehicle networks. *IEEE Access* 6:68288–68301. Digital Object Identifier. <https://doi.org/10.1109/ACCESS.2018.2871132>
19. Li J, Chen M, Dai F, Wang H (2018) Prioritizing-based message scheduling for reliable unmanned aerial vehicles ad hoc network. *Int J Perform Eng* 14(9):2021–2029. <https://doi.org/10.23940/ijpe.18.09.p10.20212029>
20. Namdev M, Goyal S, Agarwal R (2021) A BSA based communication strategy for energy efficient flying ad-hoc network (FANET). *Des Eng (Toronto)* (7):6890–6900

# Machine Learning Techniques for Intrusion Detection System: A Survey



Saby Singhal and Pradeep Yadav

**Abstract** The Intrusion Detection System (IDS) is a program that tracks a single or a malicious computer network (attacks) to copy, encrypt, or modify network protocols. Any of the strategies used in the latest IDS cannot cope with the complicated and nuanced existence of computer network cyberattacks. There will also be better identification rates, decreased false warning rates, and reasonable measurement costs and coordination costs for appropriate adaptive approaches, like the different methods of machine learning. Different methods were used in recent works to enhance efficiency. The principal function of the intrusion detection device is to examine big network traffic info. In this paper, we have studied an intelligent intrusion system taking the types of intrusion detection systems and also defined the limitations based on traditional IDS. It can also deal with intrusions by using ML algorithms. Various techniques of ML are described in this paper.

**Keywords** Intrusion detection · IDS · Classification · ML · ML techniques

## 1 Introduction

In the case of ID, activities that arise in a computer device or network are continuously tracked, evaluated for potential incidences, and often banned unauthenticated access [1]. This is usually achieved by extracting information dynamically from different applications and network channels and then reviewing the details for potential security vulnerabilities.

IDSs are monitoring equipment that have been attached to the protection wall to deter malicious intrusion on a device. A network administrator's intrusion detection system is an invaluable method since it will be difficult to examine every second the vast amounts of packets that travel across current networks without such a program. After more than 30 years of research in intrusion detection systems, more study on

---

S. Singhal · P. Yadav (✉)  
Institute of Technology and Management, Gwalior, India  
e-mail: [er.pradeepyadav0610@gmail.com](mailto:er.pradeepyadav0610@gmail.com)

the precision of the detection is still available. Also, versions of existing attacks and new attacks will be identified without the device [2].

Classification is a data mining method for determining group membership in data instances. While different strategies are available for machine learning, classification is the most frequently used. Classification is a valued role in machine learning, especially in the future. Classification, while well established in machine learning, suffers from problems such as managing missing details. At both preparation and classification times, missing values in the dataset may trigger issues. Some possible explanations for lost records are given in [3]; non-registration due to misinterpretation, data recognized at the time of entry meaningless, data exclusion due to divergence from other recorded details, and malfunctions of equipment.

ML is a theoretical area that focuses systematically on the theory, results, and properties of education systems and algorithms. It draws on several diverse areas of ideas: artificial intelligence, optimization theory, knowledge processing, psychology, cognitive sciences, optimal controls, and many other branches in research, science, and mathematics. It is a strongly interdisciplinary area. With the introduction of machine learning in a broad variety of applications, it has reached almost every research field, which has a significant influence on technology and society. A range of issues has been tackled, including advice engines, recognition mechanisms, IT and data mining, and independent testing systems.

The area of machine learning is usually categorized into three sub-domains: supervised, unsupervised learning, and reinforcement. Briefly, supervised learning involves the training of inputs and target outputs with labeled knowledge. Unsupervised learning does not need labeled training details, in comparison to regulated learning where the environment supplies only inputs without defined goals. Reinforcement learning allows input by experiences in an existing world to be understood. Based on these three paradigms, several theory structures and application frameworks for data tasks have been suggested [4].

## 2 Intrusion Detection

The Internet is now the most powerful platform and intelligence center in the world today. The Internet can be used as one of the key research and academic elements. Data must also remain safe across the Internet. Security on the Internet is also one of the main worries. Given that the Internet is compromised by numerous threats, the implementation of a mechanism to secure these data and people accessing these data is highly significant. Consequentially, IDS is an innovation that fills the need. IDS is designed to deter unauthorized attempts by Network Operators. The intrusion detection device has now been a core component of protection management. Intrusion detector detection technology identifies and records intrusion or network assault. IDS will identify and prevent malicious attacks on the Network and preserve regular operations throughout any malicious outbreak [5].

Signature- or anomaly-based intrusion detection may either be implemented on a stateless (per packet) or stateful (per-flow) basis. Most current IDSs are static, as the flow is “background” while packet analysis is unpredictable. The researchers must assess which approach is ideally adapted.

### 3 Intrusion Detection System

An ID is present to stop impacting the company in this case. They monitor network traffic and issue warnings in case of problems [6].

Any unauthorized activity causing damage to the information system may be defined as an intrusion. This ensures that any action which will endanger the confidentiality, integrity, or availability of the details shall be treated as an intrusion. For instance, actions that do not respond to authorized users of computer systems are considered an intrusion. An IDS is a software or hardware system that detects malicious attempts to enforce device stability on computer networks (Liao et al. 2013a). The purpose of IDS is to distinguish various forms of malicious network traffic and device use which a conventional firewall cannot recognize. This is important to ensure strong security against acts that threaten information systems’ functionality, credibility, or confidentiality [7].

#### 3.1 Types of IDS

IDS may use multiple strategies for detecting suspicious activities. It can be separated into big sections [6]:

- **Signature-based intrusion detection:** The input of traffic is contrasted with a current archive of established attack trends established as signatures. It’s tough to spot new threats. New names are regularly published by the manufacturers of the devices (Anti-virus-related software).
- **Anomaly-based intrusion detection:** It uses statistical data to construct a simple application at various intervals of the networks. To prevent unknown threats, they were added. To construct a model that simulates everyday tasks, this method utilizes deep learning.
- **Network intrusion detection (NIDS):** It is a technical method to track all network traffic (single or different locations).
- **Host intrusion detection (HIDS):** It works on all network computers that are connected to the organization’s Internet/intranet. You may detect traffic malicious from the inside (for example, when malware is attempting to propagate from a host in the enterprise to other systems).

### 3.2 *Limitations of Traditional IDS*

- Rather than the number of false alarms elevated are also actual threats. This also contributes to actual risks.
- Noise will dramatically reduce the IDS 'capability by producing large false alarm rates.
- To keep up with the latest risks, continuous security upgrades are mandatory for signature-based IDS.
- IDS tracks the whole network, such that the network administrators are exposed to the same threats. The IDS could fail through protocol-based attacks.
- Network IDS can identify only network irregularities reducing the amount of threat it is likely to track.
- As inbound and outbound traffic passes via Network IDS, a bottleneck can be formed.
- Any attacks which modify audit logs threaten the integrity of HIDS (Host IDS) that relies on audit logs [6].

### 3.3 *Survey on Intrusion Detection Method*

ID approaches are such as pattern matching, state complete pattern matching, and protocol decode-based research.

Different intrusion detection techniques are usable. The above are some of the approaches implemented [2]:

- **Pattern Matching:** The PM is based on a single packet's quest for a set byte series. It is a fairly linear but concise approach, as its name implies. In certain instances, the pattern is only matched if the suspicious packet is connected to one operation or more generally, to/from a port. If the packet is IPv4 and TCP and the target port is 2222 and the payload includes the string "foo" fire the alarm, the form of a signature based on a basic pattern matching method might be as follows. Naturally, this explanation is very plain, but the deviations from this stage are still simplistic.
- **State Full Pattern Matching:** A more complex approach is the total study of the sequence. This approach for the creation of signatures adds to the idea that a source of the network can fit within a system state, as it consists of more than single atomic packets. This ensures that applications performing these signature studies need to take into consideration packet delivery orders in the TCP stream and need to manage similar trends around packet borders [5].
- **Protocol Decode-Based Analysis:** Decode-based signatures are intelligent extensions in several respects to define complete pattern matches. This type of signature is applied in the same way that the client or server during the communication decodes the different components. The IDS applies rules established by the RFCs to assess breaches when the elements of the protocol are decided. In some cases,

pattern matches in a particular protocol field are identified, and in some instances, specialized techniques are required to take into account variables such as field duration or several arguments.

- **Fuzzy Clustering for IDS:** The fundamental principle of our intrusion detection model is to define attacks in a semantically rich language such as DAML as Ontological instances. This ontology collects attacks on details, such as the part in the framework and the effect on the position of the assailant. Thus, our intrusion detection model is two stages, and the emphasis of this form of ontology has been established under conferral. Initial data mining approaches to examine data streams that catch and identify anomalous behavior, device and Network conditions, and second or high phase reasons using data representative of the anomaly identified as an ontology case. One approach to constructing these data stream models is by utilizing a fuzzy clustering that includes various entity matrices for clustering. The objective function is focused on the collection of representative artifacts from features that decrease the overall fuzzy difference within a cluster.

## 4 Machine Learning

ML is a framework that can be applied to learning from past practices to maximize potential success (in this case previous data). Automatic learning methods are the only subject in this area. Learning means changing or enhancing algorithms automatically without any human intervention, based on previous “experiences” [8].

ML is a method used to show computers how to work with data more easily and to produce a better performance. We can't grasp the model or derive details from the data in certain situations after presenting the dataset. We use machine learning techniques in such cases to predict the results. A large number of datasets can be obtained from numerous sources, and machine learning is required. Machine learning is used in many fields, from medical to military, to derive useful knowledge from accessible datasets. Machine learning is specifically meant to learn from current records. A wide variety of algorithms is programmed to learn from computers. Often mathematicians and programmers use a variety of methods to solve this problem [9].

Machine learning is aimed at learning from the results. Several experiments on how computers should learn by themselves have been carried out. Often mathematicians and programmers use a variety of methods to solve this problem [10].

## 5 Machine Learning Techniques

ML is now one of the most important IT remains. With the growing availability of data, there is cause to think that intelligent data processing is becoming more prevalent as a critical component for technical purposes. All of ML's work is about

addressing these issues and finding a response. An algorithm can model a problem based on its experiences with the context or environment or input data in various ways. First of all, we must follow the vocabulary that can be used by an algorithm. A few key models of learning can be found in an algorithm. The method in which machine learning algorithms are structured is helpful because it allows one to think about the functions of input knowledge and the planning method model and to choose the one that is better for the goal results task. Object identification is called a function with which machine learning improves. Discuss the various learning types and their particular aspects in machine learning algorithms.

## 5.1 *Supervised Learning*

In supervised learning, we study an objective function that can be used as accepted or not authorized to estimate the values of a distinct class attribute. The ML algorithm predicts a certain sample collection and supervised learning algorithms check for trends inside data-point labels. The algorithm is a consequence variable from which a certain set of predictors, i.e. different variables, is to be expected. We create a function using these variables, which maps the input to the desired outputs. The phase of testing continues until the model is correct with the training results. This whole procedure aims to reduce manual checks and software expenses. Examples: NN, Regression, DC, KNN, LR, SVM, NB, and so on. Controlled teaching. It is then divided primarily into two sections:

- **Learning (training):** Using training data to develop a model. Testing: Evaluate the model to measure model accuracy using unrevealed test data. Using experimentation constantly. The computer draws on its prior knowledge or attempts to obtain the right insights to make correct business decisions like the Markov Decision Process. It can pick a means of optimizing payouts. The algorithm would adjust its plan in a good time to get the best choice and precision.

## 5.2 *Unsupervised Learning*

Unsupervised learning is referred to as studying valuable frameworks of out-of-label classes, design parameters, input signals, or some other knowledge outside raw data. We may not have a goal variable in this algorithm to predict forms in which we have no data-point mark or may claim the class labeling of training data is unclear. This method is used to assemble data into the cluster category and explain its organization. The clustering of data shows valuable partitions as well as hierarchies. Examples: K-means, clustering fuzzy, and clustering hierarchy. The data entry is not protocol and the outcome is not understood. By deducting structures contained in the input data, a model is created. This can contribute to general laws being drawn up. It will systemically minimize redundancy via a logical method.

### 5.3 Reinforcement Learning

The computer is conditioned to make such decisions to use algo. Based on each data point or later on how good the decision was, these algos select action. This is an area in which the computer is revealed.

## 6 Types of Machine Learning Techniques

### 6.1 Supervised Machine Learning Techniques

NN, decision-maker, and SVM provide controlled machine learning techniques that learn from the low-level picture features the high-level principle. With the aid of training data already classified, this methodology conducts the classification process. The input and the expected result are defined for the training results. The latest invisible data can be replicated if supervised learning algos are educated using the existing training data. The machine learning algorithm here forecasts the picture type, which is nothing more than the semantical picture definition. Here, the match is based on query image and query image type rather than the entire database. The effects of regeneration are purer [11].

- **Support Vector Machine:** SVM learning models are supervised with related learning algorithms that evaluate data for classification. Classification implies which pictures refer to a certain category class or data collection or package. In the classification of machine learning, supervised learning is called an instance that deals with the role of inferring named training data. Picture retrieval systems teaching photos that are classified in a specific class may be properly categorized, in which each class corresponds to numerous image groups. The model for SVM training algorithms consists of the latest examples for one class or another. In this model, examples are provided in groups with consistent and as large as possible differences.
- **Neural Network:** NN are models inspired by biological NN structure or function. They are types of trends widely used for the issue of regression and classification. They are focused fundamentally on the basic neuron model. A neural system reflects the human brain digitally and aims to replicate the learning process. Sometimes, this is named NN. The term NN traditionally indicates a network of nervous system biological neurons processing and distributing knowledge. The three pieces are mainly neuronal: 1. A dendrite that collects other neurons' feedback 2. A significant nonlinear treatment stage 3. A soma. An axon is a cable wire that is transmitting the feedback signal down the transmission chain to other neurons. However, ANN is an interrelated group of artificial neurons that use an information processing conceptual model focused on the connectionist algorithm approach.



- **Decision Trees:** A DC is a schematic model that utilizes the technique of branching to show all potential decisions in some conditions. This internal tree node is an attribute test; the node is the result of the test and a leaf node is a special class name. Via root to a node, classification rules are shown. It is an algorithm used mostly for classification, supervised by learners. It functions for categorical and continuous variables. DC approaches to create a decision-making model dependent on individual data attribute values. Decisions are separated into tree systems before decisions are taken for analysis of a specific document. DC area is based on classification and regression issues results. Sometimes, they are quick, precise, and a great favorite of ML. This method divides data into two or more homogeneous collections. The most critical feature is to have groups as diverse as possible. This is achieved. It uses numerous strategies such as Gini, knowledge gain, chi-square, and entropy to separate details into different heterogeneous classes.
- **Naive Bayes:** It is a procedure classification based on Bayes' theorem following the recommendation of Thomas Bayes. If the predictor classification NB asserts its freedom, it implies that the appearance of a certain feature is irrelevant to that of another feature in a similar type. As if it is circular, circular, and around 3 inches in diameter, the fruit can be called an apple. Even if these characteristics depend on each other or other characteristics, both of these characteristics will be considered by this classification to lead individuals to the likelihood that this fruit is an apple. This model can be developed quickly and is used primarily for large datasets. It covers highly advanced classification methods and functions exceptionally well, along with its simplified portion. The Bayesian approaches relate Bayes' theorem specifically to issues like classification and regression.

## 6.2 *Unsupervised Machine Learning Techniques*

K-mean, clustering, and genetic algorithms are some of the unsupervised strategies. Data are not labeled for this learning technique, and the outcomes are not established. A model is designed with the deductions of the structures in the data entered or general laws are removed. It may go through a math phase to minimize complexity systemically or to arrange data through similarity. Data trends/classes are the primary objectives of a model. We want to investigate the data to discover an intrinsic structure. The data have no goal attribute. Clusters are a tool for looking for correlation classes of data identified as clusters. Means grouping in clusters data instances which are identical to each other, in positions that vary considerably from each other. Clustering is also called unattended computing since class values are not denoted and data instances are clustered randomly. It refers to the classified details of the complexity of identifying secret constructs. There are no performance metrics to direct the learning phase. Clustering pictures is usually an unsupervised strategy of studying. The collection of photos should be clustered in a manner that minimizes associations between numerous clusters. Apriori and K-mean are the algorithms used [11].

- **K-MEAN ALGORITHM:** K-mean is a partial algorithm for clustering. It is intended to divide the provided n comments into K clusters. Average of each cluster is identified and the picture is put in a cluster whose average distance from the Euclidean vector is the least. Since image data are complexly distributed, the clustering k-mean cannot always adequately distinguish images from various definitions. Regression clustering defines the problem class and infrastructure needs. Clustering techniques usually are structured as Centroid-based or Hierarchical into two computational approaches. K-mean, which is ultimately in the class of clusterings in unsupervised learning, is the most common of all. K-mean is a kind of unsupervised algo that solves the issue of clustering. The protocol of this cluster follows a straightforward and simple way to identify a single dataset across several clusters (taken as K clusters).

## 7 Literature Survey

**S. Sharma et al. [2020]** tried the actual issue of false-positive and false-negative outcomes and sought to define and resolve it. In our experiment, we used the HTTP dataset CSIC 2010 which contains created traffic for a Web application for e-commerce. The experimental findings indicate that using the feature set extraction suggested findings in enhanced web-based attack detection and classification for all ML algorithms evaluated. Precision, retrieval, precision, and F-measurement metrics measured the efficiency of the ML algorithm in the identification of violence. The J48 decision tree algorithm generated the highest real positive rate, precision, and recall among the three tested algorithms [12].

**A. Rashid et al. [2020]** complete a comparative study of the NSL-KDD and CIDD-001 comparison datasets. To obtain maximal efficiency, before implementing algorithmic classification methods like SVM, NB, k-NN, NNs, DNN, and DAE, we have used a hybrid set of features and rankings. In some prominent performance measures like Accuracy, Precision, Recall, and F1-Score, we evaluated IDS results. The experimental evidence indicates that classifiers k-NN, SVM, NN, and DNN function roughly. 100% specificity concerns NSL-KDD data collection value appraisal parameters, while k-NN and NB classifications achieve approx. CIDD-001 dataset performance of 99% [13].

**Kunal and M. Dua [2019]** In this case, intrusion prevention remains vital to network protection and machine learning frameworks which have significantly intensified the quest for new threats. The application of hybrid systems, ensemble learning methods, and multi-classifier in recent years has also given a significant boost to the exactness of attack detection methods. Although it is also important to resolve the incidence of false-positives and false-negatives. We encourage scientists to recognize the possibility that further approaches that have a better accuracy rating are implemented [14].

**H. Zhang et al. [2019]** In this paper, two separate ML algorithms are tested based on the BIT19 datasets. These include the k-mean clustering algorithm and the vector classification support algorithm. Both strategies offer a detection rate of greater than 98 percent from the first experiment with established attack forms. The identification of vector support technologies provides the same high detection rate as the first experiment with a cumulative detection rate of 97 percent while the k-medium clustering deteriorates to around 80 percent on average [15].

**F. Yihunie et al. [2019]** It aims at an effective classifier to identify NSL-KDD dataset anomaly by experimenting with five devices with a high degree of precision and minimal error rates. Five binary classification schemes are evaluated and checked to generate results: Stochastic Turnaround, Random Woods, LR, SVM, and Sequential model. The result reveals that the Random Forest Classifier is superior to the other four classifiers without but with normalization [16].

**I. Lafram et al. [2019]** A modern intrusion detection method was implemented to cover the entire traffic recognition mechanism focused on the integration of both unsupervised and supervised ML techniques. In comparison to these strong results, a different sample and thus further representations of existing computer networks were used to validate the suggested system. The structured model efficiency system is coupled with our previous work[29] to create a comprehensive total network traffic classification architecture. Both experiments have produced promising results and enable us to create an optimized model focused on a combination of advanced techniques of ML [17].

**D. Subramanyam [2018]** The disruption detection method used is represented. IDS is primarily utilizing methods to identify many cyberthreats as well as a right to retain multiple algorithms of decision-making and to consider the concept of information flow for representation. In the present article, we will infer IDS methods to right dos and sample form attacks dependent on classification. The different attacks help to understand the usage of information exploration obtained across numerous data. But in this paper, we will conclude for attacks and stop them for ML approaches [18].

**L. Haripriya and M. A. Jabbar [2017]** Several strategies for IDS identification ML were discussed. Details of the comparison of different IDS approaches using ML are also given in this paper. In contrast to other algos, each algo has its value for improving IDS. When certain traffic data are not available, it is very difficult to train the algorithms. This is the restriction that must be strengthened. Therefore, we cannot pick a single IDS deployment strategy. Improvements would then be made in ML techniques to decrease the false alarm rate and raise the identification rate [19].

## 8 Conclusion

IDS is developed to provide the basic technologies for identifying devices that are explicitly or implicitly linked to the Internet inside the networks. However, it's up

to the network owner at the end of the day to make sure his network is secure. This does not deter the intruder's network but IDS allows the network administrator to detect bad people on the Web to keep the network vulnerable and exploit. In this survey paper, we have observed the basic concept of intrusion recognition system as well as types of an intrusion detection system. The limitations of traditional IDS are defined in this paper. Based on this, we preferred different ML techniques. Different learning styles in machine learning algorithms are concluded in this work and defined the sub-techniques of machine learning.

## References

1. [https://www.researchgate.net/publication/259212150\\_Machine\\_Learning\\_Techniques\\_for\\_Intrusion\\_Detection](https://www.researchgate.net/publication/259212150_Machine_Learning_Techniques_for_Intrusion_Detection)
2. [https://www.researchgate.net/publication/324986956\\_Survey\\_paper\\_on\\_intrusion\\_detection\\_techniques](https://www.researchgate.net/publication/324986956_Survey_paper_on_intrusion_detection_techniques)
3. Aized Amin Soofi and Arshad Awan (2017) Classification techniques in machine learning: applications and issues. *J Basic Appl Sci* 13:459–465
4. Qiu J et al (2016) A survey of machine learning for big data processing. *EURASIP J Adv Signal Process*
5. Haq NF et al (2015) Application of machine learning approaches in intrusion detection system: a survey. (*IJARAI*) *Int J Adv Res Artif Intell* 4(3) (2015)
6. <https://www.medium.com/cuelogic-technologies/evaluation-of-machine-learning-algorithms-for-intrusion-detection-system-6854645f9211>
7. <https://doi.org/10.1186/s42400-019-0038-7>
8. Das K, Behera RN (2017) A survey on machine learning: concept, algorithms, and applications. *Int J Innov Res Comput Commun Eng* 5(2)
9. [http://www.ijarse.com/images/fullpdf/1521195654\\_Vedant612ijarse.pdf](http://www.ijarse.com/images/fullpdf/1521195654_Vedant612ijarse.pdf)
10. Dey A (2016) Machine learning algorithms: a review. *Int J Comput Sci Inf Technol* 7(3):1174–1179 (2016)
11. Kaur S, Jindal S (2016) A survey on machine learning algorithms. *Int J Innov Res Adv Eng (IJIRAE)* 3(11). ISSN: 2349-2763
12. Sharma S, Zavarsky P, Butakov S (2020) Machine learning-based intrusion detection system for web-based attacks. In: 2020 IEEE 6th International conference on big data security on cloud (BigDataSecurity), IEEE International conference on high performance and smart computing, (HPSC) and IEEE International conference on intelligent data and security (IDS), Baltimore, MD, USA, 2020, pp 227–230. <https://doi.org/10.1109/BigDataSecurity-HPSC-IDS49724.2020.00048>
13. Rashid MJS, Ahmed SM (2020) Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system. In: 2020 3rd International conference on advancements in computational sciences (ICACS), Lahore, Pakistan, pp 1–9. <https://doi.org/10.1109/ICACS47775.2020.9055946>
14. Kunal, Dua M (2019) Machine learning approach to IDS: a comprehensive review. In: 2019 3rd International conference on electronics, communication, and aerospace technology (ICECA), Coimbatore, India, 2019, pp 117–121. <https://doi.org/10.1109/ICECA.2019.8822120>
15. Zhang H, Lin K, Chen W, Genyuan L (2019) Using machine learning techniques to improve intrusion detection accuracy. In: 2019 IEEE 2nd International conference on knowledge innovation and invention (ICKII), Seoul, Korea (South), 2019, pp 308–310. <https://doi.org/10.1109/ICKII46306.2019.9042621>

16. Yihunie F, Abdelfattah E, Regmi A, “Applying Machine Learning to Anomaly-Based Intrusion Detection Systems,” (2019) IEEE long island systems, applications, and technology conference (LISAT). Farmingdale, NY, USA 2019:1–5. <https://doi.org/10.1109/LISAT.2019.8817340>
17. Lafram NB, El Alami J (2019) Artificial neural networks optimized with unsupervised clustering for IDS classification. In: 2019 1st International conference on smart systems and data science (ICSSD), Rabat, Morocco, 2019, pp 1–7. <https://doi.org/10.1109/ICSSD47982.2019.9002827>
18. Subramanyam D (2018) Classification of intrusion detection dataset using machine learning approaches. In: 2018 International conference on computational techniques, electronics and mechanical systems (CTEMS), Belgaum, India, 2018, pp 280–283. <https://doi.org/10.1109/CTEMS.2018.8769270>
19. HariPriya L, Jabbar MA (2018) Role of machine learning in intrusion detection system: review. In: 2018 Second International conference on electronics, communication and aerospace technology (ICECA), Coimbatore, 2018, pp 925–929. <https://doi.org/10.1109/ICECA.2018.8474576>

# Software Fault Detection by Using Rider Optimization Algorithm (ROA)-Based Deep Neural Network (DNN)



Shilpa Garg, Deepak Kumar, Suresh Chand Gupta,  
and Vijay Anant Athavale

**Abstract** Rider Optimization Algorithm (ROA)-based Deep Neural Network (DNN) model is suggested for software fault detection purposes. The entire work is divided into three phases: feature selection, training and testing, and fault detection. The ROA-based k-nearest neighbor (KNN) classifier performs the features selection task. The output of the ROA KNN classifier is fed as the input to the DNN. The DNN layers detected the faulty section. The features selection task is performed by the ROA-based KNN classifier that classifies the optimal features from all dataset features. The DNN model is trained with the features selected by the ROA-based KNN classifier. The performance of the proposed ROA-DNN model is compared with the PSO-based DNN model of software fault detection. The ROA-based DNN model of software fault detection shows up to 99% accuracy for fault detection.

**Keywords** Rider optimization algorithm · Deep neural network · Software fault detection · KNN · Feature selection

## 1 Introduction

Software is the programs and routines for a computer or the program material for an electronic device which makes it run. Software defect detection is mainly used for identifying the defective modules that are present in the software so that it helps in improving the quality of the software system. The instructions govern the entire functioning of the computer along with the equipment, which makes the entire system and performs the actual operations. Generally, a set of data, programs, and operating systems that process information is regarded as computer software in the arena of

---

S. Garg · S. C. Gupta

Panipat Institute of Engineering and Technology, Panipat, Haryana, India

D. Kumar (✉)

State Institute of Engineering and Technology, Nilokheri, Haryana, India

e-mail: [wadhwa123deepak@gmail.com](mailto:wadhwa123deepak@gmail.com)

V. A. Athavale

Walchand Institute of Technology, Solapur, Maharashtra, India

software engineering and computer science [1–5]. Therefore, digital programming, online documents, and libraries form the whole system of computers in unison. It is better to say that computer hardware and system software work collectively in coordination with each other in a realistic way.

## 2 Software Quality

The quality of the software is the most important parameter for different applications like banking, Microsoft windows, and office. The faulty software can affect the performance of the individual and a group of people. The work prepared by a human can be deleted or crashed in case of defective software. The quality of product measuring is a challenging issue in the software. Quality can be defined as the degree of excellence of a product. The quality of the software is present in the eye of the user. The faults and errors degrade the performance of software and can be identified during the testing of software. The quality of the software must depend on the two tasks:

- The software should efficiently do their task and always do the right things.
- The specific task of the software should perform correctly.

The fault present in the software is an error or failure that produces interrupted and inaccurate results in the end. Most of the errors or faults are produced by the source code and its design. The operating system used for the software and components is also causing the software's fault. Some of the errors are produced by the compiler that produces incorrect code. The software function is seriously interrupted by the faults, and it is said to be faulty or defective software. The software industry made more efforts to overcome the faults in their code source.

## 3 Software Faults and Testing

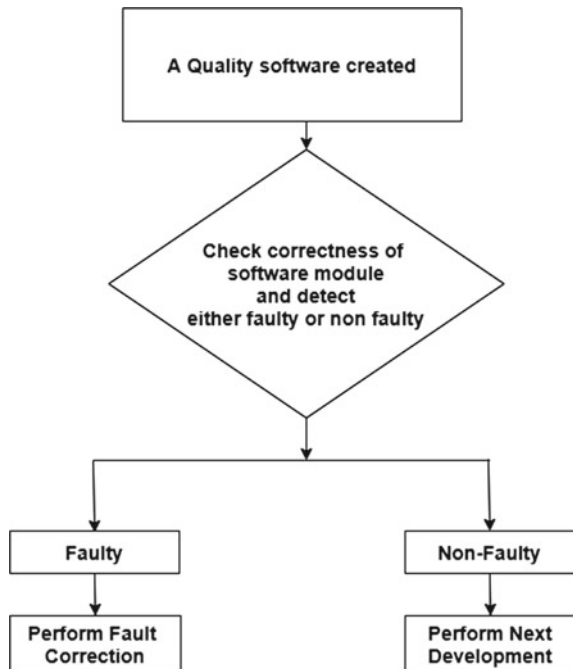
Software testing must be introduced at the beginning stage of the Software Development Life Cycle (SDLC). Software testing is done to reduce the cost of fixing the bugs in the later stages of Software Development; if it is not done in the initial stage, testing of the applications is the key factor of SDLC. Software testing is very imperative as it detects the errors and defects that took place at various stages while developing the software.

In software development, fault prediction plays a vital role. The fault prediction can minimize the efforts, time, and cost of software development. The software fault detection field is dominated by various machine learning [1–3], deep learning [4–19], Autoencoder-based [20–23], and fuzzy-based [15, 24] algorithms. The existing software techniques work in both object-oriented and real-time application environments. The demand for the software is increased day by that also enhanced the

quality of the software. The presence of faults in the software can degrade the quality and reliability that causes system failure. In software development, the fault detection task requires more effort than the early detection of faults. So early-stage fault detection is preferred in software that also minimizes the cost.

The term fault is frequently used to address all kinds of damages caused by faults [1]. This term is not precise, but it reflects the uncertainty surrounding the origins of faults and bugs creeping into programs when programmers are wrong but confident. Software Quality Engineering assures software fault prediction by an operation such as formal verification, fault tolerance, inspection, and testing. The predicted model is developed with existing software metrics and fault data. The process of fault prediction is carried out by training and prediction. In the training phase, previous software module metrics are considered for the built prediction model. This model predicts the fault proneness labels of modules located in a new software version. Figure 1 shows this fault prediction process. It works on checking the correctness and detects faults in software modules. If it is a faulty module, then perform fault correction, otherwise move to the next step.

**Fig. 1** Software fault detection





## 4 Objectives

In present days, early software fault detection is a popular topic of research. Various methods, like deep learning, machine learning, and computational, are utilized for software fault detection. The larger dataset is the primary concern for the software fault detection algorithm. The feature extraction process can increase fault detection accuracy. The studied approaches [5, 7, 20, 21] used an optimization algorithm to extract the input dataset feature. These features are reduced and used for the training phase of the classifiers. The classifiers detect the fault and evaluate the performance based on sensitivity, specificity, accuracy, etc. The unbalance and large dataset are the key challenges in a software fault detection algorithm. These challenges can be recovered in the pre-processing stage of the software fault detection model. The execution time of the software fault detection should be low to achieve optimal outcomes. In [21], a Bound Particle Swarm Optimization (BPSO) is proposed for the feature's reduction, but optimal outcomes are yet to be achieved with DNN. So, we propose a Rider Optimization Algorithm (ROA) with Deep Neural Network classifiers for software fault detection of 4 NASA projects of the PROMISE repository. The performance of the proposed method is evaluated based on AUC, specificity, and sensitivity. The following objectives are to be considered here:

1. Select the optimal dataset from the PROMISE repository (NASA) of the real-world project's dataset.
2. Select the features using the Rider Optimization Algorithm (ROA) from KNN features extractor.
3. Design a Deep Neural Network (DNN) and train it with the extracted features. The DNN model provides the detection of faults from the software.
4. Validate the performance based on the accuracy, sensitivity, and specificity evaluation.
5. Compare the outcomes with the PSO-based DNN system of software fault detection.

## 5 Related Work

We discuss the state-of-the-art methods of software fault detection. There are different schemes based on optimization, machine learning, deep learning, and the computational algorithm used for software fault detection. The studied approaches are categorized based on the dataset, features learning, and fault detection schemes.

## 6 Datasets

The software fault detection task is mostly performed on the PROMISE repository dataset. PROMISE collects publicly available datasets and tools to serve researchers in building predictive software models (PSMs) and the software engineering community. The repository encourages repeatable, verifiable, refutable, and improvable predictive software engineering [12, 15–17, 19]. In [1], 15 source projects are used for training purposes with NASA dataset [1, 3, 5, 7, 8, 10–13, 21, 24–26], APACHE projects [1–3] Academic software, and Turkish white good manufactures. 26 PROMISE projects dataset are used for the SFP task based on proprietary projects and APACHE projects [2]. Another PROMISE dataset is NASA, MDP [4, 18, 23] is also tested in several SFP tasks. The KC1 [4, 14, 18], KC2 [4, 18], PC1 [4, 18, 23], JM1 [18, 23], MC5 [11, 23], PC3 [14], CM1 [14], and PC4 [14] datasets are coming in NASA MDP dataset category. Some open-source JAVA projects [5, 17, 20, 22] are also used for the SFP [5]. In [6], firefox collected datasets DS1, DS2, and DS3 are used for the training and tested purpose of software fault prediction. A code bench dataset of C++ is utilized in [9] for the SFP. A defect 4j project database is also tested in [22]. In [24], three Eclipse database projects are tested for the SFP task. The raw data has some empty values or fewer samples that cause imbalance conditions. The data imbalance problem is avoided by the SMOTE [2, 3, 8, 20 etc.]. The noise and imbalance data is filtered in the pre-processing stage of software fault detection.

## 7 Features Learning and Classifiers

Machine learning and deep learning approaches are communally used for the feature extraction and classification task in software fault detection. Machine learning and deep learning approaches directly used raw data as a feature. Still, computational and optimization-based schemes extract the static features, software features, and imbalance ratio from the selected dataset. Multiple machine learning algorithms are combined to achieve software fault detection [1]. Naïve Bayes [NB] [1, 10] and logistic regression [LR] are combined for the universal software fault detection. Several methods like SVD with WL-ELM [2], Binary Moth Flame Optimization (BMFO) with KNN, DT, and LDA classifiers [3], MLP [4, 10] with CNN model [4, 11], PSO-GA integrated to SVM [5], ANN-based FDP and FCP model [6], Multi swarm HHO with LDA classifier and ADASYN data balance method [7], Bug Prediction using Deep representation and Ensemble learning (BPDET) with SAE [8], deep learning LSTM [9], SMOTE-based FLDA tested on different machine learning approaches KNN [10], RF [10], GSO-GA [20], BPSO and BACO-based RNN [12], BPSO-based DNN [13], GA-based DNN [14], ANN [16], Tree-based LSTM [17], ABC-based ANN [18], and DP-based CNN [19] are studied for the software fault detection process.

Some autoencoder-based techniques are also used for the software fault prediction. The autoencoder-based approaches like Stacked Denoising Autoencoder (SDAE) [21, 23, 26], Autoencoder with SBFL [22], and Fuzzy-based system computational approaches are also used for software fault detection [15, 24]. Various techniques are used to deal with the software fault detection or prediction. The features learning phase plays a vital role in fault prediction. As we studied, an optimization algorithm is used to select the raw dataset's optimal features. The performance of the tested methods is evaluated based on the Accuracy, Precision, Recall, Measures, Area under the curve (AUC), Region of Convergence (ROC), Probability of Predicted fault (PF), and imbalance ratio. The key problem of the software fault detection algorithm is extensive size data handling and data imbalance. We consider these two factors as the problem statement of our work.

## 8 Proposed Method

### 8.1 Dataset

In this paper, the four key NASA projects are considered to validate the design model of software fault detection. These projects come from the NASA IV and V facility index data programs. The four NASA projects are PC1, JM1, KC1, and KC3 to apply software fault prediction. A brief description of four NASA projects is provided in Table 1.

The datasets PC1 and JM1 are designed in the language C, whereas KC1 and KC3 are developed in the C++ language and JAVA. The raw dataset is normalized for the optimization algorithm. The raw data is normalized to remove the empty and inaccurate values.

**Table 1** Detail of NASA projects dataset [14]

Dataset name	Number of modules	Defective measurement (%)	Language of software	Line counter	Number of defects
PC1	1107	6.9	C	40,000	76
JM1	8779	19.35	C	10,885	2106
KC1	2109	15.5	C++	43,000	325
KC3	457	20.5	JAVA	10,000	44

## 9 Rider Optimization-Based Features Selection for KNN

Features selection is a binary optimization problem where a binary vector represents the features subsets. The single feature in the dataset is represented by each element present in the vector dataset. If the vector element has the value 1, then it is selected, and the optimization algorithm does not select the 0-value element. The ROA is used for the features selection task that performs a mapping function from real values to the binary space values. The binary conversion process can be achieved by various approaches in a different iteration of the optimization process. The binary ROA has several advantages over the existing approaches of optimization. The stability in the exploration and exploitation phase is more in binary ROA.

A race inspires rider optimization; few riders' groups are assumed to reach a particular location. The winner of the race is considered the optimal solution to the optimization problem. The four different rider groups select an equal number of riders in each group from a total number of riders. The categories of four riders group are attacker, bypass rider, follower, and overtaken. Each group follows the instruction to reach its optimal location.

1. The bypassing condition is followed by the bypass rider to reach the location bypassing the leading path.
2. The follower follows the leading rider on all the axis points.
3. The random position is followed by the overtaken rider for reaching the target location. The overtaken group overtakes the nearby location of the leading rider.

The attackers take the leading rider's optimal location to reach the target location. The attacker group riders use the maximum speed. The four groups are made, and their definition is most important for the algorithm. The essential definitions related to the ROA are explained below:

- **Bypass Rider**—The rider group that reaches the target bypassing the leading path is known as the Bypass Rider group. The bypass rider does not follow the leading path of the leading rider.
- **Follower**—The rider group that depends on the leading rider's position always followed the position of the leading rider.
- **Overtaker**—The unique position followed by a rider group to reach the target's location. The unique position is nearby to the leading rider position.
- **Attacker**—The attacker is utilized the maximum speed to achieve the leading rider's optimal position. The attacker groups of rider grab the position of the leading rider.

The rider group that reaches the target location is known as the race winner. The steps involved in the ROA are described below [27]. The steps involved in the rider optimization algorithm are as follows:

1. Initialization of four rider groups and their parameters gear, steering angle, accelerator, and brake.
2. Evaluate the success rate based on updated parameters.

3. Compute leading rider position.
4. Position update of four rider groups Bypass rider, overtaker rider, follower rider, and attacker rider group.
5. Again, calculate the success rate based on the update position of four rider groups.
6. Find the optimal parameter of the rider like steering angle, brake, accelerator, and gear.
7. Riding time off evaluation.

## 10 Fitness Function

The performance of the classification algorithm is enhanced by applying the features selection-based method. The effectiveness of a subset is identified based on two conditions; the number of features selected in the subset and the error rate of the KNN classifier attained by using the selected features of the subset from the training dataset by binary ROA model. The better features subset has less error rate and a minimal number of selected features. The quality of the selection procedure depends on the error rate of the KNN classifier. The error rate is considered as the objective function or fitness function for the features selection and KNN classification task. It can be formulated as

$$\text{Objectivefunction}(S_i) = (\lambda \times \text{errorrate}) + (1 - \lambda) \times [(N - T)/N] \quad (1)$$

Here,  $S_i$  is the features subset evaluated by the features selection-based binary ROA, *errorrate* is the KNN classifier error that is computed while training the KNN model based on the selected features by binary ROA,  $\lambda$  is the balance factor among the error rate and selected features size [0, 1], the number of features present in the dataset is denoted by  $N$ , and size of features subset is  $T$ . All 21 features are fed as input to the binary ROA model. The ROA algorithm is running for the several iterations set. The features selection task is completed by providing the optimal solution in each iteration case.

## 11 Deep Neural Network-Based Fault Detection

A DNN contains three main layers known as the Convolutional layer, the Pooling layer, and the Fully connected layer. A CNN is a simple neural network that provided effective results in recognition and processing tasks. This three-layer arrangement is used for both feature extraction and classification task.

The function of each layer is explained as follows:

- Convolutional Layer-The convolution layer is also known as the features learning layer of the DNN. The primary function of the convolution layer extracts the features from the input data. In the convolution layer, the kernel can remove the

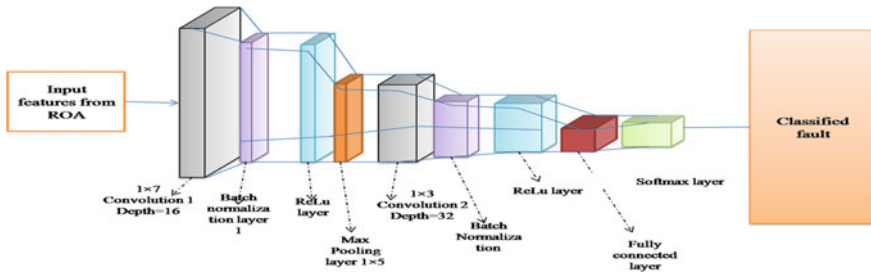


Fig. 2 DNN architecture for fault detection by using ROA selected features

features from the input signal. A features vector is extracted with the help of the first convolution layer.

- Pooling Layer-The pooling layer is used to reduce the features vector’s dimension. It takes the average and maximum value of the feature vector.
- Fully Connected Layer-The final layer is used for classification purposes, which are trained by the output of the pooling layer vector. Some additional features are added in this stage and performed classification tasks.

The features selected by the ROA-based KNN model is fed as the input of the DNN model. As discussed in the previous section, the DNN model has a convolutional layer, batch normalization layer, ReLU layer, Max pooling layer, softmax layer, and fully connected layer. The depth of the first convolutional layer is 16 and the dimension is  $1 \times 7$ , which converts it into 32 and  $1 \times 3$  dimensions in the second convolution layer. The selected features are in binary form and used for fault detection in the fully connected layer. Figure 2 shows the DNN arrangement for the software fault detection. The flowchart of the overall proposed method is shown in Fig. 3.

## 12 Results and Discussion

In this paper, a software fault detection task is performed based on the ROA integrated DNN model. The software performance is depending on the attributes present in the coding lines. We consider the four datasets of NASA projects to predict the software’s fault. The four datasets are PC1, JM1, KC1, and KC3, tested with the proposed method ROA-based DNN. The entire proposed work is designed in the MATLAB 2020 software using the optimization and deep learning toolbox.

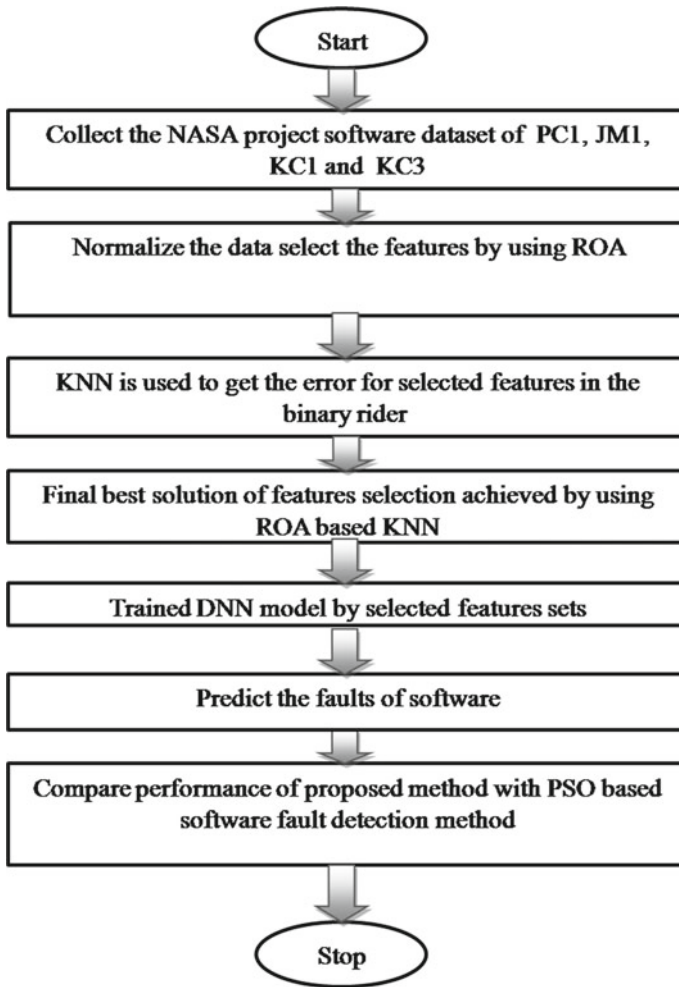


Fig. 3 Flowchart of the proposed method

### 13 Performance Evaluation

The performance of the proposed ROA-based DNN model is evaluated with the three parameters. These parameters are defined below.

**Fault Detection Accuracy**—It is the metric that is used for the evaluation of classification algorithms or models. The accurate prediction of the model in the right direction is provided with accurate performance. For the classification task, the formulation of accuracy is

$$FaultDetectionAccuracy = \frac{Number\ of\ correct\ class\ prediction}{total\ numer\ of\ classes}$$

For the binary value classifiers, the accuracy can be formulated as

$$Fault\ Detection\ Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here, the assumptions are  $TP = true\ positive$ ,  $TN = true\ negative$ ,  $FP = false\ positive$ , and  $FN = false\ negative$ .

**Sensitivity:** The ability to determine the predicted class accurately and can be calculated as

$$Sensitivity = \frac{TP}{TP + FN}$$

**Specificity:** The ability to recognize the true negative predicted classed and formulated as

$$Sensitivity = \frac{TN}{TP + FN}$$

These three evaluation terms are compared with the previously used approaches. The proposed ROA-based DNN model software fault detection is compared with the Binary PSO-based software fault detection [21] model.

## 14 Comparative Study

We compared the proposed ROA-based DNN model with the BPSO algorithm of software fault detection. The classifier model is DNN in the BPSO algorithm of software fault detection with the NASA project's dataset. We compare the outcomes of the PC1 NASA project with the ROA ad BPSO algorithms with the DNN approach. Figure 4 shows the convergence curve of the proposed ROA optimization and binary PSO algorithm. The early convergence is achieved in the ROA-based features selection and stable in the earlier stage. The ROA performance is better than the binary PSO algorithm in terms of features selection tasks.

Figure 5 shows the confusion matrix plot of the ROA- and PSO-based software fault detection method. If faults appear in the code that has high samples or features, then the ROA-based algorithm shows higher classification accuracy nearly about 94%. If less fault detection samples, ROA shows 99.9% accuracy, and PSO shows only 94.1%. This case is studied for the PC1 dataset, and remaining datasets like JM1, KC1, and KC3 are also tested with the ROA-DNN and PSO-DNN model, and ROA-based DNN shows better performance every time.



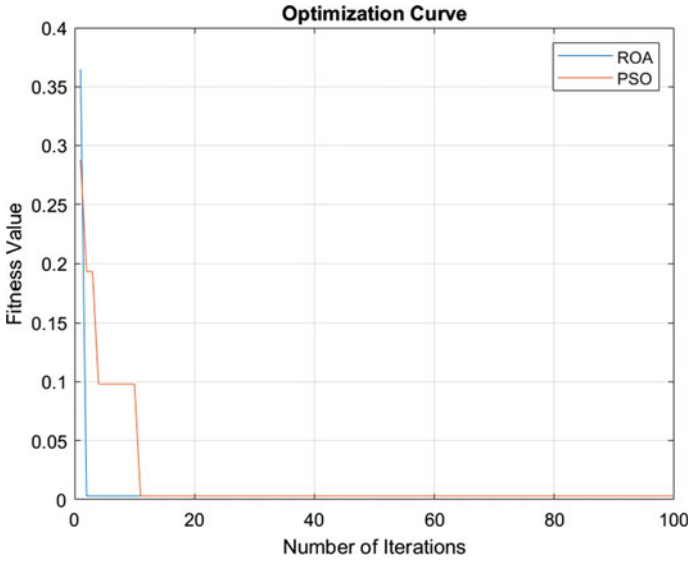


Fig. 4 Convergence curve of ROA and binary PSO algorithm

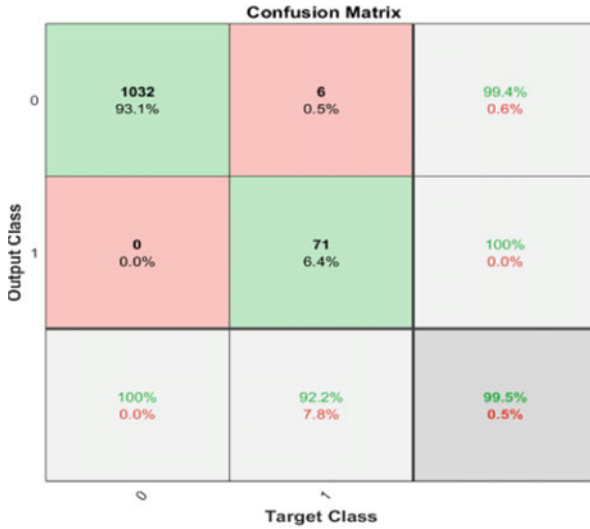
The performance of the ROA-based model is also analyzed based on the confusion matrix curve (Fig. 6).

The features selected by the ROA-based KNN approach are 5 out of 21, whereas PSO selects 14 out of 21. The ROA selected features to show higher accuracy in fault detection for the DNN model. Figure 7 shows the analysis curve among the ROA, PSO, and complete features-based software fault detection algorithm based on the specificity, sensitivity, and accuracy evaluation parameters.

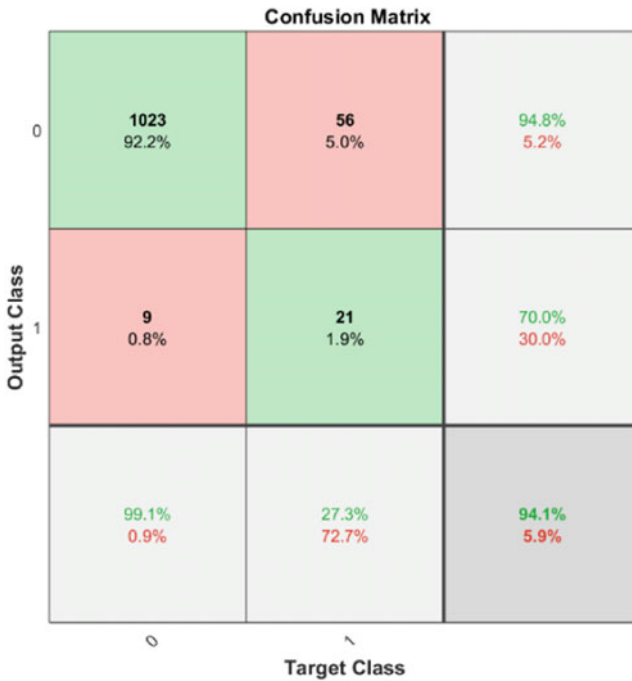
The ROA-based DNN model of software fault detection is highly sensitive, specific, and accurate than the PSO-based DNN model of software fault detection.

## 15 Conclusion

In this paper, a software fault detection model is presented based on the optimization and deep learning framework. The software fault can vitally affect the performance of the program. The software metrics like coding, defined parameters, and syntax can be the reason for the fault generation. Based on these metrics, we can develop a software fault detection model using the ROA and DNN. The entire work is divided into three phases; feature selection, training and testing, and fault detection. The ROA-based KNN classifier performs the features selection task. The output of the ROA KNN classifier is fed as the input to the DNN. The DNN architecture has three main layers; convolutional, max-pooling layer, and fully connected layer. The fully connected layer detected the faulty section. The proposed ROA-based DNN model is



(a) confusion matrix plot of ROA



(b) Confusion matrix plot of binary PSO

**Fig. 5** a Confusion matrix plot of ROA. b Confusion matrix plot of binary PSO

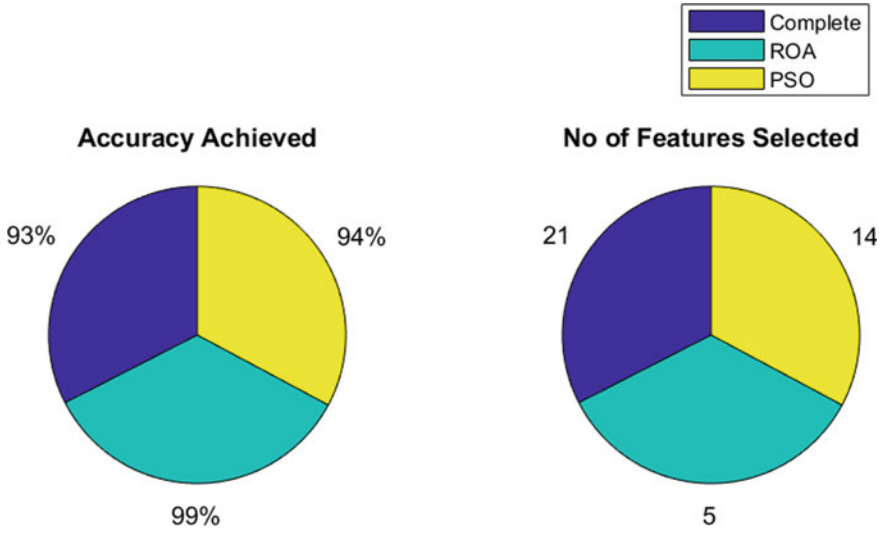


Fig. 6 Accuracy comparison curve with selected features for ROA, PSO, and total features

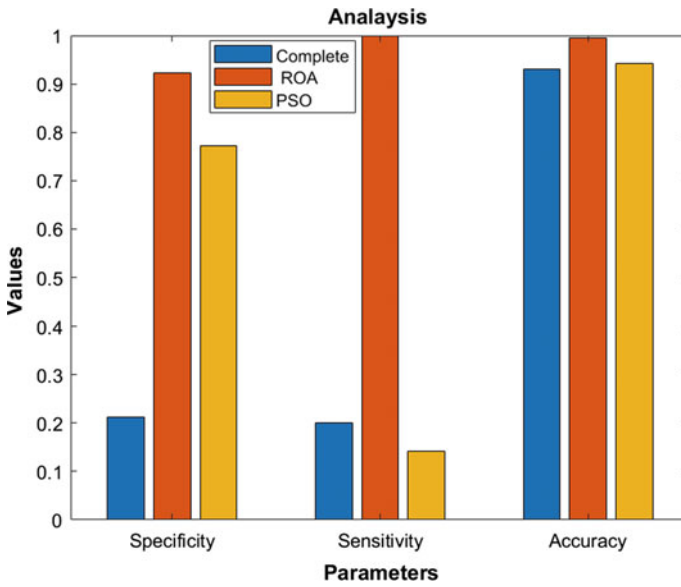


Fig. 7 Comparison among the accuracy, sensitivity, and specificity

compared with the binary PSO-based DNN model. The comparison is made based on the accuracy, sensitivity, and specificity parameters. The proposed ROA-based DNN model shows higher accuracy in software fault detection. The software maintenance cost is reduced by detecting the faults.

The proposed ROA-DNN model is tested on the four NASA datasets that are PC1, JM1, KC1, and KC3. In the future, this model can be tested on the PROMISE repository and other projects those have high dimensions. The features selection task can be improved by hybridization of the ROA algorithm with any of the local search-based meta-heuristic optimization algorithms.

## References

1. Yucalar F, Ozcift A, Borandag E, Kilinc D (2020) Multiple-classifiers in software quality engineering: combining predictors to improve software fault prediction ability. *Eng Sci Technol Int J* 23(4):938–950
2. Bal PR, Kumar S (2020) WR-ELM: weighted regularization extreme learning machine for imbalance learning in software fault prediction. *IEEE Trans Reliab*
3. Tumar I, Hassouneh Y, Turabieh H, Thaher T (2020) Enhanced binary moth flame optimization as a feature selection algorithm to predict software fault prediction. *IEEE Access* 8:8041–8055
4. Al O, Akour M, Alenezi M (2020) The influence of deep learning algorithms factors in software fault prediction. *IEEE Access* 8:63945–63960
5. Alsghaier H, Akour M (2020) Software fault prediction using particle swarm algorithm with genetic algorithm and support vector machine classifier. *Softw Pract Exp* 50(4):407–427
6. Xiao H, Cao M, Peng R (2020) Artificial neural network based software fault detection and correction prediction models considering testing effort. *Appl Soft Comput* 94:106491
7. Thaher T, Arman N (2020) Efficient multi-swarm binary harris hawks optimization as a feature selection approach for software fault prediction. In: 2020 11th International conference on information and communication systems (ICICS). IEEE, pp 249–254
8. Pandey SK, Mishra RB, Tripathi AK (2020) BPDET: an effective software bug prediction model using deep representation and ensemble learning techniques. *Expert Syst Appl* 144:113085
9. Majd A, Vahidi-Asl M, Khalilian A, Poorsarvi P, Haghghi H (2020) SLDeep: Statement-level software defect prediction using deep-learning model on static code features. *Expert Syst Appl* 147:113156
10. Kalsoom A, Maqsood M, Ghazanfar MA, Aadil F, Rho S (2018) A dimensionality reduction-based efficient software fault prediction using Fisher linear discriminant analysis (FLDA). *J Supercomput* 74(9):4568–4602
11. Bhandari GP, Gupta R (2018) Measuring the fault predictability of software using deep learning techniques with software metrics. In: 2018 5th IEEE Uttar Pradesh section international conference on electrical, electronics and computer engineering (UPCON). IEEE, pp 1–6
12. Turabieh H, Mafarja M, Li X (2019) Iterated feature selection algorithms with layered recurrent neural network for software fault prediction. *Expert Syst Appl* 122:27–42
13. Geng W (2018) Cognitive deep neural networks prediction method for software fault tendency module based on bound particle swarm optimization. *Cogn Syst Res* 52:12–20
14. Manjula C, Florence L (2019) Deep neural network based hybrid approach for software defect prediction using software metrics. *Clust Comput* 22(4):9847–9863
15. Juneja K (2019) A fuzzy-filtered neuro-fuzzy framework for software fault prediction for inter-version and inter-project evaluation. *Appl Soft Comput* 77:696–713
16. Kaur R, Sharma S (2018) An ANN based approach for software fault prediction using object oriented metrics. In: International conference on advanced informatics for computing research. Springer, Singapore, pp 341–354

17. Dam HK, Pham T, Ng SW, Tran T, Grundy J, Ghose A, Kim CJ et al (2018) A deep tree-based model for software defect prediction. [arXiv:1802.00921](https://arxiv.org/abs/1802.00921)
18. Arar ÖF, Ayan K (2015) Software defect prediction using cost-sensitive neural network. *Appl Soft Comput* 33:263–277
19. Li J, He P, Zhu J, Lyu MR (2017) Software defect prediction via convolutional neural network. In: 2017 IEEE International conference on software quality, reliability and security (QRS). IEEE, pp 318–328
20. Mohapatra Y, Ray M (2018) Software fault prediction based on GSO-GA optimization with kernel based SVM classification. *Int J Intell Eng Syst* 11(5):152
21. Tong H, Liu B, Wang S (2018) Software defect prediction using stacked denoising autoencoders and two-stage ensemble learning. *Inf Softw Technol* 96:94–111
22. Peng Z, Xiao X, Hu G, Sangaiah AK, Atiquzzaman M, Xia S (2020) ABFL: an autoencoder based practical approach for software fault localization. *Inf Sci* 510:108–121
23. Zhu Y, Yin D, Gan Y, Rui L, Xia G (2019) Software defect prediction model based on stacked denoising auto-encoder. In: International conference on artificial intelligence for communications and networks. Springer, Cham, pp 18–27
24. Arshad A, Riaz S, Jiao L, Murthy A (2018) Semi-supervised deep fuzzy c-mean clustering for software fault prediction. *IEEE Access* 6:25675–25685
25. Singh A, Bhatia R, Singhrova A (2018) Taxonomy of machine learning algorithms in software fault prediction using object oriented metrics. *Proc Comput Sci* 132:993–1001
26. Tran HD, Hanh LTM, Binh NT (2019) Combining feature selection, feature learning and ensemble learning for software fault prediction. In: 2019 11th International conference on knowledge and systems engineering (KSE). IEEE, pp 1–8
27. Binu D, Kariyappa BS (2018) RideNN: a new rider optimization algorithm-based neural network for fault diagnosis in analog circuits. *IEEE Trans Instru Measur* 1–25

# An Approach for Predicting Admissions in Post-Graduate Programme by Using Machine Learning



Shivam Sharma and Hemant Kumar Soni

**Abstract** Machine Learning (ML) has spread its wings over a wide range of fields nowadays. For all data-driven areas, machine learning is a true game-changer. Today's most powerful firms are those that have concentrated their resources and energy on understanding client data. Machine Learning is a branch of Artificial Intelligence that contains a collection of sophisticated algorithms. Machine learning aids in machine training by mapping the link between input/inputs and output. Machine Learning is used in the proposed study to estimate the probability of admission to a master's programme from the perspective of Indian students. This work is presented as a web-based interactive application. This web-application was created using the platforms 'R' and 'Shiny'. The Web-App was created to provide students with a platform to get insight into their chances of acceptance. In this study, Support Vector Regression is utilised to train the model for predicting admission probability. Since, the R-squared Value of Support Vector Regression has been greater than 0.8 for both the training and testing sets. As a result, it's a good algorithm to use in the proposed project.

**Keywords** Machine learning · Support vector regression · Random forest · Prediction · Admission · Master's programme

## 1 Introduction

The evolution of the economic structure of any country greatly banks on the education quality of its population. A good blend of theoretical and practical education aids in the creation of skilled personnel who are profitable for the company. Since the competition has reached an all-time high in recent years, it has become necessary for everyone to prove themselves. A reputable degree from a prestigious college increases an individual's chances of securing incredible possibilities. Every year, a big number of Indian students travel to a foreign country in order to acquire a decent education. Obtaining a place in a top university is not an easy challenge for

---

S. Sharma · H. K. Soni (✉)

Department of Computer Science and Engineering, Amity School of Engineering and Technology, Amity University Madhya Pradesh, Maharajpura, Gwalior, MP, India  
e-mail: [hemantsoni.gec@gmail.com](mailto:hemantsoni.gec@gmail.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022  
S. Sharma et al. (eds.), *Data, Engineering and Applications*, Lecture Notes in Electrical Engineering 907, [https://doi.org/10.1007/978-981-19-4687-5\\_5](https://doi.org/10.1007/978-981-19-4687-5_5)

57

candidates. Every year, thousands of Indian students enrol for master's programmes overseas. However, not everyone is accepted into the chosen institution since they do not satisfy certain standards. The candidate's Graduate Record Examinations (GRE) score, Test of English as a Foreign Language (TOEFL) score, International English Language Testing System (IELTS) score, cumulative grade point average (CGPA) at graduation, previous research, and other factors all play a role in this decision. Students can use the technique for 'Admission Chance Predictors' proposed in this paper to understand their odds of selection. Students can use it to get a sense of what prerequisites they'll need to get a strong chance at a selection.

## 2 Literature Review

Since a large mass of students apply for the post-graduate programme per year and the numbers are amazingly increasing every year [1]. Understanding the needs of students to predict their chances of admission, different systems have been developing [2]. Some of the works are detailed below.

In this work, a system 'GRADE' [3] is developed for graduate admission, employing the machine learning support. The system is designed using the statistical machine learning approach. This system is employed in the Computer Science Department at University of Texas. The university receives a large number of applications for the Ph.D. Program which makes the process of the human reviewing and scoring of these applications. GRADE using the previous data helps to predict the new admission probability. The researchers claim that the implementation of this system decreases the total time of review by 74%. GRADE draws out the human reviewer score by saving a lot of time. The researchers have experimented with the algorithms like 'SVM' with linear/nonlinear algorithms, Multilayer Perceptrons, and L-1 regularised Logistic Regression. In this work, logistic regression is used, since it performed the best. SVM performed comparably as that of logistic regression. Here multilayer Perceptron algorithm performed the worst. Due to capturing even the noise in the Training Data, it failed to generalise the result and tends to overfitting.

The admission for the post-graduation in abroad is not an easy process. Many new candidates don't even know the requirements for getting admission. Through this work [4], authors have developed the machine learning-based predictor of Post-Graduation which works automatically. Since, students find it difficult to decipher the university suitable as per their profiles. The work has employed three machine learning algorithms 'Logistic Regression', 'Linear Regression', and 'Decision Tree' individually for getting the best among all. The authors claim that Logistic Regression has performed best among all by providing the higher accuracy of prediction.

A model named 'UAP' [5] is developed for the prediction of the student admission. Presenters have designed a User Interface using node-red for making the model user-friendly for the non-technical people as well. This model has machine learning dependency in it for the new predictions. This model saves the time of the user and money of consultancy services. The model has certain limitations like it can

only be used for Indian students who are going to do master's in USA. The work has employed the machine learning algorithms like 'K-Nearest Neighbour', 'Linear Regression', 'Ridge Regression', and 'Random Forest'. Work claims that out of all the four algorithms, Linear Regression worked best by providing an average accuracy of 79%.

Student Admission Predictor (SAP), designed by [6], for predicting the acceptance chance of admission. This system is developed for providing the details of the universities, suitable as per the profile of the students by using Machine Learning Algorithms like 'K-Nearest Neighbour' (KNN), 'Multivariate Logistic Regression', and 'Decision Tree'. This is designed as a web-application and a standalone using Shiny and R. Here 'K-Nearest Neighbour' is employed for drawing out the chances of admission and 'Decision Tree' is employed for finding the Rank of University suitable for the profile of the student. The accuracy provided by 'KNN' for predicting the chance of admission is 76%.

The paper has implemented several regression algorithms for the prediction of the chances of admission [7]. The algorithms employed in this work are 'Linear Regression', 'Support Vector Regression', 'Decision Tree', and 'Random Forest'. After implementing each algorithm individually, the metrics R2 Value and MSE are obtained. These metrics are utilised to compare the relative accuracies of these algorithms. On the basis of these metrics, Linear Regression performed best followed by Random Forest. Since the R2 value for Linear Regression is the highest and the MSE value is the lowest among all.

### 3 Methodology

Machine Learning is the collection of manifold algorithms. Each algorithm has its own uniqueness and working methodology [8]. There are various robust algorithms like 'Support Vector Regression', 'Support Vector Machine', 'Random Forest' [9], 'Linear Regression', 'Decision Tree' [10], 'K-Nearest Neighbour', and many others. In the proposed work, Support Vector Regression (SVR) is employed. 'Support Vector' Algorithms are used for both Regression and Classification [11], where 'Support Vector Regression' is used for Regression to predict the continuous variable and 'Support Vector Machine' (SVM) is used for Classification of the categorical variables. SVR employs a similar principle as being utilised in SVM.

#### 3.1 *Support Vector Regression*

Support Vector Regression follows similar rules as 'SVM'. But there are certain differences between them as well. Some common terms between 'SVR' and 'SVM' are as follows.



### 3.2 *Hyperplane*

In SVR, a Hyperplane is a Line that Assists to Predict the Continuous Values, While for ‘SVM’, a Hyperplane is a Line that Separates Different Classes.

### 3.3 *Kernel*

Kernel is the function that is used to map the data in the lower dimension to the higher dimension.

### 3.4 *Boundary Line*

The two line that develops the margin are the boundary lines. These lines help to separate classes. ‘SVR’ follows the same concept.

### 3.5 *Support Vector*

The data points which have the least or no distance from the boundary lines are the Support Vectors. Support Vectors are in the closest proximity of the Boundary Lines.

The crucial difference between the Linear Regression and SVR is that ‘Linear Regression’ [12] employs the concept of minimising the error rate between the actual and predicted values. While Support Vector Regression aims to restrain the error within a certain threshold value.

In Fig. 1, the blue line between the yellow lines is the hyperplane. The equation of this hyperplane is manifested in Eq. (1)

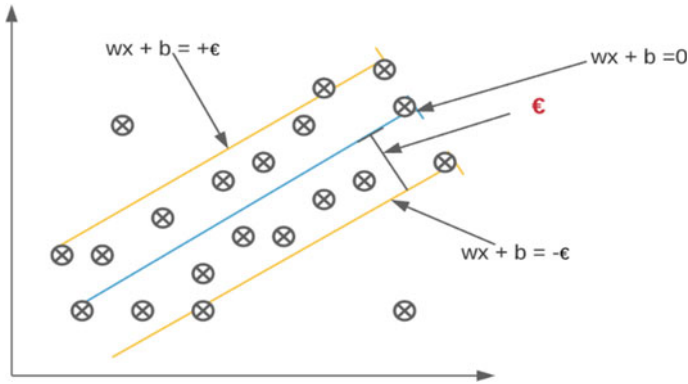
$$wx + b = 0 \quad (1)$$

The yellow lines in Fig. 1 are the boundary lines at the distance of  $-\epsilon$  and  $+\epsilon$  from the hyperplane. The Eq. (2) and Eq. (3) are the equations of the boundary lines.

$$wx + b = +\epsilon \quad (2)$$

$$wx + b = -\epsilon \quad (3)$$

In this, we are going to take those points which are having least error rate. Thus, Algorithm will take the points within the boundaries, which serves as the ‘Margin of Tolerance’.



**Fig. 1** Support vector regression’s pictorial representation

## 4 Proposed Method

Data is the paramount element of any work. The consequence of the work is subjected to the quality of data used. The dataset used in this work is garnered from Kaggle [13]. This dataset is designed from the perspective of an Indian student to predict the chances of admission for Master’s Admission. The independent attributes included in this data for predicting the response variable (Chance of Admission) are the ones, which play a crucial role during the application submission for Masters Programmes. The dataset has nine attributes, i.e., ‘Serial No.’, ‘GRE Score’, ‘TOEFL Score’, ‘University Rating’, ‘Statement of Purpose Quality (SOP)’, ‘Letter of Recommendation Quality (LOR)’, ‘CGPA’, ‘Research’, and ‘Chance of Admission’. Here, ‘Chance of Admission’ is the response variable, and the rest all are independent variables. There are 400 rows in this dataset and thence the dimension of this data is [400,9].

### 4.1 Data Pre-processing

Data Pre-Processing is a crucial step, which is employed to bring the data in the requisite format. The present work dealt with ‘Comma Separated Value’ (csv) file, [14] fetched from Kaggle. In the first step, the presence of any NA value is checked for data imputation. Since the dataset used in the proposed work doesn’t contain any ‘NA’ values, therefore, data imputation is not required. The first column in the dataset, i.e., ‘Serial No.’ is of no use, that being so, it is removed from the dataset. The columns belonging to the integer class are transformed into the numeric class for the web-application. Now the data is prepared to execute the desired operations on it.

## 4.2 Data Partitioning

After Pre-Processing, data is partitioned into two sets, ‘Training Set’ and ‘Testing Set’ [14]. This partitioning of data into two is elucidated below.

- **Training Set:** In this work, Training Set is 75% of the entire dataset. This dataset contains all the attributes, i.e., ‘Independent Variables’ and ‘Response Variable’. Here, the response variable is the ‘Chance of Admission’. This part of the dataset is employed to train the machine for building a model, which can map the relationship between independent variables and response variables.
- **Testing Set:** Rest 25% of the data is used to evaluate the accuracy of the data, since this part doesn’t contain the response variable. The model predicts the outcomes using mapped relationships, and it has drawn while training the machine.

## 4.3 Fitting Model

Support Vector Regression is used for fitting the model. Here, Training Data is used for mapping the relationship. Before fitting the model between response, and independent variables, the hyperparameters are tuned. The Tuning of the hyperparameters help in accomplishing the optimised results from the model. After tuning the parameter, the model is developed [15]. The accuracy of the fitted model for the Training Data is delineated in Fig. 2.

In the course of fitting the model, the resampling is done using tenfold cross-validation [16]. The metrics used for measuring accuracy are ‘RMSE Value’, ‘R-squared Value’, and ‘MAE Value’, where

```
Support Vector Machines with Radial Basis Function Kernel

301 samples
  7 predictor

Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 270, 271, 271, 272, 272, 271, ...
Resampling results:

   RMSE      Rsquared   MAE
0.06421741  0.8062225  0.04532498

Tuning parameter 'sigma' was held constant at a value of 0.0577
Tuning parameter 'C' was held constant at a
value of 2.21049
.
```

Fig. 2 Accuracy of model for training data

- **RMSE Value:** This is the Root Mean Square Error Value, the lower the value of RMSE, the higher will be the model’s accuracy.
- **R-squared Value:** It is the proportion of variance in the dependent variable explained by the independent variable or variables in a model. The value of R-squared ranges from 0 to 1.
- **MAE Value:** ‘MAE’ stands for Mean Absolute Error. It is one of the metrics for measuring the accuracy. The lower the MAE, the higher will be the accuracy associated with the model.

Here, R-squared Value ‘0.8062225’ suggests that the model worked pretty decent with the Training Data. The tuned parameters ‘sigma’ and ‘C’ are 0.0577 and 2.21049, respectively.

### 4.4 Accuracy Evaluation

In the testing dataset, the fitted model is employed to predict the ‘Chance of Admission’. The predicted response variables are compared to the actual responses after it has been predicted. This comparison assesses the model’s ability to predict outcomes for new variables. Figure 3 shows the results of the model prediction on the test data.

The R-squared value for the testing set is 0.80011177, which is close to the R-squared value for the training set. This indicates that the model has performed well and that there is no evidence of overfitting. When a model captures both noise and relationship, it is said to be overfitted.

Figure 4 depicts all of the steps addressed in the proposed method from subheadings 4.1 to 4.4. The way the procedure is progressing throughout the task is clearly illustrated in Fig. 4. The steps covered in Fig. 4’s process flow are detailed below.

- The data is pre-processed and transformed to the required format.
- After pre-processing, the data is divided into two parts: training (75%) and testing parts (25%).
- The SVR is used to fit the model using training data. The tweaking of hyperparameters is also done during the model fitting.
- The fitted model is now used to forecast the results for the Testing Set (In Testing Set, the response variable is not present).
- The predicted response value is then compared to the actual values. Model prediction accuracy is measured using metrics such as ‘RMSE’, ‘R-squared’, and ‘MAE’.

Rsquared	RMSE	MAE
0.80011177	0.06552382	0.04512070

Fig. 3 Prediction accuracy of the model on testing data

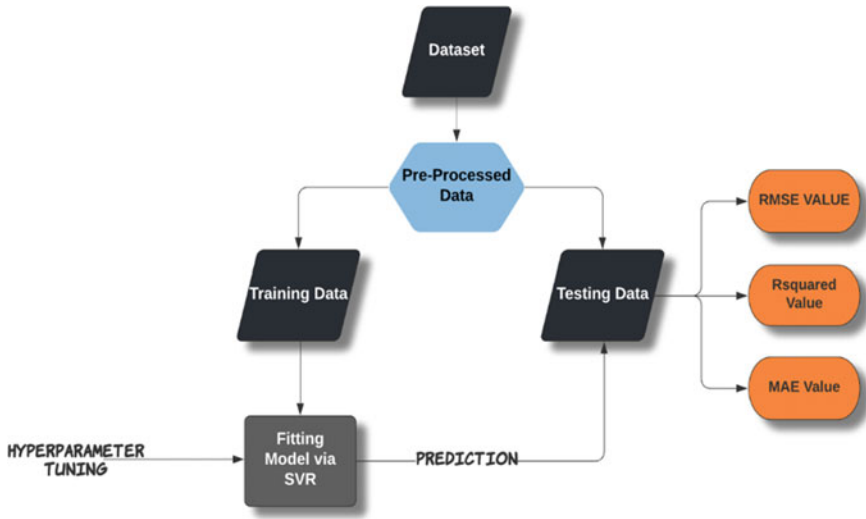


Fig. 4 The flow of process in proposed work

### 4.5 Importance of Variables

The response variable has a unique relationship with each variable. In the dataset, not all of the features are significant. The variable with the lowest relevance contributes the least to forecasting the result. Various approaches may be used to determine the importance of each factor in predicting the response variable. Random Forest [17] is employed for finding the importance of the independent variables. Figure 5 shows the relative importance of each variable.

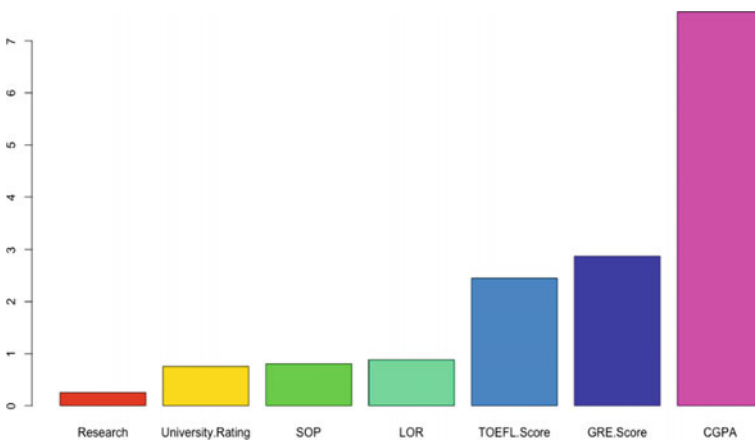


Fig. 5 Importance of variables via random forest

Figure 5 shows that the most important variable in predicting the ‘Chance of Admission’ is ‘CGPA’. It means that any change in the ‘CGPA’ will have a significant impact on the ‘Chance of Admission’. Figure 5 shows that ‘research’ is the least important variable. This behaviour explains that changes in the Research variable’s input would not result in a significant change in the output of the ‘Chance of Admission’ variable.

## 5 Implementation and Experimentation

The proposed project is implemented as an interactive R-Shiny web-application [18, 19]. The web-application would assist the users in exploring their results by spawning the percentage of admission chances via a 3-D Pie-Chart. The Architecture of this Web-Application has two significant pillars as manifested in Fig. 6, i.e., ‘User Interface’ and ‘Server’ [20].

### 5.1 User Interface (UI) Working

This is the part that is visible to users. Users are asked to input their data into the User Interface. The respective results are generated in this Interface itself. In this work, the users are asked to fill their ‘GRE Score’, ‘TOEFL Score’, ‘University Rating’,

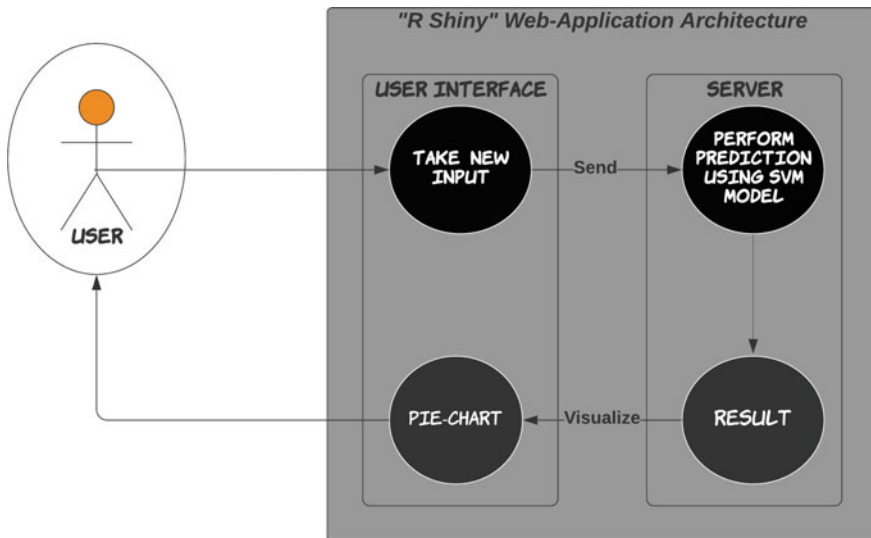


Fig. 6 Architecture of web-application

‘SOP Rating’, ‘LOR Rating’, ‘CGPA’, and ‘Research’. After submitting the data, ‘Admission Chance’ percentage is generated via a 3-D Pie-Chart in the main panel.

## 5.2 *Server Working*

This is where the input data is processed to get the final output. The input data is sent to the server through the user interface. Several procedures are carried out here in order to get the result eventually. SVM Regressor Model forecasts the outcome for the likelihood of admission in the proposed study. Following the discovery of the result, the server returns the information to the user interface in the form of a Pie-Chart.

## 5.3 *User Interface of Web-Application*

Figure 7 depicts the User Interface, or the section that is visible to the users. This application was created with the goal of offering a simple user interface that would allow any user to operate with it without any difficulties. It is required to fill out the necessary information of the new user on the left side of the Web-Application. The following are the attributes of the sidebar panel on the left:

- **GRE Score:** Slider Input, ranging from 0 to 340.
- **TOEFL Score:** Slider Input, ranging from 0 to 120.
- **University Rating:** Radio Buttons, having five categories (1, 2, 3, 4, 5).
- **SOP Rating:** Numeric Input, ranging between 0 and 5.
- **LOR Rating:** Numeric Input, ranging between 0 and 5.
- **CGPA:** Numeric Input, ranging from 0 to 10.
- **Research:** Radio Button, with two levels (‘yes’ and ‘no’).

The user only needs to hit the submit button after filling out the information. When it is pressed, the input data is transmitted to the server, which does all of the backend computations.

Figure 7 shows the primary panel on the right (white coloured). After hitting the submit button, this is where you may see your final result, which is the ‘Probability of Admission’. This section will produce the appropriate 3-D Pie-Chart. Users would eventually visualise their results with this Pie-Chart.

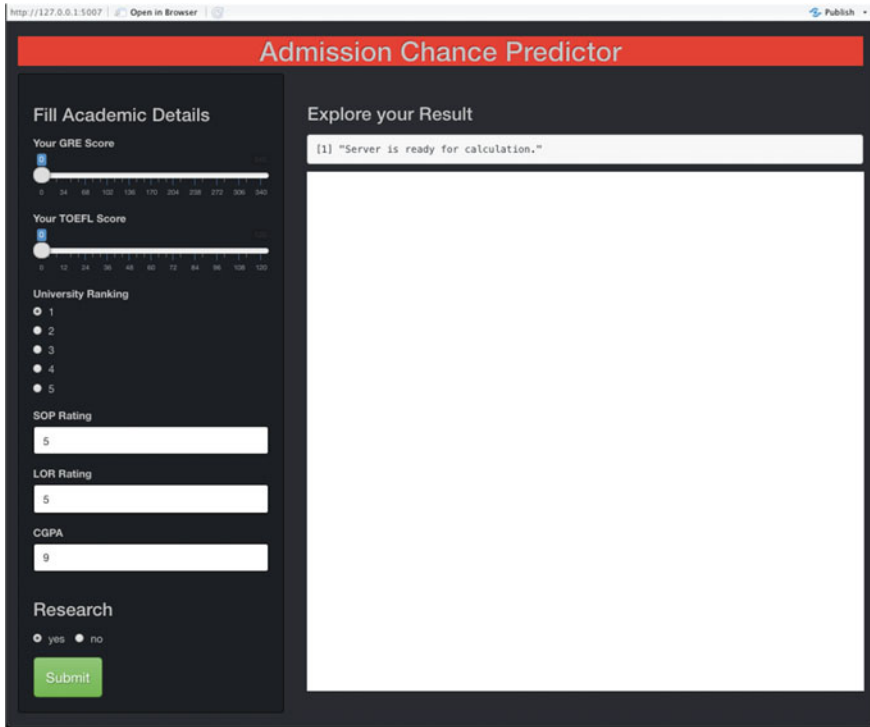


Fig. 7 The user interface of web-application

### 5.4 Working Manifestation

This is the experimentation part, here the predictive capability of the Web-Application is scrutinised along with the manifestation of how the web-application works. For the purpose of experimentation, a new user data is fed into the Web-Application for which the result is already known. The input attributes are detailed in Table 1. The **Actual** probability of getting the admission for the inputs in Table 1 is ‘82%’ (Highlighted in Blue colour).

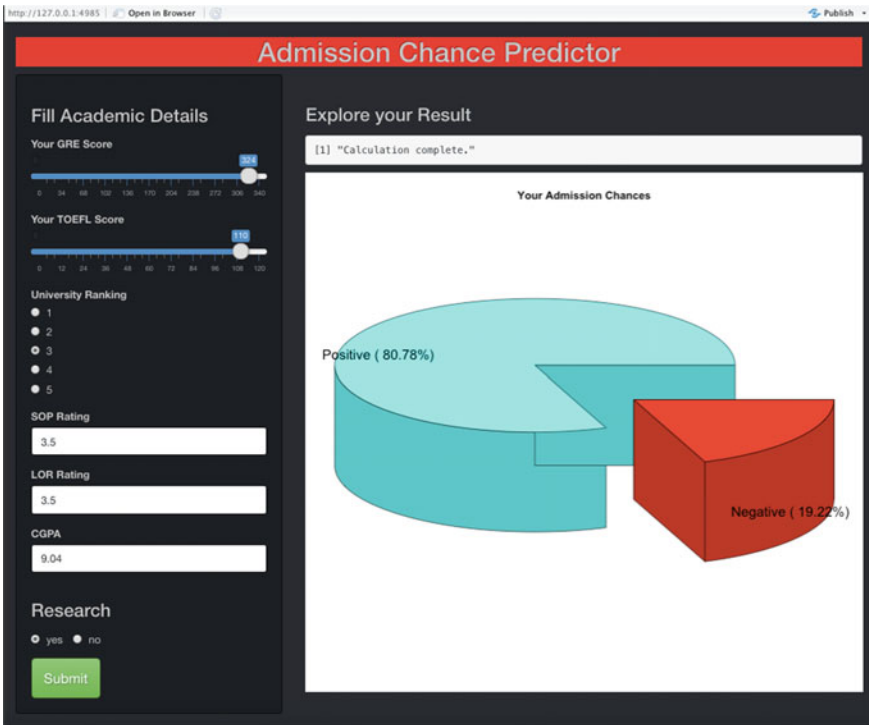
After feeding the new inputs, the result generated in Fig. 8 predicts that there are 80.78% of chances of getting an admission in the Master’s Programme, as per the ‘light-blue’ coloured Pie-Chart. This suggests that the prediction of this Web-Application is pretty close to that of the actual value. As a result, this application will provide users with valuable information on their ‘admission prospects’. Because the predictive skills of a model cannot be assessed by using only one data from the testing set, this section simply shows how the app works.

The evaluation of accuracy on the testing set was previously discussed in Sect. 4.4, where the R Squared value of ‘0.8001177’ for testing data indicates that it performs well.



**Table 1** New input values for experimentation

Attribute name	Input value
GRE score	324
TOEFL score	110
University rating	3
SOP rating	3.5
LOR rating	3.5
CGPA	9.04
Research	yes
Chances of admission (actual)	82%



**Fig. 8** Experimentation with new inputs

### 5.5 Critical Analysis

The proposed study and similar works are compared in this portion of the article. Table 2 provides a thorough analysis. This shows how important the current work is in relation to earlier ones.

**Table 2** Comparative analysis of the proposed work

Method name	Accuracy (%)
UAP [5]	79
SAP [6]	76
Proposed work	80

The proposed study is compared to two previous works in Table 2. The significance of the web-application given by the researchers in this study could be determined by analysing the accuracies.

## 6 Conclusion

This project was created with the goal of aiding students who are pondering whether or not would he/she be able to make it. The suggested project is designed as a user-friendly web-application. The web-application was created by blending the R and Shiny. Support Vector Regression is employed for the prediction of the new data by fitting a model for the previous historical data. The SVR Model performed well with both training and testing data, with R-squared values of 0.806 and 0.80 for training and testing data, respectively. The suggested project would assist students in reducing the time and money spent on consulting firms for evaluating admission possibilities. The relevance of the suggested work is defined by the critical analysis since it gives an average accuracy of 80%, which is higher than prior efforts.

## References

1. Basu K, Basu T, Buckmire R, Lal N (2019) Predictive models of student college commitment decisions using machine learning. *Data* 4:65. <https://doi.org/10.3390/data4020065>
2. Bashar A, Parr G, McClean S, Scotney B, Nauck D (2010) Machine learning based call admission control approaches: a comparative study. In: 2010 International conference on network and service management. IEEE, pp 431–434. <https://doi.org/10.1109/CNSM.2010.5691261>
3. Waters A, Miikkulainen R (2013) GRADE : machine learning support for graduate admissions. In: Proceedings of the twenty-fifth innovative applications of artificial intelligence conference, pp 1479–1486
4. AlGhamdi A, Barsheed A, AlMshjary H, AlGhamdi H (2020) A machine learning approach for graduate admission prediction. In: Proceedings of the 2020 2nd International conference on image, video and signal processing. ACM, New York, NY, USA, pp 155–158. <https://doi.org/10.1145/3388818.3393716>
5. Chithra Apoorva DA, ChanduNath M, Rohith P, Bindu Shree S (2020) Prediction for university admission using machine learning. *Int J Recent Technol Eng* 8:4922–4926. <https://doi.org/10.35940/ijrte.F9043.038620>
6. Sonawane H (2017) Student admission predictor. <http://www.trap.ncirl.ie/3102/1/himanshumahadevsonawane.pdf>

7. Acharya MS, Armaan A, Antony AS (2019) A comparison of regression models for prediction of graduate admissions. In: 2019 International conference on computational intelligence in data science (ICCIDIS). IEEE, pp 1–5. <https://doi.org/10.1109/ICCIDIS.2019.8862140>
8. Awad M, Khanna R (2015) Support vector regression. In: Efficient learning machines. Apress, Berkeley, CA, pp 67–80. [https://doi.org/10.1007/978-1-4302-5990-9\\_4](https://doi.org/10.1007/978-1-4302-5990-9_4)
9. İskenderoğlu FC, Baltacıoğlu MK, Demir MH, Baldinelli A, Barelli L, Bidini G (2020) Comparison of support vector regression and random forest algorithms for estimating the SOFC output voltage by considering hydrogen flow rates. *Int J Hydrogen Energy*. <https://doi.org/10.1016/j.ijhydene.2020.07.265>
10. Li S, Laima S, Li H (2018) Data-driven modeling of vortex-induced vibration of a long-span suspension bridge using decision tree learning and support vector regression. *J Wind Eng Ind Aerodyn* 172:196–211. <https://doi.org/10.1016/j.jweia.2017.10.022>
11. Trafalis TB, Gilbert RC (2006) Robust classification and regression using support vector machines. *Eur J Oper Res* 173:893–909. <https://doi.org/10.1016/j.ejor.2005.07.024>
12. Krmar J, Vukićević M, Kovačević A, Protić A, Zečević M, Otašević B (2020) Performance comparison of nonlinear and linear regression algorithms coupled with different attribute selection methods for quantitative structure—retention relationships modelling in micellar liquid chromatography. *J Chromatogr A* 1623:461146. <https://doi.org/10.1016/j.chroma.2020.461146>
13. Acharya MS. Graduate admission 2.
14. Afendras G, Markatou M (2019) Optimality of training/test size and resampling effectiveness in cross-validation. *J Stat Plan Inference* 199:286–301. <https://doi.org/10.1016/j.jspi.2018.07.005>
15. Ito K, Nakano R (2003) Optimizing support vector regression hyperparameters based on cross-validation. In: Proceedings of the international joint conference on neural networks. IEEE, pp 2077–2082. <https://doi.org/10.1109/IJCNN.2003.1223728>
16. Ling H, Qian C, Kang W, Liang C, Chen H (2019) Combination of support vector machine and K-fold cross validation to predict compressive strength of concrete in marine environment. *Constr Build Mater* 206:355–363. <https://doi.org/10.1016/j.conbuildmat.2019.02.071>
17. Lovatti BPO, Nascimento MHC, Neto AC, Castro EVR, Filgueiras PR (2019) Use of random forest in the identification of important variables. *Microchem J* 145:1129–1134. <https://doi.org/10.1016/j.microc.2018.12.028>
18. Yu Y, Yao W, Wang Y, Huang F (2019) shinyChromosome: an R/Shiny application for interactive creation of non-circular plots of whole genomes. *Genom Proteom Bioinform* 17:535–539. <https://doi.org/10.1016/j.gpb.2019.07.003>
19. Möller M, Boutarfa L, Strassemeyer J (2020) PhenoWin—an R Shiny application for visualization and extraction of phenological windows in Germany. *Comput Electron Agric* 175:105534. <https://doi.org/10.1016/j.compag.2020.105534>
20. Doi J, Potter G, Wong J, Alcaraz I, Chi P (2016) Web application teaching tools for statistics using R and Shiny. *Technol Innov Stat Educ* 9

# A Survey on Various Representation Learning of Hypergraph for Unsupervised Feature Selection



Rana Pratap Singh, Divyank Ojha, and Kuldeep Singh Jadon

**Abstract** Since multimedia technology is increasingly being developed, vast quantities of unlabeled data must be processed with a high dimension. Unsupervised feature selection considered to minimize the dimensionality of many machine learning and data mining activities is generally accepted as an important and challenging preliminary step. Feature selection is a big problem in machine learning and in recent years has been very important. Representation learning has become a field in itself in the community of machine learning, and both academia and industry are followed and feeding a substantial string of empirical successes by an exponential growth in research activity in the field of representation learning. Latent representation learning may help with multiple tasks in machine learning and data mining and also recently has become ever more drawn to network data particularly. A vertex set and hyperedge set are constructed of Hypergraph. In a hypergraph, each hyperedge can connect several vertices and easily extend the hyperedge, which contributes to a flexible hypergraph edge degree. In hypergraph learning, hypergraph construction is the first step for data modeling. In this survey paper, we have described various technologies like Data mining, Machine learning, and their types and also we deeply described feature selection methods and their types; classification methods explained the various learning methods such as hypergraph learning and Latent representation learning. Various optimization techniques are also discussed.

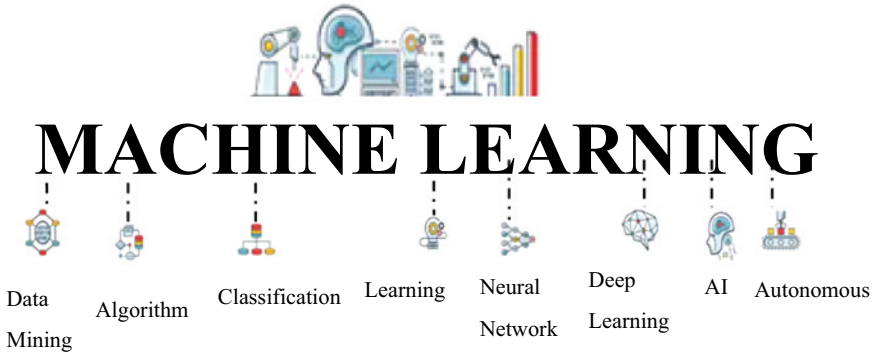
**Keywords** Data mining · Feature selection · Unsupervised feature selection · Hypergraph and representation learning · Machine learning · Optimization

## 1 Introduction

Data Mining (DM) is a means of knowledge discovery. Three key methods exist in DM: classification, regression, and clustering. The instances are combined in identified classes in these approaches. Categorization is useful for analyzing and

---

R. P. Singh (✉) · D. Ojha · K. S. Jadon  
Department of Computer Science and Engineering, ITM, Gwalior, India  
e-mail: [r.pratap5656@gmail.com](mailto:r.pratap5656@gmail.com)



**Fig. 1** Machine learning

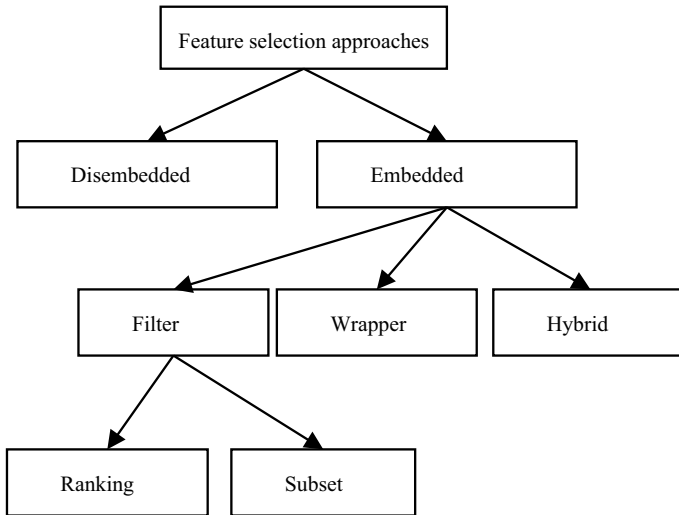
studying the existing dataset sample and for forecasting the expected behavior of the dataset. There are two phases. The first phase was learning and the second phase was data testing. The models evaluate the training data during the learning process, which generates patterns and rules. A variety of machine learning (ML) algos, such as SVM (Support Vector Machine), NBC (Naive Bayes Classifier), DT (Decision Tree), Linear Regression, LR (Logistic Regression), K-mean clustering, and ANN (Artificial Neural Networks), can be used in classification techniques [1].

Feature selection (FS) is a kind of reduction in spatial coverage measurement that essentially shows fascinating parts of the email message as a compressed feature vector. The approach is effective when the message is big and a compact feature representation is required to snappy the task of matching text or images [2]. Feature selection (called subset selection) is a method widely used in machine learning, in which the features of the data are chosen for an algorithm of learning [3].

ML [4] is a theoretical area that focuses systematically on the theory, results, and properties of education systems and algorithms. It draws on several diverse areas of ideas: optimization theory, artificial intelligence, psychology, knowledge processing, optimal controls, cognitive sciences, and also other branches of research, science, and mathematics. Often mathematicians and programmers use a variety of methods to solve this problem [5]. It is a strongly interdisciplinary area. ML area is usually categorized into three sub-domains: supervised, unsupervised [2, 6], and reinforcement learning (Fig. 1).

## 2 Feature Selection

It is one of ML's most critical preprocessing steps. It benefits by substantially reducing the dimension and removing inappropriate or redundant data and thereby increasing computational intelligence learning accuracy. It is fairly substantial so it can operate



**Fig. 2** Taxonomy of feature selection approaches

best for multiple feature subsets with the same training data. Many variables impact the success of ML. The demonstration and importance of instance data are primarily among these factors.

1. The Search Method is a selection algorithm to generate the most beneficial or appropriate feature subsets for model formation.
2. The evaluator is a search algorithm that evaluates the goodness of a feature subset and returns an assessment about search method legality [7].

The selection of features may be divided into three methods, such as filter, wrapper, and embedded methods, depending upon various searching techniques (Fig. 2 and Table 1).

## 2.1 Feature Selection Classification

The existence of data labeling allows the methods of selection of features to be usually categorized into three categories: semi-supervised feature selection, unsupervised feature selection, and supervised feature selection.

### 2.1.1 Supervised Feature Selection

For data that are labeled, supervised feature selection approaches are needed. Traditionally supervised methods like Fisher Score rank individually as per criteria that

**Table 1** Comparison of filter, wrapper, and embedded methods

Filter methods	Wrapper methods	Embedded methods
Probably less optimal filter methods	In supervised learning issues, wrapper methods are superior alternatives	If more irrelevant features are inserted into the goal set, the output of the built-in method degrades
These methods are executed quite quickly than the wrapper	Both methods are slower than the methods of sorting	This is simpler than the method of wrapping
The result of the filtering method is more popular than the wraps method	In wrapper methods, there is a loss of generality as it is linked to such classifiers	It also lacks generality because some classification algorithms also depend on it
The method of sorting has more tendency to choose a large dataset	The wrapper is more accurate than filters since it achieves better recognition speeds than filters	They are less likely to be defeated
For large data in the filter method, the device cost is less	For large datasets in the filter method, device costs are more	In comparison to wrapper methods, computation rates are lesser
Independent of classification algorithm	Dependent on the classification algorithm	Depending on an algorithm of classification [8]

cannot take into consideration the correlation between various features. Unfortunately, LDA (Linear Discriminant Analysis) has been suggested to improve the characteristics by optimizing the ratio between class dispersal and class dispersion.

### 2.1.2 Semi-supervised Feature Selections

In addition to being classified, however, semi-supervised feature selections often manage unlabeled training details. This helps semi-supervised techniques to select characteristics through the use of unlabeled data if the amount of labeled data is limited. In the graph Laplacian processes, the unlabeled data samples are used in the graph Laplacian matrix. Therefore, unsupervised feature selection is possible and important to study.

### 2.1.3 Unsupervised Feature Selection

It is the task of selecting the most relevant and best feature among large sets of data that are not labeled. There are several algorithms used to extract features among unlabelled data. Unsupervised FS selects a corresponding features subset that includes the most discriminatory data of original data. In meantime, the geometrical structure of original data without using sample label information should be retained as far

as possible [8, 9]. Feature selection which protects the best-assorted plan of unique information is a common criterion. This approach can be used in two different ways [10].

### 3 Representation Learning

The machine learning community is also a field for representation learning with frequent workshops at leading conferences including ICML & NIPS, as well as recent ICLR, 1 conference sometimes under Feature Learning and Deep Learning. While the Description is an important part of the stories and some other relevant priors also may be easily captured as a representation of learning challenge, as discussed in the next section (Fig. 3).

A noteworthy sequence of empirical successes, both in academia and industry, have followed and nurtured a quick increase in experimental activity in representation learning.

#### 3.1 Representation Learning Applications

Below, some of these high points are briefly illustrated.

##### 1. Speech Recognition and Signal Processing

The speech was one of the untimely NNs (Neural Networks) applications, particularly CNN (Convolutional Neural Networks) (or delay of time). The new interest in NNs, deep learning, and representation has revived the effect on speech recognition, by innovative results achieved by many academics and researchers in industrial laboratories that transform this algo to a bigger scale and into goods.

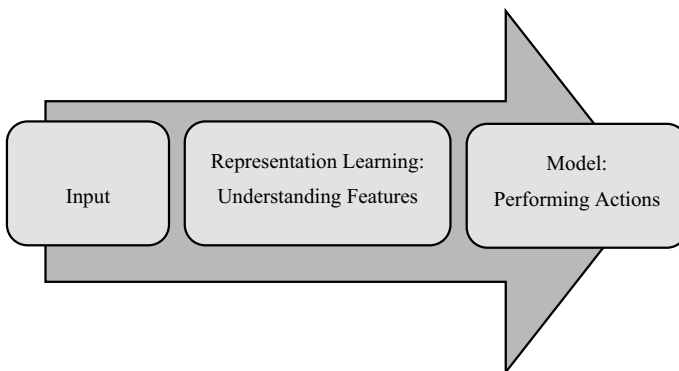


Fig. 3 Representation learning



## 2. Object Recognition

In 2006, the beginning of deep learning concentrated on the issue of digital image classification of MNIST that broke the supremacy of SVMs (1.4% error). The deep networks continue to apprehend new records: they currently hold the title of the state of the art with 0.27% error and state of the art with 0.81% error for MNIST knowledge-free variants of uncontained versions such as fully convolutional design elements. Deep learning has changed over recent years in the field of natural images to the recognition of objects, and the ImageNet dataset has accomplished the most recent breakthrough, for which the state-of-the-art error rate is 26.1–15.3%.

## 3. Natural Language Processing (NLP)

There are also several NLP representation learning applications in addition to speech recognition. Hinton implemented Distributed symbolic data representations that were developed first by Bengio et al. within the statistical language modeling framework of supposed ‘neural net language models.’ Both representations are focused on learning, in the term ‘embedding’ of a distributed representation for each word. Collobert et al. developed SENNA systems that integrate the different language models, chunking, part speech tagging, semantic role labeling, entity recognition, as well as syntax parsing tasks in the context of a convolutional architecture.

## 4. Multitask and Transfer Learning, Domain Adaptation

Learning transfer is the ability to share a learning algorithm’s statistical strength and to transfer information through tasks across commonalities across various learning tasks. As discussed below, it believes that representation learning algorithms benefit from these tasks, as they acquire representations that capture the underlying variables, which may apply to a single task under consideration. A collection of empirical findings demonstrating representation learning algos strengths in scenarios of transfer learning seem to confirm the hypothesis [11].

## 3.2 Latent Representation Learning

Latent representation learning will help many activities in DM and ML and recently has attracted more and more interest, in particular for network data. Due to various factors in networks, instances also interconnect. Some instances communicate latent representations and form data about a relation also are more likely to be compared to instances with identical latent representations than instances with different latent representations. Latent representation learning is part of unsupervised FS to make use of interconnection b/w data samples [12].

### 3.3 *Hypergraph Learning*

A vertex set and hyperedge set are constructed of Hypergraph (HG). Each hyperedge can attach any number of vertices in the hypergraph, and a hyperedge can be extended quickly, thereby leading to a flexible edge degree for HG. Therefore, in contrast to simple graphs or other linear contrast approaches, the hypergraph model may develop a high-order correlation of data. HG construction is the first step for data modeling in hypergraph learning. Existing works can be broken down into approaches based on features and representation. The hyperedge targets in characteristics based on k-NN or search radius the exploration of the nearest neighbors in the feature space. The k-NN method selects a set number of nearest neighbors to create a hyperedge for any vertex. The search radius, however, defines the predetermined search radius and is related with one hyperedge to all vertices within the radius. It is therefore difficult to find the optimal numbers of nearest neighbors or the search radius that can be noise-sensitive and limit data modeling performance [13].

## 4 Optimization Techniques

The optimization method discovers the alternative under a certain constraint, with mainly cost-effective and maximum possible performance. Therefore, optimizing implies trying to achieve the best or highest results of cost. Optimization is constrained by a lack of complete info and time to decide which data is required to improve the Optimization process. The Optimization method is used to achieve them. The best can be a single organization and objective process which models a similar entity. The process of Optimization shall be applied to factors that determine the best varies with the situation (Fig. 4).

Several examples are optimized costs, used raw materials, and time. Optimization can be done for local and global optimal achievement. Several major optimization algos are available in each field [14].

- (1) Classification of Optimization Algo(COA).
- (2) Numerical Optimization.
- (3) Integer programming (IP).
- (4) Advanced Optimization.
- (5) Simulated Annealing.
- (6) Genetic Algo.
- (7) Ant Colony Optimization.

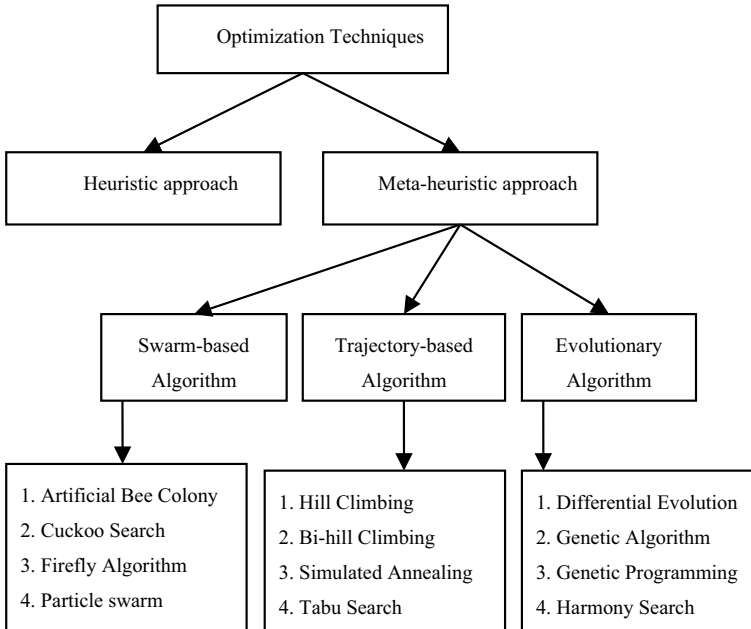


Fig. 4 Different types of optimization techniques

## 5 Literature Survey

Unsupervised FS is a significant method for reducing the dimensionality of the countless unlabeled high-dimensional data to remove the dimensionality problem. Unsupervised FS is a difficult challenge since there is a lack of label information to obtain appropriate features to improve research performance. Due to the limited size of the data, many authors in this field have been made.

Liu et al. [15] proposed the approach for unsupervised FS of Robust Neighborhood Embedding (RNE) on benchmark datasets suggesting that the RNE approach in terms of clustering efficiency is efficient and superior to the state-of-the-art unsupervised FS algos.

Tang et al. [16] introduced a robust unsupervised FS approach that integrated learning of hidden representation into FS. To optimize the proposed model, an effective alternating algo was developed. Experimental findings on eight benchmark datasets suggest that the proposed approach was efficient.

Ding et al. [17] proposed a stable, unsupervised FS approach that incorporated latent representation learning in FS. A hypergraph was studied and integrated with the following model, to capture the high order of the original data's local multiplicity geometric structure. Experimental findings on eight benchmark datasets reveal the feasibility of the approach suggested.

Yuan et al. [18] suggested nonnegative spectral analysis and joint sparse matrix regression model for 2D unsupervised FS. The solution to this proposed optimization issue was an efficient Optimization Algorithm. Extensive test findings on clustering and classification showed the proposed method’s effectiveness [18].

Yang et al. [19] to deal with the similarity matrix dilemma, a novel unsupervised approach for FS was suggested in this article, in which a coherent model was integrated into the construction of similarity matrix and feature selection. Finally, the efficacy of the proposed approach was validated by five state-of-the-art FS approaches compared (Fig. 5 and Table 2).

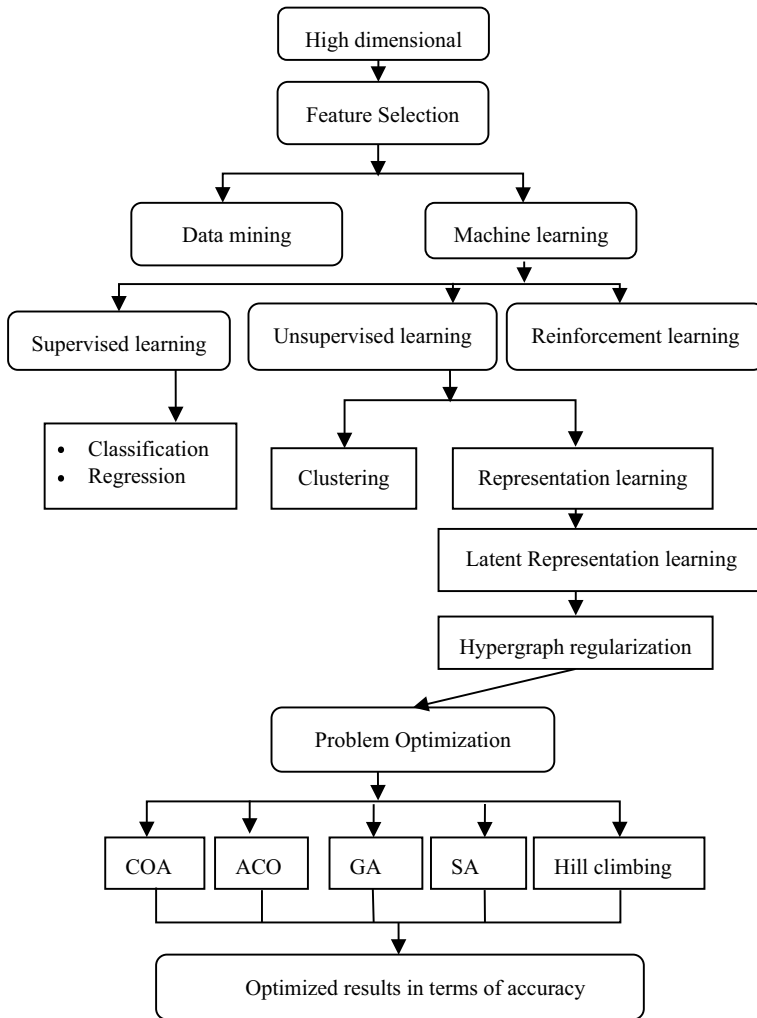


Fig. 5 Hierarchical representation of overall work

**Table 2** Comparison of various results based on a literature survey

Authors	Problems	Method	Findings	Limitations
Liu et al. [15]	dimensionality reduction	robust neighborhood embedding (RNE) method	effective and superior	examining further convex or non-convex regularizes forms
Tang et al. [16]	Massive unlabelled data	manifold regularization and latent representation learning	choose most of the discriminative features of the original data good efficiency	not the fastest
Ding et al. [17]	no dependency between data instances multifaceted structures implied in data	adaptive hypergraph regularized latent representation learning	more robust to noises strong and stable convergence behavior	cannot theoretically prove the convergence
Zhou et al. [20]	balanced k-means and feature selection ADMM (Alternating Direction Method of Multipliers)	the balanced structure of data	better clustering performance leads to a more balanced clustering structure	suffers from high time and space complexity
Yuan et al. [18]	nonnegative spectral clustering and sparse matrix regression	problem of optimization lack of the label information	ideal clustering labels effectively select the feature subset and provide more discriminative information	usually designed for vector-based approaches exploring optimum model selection parameters is required
Yang et al. [19]	unsupervised LSL-FS (Local Structure Learning Feature Selection)	FS highly based upon originally constructed similarity matrix	solve the objective function more reliable	block diagonal structure
Cui et al. [21]	robust latent representation learning (RLRL)	RLRL optimization problem	simultaneous FS and robust regression efficient convergence	RLRL incorporates learning latent representation and building expectation model into one structure for efficacious regression

(continued)

**Table 2** (continued)

Authors	Problems	Method	Findings	Limitations
Ensieh Iranmehr et al. [22]	PSO technique nearest neighbor	remove audio features like the human's ear to get better classification	remarkable enhancement of sound classification accuracy	The selection of the feature is adapted over iterations by the classifier

## 6 Conclusion

Feature selection is a significant challenge in ML and has gained substantial attention in past years. For certain labeled data, supervised feature selection methods are used as unsupervised methods. In this article, we have addressed a variety of unsupervised methods of feature selection via the combination of hypergraph and latent representation. We have to do a Selection of features in the latent space which is more sensitive to noise, rather than calculating significant features in the original data space. However, such approaches for selecting features produce reasonable results but have some problems with unlabeled high-dimensional data to pick proper features, including such space and time complexity, convergence problem, and optimum model parameters selection. The most critical challenge in existing works is to provide an accurate classification rate of features for various approaches. To overcome such issues, the concept of hypergraph will be adjusted and integrated into the proposed model for the capture in a high order of the local diverse geometric structure of original data. In this case, the suggested model is to be improved with an active alternating algorithm. For such purpose, we will try to develop a hybrid model using Multicluster feature selection and unsupervised discriminative feature selection.

## References

1. Chowdary BV (2018) A survey on applications of data mining techniques. *Int J Appl Eng Res* 13:5384–5392
2. Dey A (2016) Machine learning algorithms: a review. *Int J Comput Sci Inf Technol* 7:1174–1179
3. Mishra R, Sajja P (2018) Experimental survey of various dimensionality reduction techniques. *Int J Pure Appl Math* 119:12569–12574
4. Qiu J, Wu Q, Ding G et al (2016) A survey of machine learning for big data processing. *EURASIP J Adv Signal Process* 67:1–16
5. Reddy VK, Babu UR (2018) A review on classification techniques in machine learning. *Int J Adv Res Sci Eng* 7:42–47
6. Kaur S, Jindal S, Mark RG, Moody GB, Olson WH, Peterson PS, Schuler SK, Walters JB (2016) A survey on machine learning algorithms. *Int J Innov Res Adv Eng* 3:6–14
7. Jain K (2017) A survey on feature selection techniques. *Int J Innov Eng Res Technol* 4:1–4
8. Miruthula P, Roopa SN (2015) Unsupervised feature selection algorithms: a survey. *Int J Sci Res* 4:688–690

9. Miao J (2016) A survey on feature selection. *Information Technology and Quantitative Management (ITQM 2016)*, Proc Comput Sci 91:919–926
10. Miao J, Niu L (2016) A survey on feature selection. Proc Comput Sci 91:919–926
11. Bengio Y, Courville A, Vincent P (2013) Representation learning: a review and new perspectives. *IEEE Trans Pattern Anal Mach Intell* 35:1798–1828
12. Ding D, Yang X, Xia F, Ma T, Liu H, Tang C (2020) Unsupervised feature selection via adaptive hypergraph regularized latent representation learning. *Neurocomputing* 378:79–97
13. Zhang Z, Lin H, Gao Y (2018) Dynamic hypergraph structure learning. In: *Proceedings of the twenty-seventh international joint conference on artificial intelligence (IJCAI-18)*, pp 3162–3169
14. Pinto VD, Pottenger WM, Thompkins WT (2000) A survey of optimization techniques being used in the field. In: *Proceedings of the third international meeting on research in logistics (IMRL 2000)*, pp 1–14
15. Liu Y, Ye D, Li W, Wang H, Gao Y (2020) Robust neighborhood embedding for unsupervised feature selection. *Knowl-Based Syst* 193:1–23
16. Tang C, Bian M, Liu X, Li M, Zhou H, Wang P, Yin H (2019) Unsupervised feature selection via latent representation learning and manifold regularization. *Neural Netw* 117:163–178
17. Zhou P, Chen J, Fan M, Du L, Shen YD, Li X (2019) Unsupervised feature selection for balanced clustering. *Knowl-Based Syst* 193:1–11
18. Yuan H, Li J, Lai LL, Tang YY (2019) Joint sparse matrix regression and nonnegative spectral analysis for two-dimensional unsupervised feature selection. *Pattern Recogn* 89:119–133
19. Yang S, Nie F, Li X (2018) Unsupervised feature selection with local structure learning. In: *25th IEEE International conference on image processing (ICIP)*, pp 3398–3402
20. Cui J, Zhu Q, Wang D, Li Z (2019) Learning robust latent representation for discriminative regression. *Pattern Recogn Lett* 117:193–200
21. Iranmehr E, Shouraki SB, Faraji MM (2017) Unsupervised feature selection for phoneme sound classification using particle swarm optimization. In: *5th Iranian joint congress on fuzzy and intelligent systems (CFIS)*, pp 86–90

# Adoption of Blockchain Technology for Storage and Verification of Educational Documents



Vijay Anant Athavale, Shakti Arora, and Anagha Athavale

**Abstract** The importance of creating a reliable storage system for electronic documents in the field of education is a well-known reality. The general characteristics of the key methods of protection against counterfeiting in the field of education were analyzed. The example of a prototype of the information system named UPADHI is presented. The basis of the system is blockchain technology. This technology can be applied in a number of educational institutions to ensure the security and legality of documents. Using the IDEF methodology, the basic information flows in the system are studied. The functional blocks of the developed prototype are described and highlighted. In addition, an algorithm is proposed to validate a transaction in a socially oriented system by voting for participants who do not have equal rights.

**Keywords** Blockchain · Educational document · Storage · Verification · Fuzzy logic

## 1 Introduction

In our rapidly changing world, the requirement of education and training specialists is in demand as the non-availability of trained labor in the market is becoming a very serious problem. Due to the rapid leap in the development of telecommunications technologies, large audiences of students have new opportunities to improve their professionalism. E-learning has become one of the main innovations that is increasingly diffused in higher education institutions. The main objective of the adoption of

---

V. A. Athavale (✉)

Walchand Institute of Technology, Solapur, Maharashtra, India  
e-mail: [vijay.athavale@gmail.com](mailto:vijay.athavale@gmail.com)

S. Arora

Panipat Institute of Engineering and Technology, Samalkha, Haryana, India  
e-mail: [shakti.nagpal@gmail.com](mailto:shakti.nagpal@gmail.com)

A. Athavale

BARC, Mumbai, India  
e-mail: [anaghaathavale14@gmail.com](mailto:anaghaathavale14@gmail.com)



e-learning in educational institutions and beyond is due to the increased accessibility to the educational process without restrictions of place and time, also substantially improving the quality and content of education. This makes e-learning superior, especially for those who are employed and also interested in improving their skills. E-learning refers to the use of electronic media and technology education, also ICT such as the Internet, email and computers in the educational process [1].

The main aim of the modern educational system is to create a free and comfortable environment for the academic as well as skills development of the learner: An environment in which the learner himself determines the timing, pace of training and form of knowledge. The location of the training institution is no longer important; the main thing is the availability of access to Internet resources. This concept leads to the human formation of a new kind—"Homo Faber", that is, the man who makes and manufactures or, in a broader aspect, a person who acts, and occupies an active life position in the modern world.

The next trend of our times is Life-Long Learning. For professional growth, it becomes prestigious and useful to have several educational documents of various levels and careers. The workers collect their wallets, accumulating certificates of approval for professional tests, courses, schools and other types of training. All this influences the person's career and professional as well as personal growth.

For the growing demand for affordable training, there is a corresponding offer in the form of global online educational sites such as COURSERA, EdX and Udacity. They not only create great value for consumers, but also indicate that data is becoming a new class of asset, one that can outperform previous asset classes [2]. The e-learning portfolio is a collection through which the student, in the information technology environment, applies information media to represent and display objectives of learning, learning activities, learning achievements, learning performance, efforts to learn, progress of studies and introspection of the learning process and result, which mainly includes learning assignments, participation in learning, learning selection, learning strategy and learning introspection, etc. [3].

However, with the obvious advantages of e-learning, the following difficulties should be highlighted in its implementation. First, the educational institution must have a confirmation that the participant of the distance course and the person requesting an educational certificate are the same. Second, each year, the number of illegal educational documents is growing by leaps and bounds. In the summer of 2018, articles in University World News often report new cases, the most recent on fake Scottish degrees. Over the summer in the United Kingdom, there was an article published in The Guardian in which the UK's official service for verifying degrees, the Higher Education Degree Datacheck (HEDD), urged new graduates who take selfies with their new degrees not to share the images on social media to avoid fuelling the multimillion-pound trade in fake degrees [4]. When Chinese applicants enter American universities, it is estimated that 90% of Letters of recommendation are false, 70% of trials are not student work and 50% of the transcripts are counterfeit [5]. Third, when they move to another state or while looking for work abroad, a training diploma should be recognized as a diploma in general, and should also be comparable to the country's training standards host. To legalize foreign degrees and

diplomas, etc. there are a number of intermediary organizations conducting an assessment or verification. The intermediary company deals with the issue of authenticity of the educational certificate and issues a certificate indicating the equivalence.

Eliminating and combating these deficiencies is a problem throughout the world. Significant efforts have been made throughout Europe to identify false documents. Recently, the FRAUDOC project has issued guidelines on counterfeiting of educational documents and document fraud for credential testers. In the Russian Federation, to combat educational fraudulent documents, in August 2013, Government of Russia issued a decree [6]. This document implies the implementation of the information system “Contingent”, which would accumulate diplomas and certificates electronic devices of a citizen. The Republic of Ecuador has the National Secretariat of Higher Education, Science, Technology and Innovation (SENESCYT), whose functions are to keep a record of educational documents, including their verification and approval [7].

## 2 Developing

### 2.1 *Possibility of Using the Distributed Registry for the Document Storage*

From the above, it follows that there is an urgent need to create a technology to solve the problem of lack of trust and to provide reliable storage and secure electronic data. In our opinion, such technology can be a multifunctional and multilevel information technology called distributed registry. Distributed ledger technology is an approach for the exchange and storage of information, which has undeniable advantages such as owning a complete copy of the registration, online data synchronization, and instant access to the transaction history.

The main reason for the increased interest is the expectation that this technology will eliminate the problems and limitations inherent in traditional methods of storage, accounting and transmission of information. Blockchain is based on the distributed ledger technology concept. A blockchain is a storage mechanism designed to keep a record of transactions that occur between two parties permanently and verifiably. Blockchain is an open peer-to-peer mechanism and distributed data storage [8]. Each block points to the immediately preceding block through a reference which is essentially a hash value of the block above (parent block) and is calculated using the hash. Hash function or summary function—a function which converts a set of input data to a string of output bits of a specified length, performed by a specific algorithm. The result of the conversion (string output) is called a “hash”. It is worth noting that hashes of uncle blocks (children of the block’s ancestors) would also be stored on the blockchain.

The first block of a blockchain is called a genesis block which has no primary block [9, 10]. Each block except the primary has a link to information from the previous

block. The blocks themselves, as well as the data within them, are protected by chaining. Each entry contains a link to a previous source entry as well as a condition lock and an unlock rule. To describe the rules and conditions, a programming language is used that allows to establish complex logic and rules for the interaction of the participants. There can be various sources and results in each record, that is, a record can convert multiple source records to multiple source records results. Therefore, the blockchain leads us to “smart” contracts that allow us to formalize relationships not only between people, but also between a person and programs. As such, blockchain has the potential to build trust between independent actors and not family members who have the right to collaborate without requiring any type of central authority [11].

The construction of a common chain is based on the Merkle hash tree. The Merkle hash tree is a data structure. Its structure is not very different from the structure of usual data in tree form. The basis is the calculation of the hash based on the hashes of all child nodes. The tree root is written to the block header, which makes it impossible to remove or replace blocks of transactions. Since the block headers are adequately protected from changes, to replace all the blocks, the attacker must replace all the preceding blocks on the blockchain. In this way, a system suitable for protecting block headers makes it impossible to remove transactions from the chain of blocks registered in it relatively long ago; therefore, the blockchain satisfies the finiteness condition of transactions [12].

The structure of each transaction block is the same and it consists of the following parts [13]:

- (1) block version;
- (2) hash of the previous block (parent block);
- (3) hash of all transactions in the block;
- (4) date and time the block was created;
- (5) target, i.e. link to the current hash target in the list;
- (6) a 4-byte field that increases its value after each consecutive transaction or addition.

In the structure of the block, there is no need to store the document itself. It turns out that the data remains in user possession, and only the hash value of the specific document enters the distributed system. Keeping the password private and secret, the user can always recreate the hash value of the document being checked. The coincidence of the value obtained with the hash specified in the valid transaction determines the authenticity of the document univocally.

The security of traditional databases such as MySQL and MongoDB is guaranteed by a centralized data control, which is the responsibility of the system operator. The fact that management functions are assigned to humans makes conventional databases vulnerable. An example is the fact of compromising the biometrics database from the Indian agency UIDAI [14].

Mistakes and malicious actions of employees of the companies that process large amounts of data are dearly costing the companies and society at large. The market for analyzing voluminous user data is growing and with this the number of messages

about the commitment of hundreds of millions of user records. In June 2018, security researcher Vinny Troia [8] discovered that Exactis, a data broker based in Palm Coast, Florida, had exposed a database containing about 340 millions of individual records on the public access server. “It seems this is a database with almost all American citizens”.

The blockchain protocol and its structure, which are based on secure data storage, create an automated system of record that forms an interconnected system and fights emerging threats. Smart contracts are unprecedented methods to ensure contractual compliance, including social contracts. “If you have one large transaction with a specific control structure, you can predict the outcome at any time” Tapscotts [2]. If I have a fully signed transaction verified with a number of signatures in a multiple signature account, I can predict that transaction will be verifiable through the net. And if it is verifiable by the network, then that transaction can be redeemed irrevocably.

No central authority or third party can revoke it, no one can override network consensus. It’s a new concept both in law and in finance. The bitcoin system provides a very high degree of certainty as to the result of a contract [2]. The combination of smart contracts, as well as the idea of blockchain technology will allow you to organize a system to store information about all certificates and diplomas received by a person. The advantages provided by the combination of these methods solve the problems of an electronic document repository which is affordable, reliable and open.

## ***2.2 Blockchain Technology Offers the Following Five Benefits [8]***

- (1) **Efficiency:** Blockchain can be easily managed, and it can track complex data records.
- (2) **Security:** The security provided by a chain block is better than data management centralized. In the latter case, there is a risk of damage due to intrusion by a hacker. On blockchain, fake data is almost impossible due to the simultaneous control of the devices where data is stored, such as mobile devices. Also, the hacker needs to change all data stored on devices to fake it.
- (3) **Resilience:** All information on the blockchain is not stored in one place compared to a centralized data management, but the information is distributed equally among interacting mobile devices. There is no single point of failure. Even if multiple devices find errors or performance degradation, the possibility of malicious infrastructure finding threats from blockchain is scarce. Even if it is attacked, it can easily recover.
- (4) **Transparency:** A blockchain transparently opens all the resource status and usage data by default because it shares the resource metadata with all the participating mobile devices. The exclusive occupation of resources by specific mobile devices inside the MRM infrastructure was prevented in this research.

Therefore, the use of blockchain technology leads to the emergence of smart assets. By regulating the work of this concept through blockchain technology along with applicable law, it is possible to implement a global electronics piggy bank of educational documents, where the authenticity is based on the existence of a decentralized system. The new Europe 2020 Strategy, a strategy for smart, sustainable integration, maintains a clear objective to get out of the crisis and prepare the EU economy for the challenges of the next decade [15]. The Digital Agenda for Europe, framed within The Strategy aims to obtain the economic benefits and sustainable social networks that can derive from a single market digital on a fast and ultra-fast Internet and more interoperable applications.

Currently, some universities and institutes use blockchain technology in education, and most of them use this technology to manage their academic degree and general assessment of learning outcomes. The University of Nicosia is the first to use the blockchain technology to manage the certificates of students received from the MOOC platforms [16].

Sony Global Education has also used blockchain technology to create a global assessment platform to provide storage services and manage the title information. The Massachusetts of Technology (MIT) also issues electronically academic degrees to students who participated in projects of MIT Media Lab and those who passed the evaluation will receive a certificate to be stored in the blockchain network [17]. Another interesting project in this field is the development of BlockTac company. The blockchain-based system is oriented and works with universities, business schools, professional associations in Spain, as well as with international educational institutions in Europe and Latin America [18]. Educational institutions and students match their certificates in the system. After completing the course, the student can study at any other educational institution education and share the certificate, whose legality can be verified directly in the system, without contacting the institution that issued it.

- A. Prototype of the secure storage system. Apparently, the development and implementation of systems based on distributed ledger technology is an urgent task and requested all over the world. However, it should be noted that architectural implementations and software and its detailed descriptions are almost impossible to find. Furthermore, it is worth noting the fact that existing systems do not fully satisfy the requirements of a particular user. To resolve these problems, we propose the development of a blockchain information-based system.

### **2.2.1 The Architecture of the Secure Storage System**

To design the domain model, an approach based on the IDEF0 methodology was used. The diagram of context IDEF0 reflects an overview of the UPADHI information system activities.

The input data is data about customers, and data about documents that refer to education. The output is represented by an entry in the blockchain, which is an electronic version of the document, statistics on the system usage and the results of a

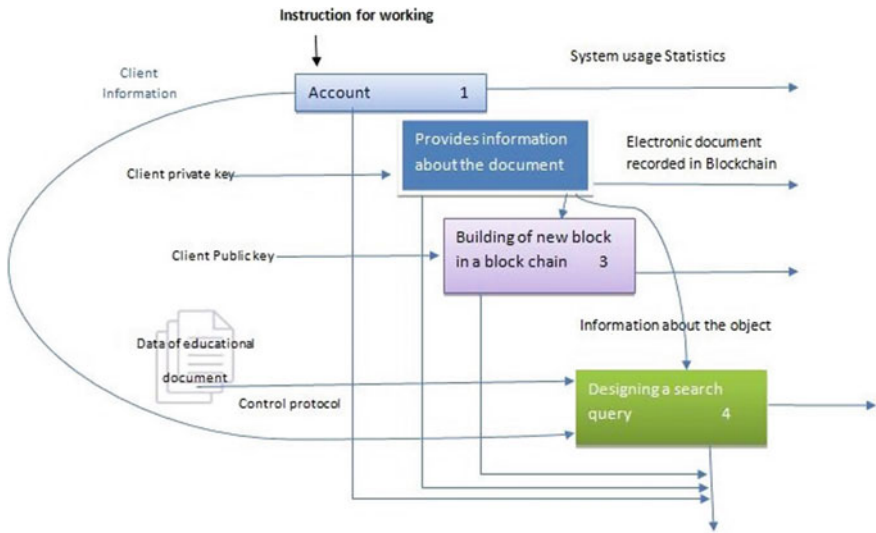


Fig. 1 Architecture

search query. The following decomposition allows us to determine the main types of work that occur in the system. According to the diagram in Fig. 1, the whole system is composed of four blocks. The data of all the users of the system is sent to the input of the first block. At the exit, we received a request to provide data in a document on education or statistics on the use of the system. In the second block, a new transaction is created that contains information about the transmitted document, as well as the data of the user that forms this transaction. In the third block, a new element of a block of the blockchain contains all the transactions created in the previous step. In the fourth block, depending on the information sent on the document, a search request is generated. As a result of the search request, the user receives information about the document in the system, if any.

Based on the above information flows, the following subsystems (modules) as part of the information system are identified: cryptographic module, interaction between networks, blockchain module and application that provides a convenient way to interact with the system, the cryptographic module should be able to generate cryptographic keys based on the RSA-2048 standard. The keys, public and private, are interconnected by unidirectional communication. A summary of the personal data obtained by the SHA-512 algorithm and the user’s private key generates a transaction signature.

The interworking module is based on P2P technology. Point-to-point technology differs from approaches to network infrastructure scaling standards. When the “peer-to-peer” approach is applied, the main thing is not communication between the client-server, but the methods of searching for other clients on the network, by which they can exchange information between them. The control module (Blockchain) is used to create transactions, and transactional and authorization blocks. Furthermore, this

module describes algorithms to verify the information received, check the local data string, store data in the local string and a protocol confirmation.

### 2.2.2 The Transaction Structure

The following block and transaction structures depicted in Figs. 2 and 3 have been developed and implemented for the UPADHI information system. A block in a chain of blocks consists of a header and a body. The body of the block of documents is a description of the number of transactions included in it and the transactions themselves. The header of the block contains specific information that is required to maintain the ideology of this technology.

The system architecture is based on the technology of distributed registry. A distributed ledger is a database of information data that is distributed through a network of computers, servers or websites. This technology does not have a geographic or administrative location. All the network users, without exception, keep an identical copy and complete details of all the records. Modifying the registry will result in a copy, and this copy is sent to changes for all network participants in a short period of time. Any asset can be an entry in a registry distributed. Supervision of truthfulness and safety of registry entries are maintained through signatures and keys based on cryptographic algorithms that allow access control to the registry.

As users of the system, both organizations perform educational activities, as well as common current users. Any of the listed user types can connect to the system and provide a list of issued educational documents. The root hash is the sum of all hashes issued for the current session. Hashes of an unlimited number of users are placed in a block. The restriction is only the block size, which in this system is 500 kb, with the capacity to accommodate up to 130 transactions. The system user, when completing a request for confirmation of the authenticity of the document issued, creates a search request. The search implemented according to the binary search

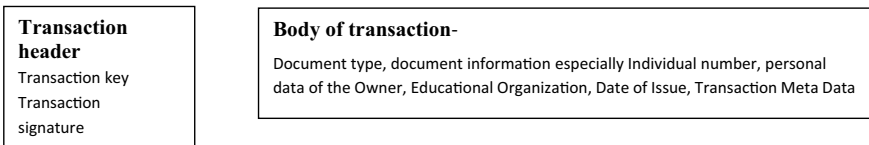


Fig. 2 Block structure

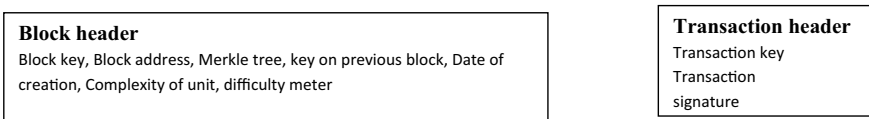


Fig. 3 Transaction structure

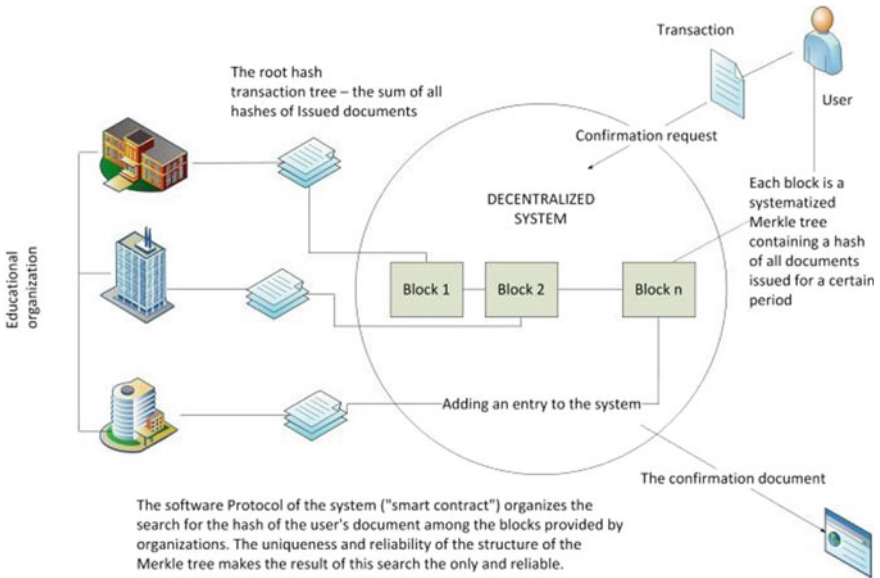


Fig. 4 Data search logic of software protocol

tree algorithm is random, since when creating the Merkle tree, the hash of the value comes in random order.

When registering in the system, the user receives a couple of keys: public and private. The public key along with registration information is recorded on the blockchain, but on a separate “tree branch” other than the one in the blocks where document data are written. In the next authorization, the program will ask the user to download a file that contains a key pair. According to his response and the software protocol algorithm as well as data stored on the blockchain, the question of grant or denial of access to your personal account. The public key is also used as the address of the block, which allows us to uniquely identify the user who has created it during the data entry process, which combines confirmed transactions in blocks. The data search is performed according to the logic of the software protocol as depicted in Fig. 4.

### 2.2.3 The System Interfaces

The navigation bar contains the logo of the application as well as menu items. The main space contains various forms of entry of information, tables and information output areas. The workspace can be divided into sections, depending on which menu item is currently selected.

Let us consider in detail the operation “Add a new document” (Fig. 5). This operation is carried out in two stages: (1) add a document; (2) formation of a block



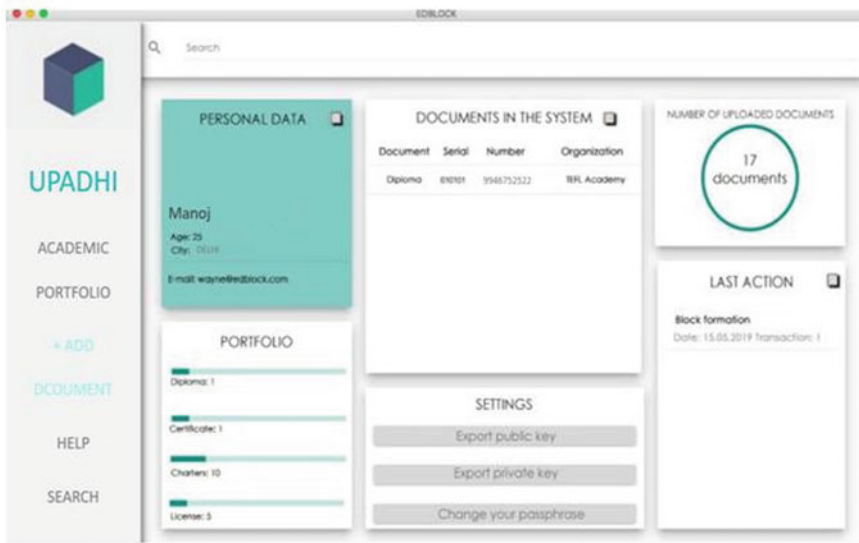


Fig. 5 The process of adding a document

with data. In the first stage, the user completes a form of sending documents. The user can place the data on the document manually using the document submission form and drag a file representing an electronic version from document to PDF backing, and the data will pass automatically, if possible, from document to form. After completing the form and taking into account all required fields, the user, by clicking on the button “Add document”, accumulates the data entered in the intermediate buffer. In the table named “Current transactions”, a new record of the archived document. To change the data about a document already sent, the user just needs to click the corresponding line in the table “Current transactions”. After entering the data for all the documents, you can proceed to the second stage of document filing.

At this stage, the user can begin to form a block or go back to the first stage to make additional changes in downloaded documents. At the beginning of the block formation, the user accepts that the data entered is correct and will not be changed later.

The following steps then occur:

Transaction formation: The temporary buffer registers are transferred to the transactional matrix. Each transaction is signed by the user’s private key, which is responsible for the accuracy of the information entered and the absence of defamation in it.

Block formation: All transactions created enter the body of the block; the public key of the person who creates this block is recorded; the hash of the previous block and other special information is entered in the block header by implementing the Merkle tree algorithm.

During the training, all the information about the job progress is displayed to the user. After the block is formed, the program will notify the user about this; at the same time the block will be sent to the network to confirm its validity. The user profile provides convenient access to personal and statistical information, a panel to export keys, a description of the last actions carried out, as well as an overview of the documents that are in the system and were submitted by this user. The user profile can be in three colors: Green, Yellow and Red. The color is set based on the level of the user trust and by default, all users have a “green profile”, that is, a profile in which none of the documents presented were doubtful.

The example of the user profile. The yellow color profile mentioned in Fig. 6 signifies that some documents sent to the user raised doubts among other members of the network. Due to the impossibility to verify the document with the organization that issued it, as they are not registered in the system, the documents like the profile are marked with a different color.

Now, let us consider an example of the untrusted user profile. If the user profile is red (Fig. 7), this means it was seen in the shipment of false documents. Your card with personal data, as well as documents is marked in the corresponding color. The users identified in the fraud are suspended from participation in the system, i.e. No permission to present documents for a period. Also, the reputation indicator drops. If the profile is yellow or red, the profile card also indicates the number of certain “unreliable” documents in the corresponding color. Distributed ledger consensus algorithm is oriented to social spheres. The validity of the block is confirmed by the algorithm of consensus. The difficulty of obtaining a common solution in the blockchain is associated with the following features:

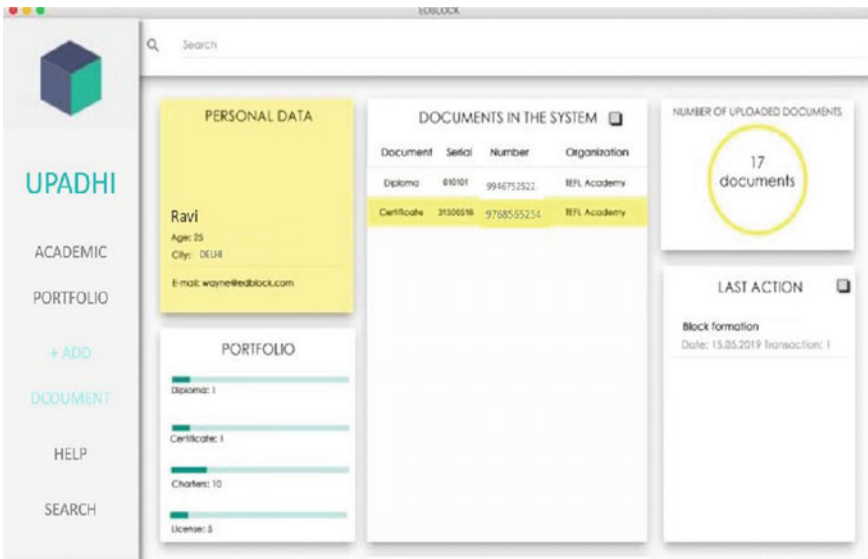


Fig. 6 The example of user profile

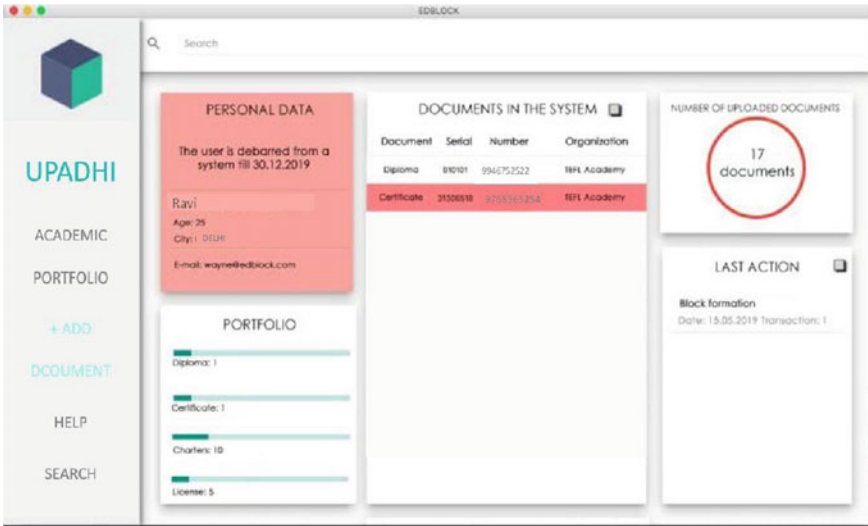


Fig. 7 Signify the untrusted user

1. Participants are strangers and can connect or disconnect freely.
2. Resistance to external influences. It is not possible to disconnect a node even if it is known in advance that the node is not trusted.
3. An authorized center is not required to confirm the transactions. The very principle of organizing a blockchain network is the root of trust.

The problem is solved using special algorithms, among which there are two categories: algorithms based on the “Proof-of-Work” work evidence and algorithms based on confirmation of the “Proof-of-Stake” part. The study of characteristics of previous working algorithms has led to understanding the inapplicability of the family of PoW algorithms to solve the class of social problems. Maintaining system performance through the efforts of the user is a key motivation for the algorithms of the PoS class. The information system “UPADHI” does not imply a monetary component. Motivation: the desire to create a unique record of legal educational documents confirming competencies and the knowledge of the students. The consensus algorithm of such a system needs a different form of remuneration because of work. As an alternative, it is proposed to use concepts such as “activity” and “right to vote”. One of a kind PoS algorithm is used as a basis: Proof of delegated participation or DPoS.

Traditional methods of systems analysis and computer modeling based on exact numerical methods are incapable of understanding the great complexity of human thought processes and the decision-making. The developed system is in every way humanist. Acceptance of this premise suggests that, for being able to make meaningful claims about the behavior of humanistic systems, it may be necessary to abandon the standards of rigor and precision [19]. That’s why the logic of the

algorithm is based on the concept of fuzzy sets and the application of fuzzy logic operations.

To determine the accumulation in the form of votes for the election of delegates, we present the set X “User of the system”. This set includes all users of the system. The right to vote in the election of a delegate is determined by the number of votes. This indicator is calculated based on the following criteria: time spent continuously in the system; total number of educational documents received; the number of educational documents with confirmation in the issuance of the relevant educational organization; account balance, estimated at the amount of the internal currency of the system. Users who express themselves actively in the system receive a reward in the form of one voice fee and form an “Active Voter” set A.

The logic of the algorithm is based on the concept of fuzzy numbers and the use of the fuzzy logic apparatus. Fuzzy set A of “Active voter” is represented by a set of pairs  $A = \{(f_A(x), x)\}$ , where x is the user of the system,

$f_A(x)$  is the membership function that determines the activity of the participant.

According to the theory of fuzzy sets, each of the criteria assumes a value from 0 to 1. These are

- $t(x)$  is the time that the user spends continuously in the system;
- $d(x)$  is the total number of educational documents received;
- $d^*(x)$  is the number of educational documents that have confirmation in the issuance of the relevant educational organization;
- $m(x)$  is the monetary balance of the account, estimated in the value of the internal currency of the system.

The voter can regulate the level of their activity increasing any of the local fuzzy criteria listed. The evaluation of the values of these criteria premises has led to the understanding of the impossibility of using the “hard” apparatus of interval mathematics diffuse. To obtain a quantitative value of the global criterion  $p_A(x)$ , L. A. Zade and R. Bellman proposed a system of “soft relationships”. The essence of soft computing is that, unlike traditional and hard computing, it is aimed at an adaptation with the general imprecision of the world real. Therefore, the guiding principle of soft computing is: “exploit tolerance for imprecision, uncertainty and partial truth to achieve traceability, robustness, the low cost of the solution and a better relationship with reality”. In the final analysis, the model to follow for soft computing is the human mind [20].

The final quantitative is determined as follows:

$$\begin{aligned}
 f_A(x) = & m(x) + d^*(x) + t(x) - m(x) \cdot d^*(x) - m(x) \cdot d(x) \\
 & - m(x) \cdot t(x) - d^*(x) \cdot d(x) - d^*(x) \cdot t(x) - d(x) \cdot t(x) \\
 & + m(x) \cdot d^*(x) \cdot d(x) + m(x) \cdot d^*(x) \cdot t(x) + m(x) \cdot d(x) \cdot t(x) \\
 & + d^*(x) \cdot d(x) \cdot t(x) - m(x) \cdot d^*(x) \cdot d(x) \cdot t(x)
 \end{aligned} \tag{1}$$

Set Y “Delegates” is formed by choosing representatives. Inside Y, a fuzzy set V “Voting delegates” is formed. The main task solved by the representatives of this

set is the regulation of the system by generating blocks. As fuzzy criterion  $rV(y)$ , a concept such as “delegate reputation and” was chosen. The accuracy of the selected judgments is confirmed by studies by other authors in this area:

“Reputation is a foundation of the new digital economy, with companies like AirBnB and Uber building the trust through ratings and reviews. Between the academics, reputation is already a commercial product, and Promotion and recruitment are partly based on reputation measured by the number of citations and the H-index metric of impact of the publication. Imagine that the trade of the academic reputation could extend beyond the world of academics and become the foundation of an educational economy” [16].

The reputation criteria  $rV(y)$  of the delegate and how evidence of initial ownership interest is determined are based on the following five local fuzzy criteria:

$l(y)$  is availability of a license to perform educational activities;

$d(y)$  is the number of educational documents issued by the organization;  $v(y)$  is the total number of votes received from users of the system;

$p(y)$  is portion of valid signed metadata in the invalid;  $s(y)$  is the total number of participations as a delegate.

The final value of the reputation criterion  $rV(y)$  is calculated similar to the fuzzy set membership function “Active voter” (1) and can be represented in the formula in the following manner:

$$\begin{aligned}
 rA(y) = & l(y) + d(y) + v(y) + p(y) - l(y) \cdot d(y) - l(y) \cdot v(y) \\
 & - l(y) \cdot p(y) - d(y) \cdot v(y) - d(y) \cdot p(y) \\
 & - v(y) \cdot p(y) + l(y) \cdot d(y) \cdot v(y) + l(y) \cdot d(y) \cdot p(y) \quad (2) \\
 & + l(y) \cdot v(y) \cdot p(y) + d(y) \\
 & \cdot v(y) \cdot p(y) - l(y) \cdot d(y) \cdot v(y) \cdot p(y) \cdot (1 - s(y)) + s(y)
 \end{aligned}$$

Each of the selected delegates is involved in the formation of a valid block of the transactions of the users. Each delegate has a fixed time to issue chain blocks within a single voting cycle. To hold an open vote on the validity of the block formed, the block is considered valid, if more than half of the delegates accept it. After completing the whole voting circle, again an auction is held to select new delegates.

### 3 Conclusion

The introduction of an accumulation system of educational documents based on the ideas of blockchain technology will solve the problems of affordable, reliable and at the same time an open repository. Due to the growing popularity of massive open online courses (MOOC) and the fast development of e-learning technologies propelled by Covid-19, the dynamic monitoring of popular learning areas will solve the problems of responding quickly to all educational level requirements. Reliability and protection of educational documents will provide distributed records. The

UPADHI system involves voluntary and free participation of universities, educational institutions, government agencies and qualified specialists. Each of them accepts the rules of the system and does not deliver money, but its own reputation.

## References

1. Al-araibi AAM, bin Mahrin MN, Yusoff RCM, Chuprat SB (2018) A model for technological aspect of elearning readiness in higher education Received: 15 March 2018. Accepted: 6 November 2018 /Published online: 26 November 2018 # Springer Science+Business Media, LLC, part of Springer Nature 2018
2. Tapscott D, Tapscott A (2016) Blockchain Revolution: How The Technology Behind Bitcoin Is Changing Money, Business, and the World. Penguin
3. Yongqiang H, Jinwu Y (2011) Study on the evaluation system of e-learning based on e-learning portfolio. In: Computing and intelligent systems: international conference, ICCIC 2011, Wuhan, China, September 17–18, 2011. Proceedings, Part 3. Springer Science & Business Media
4. Børresen LJ, Skjerven SA. Detecting fake university degrees in a digital world. <https://www.universityworldnews.com/post.php?story=20180911120249317>
5. Hesselbäck A. The modern counterfeit industry and higher education. [http://www.skvc.lt/uploads/documents/files/Naujienos/Andre\\_Hesselback\\_Vilnius\\_November\\_2016.pdf](http://www.skvc.lt/uploads/documents/files/Naujienos/Andre_Hesselback_Vilnius_November_2016.pdf)
6. Adaptive and Adaptable Learning: 11th European Conference on Technology Enhanced Learning, EC-TEL 2016, Lyon, France, September 13–16, 2016, Proceedings, pp 490–497. Springer, 2016
7. <https://www.refworld.org/docid/52a831074.html>
8. Secretaría de Educación Superior, Ciencia, Tecnología e Innovación <https://www.educacion.superior.gob.ec/>
9. Kim H-W, Jeong Y-S (2018) Secure authentication-management human-centric scheme for trusting personal resource information on mobile cloud computing with blockchain. HCIS 8(1):1–13
10. Bansal A, Athavale VA (2021) Big data and analytics in higher educational institutions. In: Mariwala N, Tripathi CC, Kumar D, Jain S (eds) Mobile radio communications and 5G networks. Lecture notes in networks and systems, vol 140. Springer, Singapore. [https://doi.org/10.1007/978-981-15-7130-5\\_15](https://doi.org/10.1007/978-981-15-7130-5_15)
11. Zheng Z, Dai H-N, Xie S (2018) Blockchain challenges and opportunities: a survey. Int J Web Grid Serv
12. Barbosa LS (2017) Digital governance for sustainable development. In Kar A et al (eds) Digital nations – smart cities, innovation, and sustainability. I3E 2017. Lecture Notes in Computer Science, vol 10595. Springer
13. BitFury Group, Jeff Garzik. Open and closed blockchains. Part 1. <http://www.forklog.com/wp-content/uploads/public-vs-private-pt1-1.0-ru.pdf>
14. Hayaat H-E, Priya A, Khatri A, Dixit P (2018) Rise of blockchain technology: beyond cryptocurrency. applications of computing and communication technologies. In: First international conference, ICACCT 2018, Delhi, India, March 9, 2018, Revised Selected Papers, pp 286–300. Springer
15. Athavale VA, Bansal A, Nalajala S, Aurelia S (2020) Integration of blockchain and IoT for data storage and management, Materials Today: Proceedings. <https://doi.org/10.1016/j.matpr.2020.09.643>
16. Rs 500, 10 minutes, and you have access to billion Aadhaar details. Tribune Investigation - Security Breach. Posted at: Jan 4, 2018, 2:07 AM. <https://www.tribuneindia.com/news/nation/rs-500-10-minutes-andyou-have-access-to-billion-aadhaar-details/523361.html>

17. Sharples M, Domingue J (2016) The Blockchain and Kudos: A distributed system for educational record, reputation and reward. adaptive and adaptable learning. In: 11th European conference on technology enhanced learning, EC-TEL 2016, Lyon, France, September, 13–16, 2016, Proceedings, pp 490–497. Springer
18. Meier A, Stormer H (2018) Blockchain = Distributed Ledger + Consensus. HMD Praxis der Wirtschaftsinformatik December 2018, vol 55, issue 6, pp 1139–1154. <https://doi.org/10.1365/s40702-018-00457-7>
19. Grigg I. Seeking consensus on consensus—DPOS or delegated proof of stake and the two generals' problem <https://www.steemit.com/eos/iang/seeking-consensus-on-consensus-dpos-or-delegated-proof-of-stake-andthe-two-generals-problem>
20. Zadeh LA (1974) The concept of a linguistic variable and its application to approximate reasoning. In: Fu K S, Tou JT (eds) Learning systems and intelligent robots. Springer, Boston, MA

# Obstacle Collision Prediction Model for Path Planning Using Obstacle Trajectory Clustering



Samir Ajani and Salim Y. Amdani

**Abstract** A probabilistic path planning is one of the major areas of interest in the field of navigation. Toward the goal of path planning, collision prediction is one of the key research challenges in order to achieve the high accuracy of a probabilistic path. In order to build a collision forecasting mode, we have proposed a multilayer collision prediction model for path planning in the dynamic environment. The binary classifier is used to classify the obstacles, that is, whether the obstacle is a dynamic or a static obstacle. The trajectory clustering is done with the help of predicting the obstacle trajectory and identifying the turning points of an obstacle over the same trajectory. The points that are generated by the obstacle trajectory clustering model are used by the time series location forecasting model after performing data de-noising and outlier detection, to generate the obstacle-free points over the working space in the dynamic environment. This time series location forecasting model provides the location of all the obstacles over the time series model that is used as an independent variable for the collision prediction model. Finally, the collision prediction model provides the most appropriate points that can be considered in the collision-free path and generate an obstacle-free probabilistic path over the dynamic environment.

**Keywords** Obstacle · Collision · Navigation · Trajectory · Dynamic · Location · Forecasting · Prediction

## 1 Introduction

Path planning is the task of finding a path for a moving robot from source to a destination. In the different application areas where the robots are working, it is necessary to build a collision detection and avoidance mechanism. The collision detection mechanism has to consider the movement of obstacles in the environment.

---

S. Ajani (✉) · S. Y. Amdani  
Babasaheb Naik College of Engineering, Pusad, Maharashtra, India  
e-mail: [samir.ajani@gmail.com](mailto:samir.ajani@gmail.com)

S. Ajani  
St. Vincent Pallotti College of Engineering and Technology, Nagpur, India



There are two types of environments: static environment and dynamic environment. A static environment is that where the obstacles are static which means their positions are static. The second environment is a dynamic environment [1] where the obstacles are moving in the environment and their positions are changing during the course of time. So, it is necessary to identify the locations of the moving obstacles in the dynamic environment.

Nowadays due to the introduction of emerging technologies like machine learning, deep learning, IoT, etc., it has been much easier to build a prediction model which will predict the future location of an obstacle for collision detection and prediction. In recent years, the machine learning and artificial intelligence mechanisms emerged as key techniques in the field of navigation and path planning. These methods are more suitable for dynamic environments where the shape and size of the working environment have been changing over the period of time. The time series analysis will be performed over the environment to get the environment parameters and perform the motion planning over the working space.

The random exploring tree (RRT) which is used for motion planning uses the Gaussian mixture model (GMM's) [2] which presents the collision-free space along with the points of collision. The RRT algorithm with GMM will generate the probabilistic path more faster than the other variants of the RRT, as it uses the proxy collision detector and sample distribution to find the node of the tree. Neural networks are also used to find collision detection in the physical environment having a rectangular shape. Neural networks show less error in calculating collision detection and provide better accuracy in the defined space. Defining neural network parameters for finding collision detection is difficult and that can be eased by increasing the number of trained neurons in the network. The artificial potential field is also a technique which is used for motion planning and collision detection. The idea behind this technique is to automatically generate the obstacle waypoints during the navigation over the defined environment. These generated waypoints can be used in training the neurons in the neural networks nearest neighbor (KNN) model [3] have been used in the static environment to generate a probabilistic path and a collision detection mechanism over a defined C-Space. The KNN uses a sampling-based technique for collision detection which helps to generate a probabilistic path much faster than the existing benchmark techniques.

In Fig. 1, a scene from the ETH multi-human trajectory dataset evolves over time. An undirected graph representation of the same scene is also visualized, illustrating how its structure varies through time. Nodes and edges are represented as white circles and solid black lines, respectively. Arrows depict potential future agent velocities, with colors representing different high-level behavior modes. They are only shown once for clarity.

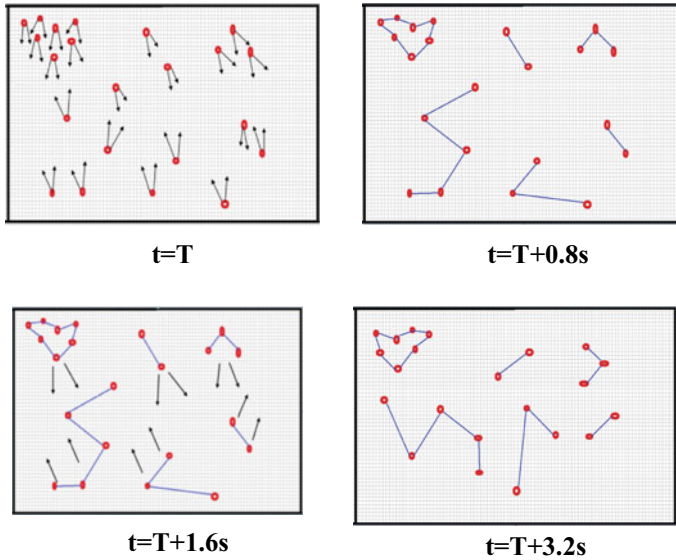


Fig. 1 ETH multi-human trajectory dataset as it evolves over time

## 2 Related Work

The research community for the past few decades is working on a similar field of navigation to solve the fundamental problems in this field like obstacle trajectory prediction, obstacle collision detection, obstacle motion prediction, etc. There has been a lot of research carried out and the result finds have been published in reputed communication in digital libraries. In this section, we try to include and discuss some of the research findings in the area of collision detection and prediction.

In this work [4], the author presented the collision detection mechanism for very fast moving objects. While presenting the collision detection mechanism for very fast moving objects, the major role is of the control unit. In this, the automatic control unit has been presented which is responsible for the prediction of collision in fast moving objects. The major problem in this fast moving collision detection algorithm is that it needs to consider the situations that are even accrued between two frames of the input stream which is used for the collision detection. To address this issue, the author proposed an assumption that if the two objects are passed through the same region within the speculated time window that would be considered as a collision. This can be mathematically solved by considering the convex hull in a vertical direction of the two frames considered to identify the collision and then applying the inter-frame collision detection algorithm to predict the collision.

In the article [5], the collision detection and prediction model has been presented, which is based on a geometrical model. In this, the field-based robot is used as an object in which collision detection and prediction have been performed. The robot is

working in the nuclear power plant in the section of steam generation. In the steam generation section, the temperature is too high. The uninterrupted functioning of this steam section is largely depending upon the working of the robot. So it is important to provide the mechanism which helps to predict the future collision of a robot. The collision prediction model uses the master–slave computer system along with a geometrical model to predict the collision.

In this research work [6, 7], the author presented the offline system used to detect the collision in a robotic environment. The collision detection mechanism is very important for the environment where the system working completely depends upon the robots and their functioning. The functioning of the robots needs to be controlled whether by using real-time collision detection mechanism or by using feed forward offline mechanism. In the feed forward offline mechanism, it is very important to provide a robust and accurate collision detection mechanism which helps the robot to work effectively and accurately to generate a correct moving trajectory. In order to generate a correct moving collision-free trajectory, the author proposes an offline feed forward space subdivision method. To increase the efficiency of robot trajectory, it uses space decomposition technique along with hierarchical boundary box. To make the collision detection more accurately, the distance threshold is set, to meet the desired accuracy.

In this work [8, 9], the author has presented that Collision detection is one of the important parameters in the field of gaming. In this paper, the author explores the importance of collision detection in gaming; they also presented the improved algorithm which is used to detect the collision in a grapple game. The objects that are considered for collision are bounded with the rectangle which is axis aligned and there is a circle to fit the object in order to recognize object correctly and effectively. In the presented improved algorithm for collision detection, the coordinate value of the object axis aligned rectangle and center marked effectively. The rectangle is marked with the center and the four corner points and the circle is marked with the center and radius. These coordinate points of rectangle and circle help to detect the collision using some distance measure to identify the collision. The results of this improved algorithm are a binary results which means collision may or may not occur.

The author [10] has presented that Collision detection and prediction can be performed with the help of sensor based or non-sensor based systems; in this paper, the author presented the sensorless mechanism to predict and detect the collision. The sensorless mechanism uses the innovative collision detection mechanism using velocity deviations in the robots as a key ingredient. Here in this mechanism, to calculate the velocity of a robot in the sensor less environment, it is important to consider all the components used to create a robot. There may be some delay in the calculated velocity and the actual velocity of a robot in this research. The gap between the actual velocity and the calculated velocity is minimized by using a low pass filter. In the complete mechanism basically, three controllers are used named positional controller, velocity controller and current controller, which will help to detect and predict the collision effectively, accurately and sensitively.

In this article [11, 12], the author has presented a quick collision detection algorithm using surface feature extraction and projection feedback is proposed. The proposed algorithm needs pre-processing; during this step, the moving area is divided into surface cells using projection feedback and extracting the surface feature. The major step in the quick collision detection algorithm is the precision collision detection step which uses the boundary box OBB model which makes collision detection precise. The speed of the proposed algorithm is improved and increased using parallel computing. The presented and proposed algorithm shows great improvement over the existing benchmark algorithms which are available to predict and detect the collision in real time which is described and shown using experimental results presented by the author.

In this paper [13, 14], the author addresses the issue of the high rate of accidents occurring at intersection points on the roadside. The intelligent vehicle trajectory prediction model is presented which is used as a warning system for drivers which are not following the correct trajectory because of any reason. This vehicle trajectory-based collision detection model is based on the integrated system which uses the vehicular infrastructure available for the communication and signaling to the vehicles. This system uses Time-To-Collision as a collision risk indicator parameter and a trigger to generate warning system for collision detection and prediction.

### 3 Proposed System

- a. **Binary Classifier.** It is used to categorize the obstacle in the static or dynamic environment. The binary classifier has two states: one is static (false state) and the other is dynamic (true state) classification (Fig. 2).
- b. **Obstacle Tracking:** the problem of obstacle tracking can be defined as the motion state of an obstacle including angle, velocity and position in the dynamic environment with the impact on obstacle trajectory detection and motion planning (Fig. 3; Table 1).

$$X = \iiint_C f(x; \mu, \Sigma, r) dx$$

$\mu$  is a mean.

$\Sigma$  is Variance.

$r$  is a radius.

The probabilistic prediction function along the axis can be represented as

$$F = R_z(\theta_1) * T_{x1}(l_1) * R_z(\theta_2) * T_{x2}(l_2) * R_z(\theta_3)$$

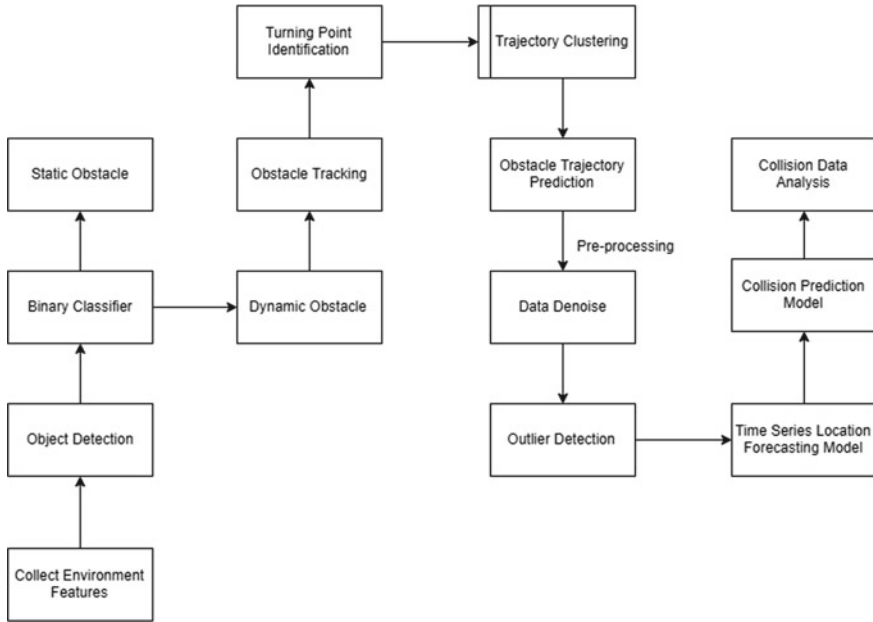


Fig. 2 Proposed system design

Fig. 3 Obstacle motion tracking

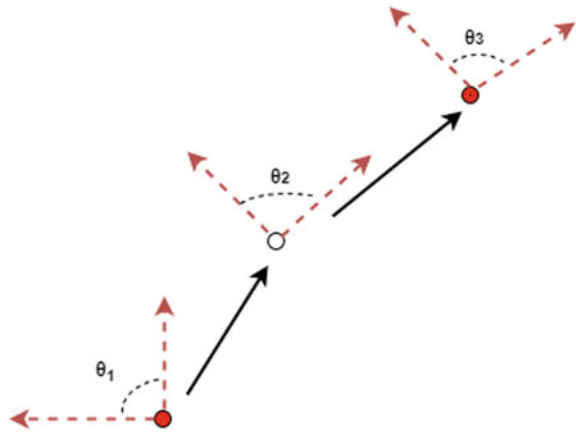


Table 1 Binary classifier prediction

Actual obstacles	Predicted obstacles	
	Static obstacle	Dynamic obstacle
Static obstacle (50)	42	8
Dynamic obstacle (80)	6	74

where R is the rotation to put the point in the next position, and T is a translation along the x-axis.

- c. **Trajectory Clustering:** Clustering is one of the useful methods of grouping similar class elements together. Identify and create a set of all the points that belong to the same trajectory in the current navigation.

$$t_i = \{c_1, c_2, \dots, c_n\}$$

$$t_1 = \{c_1^1, c_2^1, c_3^1, \dots, c_n^1\}$$

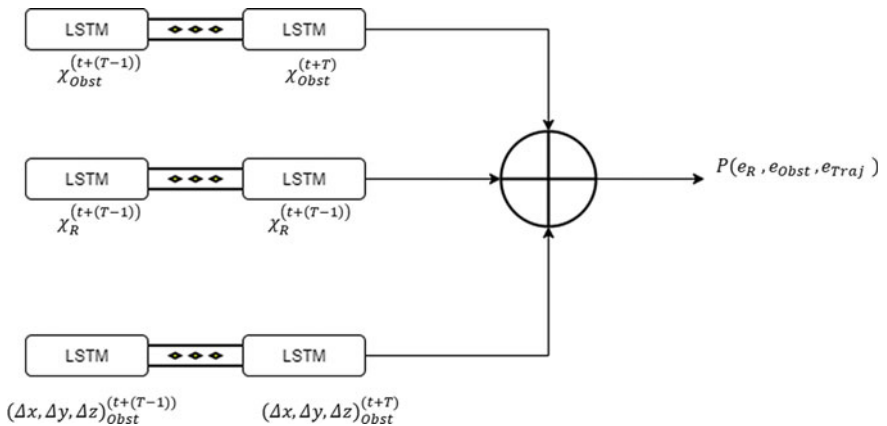
$$t_2 = \{c_1^2, c_2^2, c_3^2, \dots, c_n^2\}$$

$$t_n = \{c_1^n, c_2^n, c_3^n, \dots, c_n^n\}$$

$$\begin{bmatrix} C_1^1 & C_2^1 & \dots & C_n^1 \\ C_1^2 & C_2^2 & \dots & C_n^2 \\ \vdots & \vdots & \dots & \vdots \\ C_1^n & C_2^n & \dots & C_n^n \end{bmatrix}$$

- d. **Obstacle Trajectory Prediction:** Trajectory prediction is a time series forecasting problem over the working space and a real-time environment. Long short-term memory (LSTM) is a network that works effectively for time series forecasting and generates the probabilistic trajectory of moving obstacles in the dynamic environment.

$$P(e_R, e_{Obst}, e_{Traj}) = \text{LSTM}((\Delta x, \Delta y, \Delta z)_{Obst}^{(t+(T-1))}, (\Delta x, \Delta y, \Delta z)_{Obst}^{(t+T)}; \chi_R^{(t+(T-1))}, \chi_R^{(t+(T-1))}; \chi_{Obst}^{(t+(T-1))}, \chi_{Obst}^{(t+T)}) | \Delta T$$



### e. Pre-processing

**Data De-noise.** It helps to process the zero values present in the processing data.

$$f(g)(t) = \int f(x) * g(x - \tau) d\tau$$

**Outlier Detection.** It helps to remove the unwanted data points which are outside the trajectory and are of no use in the prediction model.

### f. Time series location forecasting model

For the time series  $T_1, T_2, T_3, \dots, T_s$ , the predicted result is  $r_1, r_2, r_3, \dots, r_s$  therefore final probability can be calculated as

$$P(R|C, T_d) = \prod_{i=1}^s P(r_i|C, T_d)$$

g. **Collision Prediction Model:** It generates all the possible points in the space with no point of collision using the time series location forecasting model.

$$P(C|R, T_d) = \frac{P(C|T_d)P(R|C, T_d)}{P(R|T_d)}$$

$P(C|T_d)$ —Obstacle collision probability based on Trajectory data.

$P(R|C, T_d)$ —Probability of obstacle collision with robot based on trajectory data.

$$P(R|T_d) = \sum_c P(C|T_d)P(R|C, T_d)$$

## 4 Experimental Results

In this section, we have shown the implementation results obtained based on the proposed various classifiers and the predictors designed to provide the input to the collision prediction model. The binary classifier classifies the environmental obstacles in two classes. i.e., static and dynamic obstacles. Figure 4 shows more than 75% of obstacles considered in the environments are dynamic obstacles. In Fig. 5, we have presented the performance results of the binary classifier. In this, the accuracy and the other evaluation parameters show that the throughput is more than 88%.

In Fig 6, we have presented the graphical error plot on the location forecasting versus actual locations of an obstacle during the real run of the system. The error reported on the forecasting of predicted location and the actual location is also measured and is presented in Fig.7, which shows the error mapping in the location forecasting and actual location observed during the real run.

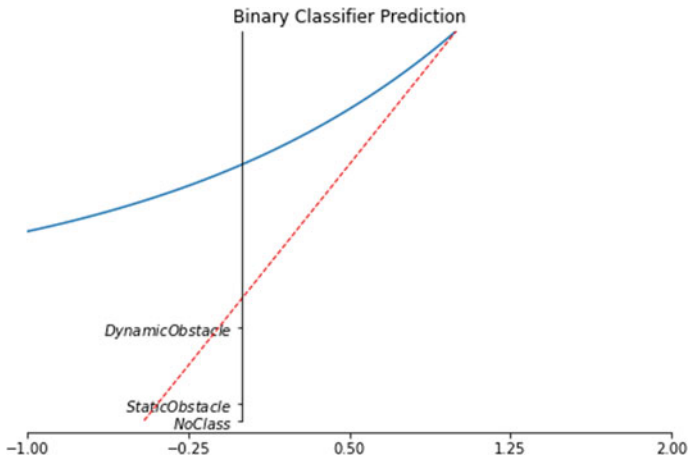


Fig. 4 Binary classifier prediction

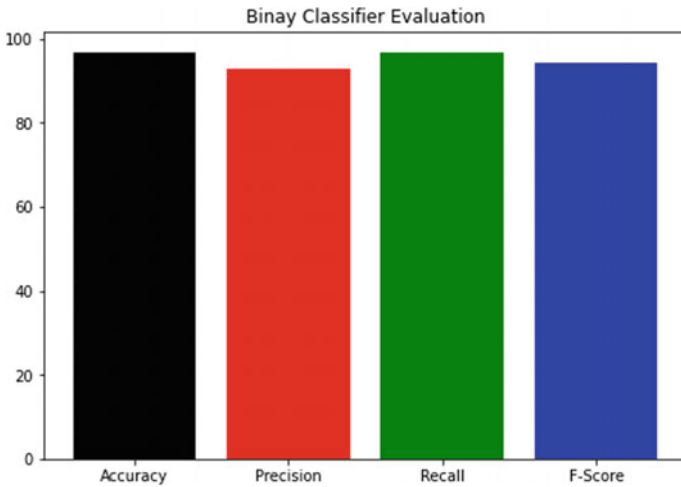


Fig. 5 Binary classifier evaluation

In order to evaluate the trajectory prediction model, the actual trajectory versus predicted trajectory graph has been plotted and based on it the evaluation of the trajectory prediction model has been done (Fig. 8). Figure 9 shows the trajectory prediction model evaluation results which are 80% accurate in the trajectory prediction of obstacles in the dynamic environment.

Finally in Fig. 10, the collision prediction model has been evaluated and the obtained results have been presented which show the accuracy, precision, recall and F-score of the model which is around 80%.



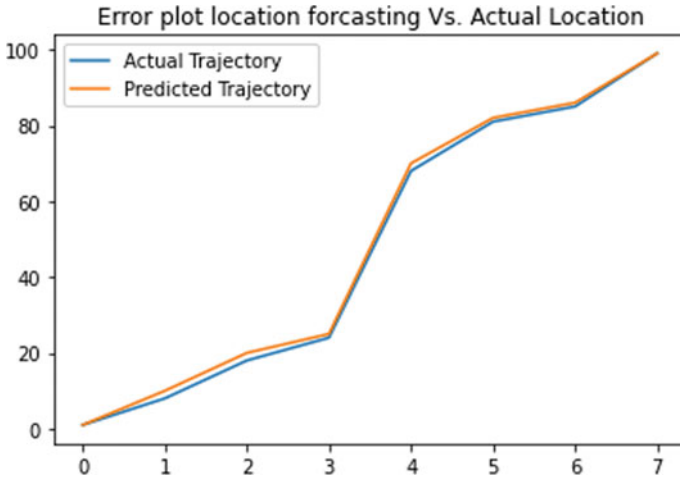


Fig. 6 Error plot in location forecasting versus actual location

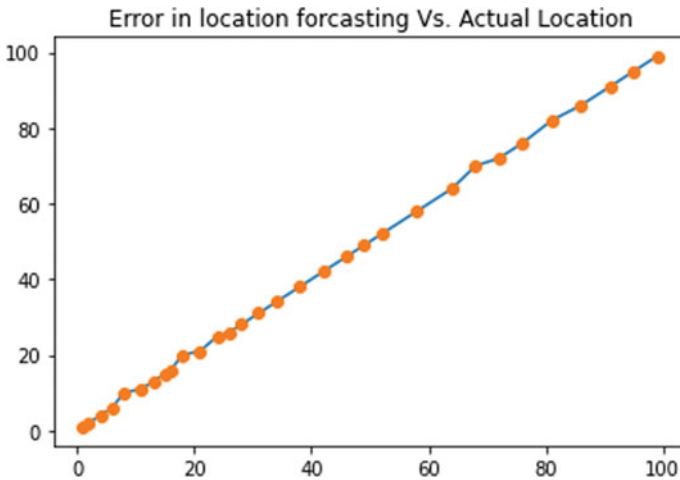


Fig. 7 Error points in location forecasting

## 5 Conclusion

This research work introduces the collision prediction model over the dynamic environment. The collision prediction model is supported by the various classifiers and the prediction models that provide an input to the proposed prediction model. This model shows accuracy of over 85% with good precision and recall outcomes. The experimental result shows that the supported classifiers and the models also perform well and have an accuracy of around 90% with excellent precision and recall. The

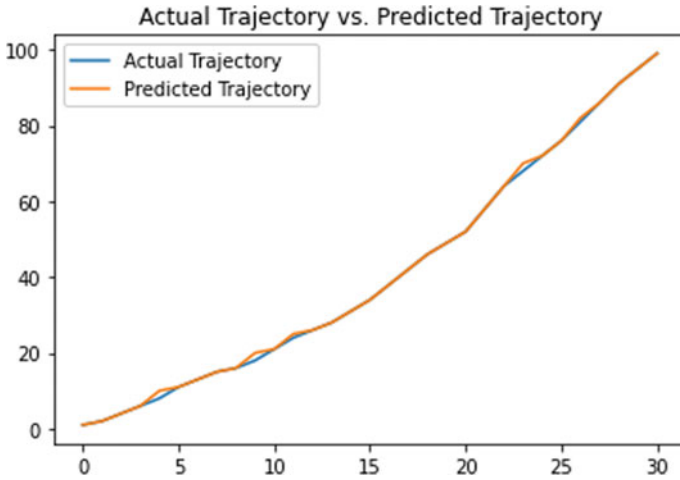


Fig. 8 Actual trajectory versus predicted trajectory

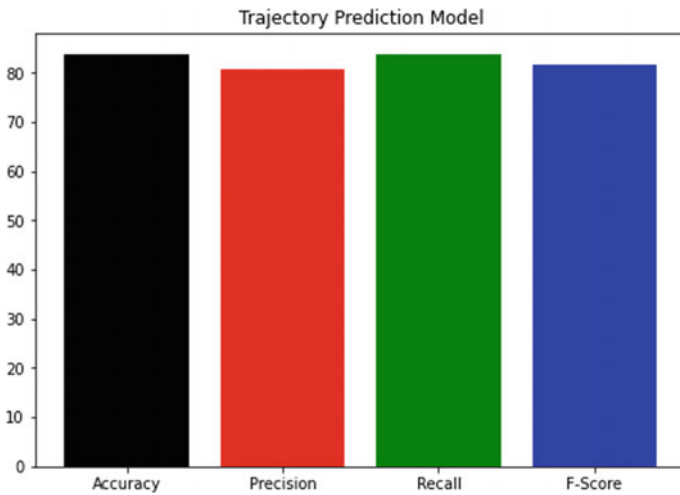
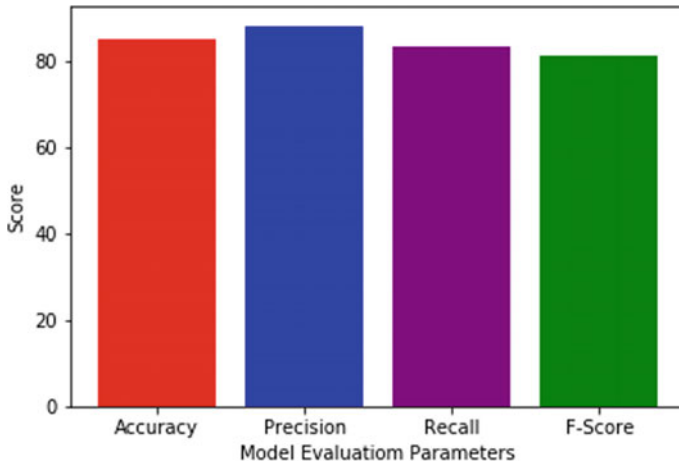


Fig. 9 Trajectory prediction model evaluation parameters

classifiers in predicting the trajectory and the location of obstacles show the better result as there is minimum error in the actual and the predicted locations and the trajectories of an obstacle. The binary classifier is one of the initial level classifiers used to identify the environmental obstacles and predict them in the static obstacle or the dynamic obstacle. The obtained results generated by the binary classifier in this obstacle classification have an accuracy of over 90%. The overall proposed system gives an accuracy of over 85% which is comparatively higher if compared with the already existing benchmark. This proposed system is further being used as a base



**Fig. 10** Collision prediction model evaluation parameters

for the probabilistic path generator to generate a navigation path over the dynamic environment with moving obstacles.

## References

1. Shen Y, Jia Q, Chen G, Wang Y, Sun H (2015) Study of rapid collision detection algorithm for manipulator. In: 2015 IEEE 10th conference on industrial electronics and applications (ICIEA), Auckland, New Zealand, pp 934–938. <https://doi.org/10.1109/ICIEA.2015.7334244>
2. Park J, Kang S, Ahmad N, Kang G (2006) Detecting collisions in an unstructured environment through path anticipation. In: 2006 international conference on hybrid information technology, Cheju, Korea (South), pp 115–119. <https://doi.org/10.1109/ICHIT.2006.253474>
3. Rawat S, Faridi ZA, Kumar P (2016) Analysis and proposal of a novel approach to collision detection and avoidance between moving objects using artificial intelligence. In: 2016 international conference system modeling & advancement in research trends (SMART), Moradabad, pp 135–138. <https://doi.org/10.1109/SYSMART.2016.7894505>
4. Feng S, Zhang J, Zeng B (2010) Collision detection algorithm of successive frames for automated air traffic control. In: 2010 3rd international conference on information management, innovation management and industrial engineering, Kunming, China, pp 265–268. <https://doi.org/10.1109/ICIM.2010.542>
5. Liqun W, Jianrong W, Yulong L (2008) Collision detection on robot applied in steam generator serving. In: 2008 27th Chinese control conference, Kunming, China, 2008, pp 305–309. <https://doi.org/10.1109/CHICC.2008.4605022>
6. Kwak H, Park G (2009) Module-based efficient self-collision detection method for humanoid robots. In: 2009 IEEE student conference on research and development (SCORED), Serdang, Malaysia, pp 483–486. <https://doi.org/10.1109/SCORED.2009.5442960>
7. Huang R, Tang T, Lou Y, Xiao M (2014) A collision detection algorithm of Robot in off-line programming system. In: 2014 4th IEEE international conference on information science and technology, Shenzhen, China, pp 349–353. <https://doi.org/10.1109/ICIST.2014.6920400>

8. Son J, Kwak H, Park G (2011) Back propagation neural network based real-time self-collision detection method for humanoid robot. In: 2011 11th international conference on control, automation and systems, Gyeonggi-do, Korea (South), pp 1505–1508
9. Guo K, Xia J (2010) An improved algorithm of collision detection in 2D grapple games. In: 2010 third international symposium on intelligent information technology and security informatics, Jian, China, pp 328–331. <https://doi.org/10.1109/IITSI.2010.176>
10. Wenzhong X, Xi X, Xinjian J (2017) Sensorless robot collision detection based on optimized velocity deviation. In: 2017 Chinese automation congress (CAC), Jinan, pp 6200–6204. <https://doi.org/10.1109/CAC.2017.8243894>
11. Wei Z, Laotao W (2011) A fast collision detection algorithm suitable for complex virtual environment. In: Proceedings 2011 international conference on transportation, mechanical, and electrical engineering (TMEE), Changchun, China, pp. 502–505. <https://doi.org/10.1109/TMEE.2011.6199251>
12. Jung B, Koo JC, Choi HR, Moon H (2013) Enhanced collision detection method using frequency boundary of dynamic model. In: IEEE ISR 2013, Seoul, Korea (South), pp 1–3. <https://doi.org/10.1109/ISR.2013.6695700>
13. Xu X, Gan Y, Xu C, Dai X (2017) Robot collision detection based on dynamic model. In: 2017 Chinese automation congress (CAC), Jinan, pp. 6578–6582. <https://doi.org/10.1109/CAC.2017.8243962>
14. Wang Y, Wenjuan E, Tian D, Lu G, Yu G, Wang Y (2011) Vehicle collision warning system and collision detection algorithm based on vehicle infrastructure integration. In: 7th advanced forum on transportation of China (AFTC 2011), Beijing, pp 216–220. <https://doi.org/10.1049/cp.2011.1407>

# Human Saliency Based Object Detection from Natural Images



Naveen Chandra and Himadri Vaidya

**Abstract** In natural images, there are a few regions and objects, which catch the attention of human beings and are called “saliency”. The saliency algorithm is used to identify fixation points from the natural images. Saliency detection method has various applications, namely object segmentation and object recognition. In this paper, the task of object detection is performed through saliency detection algorithms. Lastly, the implemented models are tested and compared using iCoSeg data.

**Keywords** Attention · Context · Feature · Saliency · Spectral

## 1 Introduction

Saliency is a primary concept in terms of machine learning/artificial intelligence/deep learning approaches. The key objective of these approaches is to consider a stream of input images/videos to deliver an effective decision for predicting meaningful information from big data. Saliency-based object detection (SBOD) concentrates on identifying and locating the most key parts of the natural image/scene [1]. Recognizing the interesting regions in a regular scene is significant in the human vision framework. In a complex visual scene, items are transformed into a serialized component termed as visual attention. Selecting a center of consideration in preparing stand-out item at a given time is attained by visual saliency and proper choice is made by the human brain. Visual saliency is a consequence of communication between two distinctive stimuli in organic and simulated visual environment. Saliency of an item can be the yield of the cognitive parameter, for example, attention. The

---

N. Chandra (✉) · H. Vaidya  
Wadia Institute of Himalayan Geology, Dehradun, India  
e-mail: [naveenchandra@wihg.res.in](mailto:naveenchandra@wihg.res.in)

H. Vaidya  
e-mail: [himadrivaidya@gehu.ac.in](mailto:himadrivaidya@gehu.ac.in)

H. Vaidya  
Computer Science and Engineering, Graphic Era Hill University, Dehradun, India

studies for transferring attention to perception are being applied in human–computer-interaction have gained significant progress [2]. In recent years, SBOD has attracted many researchers/scientists/academicians due to its wide applications such as image captioning, image classification, object segmentation [3], image understanding [4], and geospatial studies [5, 6]. The performance and capability of SBOD is evaluated using binary ground truth images against the background of the scene [4]. The available data set for SBOD are SED 100 [7], PASCAL-S [8], DUT-OMRON [9], SOC [10], SOD [11], and MSRA-1000 [12]. The dataset contains the input mages along with the corresponding ground truth [4]. The remainder of the paper is as follows: Sect. 2 describes the basics of saliency detection models, Sect. 3 provides the methodology, Sect. 4 presents the experimental results and discussion, and Sect. 5 illustrates the conclusion.

## 2 Background of Saliency Detection Models

Salient objects are always different from the background and the neighboring context. In an image, salient object are mostly placed near the center of an image. Salient object always has a closed and well-defined boundary. Saliency is a concept that is described in feature integration theory and is a bottom-up approach [4]. In this research, four saliency detection models have been implemented for object detection.

### 2.1 Context-Aware Saliency Detection Model

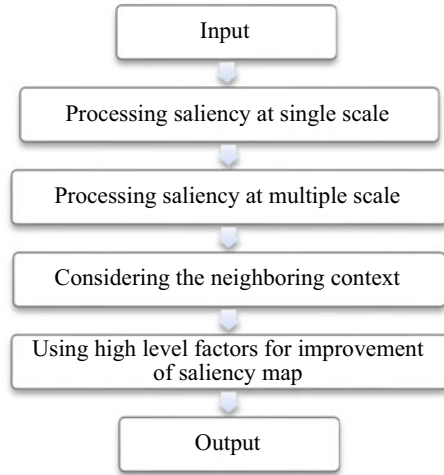
This model uses the hypotheses which have been demonstrated mentally [13–16]. The architecture of the model is shown in Fig. 1.

#### 2.1.1 Processing Saliency at a Single Scale

Some issues experienced while deciding the saliency are defining the uniqueness both locally, globally and combining the location of the salient object. Consider a solitary patch of scale  $W$  at every pixel. A pixel  $\times 1$  will be remarkable if the presence of the patch  $p_{x1}$  focused at pixel  $\times 1$  is diverse as for all other picture patches [17]. Numerically, if  $D_{(p_{x1}, p_{y1})}$  is the Euclidean separation between the patches  $p_{x1}$  and  $p_{y1}$  in CIE  $L^*a^*b$  color space, standardized to the reach  $[0, 1]$  then pixel  $\times 1$  is said to be salient when  $d_{(p_{x1}, p_{y1})}$  is high  $\forall y1$  [17]. Accordingly, patch  $p_{x1}$  is considered to be salient if the patches indistinguishable from it are closer. Assume thatthe Euclidean separation between the location of patches  $p_{x1} \wedge p_{y1}$  then difference measure between the patches is given by (1).

$$D(p_{x1}, p_{y1}) = D_{(p_{x1}, p_{y1})}^{1+c} * D_{position}(p_{x1}, p_{y1}); \quad (1)$$

**Fig. 1** Architecture of context-based saliency detection model [17]



If  $D(p_{x1}, p_{y1})$  is high  $\forall n \in [1, N]$  then pixel  $i$  is said to be salient. The saliency value of pixel  $x$  at scale  $t$  is given by (2).

$$SV_x^t = 1 - \exp \left\{ \frac{-1}{N} \sum_{n=1}^N D(p_x^t, q_k^t) \right\} \quad (2)$$

### 2.1.2 Processing Saliency at Multiple Scales

Each pixel in an image is defined as the mixture of multiple scale image patches [17]. Suppose ‘ $m$ ’ is the scale of patch  $p_x$  and scale of its neighboring patches is  $M_q = (m, \frac{1}{2}m, \frac{1}{4}m)$  then saliency value is calculated as:

$$SV_x^m = 1 - \exp \left\{ -\frac{1}{N} \sum_{n=1}^N D(p_x^m, q_k^{mk}) \right\} \quad (3)$$

### 2.1.3 Considering the Neighboring Context

According to Gestalt hypothesis, areas that are near to the center of consideration are given priority. Now, the areas whose focus of attention is high are extracted from the obtained saliency map [17]. A pixel is said to be salient if it exceeds a given threshold value. It was noticed that homogenous areas display lower saliency while the saliency of interesting regions stays high. This saliency map can be improved further by utilizing high-level factors [17] such as face detection [18].

## 2.2 Graph-Based Visual Saliency

This is a bottom-up saliency model and is built using graphical computation [19]. Saliency is obtained in two steps. Initially, activation maps are computed by using feature vectors, and then they are normalized and combined so that the conspicuity can be highlighted. An activation map  $A1 : [b]^2 \rightarrow R$  is processed from the given feature map  $F1 : [b]^2 \rightarrow R$  in a manner such that the regions  $(i, j) \in [b]^2$ , where  $F1(i, j)$  is distinct from the neighboring regions [19] and it can represent higher values for the obtained activation map. Markovian method defines the dissimilarity of  $R(x, y)$  and  $R(p, q)$  as:

In a directed graph weights assigned to each node is calculated by (4).

$$w_2((x, y)(p, q)) \triangleq (x, y)II(p, q) \cdot F(x - p, y - q) \quad (4)$$

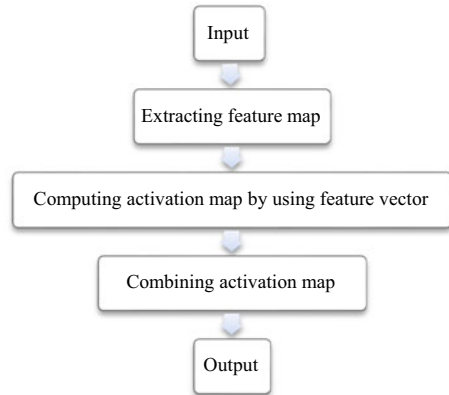
where,  $F(m, n) \triangleq \exp\left(\frac{-m^2+n^2}{2\sigma^2}\right)$ .

A graph  $G_z$  is constructed for the normalization of an activation map and weights between nodes  $(x, y)$ ,  $(p, q)$  is given by (5).

$$w_2((x, y)(p, q)) \triangleq A1(p, q) \times F(x - p, y - q) \quad (5)$$

Equilibrium distribution can be estimated by normalizing the weights of the outbound edges [19]. Mass moves through the nodes having higher activation therefore it is called as mass concentration algorithm. The overall architecture of the model is shown in Fig. 2.

**Fig. 2** Architecture graph-based visual saliency model [19]





### 2.3 Spectral Residual Approach

It is a fast and reliable saliency detection model, which does not depend on prior knowledge about the objects [20]. Figure 3 illustrates the flow process of the model. It consists of the following steps:

1. An analysis of the logarithmic spectrum for the given input image.
2. Extraction of spectral residual for the input image in the spectral domain.

Barlow Efficient coding [21] is used for removing the redundancies from the input images. It divides the image into two parts Innovation and prior knowledge (Eq. 6).

$$I(\text{Image}) = I(\text{Innovation}) + I(\text{priorknowledge}) \quad (6)$$

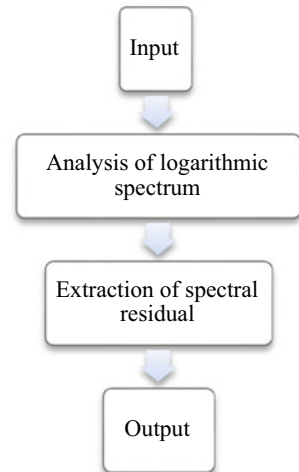
#### 2.3.1 Representation of Logarithmic Spectrum

Scale invariance is an important parameter of natural image statistics. It is also called as  $1/f$  law (Eq. 7). [21].

$$E\{A(f)\} \propto \frac{1}{f} \quad (7)$$

where  $A(f)$  is the amplitude. For an input image, log spectrum  $L(f)$  is determined using equation  $L(f) = \log(A(f))$  [21]. Log spectrum representation is widely used in many of the statistical scene analysis literature [16–19].

**Fig. 3** Architecture of spectral residual based model [21]



### 2.3.2 Extracting Spectral Residual

If a model is focusing on minimizing the redundant information then it is necessary to identify the statistical similarities of the input stimuli [3]. Computing a log spectrum  $L(f)$  for an input down-sampled image with a width of 64 pixels. Now, the spectral residual for the given image is calculated using (8).

$$I(SR(f)) = I(L(f)|A(f)) \quad (8)$$

where  $SR(f)$  is the spectral residual of an image  $A(f)$  is the shape of log spectra. To approximate the shape of log spectra an average filter  $h_n(f)$  is used for the experiment. Here,  $n = 3$ . The averaged spectrum  $A(f)$  of the input image is determined by (9).

$$A(f) = h_n(f) * L(f) \quad (9)$$

where  $h_n(f)$  is a matrix of size  $n \times n$ .

$$h_n(f) = \frac{1}{n^2} \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

Hence Spectral residual is calculated using (10).

$$SR(f) = L(f) - A(f) \quad (10)$$

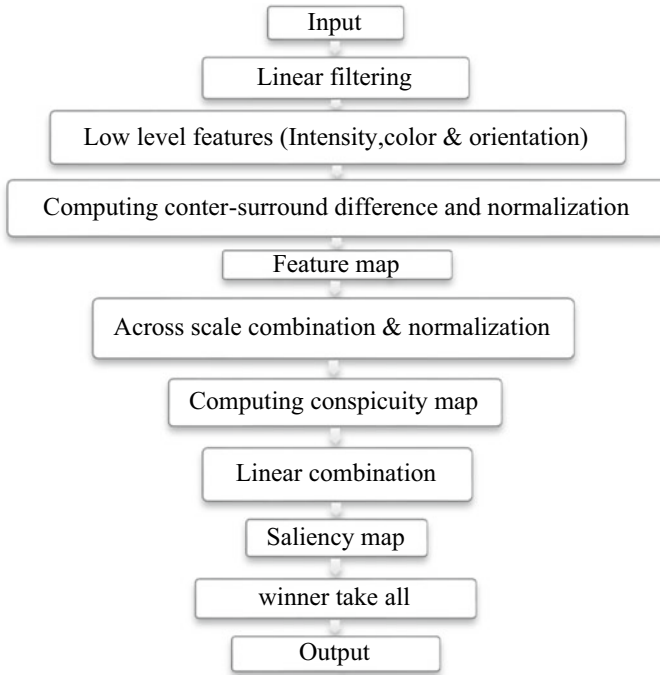
Finally, saliency map is computed by using the inverse Fourier transform in the spatial domain [21]. The obtained saliency map is smoothed using the Gaussian filter for better visual effects.

## 2.4 Itti and Koch Model

This model is based on the visual search method, which is described in feature integration theory [13, 15]. Figure 4 illustrates the working of the model. The model consists of the following steps [22]:

1. Dividing the input image into topographic feature maps.
2. Spatial location then competes for saliency within each map.
3. Applying bottom-up method for forming saliency map from master maps.

Color image I is used as an input which is digitized at  $640 * 480$  resolutions. Nine spatial scales were constructed operating dyadic Gaussian pyramids [23]. Different



**Fig. 4** Architecture of Itti and Koch model [22]

features in the images are estimated using center-surround method, which is also utilized in distinguishing between the fine and coarse scales [24].

#### 2.4.1 Extracting Early Visual Features from Input Image:

If  $r$ ,  $g$ , and  $b$  are the red green, and blue values of the color image then the intensity image is determined as (Eq. 11)

$$\mathcal{J} = \frac{(r + g + b)}{3} \quad (11)$$

This process is repeated to compute the intensity pyramid with levels  $\mathcal{J}(\sigma)$ , where  $(\sigma = 0 \dots 8)$ . To avoid the fluctuation of color opponency values at low luminance four color-tuned channels are created  $R' = r - \frac{(g+b)}{2}$  for red,  $G' = g - \frac{(r+b)}{2}$  for green and  $B' = b - \frac{(r+g)}{2}$  for blue and  $Y' = \frac{r+g}{2} - \frac{|r-g|}{2} - b$  for yellow which results in four Gaussian pyramid  $R'(\sigma)$ ,  $G'(\sigma)$ ,  $B'(\sigma)$ ,  $Y'(\sigma)$ . Center-surround differences between a fine scale ( $c'$ ) and coarser scale ( $s'$ ) are computed using across-scale subtraction between two maps which results in feature maps (given in 12) [24].

$$FM(c', s') = |\mathcal{J}(c') \ominus \mathcal{J}(s')| \quad (12)$$

Similarly, feature maps for the color channel are also constructed [24]. Therefore, red-green and blue-yellow opponencies are computed using (13) and (14).

$$R'G'(c', s') = |(R'(c')) - G'(c') \ominus (G'(s') - R'(s'))| \quad (13)$$

$$B'Y'(c', s') = |(B'(c')) - Y'(c') \ominus (Y'(s') - B'(s'))| \quad (14)$$

Information regarding the orientation is obtained from  $\mathcal{J}$  by introducing Gabor filters  $O(\sigma, \theta)$ , where  $(\sigma = 0..8)$  and  $\theta = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ . Mathematically (Eq. 15), Gabor filter is defined as

$$G_\psi(m, n, \theta) = \exp\left(\frac{-m'^2 + \alpha^2 n'^2}{2\delta^2}\right) \cos\left(2\pi \frac{m'}{\lambda} + \psi\right) \quad (15)$$

where  $\psi$  is the phase,  $\alpha$  is the aspect ratio,  $\delta$  is the standard deviation,  $\lambda$  is the wavelength and the coordinates  $(m', n')$  which are transformed with respect to orientation  $\theta$ . Center-surround scale is again used for computing the oriented feature map as given in (16).

$$\mathcal{J}(c', s') = |\mathcal{J}(c') \ominus \mathcal{J}(s')| \quad (16)$$

## 2.4.2 Conspicuity Map

Feature maps of intensity, color, and orientation (Eqs. 17–19) are summed and normalized again which results in a conspicuity map [24].

$$CM_I = \oplus_{c'=2}^4 \oplus_{s'=c+3}^{c'+4} N \quad (17)$$

$$CM_C = \oplus_{c'=2}^4 \oplus_{s'=c+3}^{c'+4} [N(R'G'(c', s')) + N(B'Y'(c', s'))] \quad (18)$$

$$CM_O = \sum_{\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}} N\left(\oplus_{c'=2}^4 \oplus_{s'=c+3}^{c'+4} N(O(c', s', \theta))\right) \quad (19)$$

where  $N$  is a normalizing operator.

Lastly, three different conspicuity maps are summed and normalized to obtain a saliency map (Eq. 20).

$$SM = \frac{1}{3}(N(CM_I) + N(CM_C) + N(CM_O)) \quad (20)$$

The location in the saliency map with the maximum saliency value is determined by the winner-take-all method [24, 25].

### 3 Methodology

In this research, the process of object detection is performed using the four saliency-based models namely context-aware, graph-based, spectral residual based, and Itti's and Koch models. The method for object detection of each model has been described in Sect. 2; however, the overall framework of the methodology is given in Fig. 5.

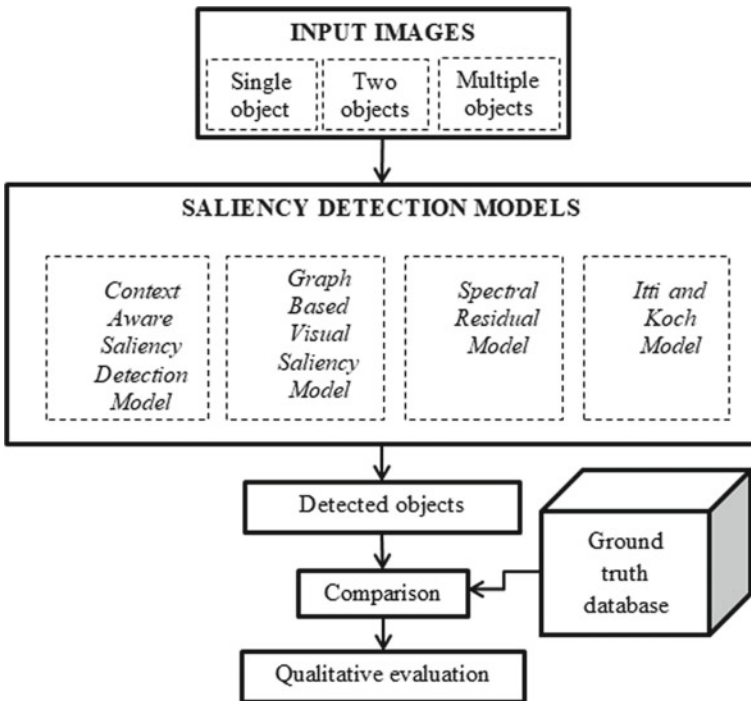


Fig. 5 Overall framework of the adopted methodology

## 4 Experimental Results and Discussion

### 4.1 Dataset

To evaluate the performance of the above-implemented models, the “iCoSeg” [26, 27] freely available dataset is used. It consists of 643 images along with pixel-wise ground truth. These images are divided into 38 groups. Here, the images used for the experimentation are divided into three categories. Type 1 consists of images which have a single salient object, type 2 consists of images with two salient objects, and type 3 consists of images with multiple salient objects. The “iCoSeg” has been used by other researchers [28–32] for evaluating their proposed models.

### 4.2 System Specification

The experiment is performed in MATLAB R2015a with the following system specifications: processor-Intel-(R)-Core-(TM)-i7, 2.80 GHz, RAM-4 GB, and 64-bit-operating system.

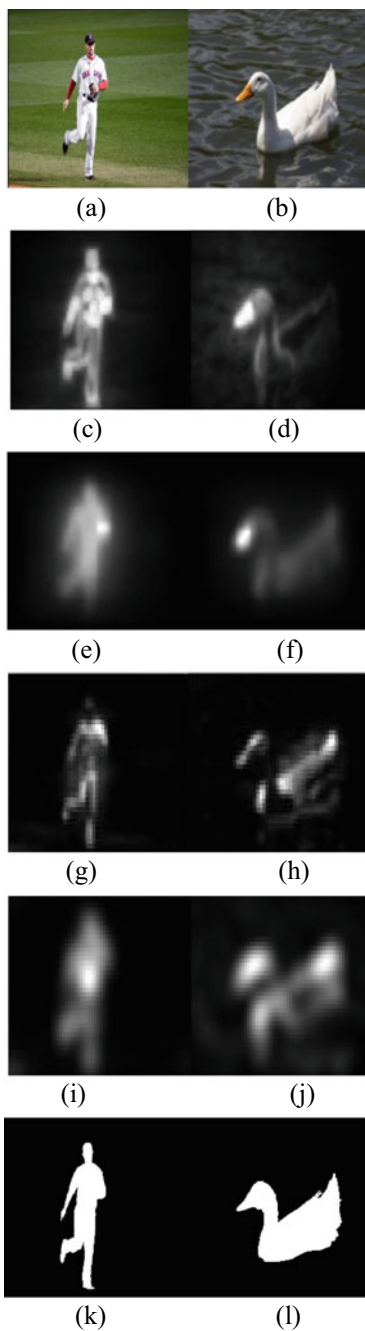
### 4.3 Discussion

In Fig. 6, context-based model detects the player and the goose clearly with the well-defined boundaries. GBVS and Itti’s approach detects the salient object but the boundaries are not detected accurately because these methods do not consider global saliency and the spectral residual approach is unable to detect objects accurately because it does not consider local saliency. In Fig. 7, context-based model detects the two salient objects along with the background whereas the background is not detected in, GBVS Itti’s and spectral residual model. Figure 8 consists of multiple salient objects and the context-based model detects the different objects clearly, GBVS, Itti’s and spectral residual model fails in distinguishing between multiple objects. Table 1 illustrates the comparative evaluation of the four saliency models.

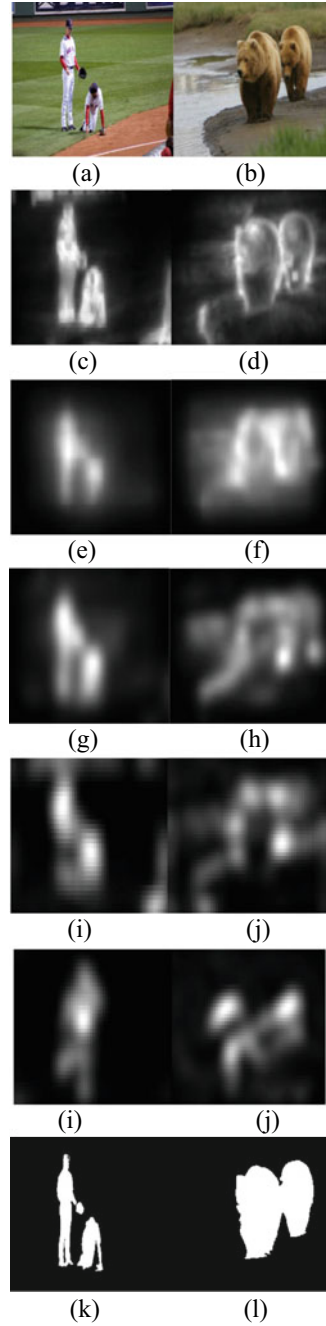
Context-aware model determines the salient object and their surroundings accurately. Saliency maps are necessary input in many image processing techniques. Two important applications of context-aware saliency detection model are image retargeting and collage summarization. Image retargeting is a process in which an image is resized by shrinking and expanding the regions in an image which does not contain any relevant information. Collage creation is a time-consuming work; therefore, many software tools are available for creating a collage in which only the salient objects and surroundings which have relevant information are included.

GBVS is a simple method that exhibits maximum saliency values in the center of the image plane and predicts human fixations points efficiently. GBVS uses a

**Fig. 6** Images with single salient object and their saliency. **a,b** input, **c,d** context based, **e,f** GBVS, **g,h** Itti, **i,j** SR, **k,l** Ground truth

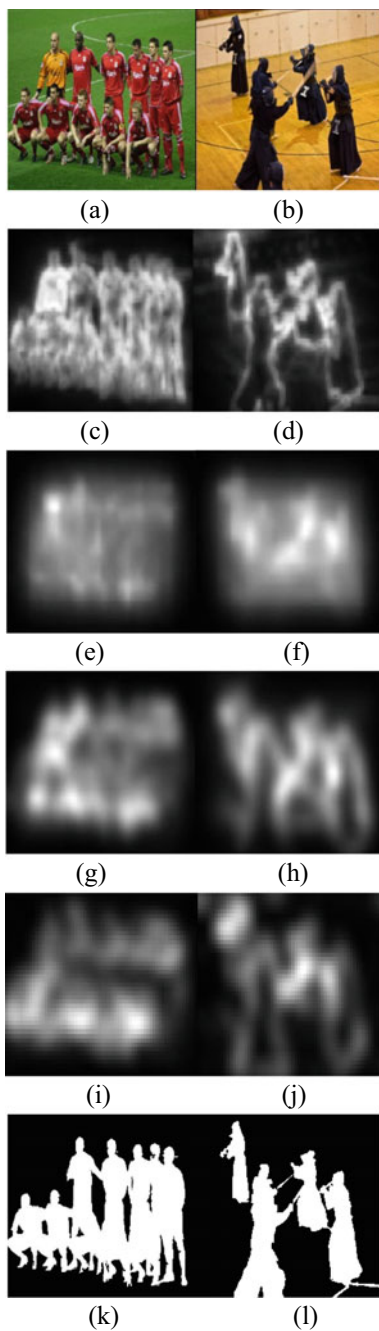


**Fig. 7** Images with two salient objects and their saliency. **a,b** input, **c,d** context based, **e,f** GBVS, **g,h** Itti, **i,j** SR, **k,l** Ground truth





**Fig. 8** Images with multiple salient objects and their saliency. **a,b** input, **c,d** context based, **e,f** GBVS, **g,h** Itti, **i,j** SR, **k,l** Ground truth



**Table 1** Comparative evaluation of saliency detection models

Model	Capabilities	Limitations	Application
CA	Determines saliency locally and globally High accuracy Improved contrast between salient and non-salient regions	High time complexity	Image retargeting, image summarization
GBVS	Bottom-up approach Biologically acceptable Salient locations away from object borders are also spotted	Goal of normalization process is not clear Needs to be extended for multi-resolution representation of map	Predicting human eye movements
SR	Global approach No prior knowledge required Problem of weighting features from various channels is resolved Delivers effective solution to real time system	Object boundaries are not detected accurately Limited to static images	Segmentation and object detection
ITTI	Bottom up approach Biological inspired Local contrast approach	Detects non interesting background	Computer vision and object recognition

center bias approach by activating and normalizing the image. It is noticed that while using the center-surround algorithm it is difficult to activate the saliency of the locations away from the object boundaries. GBVS can be extended to the multiresolution version by creating edges between nodes among the weights among the edges are also computed. GBVS is a biologically acceptable model which uses bottom-up approach for computing saliency maps and motivated by graph theory. Spectral residual is a model which is determined from the logarithmic representation of image data. It identifies the spectral residual of the image which is used in detecting objects. A major advantage of this model is that it does not require any prior knowledge for computing the saliency of a visual scene. The spectral residual method delivers the efficient output to the psychological patterns and solves the problem of weighting low-level features in an image. This model is tested using only static images; therefore, it needs to be validated for a sequence of images and motion. This model has its application in segmentation algorithms. Itti and Koch model uses feed-forward feature extraction stages. Computationally, the major capability of this model is in parallel implementation, i.e., feature extraction and attention mechanism. The saliency detection depends on a factor that an object should have appeared in at least of the computed feature maps. The model is not capable of reproducing the process such as closure and contour completion because of the lack of repetitive processes inside the feature maps. This model does not include magnocellular motion which is

required for computing saliency. The model uses the normalization operator which easily determines the saliency of a visual scene in any circumstances.

## 5 Conclusion

This paper provides the implementation of four saliency detection models for object detection from natural images. These models concentrate on the feature maps for detecting the salient objects from the image. Lastly, a comparative evaluation of the models focusing on the advantages and disadvantages has been provided. These models have various applications such as image retargeting, image segmentation, and image classification. The key contributions of the paper are summarized below.

- (a) Detecting objects from natural images exploring four saliency detection models.
- (b) Testing the four models with the different number of objects in the images.
- (c) Identifying the capabilities and limitations of the saliency detection models.

In the future, these models are needed to be tested for the image sequence, and some more applications of the saliency map are to be discovered through deep saliency-based approaches.

## References

1. Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3085–3094
2. Zhang Q (2018) A survey on approaches for saliency detection with visual attention. In: MATEC web of conferences, vol 232, p 02007. EDP Sciences
3. Zeng Y, Zhang P, Zhang J, Lin Z, Lu H (2019) Towards high-resolution salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 7234–7243
4. Zhu, J-Y, Wu J, Xu Y, Chang E, Tu Z (2014) Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Trans Pattern Anal Mach Intell* 37(4):862–875
5. Sharma A, Ghosh JK (2018) A bottom-up saliency-based segmentation for high-resolution satellite images. In: Proceedings of 2nd international conference on computer vision & image processing, pp 169–180. Springer, Singapore
6. Sharma A, Ghosh JK (2015) Saliency based segmentation of satellite images. In: *ISPRS Annals of photogrammetry, remote sensing & spatial information sciences* 2
7. Alpert S, Galun M, Brandt A, Basri R (2011) Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Trans Pattern Anal Mach Intell* 34(2):315–327
8. Li Y, Hou X, Koch C, Rehg JM, Yuille AL (2014) The secrets of salient object segmentation. In: *IEEE conference on computer vision and pattern recognition*, pp 280–287
9. Yang C, Zhang L, Lu H, Ruan X, Yang M (2013) Saliency detection via graph based manifold ranking. In: *Proceedings of IEEE CVPR*, pp 3166–3173
10. Fan DP, Cheng MM, Liu JJ, Gao SH, Hou Q, Borji A (2018) Salient objects in clutter: bringing salient object detection to the foreground. In: *European conference on computer vision (ECCV)*. Springer
11. Movahedi V, Elder JH (2010) Design and perceptual validation of performance measures for salient object segmentation. In: *Proceedings of IEEE CVPR workshops*, pp 49–56

12. Achanta R, Hemami S, Estrada F, Susstrunk S (2004) Frequency-tuned salient region detection. In: Proceedings of IEEE CVPR, pp 1597–1604
13. Wolfe JM (1994) Guided search 2.0 a revised model of visual search. *Psychon Bull Rev* 1(2):202–238
14. Koffka K (1935) Principles of gestalt psychology
15. Koch C, Poggio T (1999) Predicting the visual world: silence is golden, *nature neuroscience* 2
16. Treisman A, Gelade G (1980) A feature-integration theory of attention. *Cogn Psychol* 12(1):7–136
17. Goferman S, Manor LZ, Tal A (2012) Context-aware saliency detection. *Pattern Anal Mach Intell, IEEE Trans* 34(10). In: Breckling J (Ed) The analysis of directional time series: applications to wind speed and direction, ser. Lecture Notes in Statistics. Berlin, Germany: Springer, vol 61, 1989
18. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: Computer vision and pattern recognition, 2001. CVPR 2001. In: Proceedings of the 2001 IEEE computer society conference on, vol 1, pp 1–511. IEEE
19. Harel J, Koch C, Perona P (2006) Graph-based visual saliency. In: Advances in neural information processing systems, pp 545–552
20. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: Computer vision and pattern recognition. CVPR'07. IEEE conference on, pp 1–8. IEEE
21. Barlow HB (1961) Possible principles underlying the transformation of sensory messages. *Sensory Commun*
22. Walther D, Koch C (2006) Modeling attention to salient proto-objects. *Neural Netw* 19(9):1395–1407
23. Torralba A, Oliva A (2003) Statistics of natural image categories. *Netw Comput Neural Syst* 14(3):391–412
24. Itti L, Koch C (2000) A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res* 40(10–12):1489–1506
25. Wolfe JM, Cave KR, Franzel SL (1989) Guided search: an alternative to the feature integration model for visual search. *J Exp Psychol: Human Percept Perform* 15(3)
26. Batra D, Kowdle A, Parikh D, Luo J, Chen T (2011) Interactively co-segmenting topically related images with intelligent scribble guidance. *Int J Comput Vision* 93(3):273–292
27. Batra D, Kowdle A, Parikh D, Luo J, Chen T (2010) icoseg: Interactive co-segmentation with intelligent scribble guidance. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 3169–3176. IEEE
28. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: Computer vision and pattern recognition (CVPR), IEEE conference on, pp 2083–2090. IEEE
29. Kompella A, Kulkarni RV (2019) Weakly supervised multi-scale recurrent convolutional neural network for co-saliency detection and co-segmentation. *Neural Comput Appl* 1–18
30. Dong X, Shen J, Shao L, Yang MH (2015) Interactive cosegmentation using global and local energy optimization. *IEEE Trans Image Process* 24(11):3966–3977
31. Wang F, Huang Q, Guibas LJ (2013) Image co-segmentation via consistent functional maps. In: Proceedings of the IEEE international conference on computer vision, pp 849–856
32. Fan DP, Lin Z, Ji GP, Zhang D, Fu H, Cheng MM (2020) Taking a deeper look at co-salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2919–2929

# Dynamic Education Background: Procure the Maximum Initiation from PBL for Education Naïve Bayes Algorithm for Machine Learning



Vishnu Kumar Mishra, Megha Mishra, Jitendra Sheetlani, Rahul Deo Sah,  
Anup Mishra, and Satyendra Kurariya

**Abstract** The main advancement in the field of engineering education is learning somewhat which is based on the problem. This term surely applied to any learning environment in which students drive learning. It is presented in such a way that students understand the problem before moving toward its solution and accordingly they need to gain new information. For the fourth year of the Computer Science and Engineering degree curriculum, this research introduces the knowledge of a formal problem-based learning procedure for educating a preliminary research component in the naive Bayesian method of machine learning. At the beginning of the module, the Naïve Bayes Algorithm design problem was introduced to students. For seven weeks, a small crowd of undergraduates were operated for this assignment, at the same time the instructor served as the information acquisition facilitator. Every week, brief, written information was composed as the learner evaluation, so that the learning environments were ensured. Due to COVID-19, most of the offline classes were suspended so this PBL experiment and the written reports are conducted /collected through the online mode. A list of guidelines to assist academic interest in pursuing PBL with a similar strategy is outlined in the report.

---

V. K. Mishra (✉) · M. Mishra  
SSGI-FET, Bhilai, Chhattisgarh, India  
e-mail: [vshn123msmr@gmail.com](mailto:vshn123msmr@gmail.com)

M. Mishra  
e-mail: [megha16shukla@gmail.com](mailto:megha16shukla@gmail.com)

M. Mishra  
Mata Gujri Womens (Autonomus) College, Jabalpur, India

J. Sheetlani  
shri Shankarcharya Group of Institution, Bhilai, Chhattisgarh, India  
e-mail: [jsheetlani@gmail.com](mailto:jsheetlani@gmail.com)

R. D. Sah  
Sri Satya Sai University of Technology and Medical Science, Sheore, India

A. Mishra  
Dr. Shyama Prasad Mukherjee Univeristy, Rachi, India

S. Kurariya  
Bhilai Institute of Technology Durg, Durg, India

**Keywords** Self-learning · Naïve Bayes · Problem-based learning (PBL) · Dynamic education

## 1 Introduction

Promptly evolving expertise implies that technical program students are required to develop ongoing knowledge and continue-learning ability, which are essential components of a constant capacity to adapt the technical progress. One of the prime qualities that promote ongoing learning apart from teamwork, innovative opinion, transmission of knowledge, or significant self-consciousness, is the dedication of the student to self-assertive education. Furthermore, the peak level of technological aptitude, the business [1] requires and anticipates a broad variety of standard ability from engineers.

Also, certain skills are evaluated in reports by several engineering organizations and corporations [2–6]. This article presents the interactions, student accomplishments, and assessments of teachers for a research module entitled “Naïve Bayes Algorithm”. It also addresses the basics of the learning approach, implemented in order to improve students’ lifelong learning skills. Subsequently, the background and confront appearance from the records, the fundamentals of the anticipated education tricks, and the accomplishments with assessments of the students are identified. Finally, a list of suggestions that is the proof of interest of researchers for using the PBL strategy is provided from the gained experience.

## 2 Knowledge Policy in Engineering Teaching: The Problem-Based Learning Approach (PBL)

The study has exposed that student’s move toward learning any specific task is a significant factor in deciding education results [7]. A significant point to remember here is that the advances in knowledge are not an individual’s permanent quality [8], so speaking about deep and surface learners is incorrect. General exercise of official talk does not build this feasible, however little assembly effort can facilitate student autonomy if designed to allow cooperative learning [9–11]. The exercise of a deep approach and expert knowledge result is correlated with such work.

PBL is a solitary case of this kind of effort, where the assembly of learners put effort together to resolve a specific obstacle [12, 13]. Here, we are having sufficient confirmation of the research to sustain the belief that problem oriented approach, toward engineering education is next to the center of major improvement and also a trend in current days. It helps to improve analytics and ongoing training skills. In addition, the comparison among problem-oriented approaches and industry engineering administration has been recognized [14]. This approach is significant and provided an improvement in teaching practices, changes the traditional way of teaching, and

also transforms the teacher as a core of awareness to the foundation of the entire information, to be the instructor and organizer of the process of gaining the knowledge. Instead of being teacher-centered, learning becomes student-centered [15].

Subjected to an undergraduate line, due to the diverseness and variety of the logical and methodological awareness of every scholar's context, an extra organized approach is desirable [16]. Some researchers have given their stress to the PBL approach that it should include broad, unrestricted problems right from the beginning, and this is the standard for any degree holder [17]. There is evidence that for junior-level courses, a problem-oriented methodology that originally has an ordered structure, through obviously identified difficulty, and then progressively moving toward a broad problem is a constructive method [18]. Recently experiences have been documented in different areas, for example, electrical network theory, computer network premises, or mechanical controls [19–21]. There is also a progressive development of Problem-Based Learning advances in the general training carried out by the Department of Computer Science & Engineering.

### 3 Case Study: Issues and Context PBL

During the 2018/2019 and 2019/2020 academic years, a formal PBL approach was applied in the research unit named "Naïve Bayes Algorithm". The topic is a part of the fourth year of Artificial Intelligence and Machine Learning subject, a four-year graduate course directing to an undergraduate degree in Computer Science & Engineering at Shri Shankaracharya Group of Institutions of Chhatisgarh Swami Vivekanand Technical University (India). Within the Artificial Intelligence & Machine Learning subject, the "Naïve Bayes Algorithm" module was instructed for a span of six weeks and included four hours of timetable for each week (two lecture room/hypothesis hours, two tutorial/workshop hours). This mandatory unit aspires to communicate a basic understanding of the Naïve Bayes Algorithm, with a special emphasis on the world of machine learning. From the perspective of learner series, the component is essentially associated with: (i) artificial intelligence methodology and (ii) machine learning concerns.

The learning component was instructed with a formal PBL method during the 2018/2019 and 2019/2020 academic years. Overall 50 undergraduates joined in 2019/2020, whereas 42 did so in 2018/2019. Scattered groups of four to five individuals, the scholars began functioning through extremely basic, obviously distinct problems and were eventually motivated toward the extra complicated ones. There was no chance to keep away from the conventional daily schedule due to academic structures and directives. Therefore, the students were involved as a single group for the formal lectures and participated in the group throughout the tutorial hours, spitted into two assemblies of 4 to 5 teams. The PBL methodology was introduced with the contract of instructors for the undergraduate curriculum in order to inspire the scholars' awareness of computer science and engineering and machine learning

associated theory. The overall PBL approach is divided into four different tasks, which are explained below.

Task 1, which was to set the curriculum for the technical subjects of such lecture, was set for the scholars familiar with official lectures in the study room. The only guideline was that the classes should be constructive for grasping the theoretical knowledge of the problem of design in the subject, and the instructor should plan and teach them. Every team through the aid of the handbooks and course books projected their individual plan of content for every theoretical lecture. After review and consensus about the plan, the final curriculum was agreed and contains lectures on:

- (1) Introduction of Naïve Bayes;
- (2) Provisional Probability;
- (3) The Rules applied to Bayes Theorem;
- (4) The Naive Bayes;
- (5) Naive Bayes Manual Examples;
- (6) Laplace Rectification;
- (7) Gaussian Naive Bayes;
- (8) Structuring a Naive Bayes Classifier;
- (9) Implementation Work-out: Individual Movement Identification forecasting;
- (10) Guidelines to enhance the model. Therefore, scholars get authority for the learning of their individuals and for the assembly also. Following is the summary of the given assignments:

- Task2: The scholars were assigned to write a short manuscript on the working of the Naïve Bayes algorithm, classification strategy, variations of the Naïve Bayes Algorithm, and derivation of the Bayes theorem. Diagrams of Naïve Bayes working are represented in Fig. 1.
- Task3: Commencing from the common problem of Naïve Bayes and machine learning the scholars were assigned to reach the process to generate a classifier using any machine learning platform: (i) Create a text classifier. (ii) Select “Topic Classification”. (iii) Upload your training data. (iv) Create your tags. (v) Train your classifier. (vi) Change to Naive Bayes. (vii) Test your Naive Bayes classifier. (viii) Start working with your model, steps are shown in Fig. 2.
- Task4: By using some quick and simple coding, students were asked to integrate a machine learning platform into the code-base and to run their classification model automatically, snap of coding is shown in Fig. 3.

Tasks 2 to 4 were solved by each team; Task 3 was extra geared with respect to individual work. As the students can select any method for classification, the methods used in Task3 were written in a diverse color (for example, red, yellow, blue, and black). Every associate of the team has to pick single color along with the other teams, particular students who had selected a similar color (i.e., a similar problem). They met for 10 min in this ‘new team’ to explore the problem’s solution, and then returned to the independent panel and resolve the problem alone. The aim of this



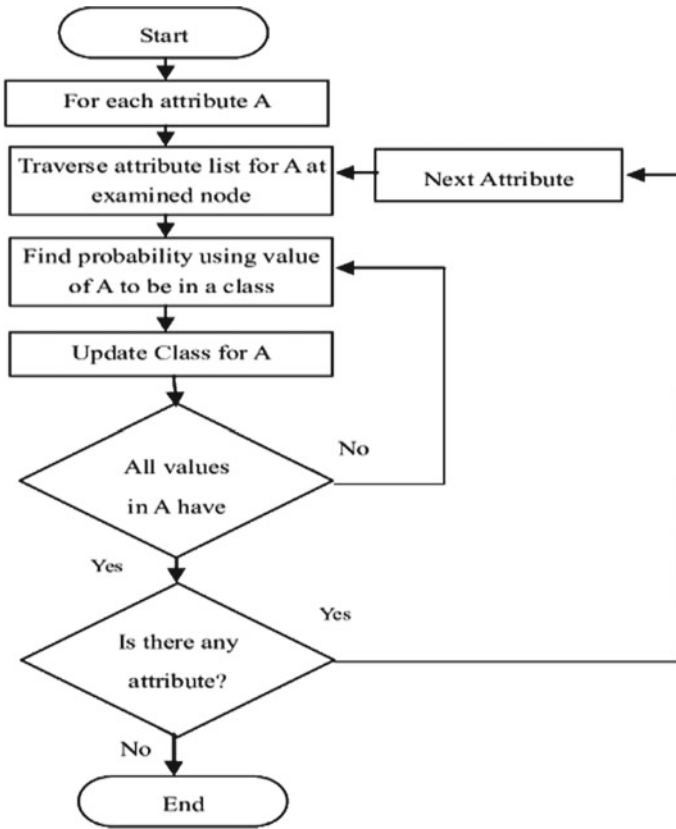


Fig. 1 Flowchart for Naïve Bayes algorithm

approach is to encourage the versatile ability for teamwork and to distinguish the individuals inside the panel.

The final design issue was dedicated to the last two weeks. The continuous and progressive appraisal was introduced into the evaluation system. Every panel of scholars has to mention their part for the subsequent challenge every six days, and the concluding solution might be offered in six days subsequent to the closing stage of the topic. As mentioned above, in weeks 5–6, a small solitary test was conducted to differentiate among individuals within the panel. A description of the organizational scheme is shown in Table 1. The progressive assessment means that the afterward assignment had more influence than the previous ones, with 50% of the overall mark going to the final issue. The hypothesis was that knowledge acquisition and general skill growth are cumulative such that the appraisal requirement increases through the scope and importance of the issues.

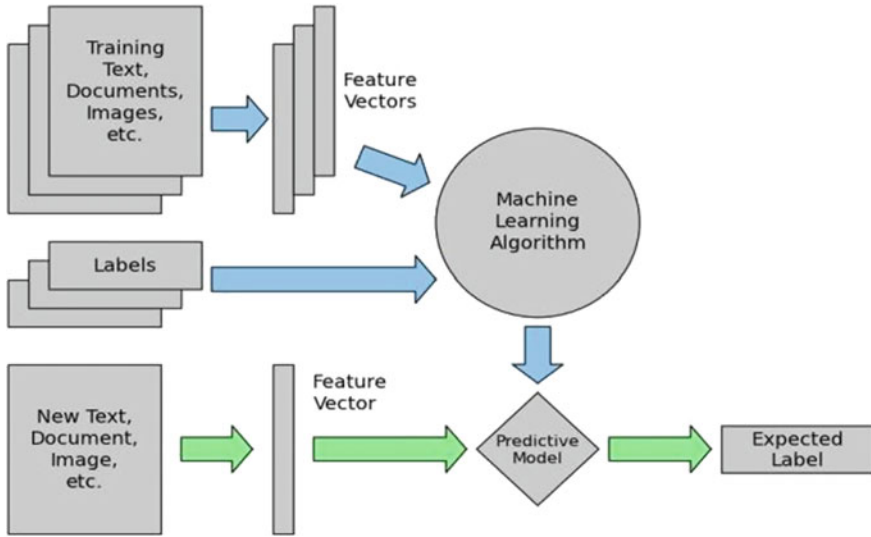


Fig. 2 Basics steps of text classification using Naïve Bayes

```

Request Example
Curl Python Ruby PHP Node.js Java
1 from monkeylearn import MonkeyLearn
2
3 ml = MonkeyLearn('<<insert your token here>>')
4 data = ["This is a great tool!"]
5 model_id = 'cl_p13C7Jil'
6 result = ml.classifiers.classify(model_id, data)
7 print(result.body)
  
```

Fig. 3 Coding snap for integrating the machine learning platform

Table 1 Naïve Bayes algorithm design schemes and assessment for the problem base learning technique

Week No	Lecture	Assignment in groups	Assessment
First	Overview	Task 1:Agenda regarding contents	Group review 0%
Second	Presentation1	Task 2	Group review 7%
Third	Presentation 2	Task 3	Team report 25%
Fourth	Presentation 3	Task 4	Individual exam 35%
Fifth and Sixth	Lecture 4,5	Last assignment	Group review and verbal demonstration 50%

## 4 Naïve Bayes Machine Learning Platform Analysis and Discussion (PBL)

In the 2018/2019 and 2019/2020 academic years, the students participating in this PBL method were surveyed. To test the influence of PBL on student learning, several methods were used.

First, an appraisal was carried out for the technical blueprint and findings provided by the scholars. During the teaching sessions, observations of student actions were made, and after that, the expectations and attitudes of students towards this form of learning technique were assessed by an unidentified survey. The concluding scores achieved by the scholars listed in Table 2 were higher than the previous conventional approach outcome acquired.

The proportion of “Bad Marks” has reduced substantially when comparing the outcomes of two previous academic years. Through each party, the instructor has a great chance to observe the approaches and actions of the scholars over the seven 3-h seminar sessions. Some of the findings that are most important are as seen. (1) Obviously, some students were confused by the thought of not understanding “exactly what they had to do” instructor team have to make effort to construct their success and they began taking responsibility for it. (2) There were, however, gaps between teams in terms of the consistency of both the procedure used and the teams’ solutions achieved. A chief team member emerged at the beginning phase in some teams and cooperation was successful from start to its completion, even if a few associates looked additional inactive than others.

A further supportive and unbiased regularity of effort among the participants was established in other teams. In both occurrences, the attainment and scores of the students are very satisfactory. There seemed to be some misunderstanding about

**Table 2** Naïve Bayes algorithm blueprint concluding scores achieved with the conventional method and with the problem based learning method scales over 10 points

Scale	Marks	Traditional approach		PBL approach	
		2016–17	2017–18	2018–19	2019–20
10 9	Excellent	6%	7%	8%	9%
8 7	Very Good	21%	13%	27%	23%
6 5	Good	20%	30%	35%	40%
4 3 2 1	Bad	53%	50%	30%	28%

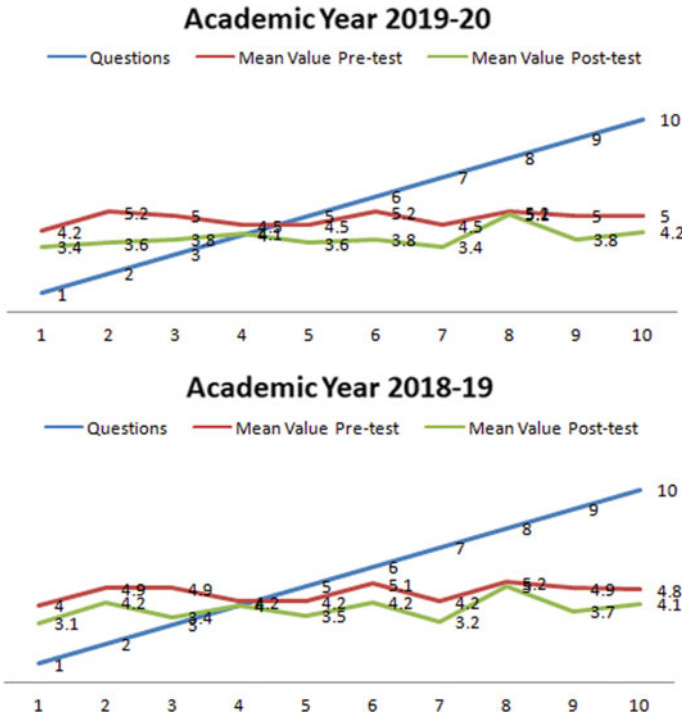
the proposed task in the other groups that received bad scores, and each participant worked individually, although the team faced problems in managing its rhythm adequately throughout the work period. (3) The extent of queries presented to the instructor throughout the work period was significantly excessive as compared to the conventional approach. A lot of queries lectured to the exercise of Naïve Bayes (e.g., its classification strategy, its accuracy, the area of use, and many more).

There was sufficient time for the instructor to communicate with each team at an individual level and to listen to the request and queries regarding the mission at hand. (4) Compared to the conventional methodology, where the scholars' study was additionally focused on notes extracted throughout the lectures, the exercise of required reading and a manual has increased dramatically. Most scholars have revealed an immense capability to discover, evaluate, and utilize the better relevant books for the intended reason.

This method offered students the chance to explore the library's great potential. Few scholars were unable to utilize books of English dialect; class explanations were helpful for those students. The Internet has also been used to explore new developments in the sense of machine learning and the usability of Naïve Bayes applications. Anonymous questioners of the 10 question statements represented in Table 3 surveyed student attitudes and expectations of this learning strategy. Based on the discussion and the review from the students, information was collected on the basis of a five marks measurement varying from "strongly agree" to "strongly disagree". The first survey was conducted at the commencement of the classes, when students were provided with the initial clarification of the PBL method (pre-test). Second, the questioner performed immediately subsequent to the verbal demonstration of the concluding assignment on the last day of the semester (post-test).

**Table 3** Survey concerning the problem base learning with the training education approach

Question	Statement
Q.1	I obtained a great extent of notes throughout the tutorial appraisal
Q.2	There is enough chance to contribute and post questions throughout the sessions
Q.3	Talking with group associates and asking questions throughout the session is effortless
Q.4	I have been skilled to recommend suggestions based on my previous understanding
Q.5	Having the chance to present my individual thought to propose the algorithm
Q.6	Listened cautiously to the statement and suggestions prepared by others during the group discussions
Q.7	I have grasped additional through group conversation that has never been done earlier at all
Q.8	Grasping during teamwork is better than through individual ones
Q.9	I have enhanced my planning ability
Q.10	Whether the project was of interest



**Fig. 4** Scholars’ feedback (mean-values) (question statements represented in pre-test and post-test questionnaire)

For the academic year 2018–19 and 2019–20, the outcome of the pre-experiment and the post-experiment is presented in Fig. 4. The outcome of the pre-tests is somewhat close, although, in almost all questions, the post-test of 2019/2020 poses higher mean values than those for 2018/2019.

The discrepancy may be attributed to slide issues with the implementation of the evaluation criteria during 2018/2019 from the opinion of the instructor.

In the beginning, scholars thought that training knowledge was important as compared to the pre-test results. The planned assignment proposal was exposed to be a difficult concern for scholar’s training (pre-tests; Question 2, mean value 5.2/5.1; Question 6, mean value 5.2/5.1). Indicated post-test evaluations are minor as compared to these questions in the pre-test. The depletion of the average values as of 5.2 in two quizzes (pre-test) to 4.2(post-test) in 2018/2019 is superior to those of 2019/2020, presumably suitable to the problem with the evaluation described above.

Teamwork is regarded as the session’s most important function (Question 8, mean value. 5.1/5.2 in pre-test; 5.2/5.0 in post-test) and the scholars believe so as to their training is strengthened by this. While the significant approach of scholars regarding other colleagues in the team is changeable, in the sessions to develop social abilities (Question 9 and Question 10, mean values about 5.0), cooperation is recognized as

being of great importance, by means of modal answers of comprehensive acceptance as elevated as 75%. The understanding of successful training by students is reasonably due to the attempt necessary to mark information (Question 7, mean value about 4.5), as compared to the knowledge of the instructor in this regard. Likewise, constructive behaviors were almost neutral during lectures or group conversation (Q1). It has also been noticed that the ability to participate and ask questions is of huge concern (Question 3, mean value about 5.0), but the scholars are not used to picking the benefit of these circumstances (Question 4, mean value varying from 4.1 to 4.5). The suggestion of theories to be exposed as reasonable attention for learner training based on their previous experience (Q5.m.v. around 4.5).

Two extra statements were used in the 2019/2020 questionnaires. The primary declaration confirmed that “I have improved my design/planning skills/abilities” and the average value of the answers was 5.4 in the pre-evaluation and 5.0 in the post-evaluation. A useful advantage of the PBL advances in technical education is the perceived benefit of students’ general capacity. The subsequent declaration on whether the plan was of concern generally, acquired as a core in both the pre-test and the post-test, which is superior for a computer science engineering curriculum, Artificial Intelligence and Machine Learning course, as an average value of 5.20.

## 5 Conclusion

For the fourth year of Computer Engineering, this paper explored a traditional PBL method as a preliminary research component on the Naïve Bayes Algorithm in Artificial Intelligence and Machine Learning. While there is a huge amount of research publications on the implementation of PBL, very modest of this applied to engineering, and still fewer to engineering subjects complementing engineering programs. Analysis of Problem Base Learning is when the scope of subjects is smaller than that obtained by the standard classes and seminar conveyance of subject material. This probable drawback is balanced by the information in contrast to the additional popular and exterior approach of the conventional learning technique; the Problem Base Learning approach encourages the profound method of knowledge learning in scholars, which ensures that the education- training practice is extra robust and successful. The strategy can be supposed to be booming, if it is measured with the previous two academic years, by enhancing academic results. Moreover, any shortage of background information and expertise required for solving composite challenges must be given careful attention. To resolve any issues that scholars possibly might have in their past information, complementary efforts should be made.

## References

1. Fauziyah N, Lant CL, Budayasa IK, Juniati D (2019) Cognition processes of students with high functioning autism spectrum disorder in solving mathematical problems. *Int J Instr* 12(1):457478. <https://doi.org/10.29333/iji.2019.12130a>
2. Geller EH, Son JY, Stigler JW (2017) Conceptual explanations and understanding fraction comparisons. *Learn Instr* 52:122–129. <https://doi.org/10.1016/j.learninstruc.2017.05.006>
3. Gultepe N, Celik AY, Kilic Z (2013) Exploring effects of high school students' mathematical processing skills and conceptual understanding of chemical concepts on algorithmic problem solving. *Aust J Teacher Educ* 38(10):106–122. <https://doi.org/10.14221/ajte.2013v38n10.1>
4. Jena PC (2014) Cognitive styles and problem solving ability of under graduate students. *Int J Educ Psychol Res* 3(2):71–76.10
5. ZA, Abdullah NH, Anthony E, Salleh BM, Kamarulzaman R (2016) Does problem-based learning improve problem solving skills?—A study among business undergraduates at Malaysian Premier Technical University. *Int Educ Stud* 9(5):166. <https://doi.org/10.5539/ies.v9n5p166.11>
6. Karaçam S, Digilli Baran A (2015) The effects of field dependent/field independent cognitive styles and motivational styles on students' conceptual understanding about direct current circuits. *Asia- Pac Forum Sci Learn Teach* 16(2), ar.6
7. Lopez BP (2014) Theme The 21st century adult learner. *Educ Res Rev* 12(8):540–548. <https://doi.org/10.5897/ERR2016.2928>
8. Loyens SMM, Jones SH, Mikkers J, van Gog T (2015) Problem-based learning as a facilitator of conceptual change. *Learn Instr* 38:34–42. <https://doi.org/10.1016/j.learninstruc.2015.03.002>
9. Margunayasa IG, Dantes N, Marhaeni AAIN, Suastra IW (2019) The effect of guided inquiry learning and cognitive style on science learning achievement. *Int J Instruc* 12(1):737–750. <https://doi.org/10.29333/iji.2019.12147a>
10. Mustofa RF, Hidayah Y (2020) The effect of problem-based learning on lateral thinking skills. *Int J Instr* 13(1):463–474
11. Palupi BS, Subiyantoro S, Rukayah, Triyanto (2020) The effectiveness of Guided Inquiry Learning (GIL) and Problem-Based Learning (PBL) for explanatory writing skill. *Int J Instr* 13(1):713–730. <https://doi.org/10.29333/iji.2020.13146a>
12. Reigeluth CM, Cheliman MC (2009) Instructional-design theories and model: building a common knowledge base, vol III. Taylor and Francis Publiser, New York
13. Rodzalan SA, Saat MM (2015) The perception of critical thinking and problem solving skill among malaysian undergraduate students. *Proc Soc Behav Sci* 172(2012):725–732. <https://doi.org/10.1016/j.sbspro.2015.01.425>
14. Sangestani G, Khatiban M (2013) Comparison of problem-based learning and lecture-based learning in midwifery. *Nurse Educ Today* 33(8):791–795. <https://doi.org/10.1016/j.nedt.2012.03.010>
15. Saricayir H, Ay S, Comek A, Cansiz G, Uce M (2016) Determining students' conceptual understanding level of thermodynamics. *J Educ Train Stud* 4(6):69–79. <https://doi.org/10.11114/jets.v4i6.1421>
16. Sellah L, Jacinta K, Helen M (2017) Analysis of student- teacher cognitive styles interaction: an approach to understanding learner performance. *J Educ Pract* 8(14):10–20
17. Strobel J, van Barneveld A (2009) When is PBL more effective? a meta- synthesis of meta-analyses comparing PBL to conventional classrooms. *Interdiscip J Problem-Based Learn* 3(1). <https://doi.org/10.7771/1541-5015.1046>
18. Sudarman, Setyosari P, Kuswandi D, Dwiyoogo WD (2016) The effect of learning strategy and cognitive style toward mathematical
19. Cuschieri S, Grech V, Savona-Ventura C (2018) WASP (Write a Scientific Paper): how to write a scientific thesis. *Early Human Dev* 127:101–105. <https://doi.org/10.1016/J.EARLHUMDEV.2018.07.012>

20. Dahl B (2018) What is the problem in problem-based learning in higher education mathematics. *Eur J Eng Educ* 43(1):112–125. <https://doi.org/10.1080/03043797.2017.1320354>
21. Mack C (2016) How to write a good scientific paper: review articles. *J Micro/Nanolith, MEMS, MOEMS* 15. <https://doi.org/10.1117/1.JMM.15.2.020101>



# Hybrid Microarray Gene Data Classification Based on GA-ACO Optimization



N. Karunya, V. A. Kanimozhi, N. Muthumani, M. P. Karthikeyan,  
and Dac-Nhuong Le

**Abstract** The advancement in the computerized medical scenario has grabbed the attention of researchers working in the field of medicine. Since clinical decision-making demands the utmost accuracy of diagnosis, it is a tedious and challenging task for physicians. An automated system that helps in disease diagnosis will benefit the medical industry. Microarray gene data classification is currently the most common cancer among world- wide. In this research work, a hybrid classification of genetic algorithm with Ant Colony Optimization (HGAACO) model for knowledge mining from microarray gene dataset is proposed and evaluated. The proposed HGAACO Classifier is one of the data mining techniques used to make decisions on real gene data. Using data directly from the database may affect the performance of the system. Pre-processing provides the data in a desirable form by handling missing values, selecting features, and scaling of datasets. Classification of the processed data provides a better decision on diagnosis. ACO optimizes and evaporation gradually reduces the pheromone level of all the trails. Subsequently, the pheromone levels of the trails that are not followed gradually decrease, which in turn lowers the probability of the trail being chosen by subsequent ants. HGAACO accuracy of 94.6% and a maximum accuracy improvement of 3.0% of other conventional approaches.

**Keywords** Microarray gene dataset · Genetic algorithm · Ant colony optimization · Cancer prediction

---

N. Karunya

Sri Ramakrishna College of Arts and Science, Coimbatore, Tamilnadu, India

V. A. Kanimozhi · N. Muthumani

Muthumani PPG College of Arts and Science, Coimbatore, Tamilnadu, India

M. P. Karthikeyan

School of Computer Science and IT, Jain (Deemed-to-be University), Bengaluru, India

D.-N. Le (✉)

Faculty of Information Technology, Haiphong University, Haiphong, Vietnam

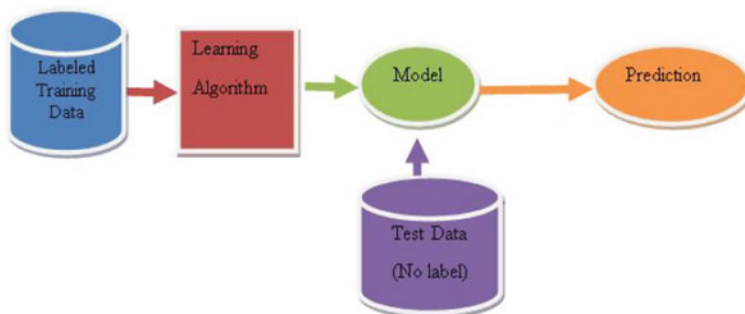
e-mail: [nhuongld@dhhp.edu.vn](mailto:nhuongld@dhhp.edu.vn)

## 1 Introduction

Artificial Intelligence based classification is a long-standing field showing continuous and vigorous growth. It incorporates intelligent behavior into machines and software. It is an interdisciplinary field that includes a number of sciences, professions, and specialized areas of technology. Artificial Intelligence assists in the decision-making process by performing data collection, treatment, processing, presentation, testing, and simulating new treatments, scenarios, and devices.

The medical data classification of data feature selection is optimization problem, which is based on the principle of picking a subset of attributes that are most significant in deciding the class label. It reduces the dimension of the data. During the training process, the presence of instances with missing values can lead to the degradation of the accuracy and performance of the classification model. By dealing with these missing values suitably, the performance of the model can be improved. Case Deletion is a simple and commonly used missing value handling technique which is used to delete the instances with missing values [1, 2]. Ant Colony Optimization (ACO) [3], a non-greedy local heuristic approach, is used to solve optimization issues. Because of its non-greedy nature, this algorithm can achieve the global maxima without getting stuck into local ones. It derives its name from the metallurgical annealing process, a technique that involves heating followed by controlled cooling of a material like steel so as to increase the size of the crystals (Fig. 1).

There is a rapid increase in the death rate of patients dying from different types of cancer. Early diagnosis can reduce the death count. Effective diagnostic methods and tools like medical expert systems and improvements in treatment methodologies have saved a considerable number of lives. The survival duration of women after proper diagnosis and treatment has increased [1, 2]. The researcher has proposed the integrated method of hybrid genetic algorithm with Ant Colony Optimization (HGAACO) for overcoming the challenge and providing an effective classification of microarray gene datasets by inclusive of efficiency maintenance [3–5].



**Fig. 1** General microarray gene prediction for machine learning approach

In the medical field, the accuracy of the disease diagnosis plays a vital role as it leads to further treatment of the patient. So, the prime objective dissertation is to improve the diagnosing accuracy of the medical expert system by

- Employing feature optimization techniques to select the most significant feature subset in the medical data.
- Constructing various classifier models (two-class) to train and test the clinical data.
- Optimizing classifier parameters and fuzzy rules by using single and hybrid optimization techniques.

The remainder of this paper is organized as follows. Section 2, microarray dataset prediction and its related work, Sect. 3 discussed Hybrid Genetic Algorithm with Ant Colony Optimization algorithm (HGAACO), and Sect. 4 presents the proposed system and existing systems experimental results comparison. Finally, Sect. 5 provides the concluding remarks and future scope of the work.

## 2 Literature Review

Many recent types of research are focused on secure classification due to the vastly rising development of internet usage throughout the information world.

The pattern recognition and data mining methods employed in risk prediction systems in the domain of cardiovascular medicine were introduced by Peter and Somasundaram [6]. A few restrictions in the usual medical scoring systems, there was an occurrence of intrinsic linear variable input set combinations and therefore were not adapted to model nonlinear difficult medical domains interactions. These restrictions had been tackled here by means of a classification pattern that indirectly identifies difficult nonlinear affiliations among dependent and independent variables and the capability to identify each probable interaction among predictor variables.

Xu et al. [7] discussed the realistic issue of Chinese hospitals handling with cardiovascular patients' data to create an early detection and prediction risk. To consider entire multi-techniques benefits and minimized bias, the top 6 sub-classifiers had been chosen to structure an ensemble system; a regulated voting system had been employed to create final consequences that were composed of risk prediction and poise. The system revealed a great accuracy of 79.3% for 2628 instances experiments with authentic patients. The risk prediction confidence and algorithm precision had revealed greater importance in practical usage for doctors' diagnosing.

A framework had been incorporated for easy prediction of cancer disease by Kavitha and Kannan [8]. The framework was generated with the Principal Component Analysis (PCA) to mine the characteristics and mathematical pattern from which to choose the related restraint. The projected work assisted in enhancing the efficacy, precision, and process speed. This could be applied in the applications like information retrieval, image processing, and pattern matching.

Saxena and Sharma [9] intended a system to identify the rules effectively to predict patients' risk level on the basis of a provided parameter regarding their health. The rules had been prioritized on the basis of the user's prerequisite. The system performance was assessed based on precision classification and the consequences revealed this system had greater potential to a level more precisely. Radhimeenakshi [10] integrated disease dataset classes by exploiting SVM and ANN. The investigation had been completed between two schemes based on accuracy and training time premise. The dataset exploited were the Cleveland Heart Database and Statlog Database obtained from UCI Machine learning dataset vault. Here, the data were deployed into two classes in SVM and ANN. In addition, it examined both dataset performances.

Wijaya and Prihatmanto [11] focused on cancer disease prediction development by means of machine learning. Data had been obtained using smartphones and smart chairs. Cancer rate data was obtained via the Internet and gathered in a server. System approaches had been performed for 1-year period to attain adequate data for predictions. Potential heart disease predictions over a period of one year had improved a person's heart disease awareness. This system was also supposed to minimize patient deaths from heart disease.

Sabab et al. [12] objectivized to optimize the study of cardiovascular disease prognosis by using multiple data mining methods. The authors have offered a method to enhance the projected classifier pattern by feature selection. A feature selection approaches assisted to enhance the precision of every through minimizing a few low-ranked attributes that aided in attaining precision of 87.8%, 86.80%, and 79.9% in the case of SMO, Naive Bayes, and C4.5 Decision Tree algorithms correspondingly.

The kNN algorithm efficiency was incorporated with Ant Colony Optimization (ACO) by Rajathi and Radhamani [13] to predict heart disease. The investigation had been carried out in two phases. The dataset employed here was Streptococcus Pyogenes bacteria that cause Rheumatic Fever, as Acute Rheumatic Fever (ARF). A novel algorithm kNNACO that had been incorporated into the present approach and the same was examined on the basis of precision and error rate. Amin et al. [14] explained a system by using significant risk factors. The projected method is composed of the two most unsurpassed data mining tools, namely neural networks and genetic algorithms. The executed hybrid system employed the global optimization merits of a genetic algorithm to initialize neural network weights. The learning was rapid, constant, and precise when evaluated over back propagation. The system had executed in MATLAB and predicted heart disease with a precision of 89%. This literature summarized disease prediction data mining techniques, feature selection techniques, classifiers techniques, and optimization techniques.

### 3 System Design

Classification is the finding of a model for describing as well as distinguishing the classes or the concepts of data for being able to utilize the model for predicting of the class of the object class labels that are unknown. This model is further based on data object analysis that has known class labels. There are various techniques of classification in data mining and GA is one of them. The GA architecture, the actual number of codes to be chosen, and how the weights have to be set between the nodes at the time of training and evaluation of results are all completely covered. The function of activation is mentioned together with the rate of learning, the momentum, and the pruning. The BP algorithm is a very popular GA algorithm. The GA can work on errors better than that of the traditional computer programs (like in a scenario of a faulty statement in the program which can halt everything when the GA will handle errors better). Here in this work, the optimized ACO along with the GA and the ACO are proposed.

#### 3.1 Genetic Algorithm

The Genetic Algorithm (GA) [4] is population based and makes use of the producer-scrounger model and data mining. The producer-scrounger is the design of an optimal search scheme, which owes its inspiration to animal searching behavior and also group living theories. To ensure that it is not forced into the local minimum, the Genetic Algorithm uses the ranger foraging method. The Genetic Algorithm protocol is referred to as a group and all individuals are members.

#### 3.2 Ant Colony Optimization

The ant mechanism is the first algorithm that uses the principle of the ACO heuristic. The ant forage for food by iteratively constructing answers and adding appropriate the paths suitable to these answers as in the stochastic procedure of path selection is based on two parameters namely, the heuristic values and pheromone. There are two foraging methods that produce (search for food) and scrounging (resource combination that is discovered by others) that have been adopted by this protocol. To ensure that it is not forced into the local minimum, the ACO uses the ranger foraging method [3, 5, 15, 16].

Once the ant reaches its final destination, the path followed by the ant is calculated and the appropriate measures are expanded appropriately. Evaporation gradually reduces the pheromone level of all the trails. Subsequently, the pheromone levels of the trails that are not followed gradually decrease, which in turn lowers the probability of the trail being chosen by subsequent ants. If a searching procedure in the ACO

is executed, the ranger or the scrounger will have chances of discovering a location that is better and the current producer of other members will fail to discover a better location. The ranger or the scrounger has a better location in the next session and the producer and the other members in the previous search session carry out the activity of scrounging.

There are two foraging methods that produce (search for food) and scrounging (resource combination that is discovered by others) that have been adopted by this protocol. The Scout bees: The food source which is evaluated becomes scout bees. At least 5 to 10% of the employee bees will turn into scout bees. These bees evaluate the source of food using the fitness value of a certain source of food that is greater than the previous one. The bee will delete the memory of the previous one and will memorize the existing one.

### 3.3 Hybrid Algorithm

The technique will merge the global scheme which is of opposition-based ACO having a local search capacity of a GA protocol that is conventionally having a momentum term. The opposition based and arbitrary perturbation techniques are two types of elements in the protocol. A time-variant social and cognitive element will enhance the capacity of this protocol for a search. The factor of constriction is that one other variable that ensures convergence and overfitting being the problem acquires more specifications while training (see Fig. 2).

Classifications tend to make life easier in the case of a supermarket when things are placed on a shelf randomly it can make it an unpleasant experience to shop. The GA architecture, the actual number of codes to be chosen, and how the weights have to be set between the nodes at the time of training and evaluation of results are all completely covered. The function of activation is mentioned together with the rate of learning, the momentum, and the pruning. The GA algorithm is a very popular ACO algorithm that was demonstrated in [4, 17, 18]. The GA can work on errors better than that of the traditional computer programs (like in a scenario of a faulty statement in the program which can halt everything when the ACO will handle errors better). Here in this work, the optimized GA-ACO is proposed.

If a searching procedure in the GA is executed, the ranger or the scrounger will have chances of discovering a location that is better and the current producer of other members will fail to discover a better location. The ranger or the scrounger having a better location in the next session and the producer and the other members in the previous search session carries out the activity of scrounging. This fitness function is designated to the  $i$ th individual is a least-squared error function as per equation (Fig. 3).

$$F_i = \frac{1}{2} \sum_{p=1}^p \sum_{k=1}^K (d_{kp} - y_{kp}^i)^2 \quad (1)$$

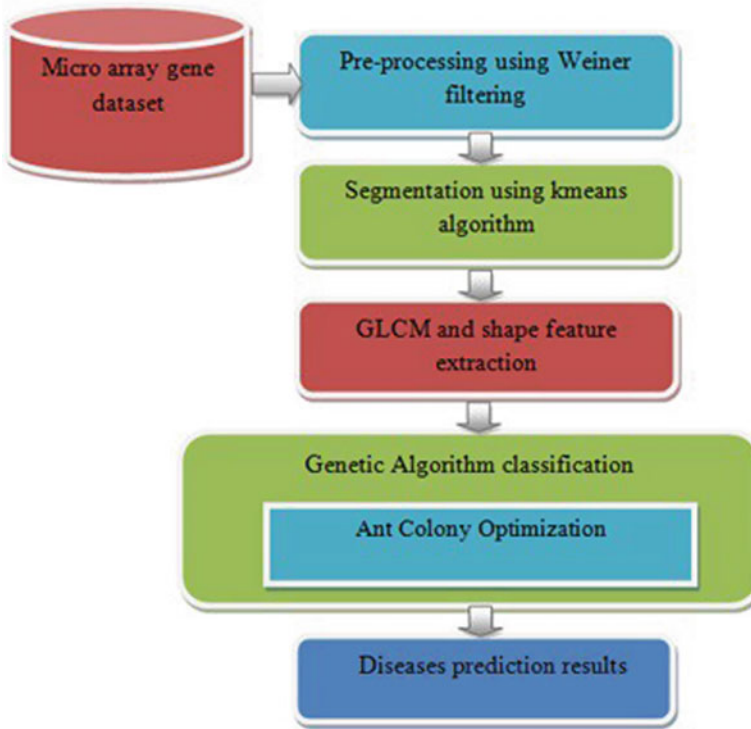
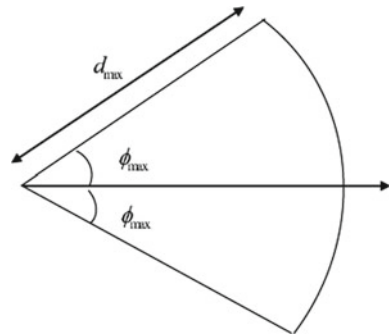


Fig. 2 Flow diagram of the proposed system

The error in the training set can be driven to a small value by means of minimizing the error function but has its side effect, the problems of over-fitting may sometimes occur and result in a generalization error that may be large. So, for improving the performance of the ACO, the previous stopping strategies are suggested. The rate of error validation has been watched during the training period.

Fig. 3 Producer scanning field



## 4 Result and Discussion

The proposed methodology is applied by making use of Python3.6IDE on Intel(R) Core (TP) i3-2410 M CPU @ 3.20 GHz and 8 GB RAM. Two different models GA and ACO have achieved an average accuracy of 95.6%, which were developed through the diagnosis of heart disease prediction. Also compare the performance of the hybrid genetic algorithm with ant colony optimization (HGAACO) method to other standard classifiers: Support Vector Machine (SVM), Decision Tree (DT), and genetic algorithm. A set of experiments is conducted on the dataset by different number of genes chosen to receive the highest classification accuracy (Table 1).

The dataset of ALL/AML considered has 144 samples 14,258 types of genes are classified into the class 3 and class 4 types. The trained dataset of AML/ALL 28 feature attributes are considered as a genetic parameter and the Ant Colony Optimization algorithm is used to optimize 28 feature parameters and train data stored. Test data s input data of 7 features that are taken as AML/ALL parameters and test data are stored.

Table 2 explains the model based on the proposed HGAACO that yields the maximum false positive rate Fscore, and true positive rate for ALL/AML dataset (81.32, 91.25, and 84.12), at nearer Genetic Algorithm provides the only 54.85,87.54, and 71.78 values comparatively. The proposed approach provides a better result in the different terms between 27.12, 4.25, and 12.87 values.

The SVM offers the least false positive values of 31.02%, true positive values of 78.45, and Fscore values of 54.12. While the proposed hybrid genetic algorithm with ant colony optimization (HGAACO) yields the quality matrix values for the micro-array gene dataset explained in Fig. 4. The quality performance values exposed to the comparatively HGAACO are better than SVM, DT, and GA.

Table 3 explains the model based on the proposed HGAACO yields the maximum efficiency, precision, and error rate for ALL/AML Dataset of (91.75, 94.65, and 6.75), at nearer Genetic Algorithm, provides the only 86.50 of efficiency, 90.85 of precision,

**Table 1** Summary of the ALL/AML dataset

Dataset	Samples		Classes
ALL-AML-3	72	7129	3
ALL-AML-4	72	7129	4

**Table 2** Performance comparison ALL/AML dataset

Methods	False positive	True positive	Fscore
Support vector machine	31.01	78.45	54.15
Decision tree	43.65	81.87	61.78
Genetic algorithm	54.85	87.54	71.78
Proposed HGAACO system	81.32	91.25	84.12



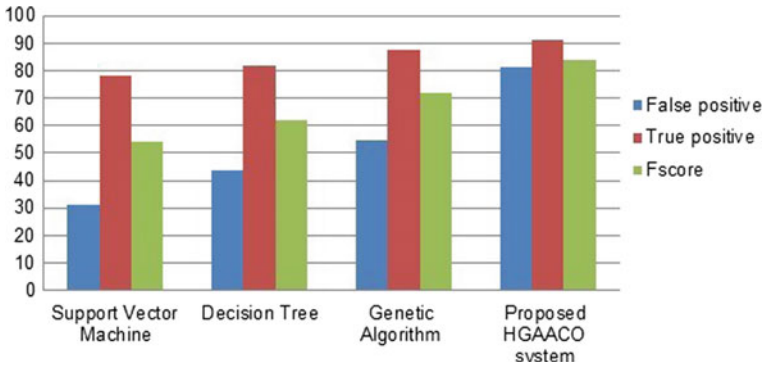


Fig. 4 Performance comparisons for ALL/AML dataset

Table 3 Efficiency comparison ALL/AML dataset

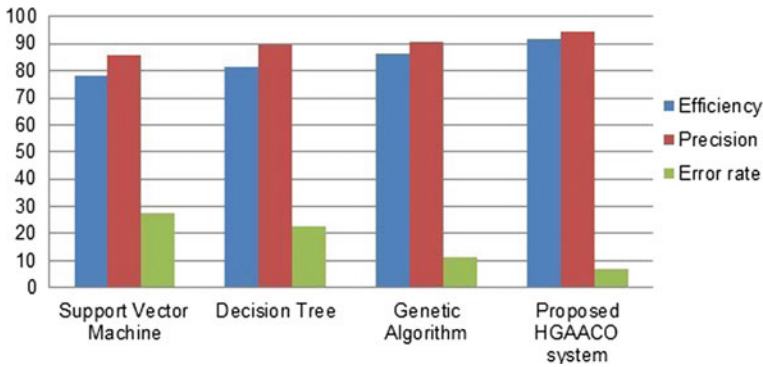
Methods	Efficiency	Precision	Error rate
Support vector machine	78.35	85.60	27.65
Decision tree	81.25	89.70	22.68
Genetic algorithm	86.50	90.85	11.24
Proposed HGAACO system	91.75	94.65	6.75

and 11.24 of error rate values. Comparatively, the proposed approach provides the better result in the different terms between 5.25, 3.80, and 4.49 values.

The SVM offers the least efficiency values of 78.35%, precision of 85.6, and error rate values of 27.65 (Table 3). While the proposed Hybrid Genetic Algorithm with Ant Colony Optimization (HGAACO) yields the quality matrix values for the micro array gene dataset explained in Fig. 5. The quality efficiency values exposed to the comparatively HGAACO are better than SVM, DT, and GA. In the proposed Hybrid genetic algorithm with Ant Colony Optimization (HGAACO) system, unrelated and redundant structures are removed from the data, the choice of the structure will help in the enhancement of the presentation of the learning models if the decreased data goes for its classification.

## 5 Conclusion

The prediction of diseases of cancer is intended to help oncologists in diagnosis. This method is proposed for classifying the data on ailments in cancer. The medical history of the patients and symptoms by data mining as the data will have a major trait in selecting the method used for dataset plummeting. For this work, the method known as GA-ACO with the microarray gene datasets is used. A population-based system of stochastic optimization will be a healthy and active ACO that is based



**Fig. 5** Efficiency comparisons for ALL/AML dataset

on the movement of swarms. The GA training will be facilitated using the ACO for obtaining a real outcome in real-world standards. The device helps in addressing issues of uninterrupted augmentation. Results have proved that the HGA-ACO tends to have a higher accuracy of classification by about 5.25% for the GA, by about 9.75% for the DT, and by about 13.25% for the SVM.

## References

1. Polat K, Gunes S (2007) An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. *Digital Signal Process* 17(4):702–710
2. Sarkhel D, Deka D, Samanta D, Kumudavalli MV, Le DN (2020) GUI-Based Percentage Analysis for Curing Breast Cancer Survivors. In: *Frontiers in intelligent computing: theory and applications*, pp 319–327. Springer
3. Le DN (2017) A new ant algorithm for optimal service selection with end-to-end QoS constraints. *J Internet Technol* 18(5):1017–1030
4. Nayyar A, Le DN, Nguyen NG (Eds) (2018) *Advances in swarm intelligence for optimizing problems in computer science*. CRC Press
5. Le DN (2015) Evaluation of pheromone update in min-max ant system algorithm to optimizing QoS for multimedia services in NGNs. In: *Emerging ICT for bridging the future-proceedings of the 49th annual convention of the computer society of India CSI, Vol 2*, pp 9–17. Springer, Cham
6. Peter TJ, Somasundaram K (2012) An empirical study on prediction of heart disease using classification data mining techniques. In: *Advances in engineering, science and management (ICAESM 2012)*, pp 514–518
7. Xu S, Shi H, Duan X, Zhu T, Wu P, Liu D (2016) Cardiovascular risk prediction method based on test analysis and data mining ensemble system. In: *ICBDA-2016*, pp 1–5
8. Kavitha R, Kannan E (2016) An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining, In *Emerging trends in engineering, technology and science (ICETETS)*, pp 1–5
9. Saxena K, Sharma R (2015) Efficient heart disease prediction system using decision tree. In: *Computing, communication & automation (ICCCA), 2015 international conference on IEEE*, pp 72–77

10. Radhimeenakshi S (2016) Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural network. In: Computing for sustainable global development (INDIACom), 2016 3rd international conference on IEEE, pp 3107–3111
11. Wijaya R, Prihatmanto AS (2013) Preliminary design of estimation heart disease by using machine learning ANN within one year. In: Rural information & communication technology and electric-vehicle technology (rICT & ICeV-T), 2013 joint international conference on IEEE, pp 1–4
12. Sabab SA, Munshi MAR, Pritom AI (2016) Cardiovascular disease prognosis using effective classification and feature selection technique. In: Medical engineering, health informatics and technology (MediTec), pp 1–6
13. Rajathi S, Radhamani G (2016) Prediction and analysis of Rheumatic heart disease using kNN classification with ACO. In: Data mining and advanced computing (SAPIENCE), international conference on IEEE, pp 68–73
14. Amin SU, Agarwal K, Beg R (2013) Genetic neural network-based data mining in prediction of heart disease using risk factors. In: Information & communication technologies (ICT), 2013 IEEE conference on IEEE, pp 1227–1231
15. Bhateja V, Tripathi A, Sharma A, Le BN, Satapathy SC, Nguyen GN, Le DN (2016) Ant colony optimization based anisotropic diffusion approach for despeckling of SAR images. In: International symposium on integrated uncertainty in knowledge modelling and decision making, pp 389–396. Springer
16. Le DN, Nguyen GN, Bhateja V, Satapathy SC (2017) Optimizing feature selection in video-based recognition using Max-Min ant system for the online video contextual advertisement user-oriented system. *J Comput Sci* 21:361–370
17. Le DN (2015) Performance evaluation of heuristic algorithms for optimal location of controllers in wireless networks. In: Information systems design and intelligent applications, pp 843–853. Springer, New Delhi
18. Bhateja V, Sharma A, Tripathi A, Satapathy SC, Le DN (2016) An optimized anisotropic diffusion approach for despeckling of SAR images. In: Annual convention of the computer society of India, pp 134–140. Springer.

# An In-Field Real-Time Automatic Weed Detection Using Deep Learning Techniques



Siddharth Dutt Choubey and Rohit Singh Thakur

**Abstract** According to various studies, the world population will exceed 10 billion in 2050, and with the increasing demand for food quantity, maintaining quality will become more difficult. On-field, some undesirable plants grow along with the crops, which affects the crop yields. They share the crop's resources; therefore, the actual crop generally lacks the necessary nutrients and other factors. To treat such weed plants, farmers use herbicides, which affect the food quality, soil condition, and environment. A real-time weed identification in the field is required at this time so that a proper diagnosis can be made to stop the growth of such weed plants. In this work, soybean images are captured using a UAV device, and weed images are labelled to generate a dataset. Total, 3324 images are labelled in the form of bounding boxes to localize the weed area. To evaluate the performance, two more datasets are taken into consideration. A comparison of three different object detection models, YOLO v3, YOLO v4, and Faster RCNN, has been performed. As per the results, YOLO v4 achieved the best results among the three methods on all three datasets, with 90.00% mAP on the soybean dataset.

**Keywords** Deep learning · Object detection · Weed detection · YOLO · Faster RCNN

## 1 Introduction

India is a farming nation. It owns a diversity of agricultural-friendly climates and agriculture plays a significant part in India's economy. It contributes to around 16% of India's GDP and is responsible for 49% of employment. India has approximately 179.8 MAh of net cropland, which is 9.6% of the global, making it the number one agricultural production country [1]. In India, approximately 65% of the population

---

S. D. Choubey (✉)  
Shri Ram Institute of Technology, Jabalpur, India  
e-mail: [siddharth.choubey@gmail.com](mailto:siddharth.choubey@gmail.com)

R. S. Thakur  
Rajiv Gandhi Proudyogiki Vishwavidyalaya, Bhopal, India

resides in rural areas where nearly 60% are still totally dependent on agriculture for their survival [2].

However, India has an extremely diverse agriculture system surrounded by several types of agricultural problems associated with weeds. Weeds are unnecessary plants that grow with crops and share their nutrients and other supplements. They affect productivity, thus minimizing the production profits for farmers. As per studies, weeds are responsible for approximately 20–80% of crop yield losses, decreasing the quality of the final product. It is also not good for health and the environment [3].

Weeds are affecting agriculture production on a large scale. Weeds have always been a severe problem for the farmers of India. They cause significant harm to the yield and quality of crops. Around 33% of the crop yield loss in India is due to weeds. If weeds are untreated, the possible yield loss can range from 20 to 100% [4], which is a significant loss for the country's economy. In Fig. 1, an example image from a soybean field is presented with the area covered by weeds highlighted.

Indian farmers spend considerable resources to control weed spread, in which they often fail. Losses in the crop can vary depending on the type of weed affecting it. Diagnosis and cure of these plants also require ample human resources and herbicides, which directly increase expenses. Also, such herbicides sprayed over the entire field affect crop conditions. Usage of such herbicides affects human health and worsens environmental conditions. An automatic weed identification system is required to tackle such problems that can result in timely action and control. Hence, maintaining crop conditions and the cost to the farmer will be reduced.

For the past decade, deep learning has been showing outstanding results in most real-world problems. Notably, in image-based problems, deep learning outperforms human object identification capabilities. It automatically extracts a particular object's features and distinguishes it from others when a sufficient number of examples are



**Fig. 1** Soybean field with weeds

presented. For supervised object detection tasks, object ground truths are fed to the deep learning model, which helps in understanding the unique features of the object in all the ground truths.

Various works have taken into consideration weed identification with vision-based approaches using image processing machine learning and deep learning techniques. In 2012, Ahmed et al. [5] classified weed into six categories with the help of SVM and achieved 97.3% accuracy on 224 images. Tang et al. [6] applied vertical projection and a linear scanning approach for crop row identification along with horizontal projection. They reported 92.5% classification accuracy with the SVM classifier with 1300 images. Many image processing and machine learning methods have shown promising results for weed identification [7–13].

Identification of weeds in the field is a challenging task as the features of weeds are similar to those of typical crop plants. The weed-affected area is covered by the bounding boxes to work as ground truth for the model. In this work, three CNNs are deployed to identify the best for weed detection work. Figure 2 depicts the overall workflow for the weed detection process. Here, YOLO v3 [14], YOLO v4 [15], and Faster RCNN [16] methods are used for the weed localization task. Three different datasets are used for the weed detection task. The soybean dataset is collected and labelled for the in-field weed detection task. The main contributions of the present work are as follows:

- In-field weed identification and localization using CNN-based object detection methods.
- Labeled dataset of soybean with bounding box annotation.
- Comparison of YOLO v3, YOLO v4, and Faster RCNN architectures.

The paper is organized as follows: Sect. 2 provides a brief overview of the existing work in the area of weed identification. Section 3 is dedicated to real-time object detection algorithms and the proposed work. Section 4 discusses the results and discussion. Section 5 provides the conclusion and possible future scope.

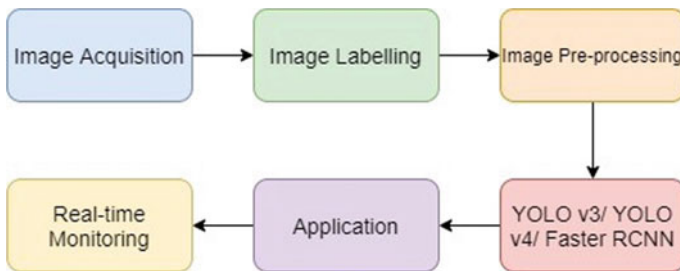


Fig. 2 Workflow of the weed detection system

## 2 Related Work

With the help of computer vision and deep learning techniques, several approaches have been introduced by the researchers in the area of weed detection tasks. Initially, Ahmed et al. [17] suggested a texture-based weed classification scheme. They are adopting the local binary pattern (LBP) as a feature descriptor in their scheme. Furthermore, for the classification rules, theories of SVM and template matching are employed. The LBP operator is used for extracting the features from the farm images. The generated feature representations helped in the classification of multiple varieties of weeds. For experimenting, the authors used two hundred sample images.

Lee et al. [18] recommended a Deep Conventional Neural Network based plant identification system called DEEP-Plant [18]. In the proposed system, the CNN model was used as a classifier that learns the features of a leaf to make plant classification, and DN was used as a visualization method for the learned features. The suggested system contributed the strategy to researchers for how the algorithm observes a leaf. Performance was reported on the Malayakev (MK) leaf dataset, consisting of 44 sections, developed by the Royal Botanical Gardens, Kew, England. A total of 2288 training and 528 testing images were used in the experiment after applying rotation-based data augmentation.

Pearlstein et al. [19] developed a CNN-based method for weed detection in lawn grass based on realistic synthetic imagery. A Caffe-based deployment is performed for training and testing a DCNN model using AlexNet. This includes five convolutional layers followed by three fully connected layers. They employed the ReLU and dropout in the network and evaluated categorical cross-entropy loss. The study has been evaluated on 9,000 synthetic grayscale images. The results of the proposed work suggest that realistic synthetic imagery can be effectively used for training DCNNs and that minimal deep convolutional neural networks can also be handy for simple image recognition tasks.

Barrero et al. [20] developed an artificial neural network (ANN) model for weed detection in aerial images of rice land. The images were acquired from a height of 50 m. The dataset has 250 images, which are divided into 7:0.15:0.15 ratios for training, validation, and testing, respectively. They have employed the Gray-Level Co-Occurrence Matrix with the Harlix descriptor and Normalized Difference Index for texture and color features.

Umamaheswari et al. [21] proposed a Parallelized Weed Detection System (PWDS) employing CNN to localize weed. In this system, parallel processing on the GPU has been used to enhance the in-field test performance. The system was mainly developed for the classification process. Furthermore, it identifies the position of the weed in the image. There were three modules in PWDS: the first is the image acquisition module to collect the real-time images from the carrot farm. In the second image labelling module, some supervised classification techniques were used for grouping the images into separate labelled directories, which contained the annotated images. The OverFeat network has been used for image feature extraction. The final result of the PWDS was an image with bounding boxes over the weed. The author claimed

that the proposed PWDS has an accuracy of 91.1% for carrot crops, with weeds close to crops, both in the same shape and overlapping each other.

Tiwari et al. [22] revealed an experimental study using CNN for weed detection. The Inception v2 model has been trained to reduce the time in which weeds are detected in crops. The image dataset has a total of 555 images captured using a mobile phone and a DSLR camera with varying exposure, brightness, and saturation. The collected images were labelled for weed and plant features via some labelling software. Due to the shortage of computational power, only 50 images were used as the training dataset, and ten images were used as the test dataset. To increase model predictability for precisely distinguishing a weed in the dataset, cropped weed images were used.

Sarvini et al. [23] proposed a comparative study of a machine learning-based weed detection system. The morphological features are extracted, and the performance study for three classifier algorithms, Support Vector Machine (SVM), Artificial Neural Network (ANN), and Convolutional Neural Network (CNN) was conducted. The excess green method and Otsu's thresholding were used to mask the soil and extract the region of interest. According to the study, the authors claimed that CNN gives better results as compared to SVM and ANN.

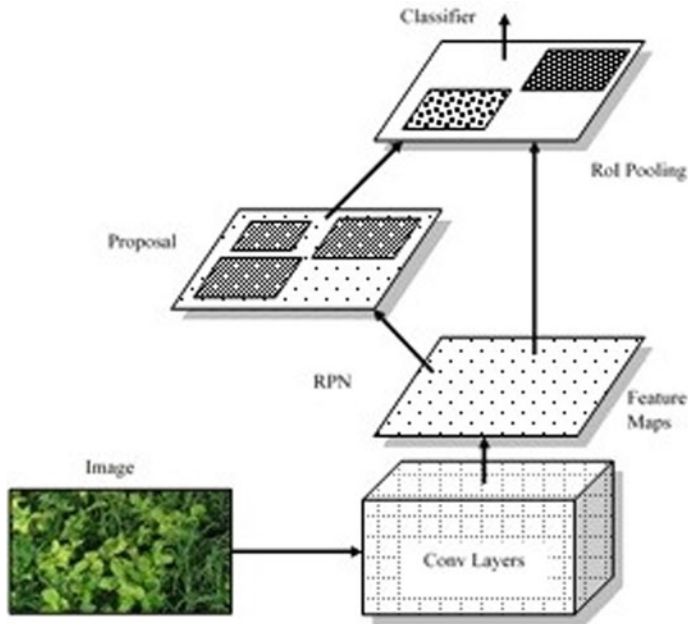
Osorio et al. [24] worked on lettuce crop weed detection using SVM with histograms of oriented gradients (HOG), YOLO v3, and Mask R-CNN methods and reported 88%, 94%, and 94% F1-score, respectively. Gao et al. [25] detected *Convolvulus sepium* weed in sugar beet plants with the help of YOLOv3 and Tiny YOLO architectures. 2271 synthetic images and 452 images were used for performance evaluation and weed detection and reported 0.76 and 0.89 average precision with YOLO v3 and Tiny YOLO, respectively.

## 3 Methodology and Approach

### 3.1 Methodology

**Faster RCNN**—The Faster R-CNN [16] is mainly used for real-time object detection tasks in computer vision. In Faster R-CNN, there are two fundamental components as shown in Fig. 3. The first is a deep convolutional network, which is for the proposed regions, and the second one is the Fast R-CNN Detector [26] that uses the region generated by the CNN. Faster R-CNN uses Region Proposal Network (RPN), that are utilized the attention mechanism [27]. It is prepared to create superior recommendations for regions, which are utilized by Fast R-CNN for the detection task. The generation of some bounding boxes termed “Region of Interests” (ROIs) is the primary objective of RPN. These boxes have a high possibility of including any object. The Faster R-CNN is faster than the former models like R-CNN and Fast R-CNN.





**Fig. 3** Architecture of Faster-RCNN

**YOLO**—YOLO (You only look once) [28] appeared in 2016, and it was a breakthrough in a real-time-based object detection task. It is a single-stage model for object detection. The Darknet53 is the backbone of YOLO, which is written in C and CUDA. After the first version of YOLO, new versions of YOLO came: YOLO9000 [27], YOLO v3 [29], and YOLO v4 [14], which were much faster and more accurate than their previous variants. YOLO uses DarkNet53 CNN for image feature extraction as shown in Fig. 4. In this study, we are applying YOLO v3 and YOLO v4 for the weed detection and localization task. YOLOv3 has 106 layers for image feature extraction. YOLOv3 detects an object from small to large with three different scales. There are nine anchor boxes used by YOLOv3, so it predicts more bounding boxes in comparison to YOLOv1. YOLOv4 is an improvement and the latest method over the previously existing methods. It out-performs the existing methods in terms of both “detection performance” and “speed”. Spatial Pyramid Pooling additional modules and PANet path-aggregation are used in the model.

### 3.2 Dataset Collection

As per the best of our knowledge, there was no dataset available for weed detection and localization tasks till 2019. So, we acquired a dataset [30] from the GitHub

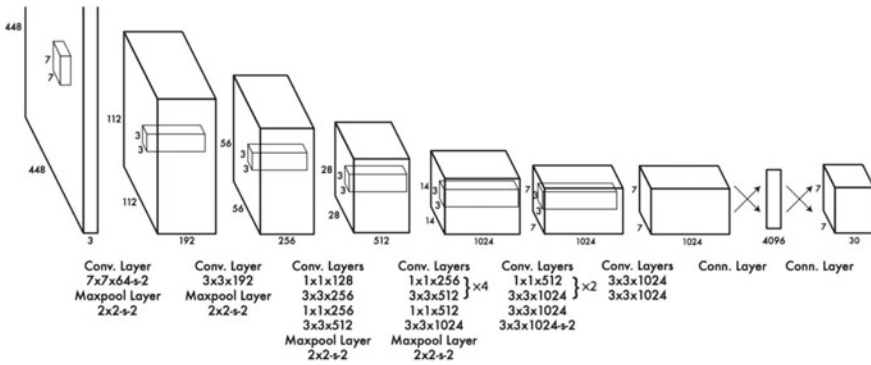


Fig. 4 Architecture of YOLO [28]

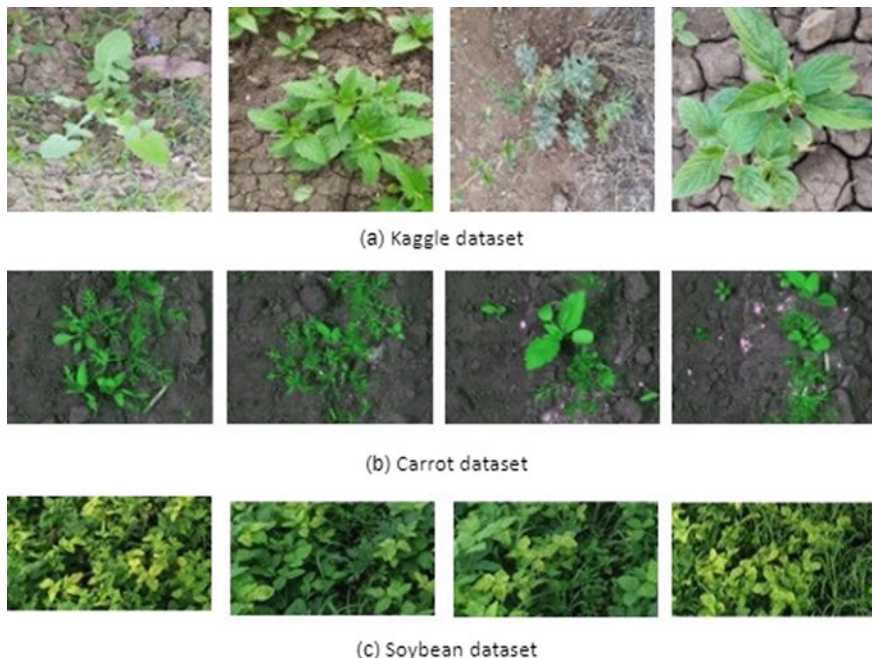
platform, which consists of carrot crop images with weeds present in the images. So, we started creating bounding boxes in all the images and with the data augmentation task due to the smaller number of images available in the dataset. After augmentation, a total of 300 images are labelled in weed and crop classes with bounding boxes. Then, in 2020, we could locate a new dataset [31] for weed detection with bounding boxes on the Kaggle platform. It consists of a total of 1300 images with labels such as “crop” and “weed”.

Simultaneously, we also collected the soybean disease dataset and started labeling the dataset. We have selected some 3324 images and labelled barnyard grass weed present in the images. Figure 5 shows some sample images from all the datasets and Table 1 presents details of all three datasets.

### 3.3 Approach

To localize the weed area in the images, we started working on the state-of-the-art approaches for object localization. As per the literature, YOLO and Faster RCNN are the most prominent methods applied in object localization. Hence, in this work, YOLO v3, YOLO v4, and Faster RCNN with VGG16 as image feature extractors are implemented for the weed identification process. All the models have been fine-tuned on the three different datasets to identify the best performing model. So that the same model can be further deployed for real-time in-field testing.

In YOLO v3 and v4, a darknet framework has been set up based on the datasets. In both the models, the pre-trained weights of COCO datasets have been fine-tuned on all three datasets. In Faster RCNN, five convolutional blocks of the VGG-16 model have been used for image feature extraction with pre-trained weights on the ImageNet dataset.



**Fig. 5** Example images from all the three datasets

**Table 1** Dataset description

Dataset	No. of images	Classes
Carrot dataset	300	2 (weed, crop)
Kaggle dataset	1300	2 (weed, crop)
Collected dataset	3324	1 (barnyard weed)

## 4 Results and Discussion

### 4.1 Training

The labelling process is performed on the local computer with the help of LabelImg tool [32]. Because image processing tasks necessitate GPU training, Google Colab with GPU Tesla K80 and 12 GB RAM is used to train faster RCNN, YOLOv3, and YOLOv4 models using the Python programming language. In the carrot dataset, only 60 images were present, so data augmentation techniques were applied to those images, and then 300 images were generated. We used horizontal flip, vertical flip, rotation, shear, and zoom augmentation. For training the approaches on various methods, all the datasets are divided into training and validation sets with the ratio of 80:20.

## 4.2 Performance Metrics

**Confusion Matrix:** A confusion matrix is defined using four other classification factors: true positive, true negative, false positive, and false negative. Either the object prediction is correct or incorrect. As per the prediction, these factors are defined as below:

**True Positive (TP):** The predicted output is 1, and the actual output is 1. **True Negatives (TN):** The predicted output is 0 and the actual output is 0. **False Positives (FP):** The predicted output is 1 but the actual output is 0. **False Negatives (FN):** The predicted output is 0 but the actual output is 1.

**Accuracy:** Classification accuracy is the sum of TP and TN divided by the total number of predictions.

**Precision:** Precision is the number of True Positives (TP) divided by the number of True Positives (TP) and False Positives (FP).

**mAP:** To measure the accuracy of object detectors, one popular metric, namely AP (average precision) is used. Average precision is used to compute the average precision value for recall value ranging from 0 to 1. Mean average precision is the average of AP and is conventionally referred to as mean average precision (mAP).

## 4.3 Result

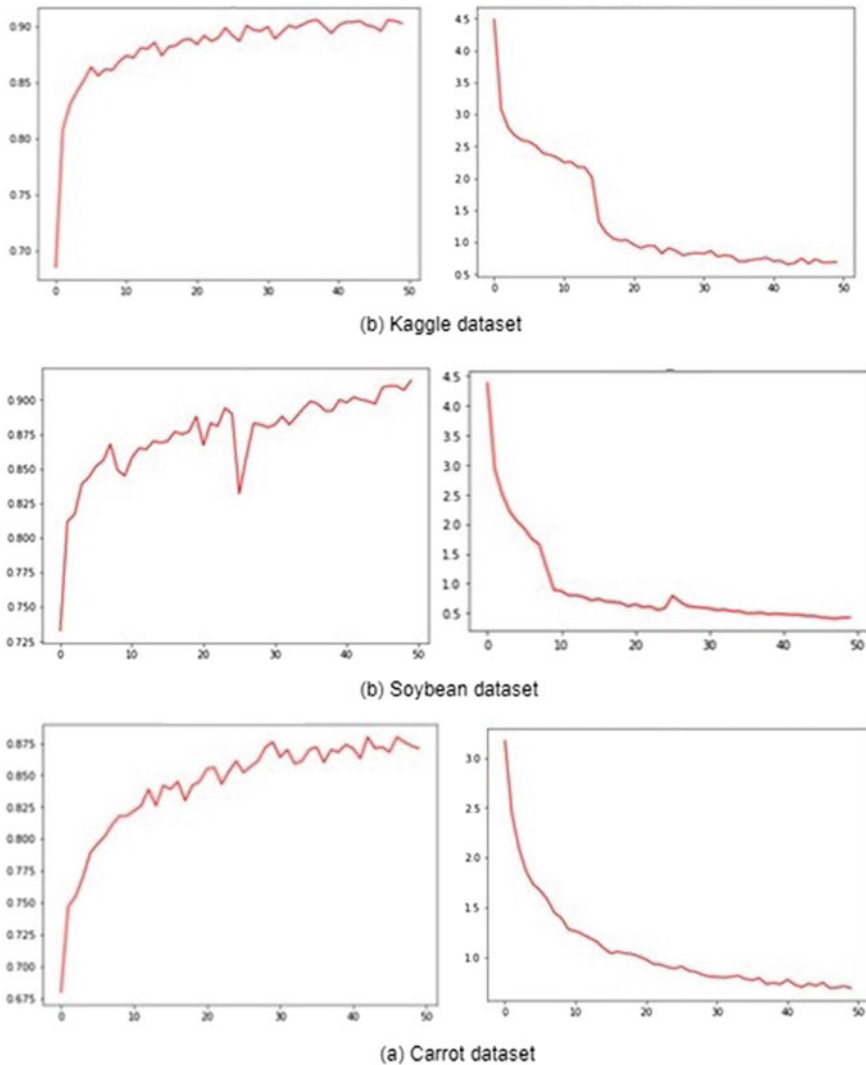
After successful training, the models are analyzed based on the performance measures known as mAP (mean average precision). YOLO v3, v4, and Faster RCNN methods are evaluated on three datasets. Table 2 consists of a comparison of all the three approaches on all datasets based on the prediction time on a single image and mAP on the test dataset.

Graphs in Fig. 6 present classification accuracy for Faster RCNN on the left and total loss, i.e., regression loss and classification loss, on the right on all three datasets. As the graphs show, the model is adequately fitted on all three datasets as the curve is nearly flattening.

The Carrot dataset was obtained from the GitHub repository and then labelled to create bounding boxes. In Table 2, the results of all the three approaches are present, and as mentioned, the most commendatory result is obtained by the YOLOv4

**Table 2** Comparative results for all the three datasets

Dataset	Faster RCNN		YOLO v3		YOLO v4	
	mAP	Time (secs)	mAP	Time (secs)	mAP	Time (secs)
Carrot	72.63%	1.50	86.31%	1.20	89.90%	0.4
Kaggle	66.7%	0.40	60.45%	0.12	78.85%	0.10
Soybean	66.7%	2.9	87.18%	0.50	90.00%	0.43



**Fig. 6** Accuracy and loss performance on the training dataset for Faster RCNN

architecture with 89.90% mAP, and the prediction time is 0.4 secs. Some qualitative results with localization in images are presented in Fig. 7.

Another dataset we have collected from Kaggle for weed detection, which already has the bounding boxes. On the second dataset, all three methods are applied for weed detection, and as mentioned in the table below, YOLOv4 out-performs among all with 78.85% mAP and 0.10 prediction time. Some qualitative results obtained from all three approaches on the Kaggle dataset are presented in Fig. 8.



Fig. 7 Results for carrot dataset

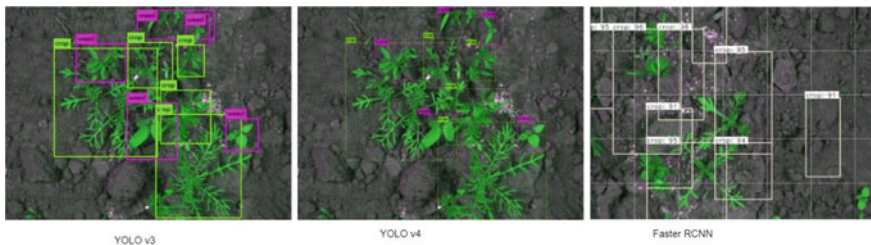


Fig. 8 Localization of kaggle dataset



Fig. 9 Localization of soybean dataset

On the soybean dataset, after labelling the weed area; all three models are trained and the results are presented in the table below. As per the results, YOLO v4 performs much better than Faster RCNN and YOLO v3. Figure 9 presented results obtained from some images from the field of soybean.

## 5 Conclusion and Future Work

Automatic and fast weed identification is necessary to increase crop yield. Three methods for weed localization are applied in this work: YOLO v3, YOLO v4, and Faster RCNN. We found that YOLO v4 performs best among the three approaches



for accurate weed identification. Our work is to implement state-of-the-art methods for weed detection and identify the best-performing approach to improve further the work for deployment on IoT and mobile-based devices. We have used three different datasets from different crops. We have performed labelling manually for two datasets and then applied three methods. As per the results, YOLO v4 outperforms faster RCNN and YOLO v3 models for all the datasets in terms of prediction time and average precision. In the future, we will implement the YOLO v4 model on UAV devices for real-time weed detection. Furthermore, the spray unit will be embedded in the device for herbicide spray with the required amount calculation.

**Acknowledgements** Rajiv Gandhi Proudyogiki Vishwavidyalaya Bhopal provided financial assistance for the research work of: An in-field real-time automatic weed detection using Deep Learning Techniques under the Collaborative Research Scheme (CRS) of TEQIP-III. Details of the CRS project grant are as follows. Grant No:RGPV/TEQIP-III/CRS/2019/17. Funding Agency: National Project Implementation Unit (NPIU), New Delhi Project: Technical Education Quality Improvement Plan-III (TEQIP-III).

## References

1. Economic survey 2017-2018. Chapter 6. Climate, climate change, and agriculture. Government of India, Ministry of Finance, Department of Economic Affairs, Economic Division. <http://www.indiaenvironmentportal.org.in/files/file/economic%20survey%202017-18%20-%20vol.1.pdf>. Accessed 08 August 2020
2. Agricultural mechanisation development in India. Indian J Agric Econ 70(1):64–82. <https://ageconsearch.umn.edu/bitstream/229967/2/08-Gajendra%20Singh%20-%20Keynote-01-n.pdf>. Accessed 14 September 2020
3. Oerke E-C (2006) Crop losses to pests. J Agric Sci 144(1):31–43
4. Joshi NC (2001) Introduction to weed science and classification of weeds. In: Manual of weed control, p 538. <https://www.isws.org.in/Documents/Proceedingsofconference/Conference-2018/FiftyYearsofWeedResearchinIndia.pdf>. Accessed 08 August 2020
5. Ahmed F, Abdullah Al-Mamun H, Hossain Bari ASM, Hossain E, Kwan P (2012) Classification of crops and weeds from digital images: a support vector machine approach. Crop Protect 40:98–104
6. Tang J-L, Chen X-Q, Miao R-H, Wang D (2016) Weed detection using image processing under different illumination for site-specific areas spraying. Comput Electron Agric 122:103–111
7. Perez AJ, Lopez F, Benlloch JV, Christensen S (2000) Colour and shape analysis techniques for weed detection in cereal fields. Comput Electron Agric 25(3):197–212
8. Desai R, Desai K, Desai S, Solanki Z, Patel D, Patel V (2015) Removal of weeds using image processing: a technical review. Int J Adv Comput Technol (IJACT) 4:27–31
9. Weis M (2010) An image analysis and classification system for automatic weed species identification in different crops for precision weed management
10. Choudhary J, Nayak S (2016) A survey on weed detection using image processing in agriculture. Int J Comput Sci Eng 4(6):97–100
11. Karimi Y, Prasher SO, Patel RM, Kim SH (2006) Application of support vector machine technology for weed and nitrogen stress detection in corn. Comput Electron Agric 51(1–2):99–109
12. Bossu J, G'ee C, Jones G, Truchetet F (2009) Wavelet transform to discriminate between crop and weed in perspective agronomic images. Comput Electron Agric 65(1):133–143

13. Juraiza Ishak A, Hussain A, Marzuki Mustafa M (2009) Weed image classification using gabor wavelet and gradient field distribution. *Comput Electron Agri* 66(1):53–61
14. Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. arXiv preprint. [arXiv:1804.02767](https://arxiv.org/abs/1804.02767)
15. Bochkovskiy A, Wang C-Y, Mark Liao H-Y (2020) Yolov4: optimal speed and accuracy of object detection. arXiv preprint. [arXiv:2004.10934](https://arxiv.org/abs/2004.10934)
16. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* 28:91–99
17. Ahmed F, Hossain Bari ASM, Shihavuddin ASM, Abdullah Al-Mamun H, Kwan P (2011) A study on local binary pattern for automated weed classification using template matching and support vector machine. In: 2011 IEEE 12th international symposium on computational intelligence and informatics (CINTI). IEEE, pp 329–334
18. Han Lee S, Seng Chan C, Wilkin P, Remagnino P (2015) Deep- plant: plant identification with convolutional neural networks. In: 2015 IEEE international conference on image processing (ICIP). IEEE, pp 452–456
19. Pearlstein L, Kim M, Seto W (2016) Convolutional neural network application to plant detection, based on synthetic imagery. In: 2016 IEEE applied imagery pattern recognition workshop (AIPR). IEEE, pp 1–4
20. Barrero O, Rojas D, Gonzalez C, Perdomo S (2016) Weed detection in rice fields using aerial images and neural networks. In: 2016 XXI symposium on signal processing, images and artificial vision (STSIVA). IEEE, pp 1–4
21. Umamaheswari S, Arjun R, Meganathan D (2018) Weed detection in farm crops using parallel image processing. In: 2018 conference on information and communication technology (CICT). IEEE, pp 1–4
22. Tiwari O, Goyal V, Kumar P, Vij S (2019) An experimental set up for utilizing convolutional neural network in automated weed detection. In: 2019 4th international conference on internet of things: smart innovation and usages (IoT-SIU). IEEE, pp 1–6
23. Sarvini T, Sneha T, Sukanya Gowthami GS, Sushmitha S, Kumaraswamy R (2019). Performance comparison of weed detection algorithms. In: 2019 international conference on communication and signal processing (ICCSP). IEEE, pp 0843–0847
24. Osorio K, Puerto A, Pedraza C, Jamaica D, Rodríguez L (2020) A deep learning approach for weed detection in lettuce crops using multispectral images. *AgriEngineering* 2(3):471–488
25. Gao J, French AP, Pound MP, He Y, Prid TP, Pieters JG (2020) Deep convolutional neural networks for image- based convolvulus sepium detection in sugar beet fields. *Plant Methods* 16(1):1–12
26. Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448
27. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning. PMLR, pp 2048–2057
28. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788
29. Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271
30. Lameski P, Zdravevski E, Trajkovik V, Kulakov A (2017) Weed detection dataset with rgb images taken under variable light conditions. In: International conference on ICT innovations. Springer, pp 112–119
31. Crop and weed detection dataset. <https://www.kaggle.com/ravirajsinh45/crop-and-weed-detection-data-with-bounding-boxes>
32. Tzutalin. Labelimg. <https://github.com/tzutalin/labelImg>. Accessed 30 May 2019



# Sentiment Analysis of Indians Due to Conflict Between India and China at the Actual Line of Control in Galwan Valley, Ladakh



Ravi Kumar and Deepak Kumar

**Abstract** Sentimental Analysis, the study of human emotion and opinion is one of the fields which keeps on progressing over the years, as a result of which, it can play a very important role in analyzing market trends and decision-making. Living in an era, when people freely express their viewpoints on social media, a lot of information can be gathered by identifying the emotions behind them. Twitter, a platform led by strong public opinion, has the power to affect the market and even the government. Following the tweets of one such powerful hashtag, #BoycottChineseProducts, we performed a sentimental analysis on this trending tweet in this paper as it showcases the response of Indians to China's aggression in Galwan Valley. By identifying particular words and giving them weights, the percentage of positive sentiments of Indian consumers toward Chinese Products is calculated which can be used for further studies as it can have a large impact on the Chinese Market in India.

**Keywords** Opinion mining · Sentimental analysis · Social media

## 1 Introduction

Twitter, Facebook, YouTube, and other microblogging and long-range interpersonal communication locales have not just contributed change to social media but also on a very basic level changed the way we utilize these destinations, what's more, is how we share our sentiments, perspectives, and our voices with the more extensive group of onlookers. Opinion mining (OM), a field that has recently emerged for information retrieval and identification of the lexical, has a wide range of applications. It is not concerned about what the document is conveying, but what the opinion every part of it conveys, like individual tweets in our case. Any opinion ranging from customers' outlook on products, review for a government, relationship with customers, etc. can

---

R. Kumar (✉) · D. Kumar  
Faculty of Mathematics and Computing, Banasthali Vidyapith, Banasthali, India  
e-mail: [godara.ravi@gmail.com](mailto:godara.ravi@gmail.com)

D. Kumar  
e-mail: [deepakkumar@banasthali.in](mailto:deepakkumar@banasthali.in)

be calculated on a scale of negative to positive using OM. Based on certain words, the polarity and subjectivity of any text can be calculated which are discussed further in the paper.

Sentimental analysis, similar to opinion mining, is becoming one of the most desired tools for businesses. With wide and varied data being available on these social media platforms, analysis of public opinion can be used to predict their opinions which can give profitable results in various case scenarios. Sentimental analysis is not only constrained to getting results for a recommendation system based on reviews but can also be extensively used for understanding sentiments toward political, health, sports, travel, brand perspective, stock market, etc.

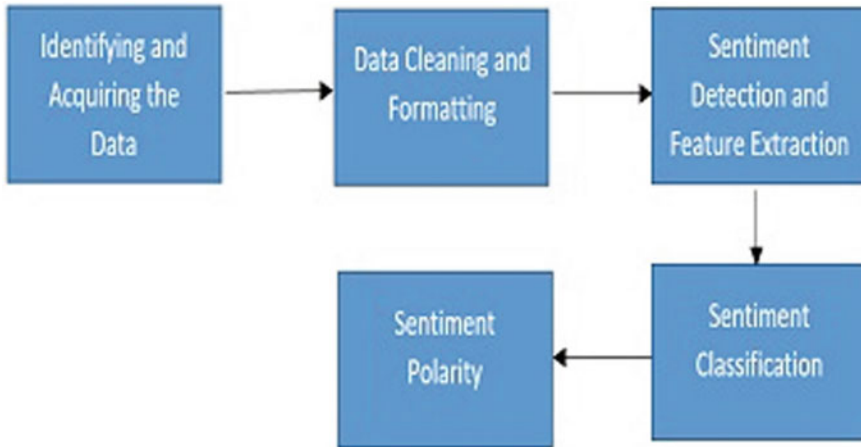
Millions of tweets are shared on Twitter every day where people reflect on the things they do in their everyday lives and scroll through the timeline to see what others have shared as well. Most importantly, Twitter becomes the first online platform to visit to know the important origin of people's assessments and estimations on any current situation. The face-off between India and China in Galwan Valley rose anger among Indians toward China. As support and feeling of patriotism, Indians decided to boycott the use of Chinese products. A further analysis of sentiments of Indians following the trending tweet #BoycottChineseProducts has been done to get the result in the form of a positive index.

## 2 Literature Review

Namrata Godbole [1] performed a large-scale sentimental analysis on news and blogs as they are the platform where people express themselves freely and give an honest opinion. The first step was to determine the semantic orientation of words, followed by sentiment lexicon generation and then finally performance evaluation. The results showed how blogs and news about influencing people and sportspeople had a positive sentiment and that about criminals have a negative one. This sentiment can vary by demographic group, news source, or geographic location. By expanding the spatial analysis of news entities to sentiment maps, we can identify geographical regions of favorable or adverse opinions for given entities. We can also analyze the degree to which our sentiment indices predict future changes in popularity or market behavior.

Harshita Pandey [2] in her paper reviewed various aspects of Sentimental Analysis. Identifying and acquiring the data, data cleansing and formatting, sentiment detection and feature extraction, sentiment classification using a lexicon-based approach or machine learning approach; all machine learning techniques such as decision tree, artificial neural network, random forest, regression, support vector machine, nearest neighbor, naive Bayes classifier, clustering, and sentiment polarity have been extensively reviewed. Sentimental analysis can be applied in customer acquisition, political areas, stock market and stock forecast, financial analysis, and retail chain (Fig. 1).

Ayeena Malik [3] analyzed the tweets of various users regarding a political leader, which showcased the responses or opinions of voters pre- and post-elections. The



**Fig. 1** Sentiment analysis process

dataset was obtained from Twitter during Prime Minister Narendra Modi's visit to the United States of America. The project work started with the extraction of tweets from Twitter, tokenizing them, passing each word from a dictionary AFINN-11, and then identifying the sentiment associated with that word. The results showcased how sentimental analysis can be used as a very powerful tool to identify the gray areas for a political leader and thus help the respective party to ennoble the status of the leader to win the elections. It can also be useful for predicting the outcome of any elections with high accuracy and thus provide help to take the required action accordingly.

Satyanarayana [4] performed sentimental analysis on voice using AWS comprehend and showed collecting audio and finding the sentiment is useful for the improvement of business and analysis. The application made can be helpful in analyzing various emotions of people in different environments such as political campaigns, customer feedback, social media, and many others.

Surya Prabha [5] used naive Bayes classifier for sentiment analysis. The classifier is highly scalable, requiring a small number of parameters. Naive Bayes is a simple classifier technique. There are many other machine learning techniques available for analysis; Naive Bayes is one of the efficient methods which provides a better level of accuracy and good results after classification. The proposed approach could be applied to any kind of review dataset. The work could also be extended to a higher level of the input set (Fig. 2).

Prosanta Kumar Chaki [6] represented nouns as sentimental words, which have a good impact on sentiment detection and words have dual sentiment based on their application, mostly those words are nouns. They introduced a possible solution for these issues and experimented on them. Finally, maximized accuracy of 3% was obtained by applying the offered solution.

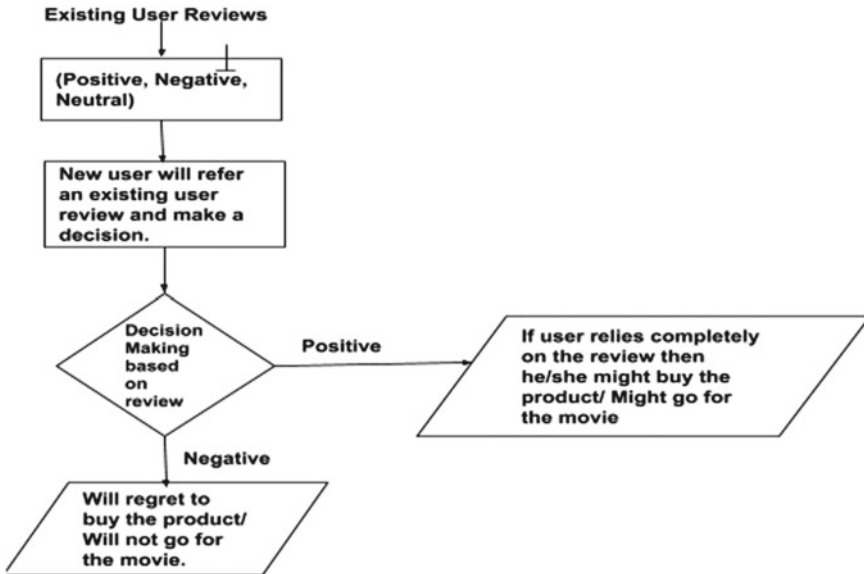


Fig. 2 Diagrammatic overview

### 3 Methodology

#### 3.1 Extracting Data

For getting the tweets, the creation of a Twitter application is required. This helps in fetching certain tokens and access keys that can be used by the user to get the required number of tweets and then based on the needs the complexity can be set according to the desired level. Ayeena Malik [3] The initial step is to extract tweets from a Twitter account, which is accomplished by making an account on Twitter. From the respective account, a Twitter development application is created on the developers' end through an application on Twitter site which gives the user access to certain keys. These keys are used in the code to extract the tweets from the database and save them in a .csv file. The .csv file generated is used as an input for the next modules in the code to generate the required result from the analysis of the sentiments. With the grant of the keys, now the extraction of the tweets starts. The latest 3000 tweets having the hashtag #BoycottChineseProducts were extracted.

### 3.2 *Data Cleaning*

Repetitive tweets, retweets, URLs, mentions, and hashtags were removed from the tweets to make them easier to work with.

### 3.3 *Natural Language Processing (NLP)*

With around 80% of the data being unstructured, it is not possible for humans to manually understand the data. The machines are thus trained to extract meaningful information from natural language text. NLP is a part of computer science and artificial intelligence which is used to deal with human languages.

-Tokenization:

The text is broken down into separate words. This step is very crucial for NLP as by having separate words, it is easier to perform separate functions on them, they are easier to work with and most importantly sentiment can be assigned to each word separately.

-Stemming:

It is the process of normalizing a word into its base form or root form so that the same word used in different tenses can be given the same sentiment. It reduces the ambiguity while assigning the sentiments as all the words are in their base form.

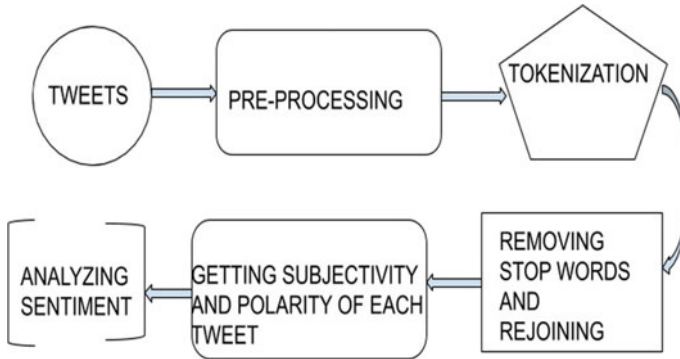
-Removing stop words

After removing the stop words, which again may pose a problem while calculating the accuracy of the sentiments, the words available now may have a chance of having their own sentiment.

### 3.4 *Getting Subjectivity and Polarity of Each Tweet*

Polarity is the analysis of how negative, positive, or neutral a text is. Based on certain words in a string that have been already assigned a sentiment weight, the total polarity score is calculated. It's a float value that lies in the range between  $[-1, 1]$ , where  $-1$  and  $1$  are the extremely negative and positive scores, respectively, and  $0$  is the neutral score.

Subjectivity is the measure of whether a text is based on personal opinions and emotions or whether it is objective that is based on facts. It is also a float value that lies in the range  $[0, 1]$ , where  $0$  defines a text to be completely objective and  $1$  defines a text to completely be subjective (Fig. 3).



**Fig. 3** Work flow diagram

### 3.5 Assigning Positive, Negative, and Neutral Sentiments

Finally, all the tweets with a polarity score  $< 0$  are assigned a negative sentiment, score  $> 0$  are assigned a positive sentiment and score  $= 0$  are assigned a neutral sentiment.

## 4 Experimental Results

A lot of tweets contained Hindi words that were written in English and possessed problems that led to them being identified as Neutral. Neutral, Positive, and Negative tweets have 60.9%, 31.5%, and 7.6% percentages respectively (Fig. 4).

The removal of neutral tweets shows that 80.6% of tweets are positive and 19.6% of tweets are negative (Fig. 5).

This shows that a wide majority of Indians are in support of the government in boycotting the Chinese products, which can positively affect the market of Made in India products.

### Applications and Future Scope of Opinion Mining:

- Business and internet business applications, for example, item surveys and motion picture evaluations [7];
- Predicting the stock costs in view of sentiments that individuals have about the organizations and assets [8];
- Determine ranges of an item that should be enhanced by outlining item audits to see what parts of the item are for the most part thought to be great or terrible by clients [9];
- Customer inclination and customer relationship management;
- Market analysis and decision-making;
- Understanding any political front.

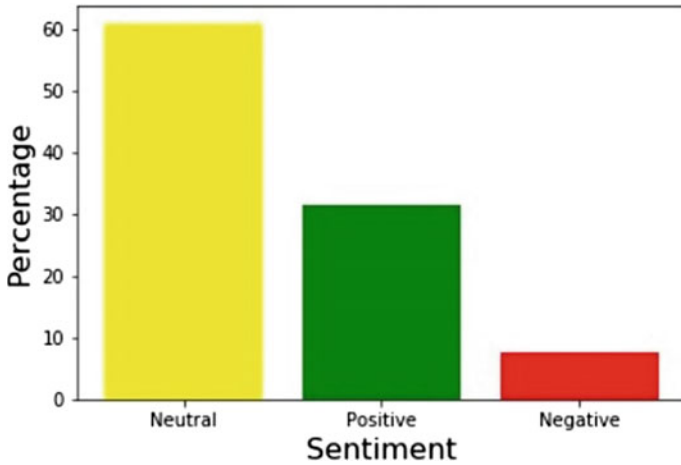
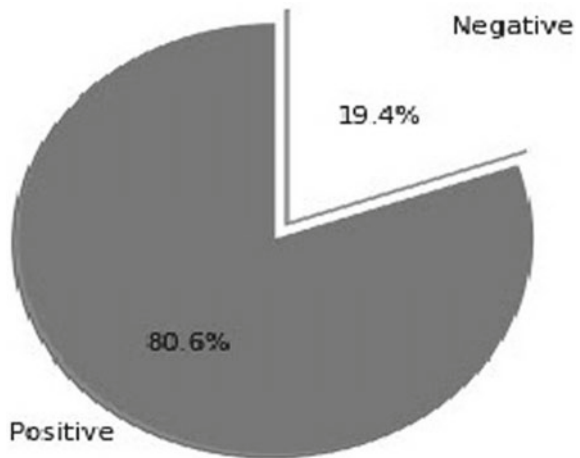


Fig. 4 Sentimental Analysis (Bar-Plot)

Fig. 5 Pie chart between positive and negative tweets



**Challenges:**

The investigation of information stream mining has brought forth a couple of open issues that request consideration. Here is a short survey of them:

- Smart information preprocessing module in the calculation can guarantee a high caliber of final products.
- Due to the utilization of restricted assets for taking care of a huge measure of information, one must guarantee that the information structures are proficient in handling operations on the circle. I/O and ordering systems are likewise basic viewpoints on the preparing time.

- The system ought to be knowledgeable to separate among clamor and idea change in the live stream.
- A positive or negative suspicion word may have a backward significance in a particular territory so it is hard to predict by its watchword meaning.

Interrogative sentences won't have positive or negative thoughts, but instead the catchphrase used as a part of the estimation may make positive or negative [10].

## 5 Conclusion

Sentimental analysis is being widely used nowadays by various organizations belonging to different fields to make the most profit as sentiments of the consumers greatly affect the market and the decision-making. Twitter, a micro-blogging site is one of the most desired social media platforms by the users and everyone feels free to give their opinion. By using proper tools and doing opinion mining, finding the polarity and subjectivity of a tweet, the positive index of the topic the user is talking about can be calculated. Twitter has the power to influence people which can serve both as negative and positive. If the sentiments are identified properly, it can give rise to better market analysis and decision-making. The 80.6% positive index for #BoycottChineseProducts shows that Made in India products will flourish and there are a lot of new-found opportunities in the Indian Market.

## References

1. Namrata Godbole MS (2007) Large-scale sentiment analysis for news and blogs. In: ICWSM'2007, pp 1–4. Boulder, Colorado, USA
2. Harshita Pandey DA (2019) Various aspects of sentiment analysis: a review. In: 2nd international conference on advanced computing and software engineering (ICACSE-2019)
3. Ayeena Malik DK (2016) Sentiment analysis on political tweets. In: Vth international symposium on "Fusion of Science & Technology", New Delhi, India, Jan 18–22, pp 359–361. ISBN
4. Satyanarayana G, DB (2020) Sentimental analysis on voice using AWS comprehend. In: 2020 international conference on computer communication and informatics (ICCCI -2020), Jan 22–24, pp 1–5. IEEE, Coimbatore
5. Surya Prabha SB (2019) Sentimental analysis using Naive Bayes. In: 2019 international conference on vision towards emerging trends in communication and networking (ViTECoN), pp 1–5. IEEE
6. Prosanta Kumar Chaki IH (2017) An aspect of sentiment analysis: sentimental noun with dual sentimental words analysis. In: International conference on current trends in computer, electrical, electronics and communication (ICCT CEEC-2017), pp 1242–1246. IEEE, Dhaka
7. Batool RK (2013) Precise tweet classification and sentiment analysis. In: International conference on computer and information science (ICIS), pp 461–466. IEEE
8. Jérôme Treboux FC (2016) Mining and visualizing social data to inform marketing decisions. In: The 30th IEEE international conference on advanced information networking and applications (AINA-2016). IEEE, CransMontana, Switzerland



9. Castillo C, MM (2011) Information credibility on twitter. In: Proceedings of 20th international conference world wide web, pp 675–684
10. PhridviRaj MSB, C (2014) Clustering text data streams—A tree based approach with ternary function and ternary feature vector. ITQM, pp 976–984

# Feature Selection Using Information Gain for Software Effort Prediction Using Neural Network Model



Sushma Khatri and Pratosh Bansal

**Abstract** An important phase in software development is the prediction of effort. It has its significance in project planning, control, and budgeting. Many researchers developed common software effort estimation models known as algorithmic models. These models need accurately estimated input parameters, namely lines of code and complexity. Accurate estimation of these features during the initial phases of the software life cycle is quite difficult. This issue of algorithmic models can be handled with non-algorithmic models. These non-algorithmic models are based on soft computing techniques such as Genetic Programming, Fuzzy Sets, and Artificial Neural Network (ANN). Many researchers proposed various models based on ANN but we did not find any estimation method focused on feature selection to remove the negative impact of irrelevant information. In this study, features with high information gain are selected using information gain to train the multilayer perceptron network FITNET. Experiment with two- and threefold cross-validation on 3 benchmark datasets shows that New ANN (NANN) trained on selected features makes effective prediction as compared with the ANN trained on all features. Our approach compared the performance using 5 performance metrics MAR, MMRE, MdmRE, PRED (25), and MSE to show that it will perform better for different metrics.

**Keywords** Artificial neural network · Feature selection · Information gain · Software effort estimation

---

S. Khatri (✉) · P. Bansal  
Devi Ahilya Vishwavidyalaya, Institute of Engineering & Technology, Khandwa Road, Indore,  
Madhya Pradesh 452017, India  
e-mail: [skhatri10@gmail.com](mailto:skhatri10@gmail.com)

P. Bansal  
e-mail: [pratosh@hotmail.com](mailto:pratosh@hotmail.com)

S. Khatri  
Acropolis Institute of Technology & Research, Bypass Rd, Behind Malwa County, Manglaya  
Sadak, Indore, Madhya Pradesh 453771, India

## 1 Introduction

One of the necessary steps of the software development process is the precise prediction of effort. Both underestimation and overestimation are undesirable for developers as well as customers. There exist several significant models and techniques in the literature for estimation. There are different ways in which they can be classified such as [1, 2].

Classification of software effort prediction methods based on Approach:

- Model-based techniques:
  - COCOMO
  - SEER-SEM
- Techniques based on expertise:
  - Delphi technique
- Techniques involving learning:
  - Artificial Neural Networks
- Dynamic-based models:
  - Dynamic based
- Regression-based techniques:
  - Robust regression
- Composite approaches

Boehm [3] has initially proposed the effort and schedule assessment model called COCOMO in 1980, which is the most famous of that time. Later versions of this model, i.e., COCOMO' 81 face problems in the estimation of software effort. The University of Southern California started their research in 1994 for COCOMO II [4], which deals with various issues, namely rapid application development, reusability, object-oriented programming, and reengineering.

Classification of software cost estimation model based on Project Type:

- Traditional Estimation Techniques [5] for Waterfall Methodology Projects:
  - Models based on Analogy
  - Effort prediction method
  - Method based on programming techniques (lines of code)
  - Expert Judgment effort estimation
  - Delphi/Wideband Delphi
  - Program Evaluation and Review Technique (Beta or normal)
  - COCOMO I/COCOMO II
  - Function Points method proposed by Albrecht

- Parametric methods (PRICE)
- Several bottom-up and top-down methods.
- Agile Estimation approaches for Agile Projects:
  - Relative sizing method
  - Wideband Delphi method
  - Planning poker method
  - Affinity estimation method
  - Ideal time and elapsed time methods
  - Disaggregation method
  - Bottom-up and top-down estimation methods.

Most of the researchers argued an urgent need for the selection of significant attributes to raise the accuracy of effort estimation and also to remove immaterial and redundant information [6–8]. Hence, in order to achieve comparable, or occasionally improved accuracy, methods for selecting less, but highly relevant features are usually applied to identify features that have a high impact on dependent variable effort and provide a brief and coherent model [9, 10]. Different researchers applied different attribute selection methods, such as Desai and Mohanty [11] considering the count for input, output, and inquiry along with the master file and interface file as relevant five attributes for predicting the effort. Fernández-Diego et al. [12] identified a potential list of factors influencing effort within the ISBSG dataset [13–18]. Rao et al. [7] and Mendes et al. [19] used Principal Component Analysis (PCA) for feature selection. Rao et al. [7] and Ahmad et al. [20] also used Correlation Based Feature Selection (CFS) to rank the attributes. Dejaeger et al. [21] performed computationally expensive approaches to select features based on a generic backward input selection wrapper. Padhy et al. [22] applied the T-test analysis-based feature selection approach, which uses a correlation coefficient to calculate the heterogeneity and homogeneity of the data elements. According to Liu et al. [23], attribute selection is an NP-hard problem so they applied a greedy method for selecting the attributes, called LNI-based Feature Selection (LFS) in which an attribute is selected for which maximum LNI value between the currently selected attributes and the cost increases.

We used information gain [24] for selecting the attribute proposed by Shannon [25] for information theory. The attributes with the higher information gain are selected to train the ANN to predict the effort. And we found that in most of the cases, performance is comparable with the Artificial Neural Network trained on all attributes for predicting the effort.

Rest of the proposed work is systematized as discussed: First, Sect. 2 extends literature review on effort prediction using ANN and attribute selection methods. In this section, we also discussed the problem to be solved. Then, in Sect. 3 the architecture, configuration, and details about the ANN employed for effort estimation in the study are discussed. Section 4 describes the experimental design which reflects datasets, overall procedure, and evaluation metrics. In Sect. 5, we predict the effort using the proposed model with selected attributes and compared it with the effort

predicted using all attributes. Section 6 concluded the paper by providing the general findings and issues that can be considered for future research.

## 2 Related Work

Mendes et al. [26] propose the classification of effort prediction techniques into 3 categories: Expert judgment, algorithmic and non-algorithmic methods. Expert judgment methods [27] such as Planning Poker, Delphi, and WBS are based on the experiences of experts on similar kind of projects but they are subjective [28]. Algorithmic models (also known as parametric models) [29] such as COCOMO [3], SLIM [30], and SEER-SEM [31] need certain parameters such as lines of code, function point, and complexity as input for effort estimation. These parameters are imprecise and uncertain; it is tough to provide a correct approximation of these parameters during the early phase of the software. These limitations of the existing models lead to the invention of soft computing-based techniques, which are mainly known as non-algorithmic such as ANN, fuzzy sets, and genetic programming for more accurate software effort estimation development [32].

### 2.1 ANN for Software Effort Estimation

Artificial Neural networks (ANN) are well known for giving accurate predictions over imprecise and uncertain input patterns. ANNs are also good at finding nonlinear relationships and patterns in data because of their generalization capabilities. Also, we can train ANN to learn from past project experiences and they can handle different numbers and types of input [29]. Hence, it can produce better effort estimates as compared to expert judgment techniques and parametric techniques [33, 34]. Due to this, ANNs are mostly used for predicting the effort of software development since long ago. Feed-forward neural network (FFNN) using backpropagation algorithm is used for software effort estimation [35–37]. Finnie et al. [38] concluded that ANN can handle the noise in the data set. They also compared two AI-based estimation models with regression models and concluded that AI-based estimation models provide better results for effort estimation. Idri et al. [39] employed radial basis function neural networks. Kumar et al. [40] employed neural network based on a wavelet. Roh et al. [41] employed fuzzy radial basis function based polynomial ANN. Rao et al. [42] first time used the Functional Link ANN (FLANN) for estimating the effort. Bardsiri et al. [43] suggested a hybrid approach of analogy, clustering based on fuzzy logic, and ANN to enhance the accuracy of cost prediction. Clustering is employed to nullify the effect of unrelated and unpredictable projects on effort prediction. Kaushik et al. [44] used a validation technique known as leave one out in place of 3-way and showed that intuitionistic fuzzy c-means (IFCM)-Functional Link ANN

provide better performance for software effort prediction as compared to the FCM-FLANN. Dave and Dutta [45] provide a review of 21 articles regarding the cost prediction of software through ANN. Nassif et al. [29] compared the 4 ANN models: multilayer perceptron (MLP), cascade correlation NN (CCNN), general regression NN (GRNN), and radial basis function NN (RBFNN) for predicting the cost of the software. Wen et al. [46] presented an analysis of machine learning approaches for predicting the cost and they showed that models based on ANN are ranked second amongst the selected publications.

## ***2.2 Feature Selection***

There are two methods for attribute selection known as filters and wrappers [47]. Wrapper approaches are expensive in terms of computation as they apply searching policy with estimation models repetitively during each iteration to search better feature subset. As compared to wrappers, filters are computationally cheap because they use the inherent features of the data to find the subsets of attributes. Commonly correlation measures are used in filter methods for evaluating the feature subsets. Two different approaches are used for calculating the correlation that exists between attributes and cost variables [48]. First approach uses classical linear correlation which is disapproved because of its linearity and Keung et al. [49] also declared that there will be a risk if we use the linear correlation between attributes for Software Effort Estimation datasets. Information theory is the basis for the second approach. Measures based on mutual information have been presented by many authors [50–55].

## ***2.3 Proposed Problem***

Many researchers successfully applied ANN models for predicting the software effort but we did not find any of them which have exploited ANN with selected features using information gained for software effort estimation in our literature survey. Also, the performance metrics used for measuring the performance of the models need to be considered. Performance of an effort estimation model may be better on one metric, e.g., Mean Absolute Residual (MAR) but can be worst depending on another metric, namely Mean Magnitude of Relative Error (MMRE).

For addressing the above issues, we selected a subset of features having higher information gain. Then we compared the performance of the ANN trained on all features with the New ANN (NANN) which is trained on the selected subset of features. We performed experiments with two- and threefold cross-validation on three benchmark datasets and compared the accuracy using five performance metrics MAR, MMRE, MdmRE, PRED (25), and MSE.

### 3 Artificial Neural Network (ANN)

An ANN is a network of artificial neurons for processing the information to simulate the human brain. The human brain has four main components: Cell becomes neuron or node or unit. Dendrites become weight or interconnection, Soma becomes net input and axon becomes the output of ANN. The human brain contains about 1011 neurons and 1015 interconnections between those neurons [56]. It is not possible to imitate the brain to such an extent, but the application under consideration and designers of the network decide the size, complications involved, and structure of the ANN. Following are the characteristics of the ANN [57].

- Mathematically implemented model.
- Network of highly interconnected processing nodes.
- The value of the interconnections holds the information.
- Neurons receive input signals through the connections and connecting weights.
- Through weight adjustment, we can make the neurons learn, recollect and theorize over the dataset under consideration.

Various models of ANN are classified on the basis of three factors: network architecture or interconnections (the way in which the neurons are interconnected with each other to structure them in the form of layers), the rules used for training or learning for updating the values over the interconnections and functions that are used for activation of neurons. The first factor used for classifying the network is network structure. There are mainly five types of ANN models based on network architecture: feed-forward networks having one layer, feed-forward network having multiple layers, feedback to the node itself consisting of only one node, recurrent network having one layer, recurrent networks having multiple layers.

Input layer just receives the input and does not perform any function except buffering the input. Output layer produces the output of the network. The hidden layer has no direct interaction with the environment and it is internal to the network and it is between the two layers, namely input and output. Hidden layers can vary from zero to several in an ANN depending upon the application. More number of hidden layers will be a complexity to the ANN.

A network having only one input and output layer is called a single-layer network. If one or more hidden layers exist along with the input and output layer, it forms a multilayer network. If the output of every neuron from the preceding layer is interconnected with each and every neuron of the next layer, then the network is termed as fully connected network. If no signal from the output layer neuron is an input to the neurons of the preceding layer or to the neurons of the same layer then the network is called the feed-forward network. On the other hand, back propagation or feedback networks are one in which the outputs of a layer's neurons are set back as an input to the neurons of the same layer or preceding layer. Feedback networks with a closed loop are called recurrent networks [57].

The second important factor for classifying the neural networks is learning or training, through which NN makes appropriate adjustments in parameters and adapts

itself to excitement for producing the desired response. ANN mainly performs two types of learning simultaneously or separately: parameter learning and structure learning. Parameter learning refers to weight adjustment and structure learning means changing the connection types and number of neurons. In addition, learning of ANN can also be classified as supervised, unsupervised, and reinforcement learning. In case of supervised learning, complete training pair (inputs in combination with the corresponding targets called as desired output) is presented to the network during training. Unsupervised learning is also called self-learning in which only the input vector is presented to the NN for training. Network organizes these patterns into clusters. One kind of supervised learning is reinforcement learning, where NN accepts critical information through the atmosphere as feedback along with the input vector.

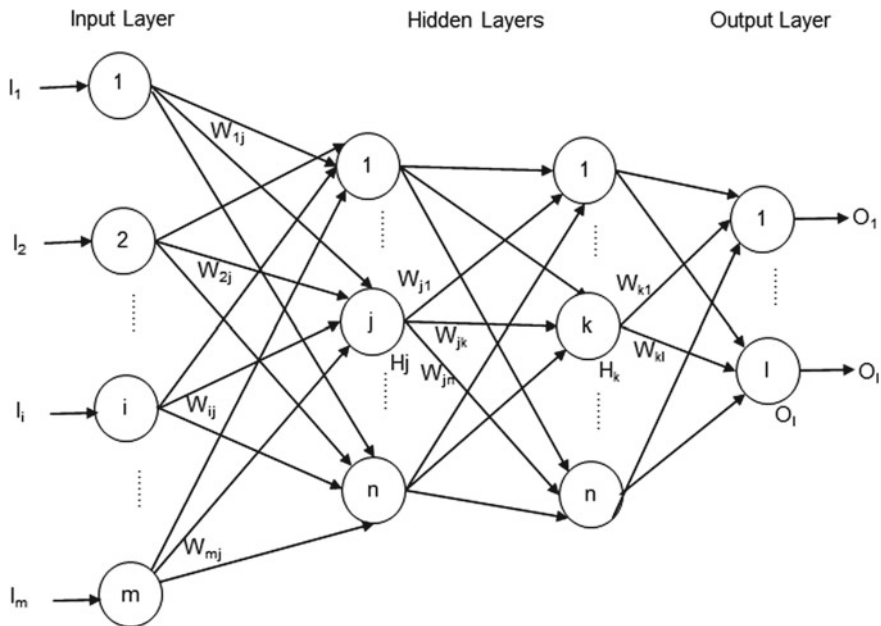
The third factor used for classifying the network is the activation function or transfer function. The output of the ANN is calculated by applying an activation function over the net input. Some of the common activation functions are: ramp, binary step, identity, sigmoidal, hyperbolic tangent, bipolar step functions. Identity function is called a linear function because its output is equal to its input. Generally, this function is used in the input layer. Binary and bipolar step functions are used in the single-layer networks. Multilayer networks generally use a nonlinear activation function. Sigmoidal and hyperbolic tangent functions are generally used in back propagation networks.

### ***3.1 FITNET: Feed-Forward Neural Network with Multiple Layers***

Our work uses FITNET [58], which is a feed-forward neural network with multiple layers. MATLAB includes this function which is trained using supervised learning. Figure 1 is showing a feed-forward NN having multiple layers in the form of input, output, and hidden layers (can be 1 or many). Most of the researchers used a multilayer feed-forward neural network with one hidden layer for effort estimation [59, 60]. The number of neurons in the input layer depends upon the number of input variables in the ANN model. In case of effort estimation, the number of independent variables or effort drivers of the given data set decides the number of neurons in the input layer. The hidden layer receives the input from the neurons of the input layer. In our study, we created the network with only one hidden layer, and the number of neurons of the hidden layer varied from 1 to 5. The output of the model decides the number of nodes in the output layer. In our case, the target is to predict the effort, which means the number of dependent variables is equal to one hence there is only one neuron at the output layer.

In our case, hidden layer is trained using the tan-sigmoid transfer function and input and output layers are trained using the linear transfer function. For parameter learning (i.e., network weights adjustment), we trained the network using the Levenberg–Marquardt algorithm (trainlm) [61]. In this study, the number of folds used for





**Fig. 1** Multilayer feed-forward neural network

cross-validation is 2 and 3 for dividing the datasets into training and test set. We used every sample only one time for testing and an equal number of times for training. Hence, it is better to compare the random subsampling approach and the holdout approach. In the holdout method and random subsampling, the prediction is not as accurate as desired because the model is trained only on a specific part of the data [43, 62]. Finally, we tested the network using 5 performance metrics: MAR, MMRE, MdMRE, PRED (25), and MSE.

## 4 Experiment Design

Experiment is designed to present that ANN trained on selected highly relevant attributes perform better as compared to ANN trained on all attributes for effort estimation. The following subsection contains the details about datasets, overall procedure, and evaluation metrics used.

**Table 1** Dataset Statistics

Name	No of attributes	No of observations	Unit of effort	Minimum effort	Maximum effort	Previously used in other studies
Kemerer	8	15	Months	23.2	1107.31	[23, 64]
Coc81	19	63	Months	5.9	11,400	[21, 44, 64, 65]
Maxwell	27	62	Hours	583	63,694	[21, 43, 44, 64]

### 4.1 Datasets

Three benchmark datasets Kemerer, Coc81, and Maxwell are collected from the PROMISE repository [63] for our study. Preprocessing is performed on the datasets, which are discussed in detail in Sect. 4.2. Table 1 contains the information about the three datasets.

Data preprocessing consists of two substeps: handling of missing values and preparing data for ANN. Missing values are handled by deleting that particular record. We used the FITNET [58] function available in MATLAB for designing the ANN model. This function requires the training pair in column \* row format. Also, the train function requires the input (independent attributes) and targets (dependent attribute i.e., effort) in a separate file. Hence, we need to prepare the data for ANN.

### 4.2 Overall Procedure

The overall procedure is implemented in MATLAB. The complete implementation of the work proposed is shown in Fig. 2. Primary step is data preprocessing which includes delete missing values and prepare data for ANN, discussed in Sect. 4.1. The second step is feature selection in which information gain of all the independent variables (cost drivers) is calculated and then features with higher information gain are selected. It is very vital to apply the feature selection method to select the highly relevant attributes for effort estimation. It will improve the estimation accuracy and nullify the effect of less relevant attributes. For this purpose, we used the measure known as information gain which was proposed in a decision tree algorithm known as ID3 as an attribute selection measure [62]. Attributes with the higher information gain were selected for training as well as testing the NN for effort prediction.

Information gain value for all the attributes of the 3 benchmark datasets used in our study is represented in Tables 2, 3, and 4. The attributes represented in italic are the selected attributes. For this purpose, we had chosen a threshold value and all the attributes having information gain above that threshold were selected. The threshold values chosen are 3.0, 2.4, and 4.0, respectively, for the datasets Kemerer, Coc81, and Maxwell. Accordingly, the selected attributes for the dataset Kemerer, Coc81, and Maxwell are listed in Table 5. Information gain measure is biased toward the test with many results. Attributes with a large number of different values are preferred

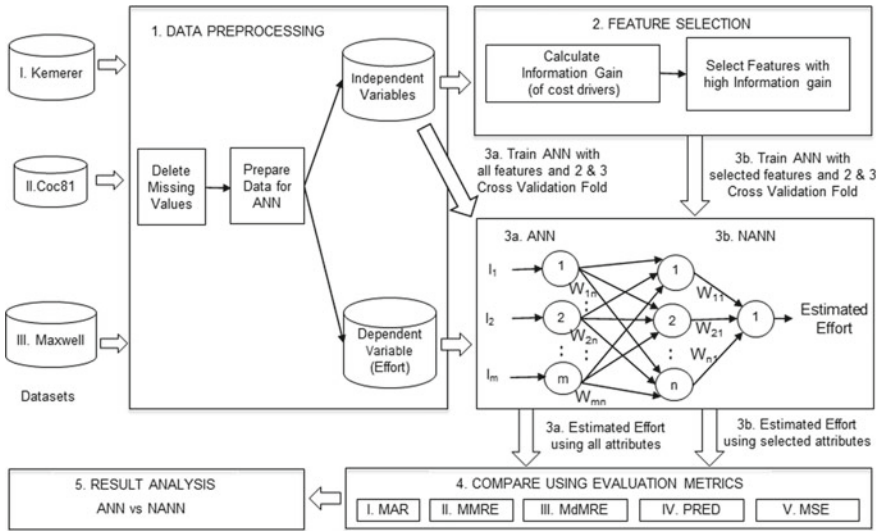


Fig. 2 Overall procedure

Table 2 Information gain for Kemerer dataset

S. No	Attribute name	Information gain
1	ID	3.91 <sup>a</sup>
2	Language	0.70
3	Hardware	2.15
4	<i>Duration</i>	3.32
5	<i>KSLOC</i>	3.91
6	<i>AdjFP</i>	3.91
7	<i>RAWFP</i>	3.91

The attributes represented in italic are the selected attributes, having info. gain  $\geq 3.0$  (threshold value chosen) for the Kemerer dataset. 3.91<sup>a</sup> not selected to remove the biasness of the information gain measure

by this measure and their information gain will also be maximum. For example, an attribute Project\_ID, which uniquely identifies a record in the dataset will have higher information gain because it will have as many outcomes as there are records. Such attributes have no effect on the dependent variable—Effort [62]. In the Kemerer dataset, attribute ID has an information gain value of 3.91 and in the Coc81 dataset, the attribute project\_id has the information gain value of 5.88. In both cases, the value is higher than the selected threshold then too we had not selected such attributes in our study to remove the biases of the information gain measure.

In the third step, network is trained and tested on all attributes, which is denoted by ANN (Artificial Neural Network). The NN is also trained and tested

**Table 3** Information gain for Coc81 dataset

S No	Attribute name	Information Gain
1	project_id	5.88 <sup>b</sup>
2	dev_mode	1.49
3	rely	2.22
4	data	2.00
5	<i>cplx</i>	2.40
6	<i>time</i>	2.47
7	stor	2.02
8	virt	1.70
9	turn	1.80
10	acap	1.97
11	aexp	1.80
12	pcap	2.08
13	vexp	1.74
14	lexp	1.67
15	modp	2.28
16	tool	1.88
17	sced	1.73
18	<i>loc</i>	5.61

The attributes represented in italic are the selected attributes, having info. gain  $\geq 2.4$  (threshold value chosen) for the Coc81 dataset. 5.88<sup>b</sup> not selected to remove the biasness of the information gain measure

on selected attributes, which is denoted by NANN (New Artificial Neural Network) as demonstrated in step 3 of Fig. 2. We used the multilayer feed-forward neural network FITNET. The overall architecture and various parameters of the network are discussed in Sect. 3.1. Step four calculates the performance metrics: MAR, MMRE, MdMRE, PRED (25), and MSE for both ANN and NANN. Finally, a comparative analysis of ANN and NANN is presented.

### 4.3 Evaluation Metrics

To compare the performance of ANN and NANN for effort estimation, we use the five metrics: Mean Absolute Residual (MAR), Mean Magnitude of Relative Error (MMRE), Mean Squared Error (MSE), Percentage of Prediction (PRED (25)), and Median Magnitude of Relative Error (MdMRE). Many performance measures were found in the research [64, 66, 67], which are presented in Table 6.

**Table 4** Information gain for Maxwell dataset

S. No	Attribute name	Information gain
1	Syear	2.95
2	App	1.81
3	Har	1.50
4	DbA	0.44
5	Ifc	0.35
6	Source	0.55
7	Telanuse	0.80
8	Nla	1.96
9	T01	1.92
10	T02	1.53
11	T03	1.78
12	T04	1.51
13	T05	1.41
14	T06	1.48
15	T07	1.87
16	T08	1.86
17	T09	1.55
10	T10	1.33
19	T11	1.93
20	T12	1.43
21	T13	1.92
22	T14	2.01
23	T15	1.54
24	<i>Duration</i>	<i>4.57</i>
25	<i>Size</i>	<i>5.89</i>
26	Time	2.95

The attributes represented in italic are the selected attributes, having info. gain  $\geq 4.0$  (threshold value chosen) for the Maxwell dataset

**Table 5** Selected attributes for datasets

Datasets	Name of selected attributes	S. No. of selected attributes
Kemerer	Duration, KSLQC, AdjFP and RAWFP	4,5,6,7
Coc81	cplx, time and loc	5,6,18
Maxwell	Duration and size	24,25

**Table 6** Performance metrics

S No	Measure	Full form	Definition	Formula	Previously used in other studies
1	AR	Absolute residual	Difference between the predicted and the actual values	$ARi =  xi - \hat{xi} $	-
2	MAR	Mean Absolute residual	Mean of individual AR values	$MAR = \frac{\sum_{i=1}^n  xi - \hat{xi} }{n}$	[23, 29, 64]
3	MRE [68, 69]	Magnitude of relative error	Measures the error ratio between the actual effort and the predicted effort	$MREi = \frac{ xi - \hat{xi} }{xi}$ $= \frac{ARI}{xi}$	-
4	MER [69]	Magnitude of Error Relative to the estimate	A related measure	$MERi = \frac{ xi - \hat{xi} }{\hat{xi}}$ $= \frac{ARI}{\hat{xi}}$	-
5	MMRE	Mean Magnitude of Relative Error	A summary of MRE	$MMRE = \frac{\sum_{i=1}^n MREi}{n}$	[43, 65], [43, 44, 64, 70]
6	MdMRE	Median Magnitude of Relative Error	A summary of MRE	$MdMRE = median(MRE1, MRE2, \dots, MREn)$	[21, 44, 64]
7	PRED (25) [68]	Percentage of Prediction	Percentage of predictions falling within 25 percent of the actual values	$PRED(25) = \frac{100}{n} \sum_{i=1}^n \begin{cases} 1, & \text{if } MREi \leq \frac{25}{100} \\ 0, & \text{otherwise.} \end{cases}$	[21, 43, 44, 64, 65, 70]
8	MBRE [69]	Mean Balanced Relative Error	Other error measures	$MBREi = \frac{ xi - \hat{xi} }{\min(\hat{xi}, xi)}$	[64, 71]

(continued)

Table 6 (continued)

S No	Measure	Full form	Definition	Formula	Previously used in other studies
9	MIBRE [69]	Mean Inverted Balanced Relative Error	Other error measures	$MIBREi = \frac{ \hat{x}_i - x_i }{\max(\hat{x}_i, x_i)}$	[64, 71]
10	MSE [62]	Mean Squared Error	Other error measures	$MSE = \frac{\sum_{i=1}^n (x_i - \hat{x}_i)^2}{n}$	[72]

$x_i$  = actual effort;  $\hat{x}_i$  = predicted effort;  $i$  = test instance;  $n$  = total number of instances

## 5 Results and Discussion

Our objective is to check that New Artificial Neural Network (NANN) trained using selected attributes having higher information gain performs better as compared to the Artificial Neural Network (ANN) trained on all attributes. For this purpose, we evaluate the performance on 3 benchmark datasets Kemerer, Coc81, and Maxwell using 5 performance metrics MAR, MMRE, MdmRE, PRED (25), and MSE as shown in Tables 7, 8, 9, 10 and 11.

Smaller values of MAR, MMRE, MdmRE, MSE, and PRED’s larger values indicate improvement in effort prediction. Bold entries in the table signify the improved result and italic values show that measures are equal. We test the networks for 1 to 6 hidden neurons (HN) using two- and threefold cross-validation and found that in most of the cases performance is improved in NANN trained on selected attributes as compared to the ANN trained on all attributes. Hence, we find that feature selection using the information gain method is effective for effort prediction using neural networks.

**Table 7** MAR comparison for ANN and NANN for 3 benchmark datasets

Dataset		Kemerer		Coc81		Maxwell	
FOLD	HN	ANN	NANN	ANN	NANN	ANN	NANN
		(All Attribute)	(4,5,6,7)	(All Attribute)	(5,6,18)	(All Attribute)	(24,25)
2	1	276.61	<b>166.71</b>	2253.83	<b>606.51</b>	6856.39	<b>4858.46</b>
	2	203.17	<b>180.39</b>	768.31	<b>757.17</b>	6384.03	<b>6318.64</b>
	3	211.60	<b>202.55</b>	<b>1663.08</b>	7358.04	7456.83	<b>3881.68</b>
	4	355.25	<b>199.49</b>	896.78	<b>646.67</b>	9234.85	<b>4974.95</b>
	5	<b>138.15</b>	228.30	1178.69	<b>618.36</b>	7193.85	<b>5261.56</b>
	6	<b>60.10</b>	229.94	1631.00	<b>392.60</b>	7041.64	<b>2781.52</b>
3	1	<b>92.45</b>	320.66	975.86	<b>739.50</b>	6294.88	<b>3212.83</b>
	2	562.22	<b>128.73</b>	1642.78	<b>1118.59</b>	<b>6930.06</b>	7604.75
	3	<b>217.60</b>	282.26	695.14	<b>505.30</b>	3583.21	<b>2797.17</b>
	4	219.77	<b>187.17</b>	525.43	<b>351.45</b>	4147.94	<b>3887.50</b>
	5	<b>361.15</b>	1092.96	1388.15	<b>345.42</b>	8677.69	<b>3794.80</b>
	6	222.68	<b>219.55</b>	1474.96	<b>320.28</b>	7328.73	<b>4711.87</b>

Lower values of MAR are represented in bold, which shows better prediction



**Table 8** MMRE comparison for ANN and NANN for 3 benchmark datasets

Data Set		Kemerer		Coc81		Maxwell	
FOLD	HN	ANN (All Attribute)	NANN (4,5,6,7)	ANN (All Attribute)	NANN (5,6,18)	ANN (All Attribute)	NANN (24,25)
2	1	3.06	<b>0.86</b>	41.85	<b>6.68</b>	2.14	<b>0.81</b>
	2	1.03	<b>0.46</b>	3.35	<b>1.82</b>	3.10	<b>2.76</b>
	3	1.38	<b>0.86</b>	<b>29.66</b>	133.99	1.28	<b>1.05</b>
	4	3.53	<b>1.22</b>	8.40	<b>2.67</b>	3.19	<b>1.25</b>
	5	<b>0.84</b>	1.97	14.35	<b>2.71</b>	1.13	<b>0.67</b>
	6	<b>0.66</b>	2.72	38.81	<b>2.27</b>	2.81	<b>0.44</b>
3	1	<b>0.47</b>	1.31	17.86	<b>7.33</b>	1.50	<b>0.62</b>
	2	4.89	<b>1.11</b>	<b>9.86</b>	10.81	0.68	<b>0.53</b>
	3	1.86	<b>1.21</b>	4.32	<b>1.71</b>	<b>0.52</b>	0.55
	4	<b>0.39</b>	0.40	4.25	<b>3.03</b>	1.71	<b>0.84</b>
	5	<b>3.66</b>	9.92	31.37	<b>2.50</b>	3.15	<b>1.00</b>
	6	<b>0.70</b>	0.78	24.11	<b>3.66</b>	2.09	<b>0.63</b>

Lower values of MMRE are represented in bold, which shows better prediction

**Table 9** MdMRE comparison for ANN and NANN for 3 Benchmark datasets

Data Set		Kemerer		Coc81		Maxwell	
FOLD	HN	ANN (All Attribute)	NANN (4,5,6,7)	ANN (All Attribute)	NANN (5,6,18)	ANN (All Attribute)	NANN (24,25)
2	1	2.48	<b>0.67</b>	22.95	<b>2.74</b>	<b>0.53</b>	0.55
	2	0.51	<b>0.38</b>	<b>0.88</b>	0.89	<b>0.61</b>	1.37
	3	0.90	<b>0.67</b>	<b>7.47</b>	92.10	0.73	<b>0.47</b>
	4	1.57	<b>1.06</b>	2.49	<b>0.74</b>	1.62	<b>0.92</b>
	5	<b>0.48</b>	0.78	7.37	<b>0.91</b>	0.73	<b>0.49</b>
	6	<b>0.33</b>	1.81	18.53	<b>0.86</b>	1.00	<b>0.36</b>
3	1	<b>0.38</b>	0.60	<b>1.60</b>	1.71	0.61	<b>0.46</b>
	2	2.02	<b>0.36</b>	<b>0.91</b>	2.20	<b>0.54</b>	0.58
	3	1.75	<b>0.49</b>	0.82	<b>0.76</b>	0.38	<b>0.34</b>
	4	0.36	<b>0.34</b>	1.20	<b>1.05</b>	0.62	<b>0.61</b>
	5	<b>4.65</b>	10.52	12.55	<b>0.88</b>	1.66	<b>0.61</b>
	6	0.69	<b>0.61</b>	14.76	<b>1.66</b>	0.59	<b>0.46</b>

Lower values of MdMRE are represented in bold, which shows better prediction

**Table 10** PRED (25) comparison for ANN and NANN for 3 Benchmark datasets

Data Set		Kemerer		Coc81		Maxwell	
FOLD	HN	ANN (All Attribute)	NANN (4,5,6,7)	ANN (All Attribute)	NANN (5,6,18)	ANN (All Attribute)	NANN (24,25)
2	1	0.00	<b>37.50</b>	<b>3.23</b>	0.00	22.58	22.58
	2	12.50	<b>25.00</b>	<b>9.68</b>	3.23	12.90	<b>19.35</b>
	3	<i>0.00</i>	<i>0.00</i>	0.00	<b>3.23</b>	16.13	<b>29.03</b>
	4	<i>12.50</i>	<i>12.50</i>	3.23	<b>25.81</b>	6.45	<b>9.68</b>
	5	<b>37.50</b>	25.00	3.23	<b>12.90</b>	16.13	<b>29.03</b>
	6	<b>37.50</b>	25.00	6.45	<b>16.13</b>	12.90	<b>35.48</b>
3	1	<b>40.00</b>	<b>0.00</b>	4.76	<b>9.52</b>	14.29	<b>33.33</b>
	2	0.00	<b>20.00</b>	14.29	<b>19.05</b>	28.57	<b>33.33</b>
	3	0.00	<b>20.00</b>	<b>19.05</b>	14.29	23.81	<b>33.33</b>
	4	<i>20.00</i>	<i>20.00</i>	4.76	<b>14.29</b>	<b>23.81</b>	19.05
	5	<i>0.00</i>	<i>0.00</i>	4.76	<i>4.76</i>	4.76	<b>19.05</b>
	6	<i>20.00</i>	<i>20.00</i>	4.76	<b>14.29</b>	19.05	<b>38.10</b>

Higher values of PRED (25) are represented in bold, which shows better prediction and equal values are represented in italic

## 6 Conclusions and Future Possibilities

Our study proposes the application of information gain for feature selection. This research is carried out to compare ANN trained on all attributes with the NANN trained on selected attributes using two- and threefold cross-validation for effort prediction. We used three datasets Kemerer, Coc81, and Maxwell. The performance metrics used are: MAR, MMRE, MdmRE, PRED (25), and MSE. As per the result discussion in segment 1.5, we propose that the accuracy of the NN model for effort prediction can be enhanced by feature selection using Information Gain. It removes the effect of irrelevant attributes on effort prediction. We also find that the proposed study works better even if we vary the number of hidden neurons from 1 to 6.

Future work will focus on investigating other neural network architectures for effort prediction. A hybrid model can be created by a combination of neural network and other soft computing-based approaches such as genetic programming and fuzzy logic.

**Table 11** MSE comparison for ANN and NANN for 3 benchmark datasets

Data Set		Kecmer		Coc81		Maxwell	
FOLD	HN	ANN (All Attribute)	NANN (4,5,6,7)	ANN (All Attribute)	NANN (5,6,18)	ANN (All Attribute)	NANN (24,25)
2	1	110,053.92	<b>83,054.32</b>	7,172,153.99	<b>2,980,618.75</b>	131,341,820.67	<b>65,716,571.05</b>
	2	133,453.05	<b>119,094.61</b>	3,234,572.76	<b>3,176,593.07</b>	149,446,823.52	<b>69,480,109.21</b>
	3	<b>99,675.23</b>	114,843.30	<b>3,912,278.86</b>	55,702,082.18	122,784,682.50	<b>49,089,248.92</b>
	4	212,070.67	<b>57,743.10</b>	4,563,974.84	<b>4,199,888.97</b>	149,200,162.64	<b>47,878,694.41</b>
	5	<b>62,792.89</b>	152,513.36	5,026,531.99	<b>2,136,017.55</b>	144,820,453.36	<b>110,263,216.02</b>
	6	<b>6078.57</b>	87,668.14	4,992,017.14	<b>2,537,033.41</b>	84,899,579.44	<b>22,496,786.12</b>
3	1	<b>12,891.51</b>	293,718.13	<b>2,153,401.57</b>	2,258,960.79	156,116,410.45	<b>20,223,422.17</b>
	2	539,220.22	<b>46,342.66</b>	7,908,742.72	<b>6,693,134.89</b>	<b>170,019,964.22</b>	223,562,245.98
	3	<b>48,651.65</b>	235,703.94	<b>1,361,692.67</b>	1,458,056.52	26,432,121.50	<b>17,515,027.42</b>
	4	173,351.99	<b>119,620.97</b>	2,012,200.58	<b>697,307.73</b>	44,186,632.76	<b>41,619,233.92</b>
	5	<b>141,070.15</b>	1,227,729.24	4,070,943.36	<b>627,161.21</b>	113,578,930.30	<b>24,803,886.85</b>
	6	124,696.61	<b>105,646.29</b>	4,656,738.86	<b>352,215.69</b>	141,932,745.84	<b>115,624,726.25</b>

Lower values of MSE are represented in bold, which shows better prediction

**Acknowledgements** The authors wish to thank Mr. J. S. Shirabad and Mr. T. J. Menzies for providing the benchmark datasets that are very necessary for this research work.

## References

1. Boehm B, Abts C, Chulani S (2000) Software development cost estimation approaches—A survey. *J Annals Softw Eng* 10:177–205. <https://doi.org/10.1023/A:1018991717352>
2. Catal C, Aktas MS (2011) A composite project effort estimation approach in an enterprise software development project. In: Proceedings of twenty third international conference software engineering and knowledge engineering (SEKE '11), pp 331–334
3. Boehm BW (2001) Software engineering economics. In: Broy M, Denert E (Eds) *Pioneers and their contributions to software engineering—Algorithms, architectures and applications*. Springer-Verlag, Berlin
4. Boehm BW et al (2000) *Cost estimation with COCOMO II*. Prentice Hall
5. Nielsen K (2013) Software estimation using a combination of techniques. Project Management Institute (PMI) Virtual Library
6. Vergara JR, Estévez PA (2014) A review of feature selection methods based on mutual information. *J Neural Comput Applic* 24(1):175–186. <https://doi.org/10.1007/s00521-013-1368-0>
7. Rao PS, Reddi KK, Rani RU (2017) Optimization of neural network for software effort estimation. In: Proceedings of IEEE international conference algorithms, methodology, models and applications in emerging technologies (ICAMMAET '17), pp 1–7, Feb. <https://doi.org/10.1109/ICAMMAET.2017.8186696>
8. Port D, Korte M (2008) Comparative studies of the model evaluation criterions MMRE and PRED in software cost estimation research. In: Proceedings of second ACM-IEEE international symposium empirical software engineering and measurement (ESEM '08), pp 61–69. <https://doi.org/10.1145/1414004.1414015>
9. Fernandes SL, Gurupur VP, Sunder NR, Arunkumar N, Kadry S (2017) A novel nonintrusive decision support approach for heart rate measurement. *J Pattern Recogn Lett*. <https://doi.org/10.1016/j.patrec.2017.07.002>
10. Martens D, De Backer M, Haesen R, Snoeck M, Vanthienen J, Baesens B (2007) Classification with ant colony optimization. *IEEE Trans Evol Comput* 11(5):651–665. <https://doi.org/10.1109/TEVC.2006.890229>
11. Desai VS, Mohanty R (2018) ANN-Cuckoo optimization technique to predict software cost estimation. In: Proceedings of IEEE conference information and communication technology (CICT '18), pp 1–6, Oct. <https://doi.org/10.1109/INFOCOMTECH.2018.8722380>
12. Fernández-Diego M, Elmouaden S, Torralba-Martínez J (2012) Software effort estimation using NBC and SWR: A comparison based on ISBSG projects. In: Proceedings of IEEE joint conference international workshop on software measurement and international conference software process and product measurement (IWSM-MENSURA '12), pp 132–136, Oct. <https://doi.org/10.1109/IWSM-MENSURA.2012.28>
13. Deng K, MacDonell SG (2008) Maximising data retention from the ISBSG repository. In: Proceedings of twelfth international conference evaluation and assessment in software engineering (EASE '08), pp 21–30, June. <https://doi.org/10.5555/2227115.2227118>
14. Jiang Z, Comstock C (2007) The factors significant to software development productivity. *Int J Comput Inf Eng, World Acad Sci Eng Technol* 1(1):160–164
15. Huang S, Chiu N, Liu Y (2008) A comparative evaluation on the accuracies of software effort estimates from clustered data. *J Inf Softw Technol* 50(9–10):879–888

16. Lokan C, Mendes E (2006) Cross-company and single-company effort models using the ISBSG database: a further replicated study. In: Proceedings of fifth ACM/IEEE international symposium empirical software engineering (ISESE '06), pp 75–84, Sep. <https://doi.org/10.1145/1159733.1159747>
17. Seo Y, Yoon K, Bae D (2008) An empirical analysis of software effort estimation with outlier elimination. In: Proceedings of fourth international conference predictive models in software engineering (PROMISE '08), pp 25–32, May. <https://doi.org/10.1145/1370788.1370796>
18. Jeffery R, Ruhe M, Wiczorek I (2001) Using public domain metrics to estimate software development effort. In: Proceedings of seventh IEEE international software metrics symposium, pp 16–27, Apr. <https://doi.org/10.1109/METRIC.2001.915512>
19. Mendes E, Mosley N, Counsell S (2003) A replicated assessment of the use of adaptation rules to improve web cost estimation. In: Proceedings of international IEEE symposium empirical software engineering (ISESE '03), pp 100–109, Sep/Oct. <https://doi.org/10.1109/ISESE.2003.1237969>
20. Ahmad I, Abdullah A, Alghamdi A, Hussain M (2013) Optimized intrusion detection mechanism using soft computing techniques. *J Telecommun Syst* 52:2187–2195. <https://doi.org/10.1007/s11235-011-9541-1>
21. Dejaeger K, Verbeke W, Martens D, Baesens B (2012) Data mining techniques for software effort estimation: a comparative study. *IEEE Trans Softw Eng* 38(2):375–397. <https://doi.org/10.1109/TSE.2011.55>
22. Padhy N, Singh RP, Satapathy SC (2019) Cost-effective and fault-resilient reusability prediction model by using adaptive genetic algorithm based neural network for web-of-service applications. *J Cluster Comput* 22:14559–14581. <https://doi.org/10.1007/s10586-018-2359-9>
23. Liu Q, Xiao J, Zhu H (2019) Feature selection for software effort estimation with localized neighborhood mutual information. *J Cluster Comput* 22:6953–6961. <https://doi.org/10.1007/s10586-018-1884-x>
24. Quinlan JR (1986) Induction of decision trees. *J Mach Learn* 1:81–106. <https://doi.org/10.1007/BF00116251>
25. Shannon CE (1948) A mathematical theory of communication. *J Bell Syst Tech* 27(3): 379–423, 623–656
26. Mendes E, Mosley N, Watson I (2002) A comparison of case based reasoning approaches. In: Proceedings of eleventh international conference world wide web (WWW '02), pp 272–280, May. <https://doi.org/10.1145/511446.511482>
27. Jørgensen M (2007) Forecasting of software development work effort: evidence on expert judgment and formal models. *Int J Forecast* 23(3):449–462. <https://doi.org/10.1016/j.ijforecast.2007.05.008>
28. Nassif AB, Ho D, Capretz LF (2013) Towards an early software estimation using log linear regression and a multilayer perceptron model. *J Syst Softw* 86(1):144–160. <https://doi.org/10.1016/j.jss.2012.07.050>
29. Nassif AB, Azzeh M, Capretz LF, Ho D (2016) Neural network models for software development effort estimation: a comparative study. *J Neural Comput Applic* 27:2369–2381. <https://doi.org/10.1007/s00521-015-2127-1>
30. Putnam LH (1978) A general empirical solution to the macro software sizing and estimating problem. *IEEE Trans Softw Eng* 4(4):345–361. <https://doi.org/10.1109/TSE.1978.231521>
31. Galorath DD, Evans MW (2006) Software sizing, estimation, and risk management. Auerbach Publications, Boston. ISBN 0849335930
32. Saliu MO, Ahmed M (2004) Soft computing based effort prediction systems—A survey. In: Damiani E, Jain LC (Eds) *Soft computing in software engineering*, Studies in fuzziness and soft computing series (STUDFUZZ), vol 159. Springer-Verlag, Berlin, pp 151–182. [https://doi.org/10.1007/978-3-540-44405-3\\_6](https://doi.org/10.1007/978-3-540-44405-3_6)
33. Jodpimai P, Sophatsathit P, Lursinsap C (2010) Estimating software effort with minimum features using neural functional approximation. In: Proceedings of international conference computational science and its applications (ICCSA '10), pp 266–273, Mar. <https://doi.org/10.1109/ICCSA.2010.63>

34. Idri A, Khoshgoftaar TM, Abran A (2002) Can neural networks be easily interpreted in software cost estimation? In: Proceedings of (Cat. No.02CH37291) IEEE world congress on computational intelligence, IEEE international conference fuzzy systems (FUZZ-IEEE '02), vol 2, pp 1162–1167, May. <https://doi.org/10.1109/FUZZ.2002.1006668>
35. Venkatachalam AR (1993) Software cost estimation using artificial neural networks. In: Proceedings of international conference neural networks (IJCNN '93), vol 1, pp 987–990, Oct. <https://doi.org/10.1109/IJCNN.1993.714077>
36. Wittig GE, Finnie GR (1994) Using artificial neural networks and function points to estimate 4GL software development effort. *Austr J Inform Syst* 1(2):87–94. <https://doi.org/10.3127/ajis.v1i2.424>
37. Tadayon N (2005) Neural network approach for software cost estimation. In: Proceedings of international conference information technology: coding and computing (ITCC '05), vol 2, pp 815–818, Apr. <https://doi.org/10.1109/ITCC.2005.210>
38. Finnie GR, Wittig GE, Desharnais JM (1997) A comparison of software effort estimation techniques: using function points with neural networks, case-based reasoning and regression models. *J Syst Softw* 39(3):281–289. [https://doi.org/10.1016/S0164-1212\(97\)00055-1](https://doi.org/10.1016/S0164-1212(97)00055-1)
39. Idri A, Zakrani A, Zahi A (2010) Design of radial basis function neural networks for software effort estimation. *Int J Comput Sci* 7:11–17
40. Kumar KV, Ravi V, Carr M, Kiran NR (2008) Software development cost estimation using wavelet neural networks. *J Syst Softw* 81(11):1853–1867. <https://doi.org/10.1016/j.jss.2007.12.793>
41. Roh S-B, Oh S-K, Pedrycz W (2011) Design of fuzzy radial basis function-based polynomial neural networks. *J Fuzzy Sets Syst* 185(1):15–37. <https://doi.org/10.1016/j.fss.2011.06.014>
42. Rao BT, Sameet B, Swathi GK, Gupta KV, RaviTeja Ch, Sumana S (2009) A novel neural network approach for software cost estimation using functional link artificial neural network (FLANN). *Int J Comput Sci Netw Sec (IJCSNS' 09)* 9(6):126–131
43. Bardsiri VK, Jawawi DNA, Hashim SZM, Khatibi E (2012) Increasing the accuracy of software development effort estimation using projects clustering. *J IET Softw* 6(6):461–473. <https://doi.org/10.1049/iet-sen.2011.0210>
44. Kaushik A, Soni AK, Soni R (2016) An improved functional link artificial neural networks with intuitionistic fuzzy clustering for software cost estimation. *Int J Syst Assur Eng Manag* 7:50–61. <https://doi.org/10.1007/s13198-014-0298-2>
45. Dave VS, Dutta K (2014) Neural network-based models for software effort estimation: a review. *J Artif Intell Rev* 42:295–307. <https://doi.org/10.1007/s10462-012-9339-x>
46. Wen J, Li S, Lin Z, Hu Y, Huang C (2012) Systematic literature review of machine learning based software development effort estimation models. *J Inf Softw Technol* 54(1):41–59. <https://doi.org/10.1016/j.infsof.2011.09.002>
47. Guyon I, Elisseeff AE (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3:1157–1182
48. Yu L, Liu H (2003) Feature selection for high-dimensional data: a fast correlation-based filter solution. In: Proceedings of twentieth international conference machine learning. (ICML '03), pp 856–863
49. Keung JW, Kitchenham BA, Jeffery DR (2008) Analogy-X: providing statistical inference to analogy-based software cost estimation. *IEEE Trans Softw Eng* 34(4):471–484. <https://doi.org/10.1109/TSE.2008.34>
50. Battiti R (1994) Using mutual information for selecting features in supervised neural net learning. *IEEE Trans Neural Netw* 5(4):537–550. <https://doi.org/10.1109/72.298224>
51. Estevez PA, Tesmer M, Perez CA, Zurada JM (2009) Normalized mutual information feature selection. *IEEE Trans Neural Netw* 20(2):189–201. <https://doi.org/10.1109/TNN.2008.2005601>
52. Liu H, Sun J, Liu L, Zhang H (2009) Feature selection with dynamic mutual information. *J Pattern Recogn* 42(7):1330–1339. <https://doi.org/10.1016/j.patcog.2008.10.028>
53. Peng H, Long F, Ding C (2005) Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell* 27(8):1226–1238. <https://doi.org/10.1109/TPAMI.2005.159>

54. Hu Q, Zhang L, Zhang D, Pan W, An S, Pedrycz W (2011) Measuring relevance between discrete and continuous features based on neighborhood mutual information. *J Expert Syst Appl* 38(9):10737–10750. <https://doi.org/10.1016/j.eswa.2011.01.023>
55. Hall MA (1999) Correlation-based feature selection for machine learning, PhD dissertation, Department of Computer Science, The University of Waikato, Hamilton, New Zealand, Apr
56. Haykin S (1999) *Neural networks a comprehensive foundation*. 2nd edn, Prentice-Hall, Englewood Cliffs, NJ
57. Deepa SN, Sivanandam SN (2011) *Principles of soft computing*. 2nd edn, Wiley India Pvt. Ltd.
58. Beale MH, Hagan MT, Demuth HB (2020) *Deep learning toolbox*. `nnet_gs`, `fitnet`, pp 1.56–1.62. [https://in.mathworks.com/help/pdf\\_doc/deeplearning/nnet\\_gs.pdf](https://in.mathworks.com/help/pdf_doc/deeplearning/nnet_gs.pdf)
59. Park H, Baek S (2008) An empirical validation of a neural network model for software effort estimation. *J Expert Syst Appl* 35(3):929–937. <https://doi.org/10.1016/j.eswa.2007.08.001>
60. Hornik K, Stinchcombe M, White H (1989) Multilayer feedforward networks are universal approximators. *J Neural Netw* 2(5):359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)
61. Hagan MT, Menhaj MB (1994) Training feedforward networks with the Marquardt algorithm. *IEEE Trans Neural Netw* 5(6):989–993. <https://doi.org/10.1109/72.329697>
62. Han J, Kamber M, Pei J (2012) *Data mining: concepts and techniques*, 3rd edn, Morgan Kaufmann, Elsevier, pp 243–278
63. Shirabad JS, Menzies TJ (2005) The promise repository of software engineering databases. Repository. <http://promisedata.org/repository/>
64. Kocaguneli E, Menzies T, Keung JW (2012) On the Value of ensemble effort estimation. *IEEE Trans Softw Eng* 38(6):1403–1416. <https://doi.org/10.1109/TSE.2011.111>
65. Resmi V, Vijayalakshmi S, Chandrabose RS (2019) An effective software project effort estimation system using optimal firefly algorithm. *J Cluster Comput* 22:11329–11338. <https://doi.org/10.1007/s10586-017-1388-0>
66. Kocaguneli E, Menzies T (2011) How to find relevant data for effort estimation? In: *Proceedings of IEEE international symposium empirical software engineering and measurement (ESEM '11)*, pp 255–264, Sep. <https://doi.org/10.1109/ESEM.2011.34>
67. Khan K (2010) The evaluation of well-known effort estimation models based on predictive accuracy indicators, Master Thesis, Master of Science, Software Engineering, School of Computing, Blekinge Institute of Technology, Ronneby Sweden, Jan
68. Shepperd M, Schofield C (1997) Estimating software project effort using analogies. *IEEE Trans Softw Eng* 23(11):736–743. <https://doi.org/10.1109/32.637387>
69. Foss T, Stensrud E, Kitchenham B, Myrvtveit I (2003) A Simulation study of the model evaluation criterion MMRE. *IEEE Trans Softw Eng* 29(11):985–995. <https://doi.org/10.1109/TSE.2003.1245300>
70. Satapathy SM, Acharya BP, Rath SK (2016) Early stage software effort estimation using random forest technique based on use case points. *J IET Softw* 10(1):10–17. <https://doi.org/10.1049/iet-sen.2014.0122>
71. Azzeh M, Nassif AB, Banitaan S (2018) Comparative analysis of soft computing techniques for predicting software effort- based use case points. *J IET Softw* 12(1):19–29. <https://doi.org/10.1049/iet-sen.2016.0322>
72. De Campos Souza PV, Guimaraes AJ, Araujo VS, Rezende TS, Araujo VJS (2019) Incremental regularized data density-based clustering neural networks to aid in the construction of effort forecasting systems in software development. *J Appl Intell* 49:3221–3234

# A Comprehensive Survey of Web and Mobile Apps for Fishermen



M. R. Jivthesh, M. R. Gaushik, N. B. Sai Shibu, Dhanesh Raj,  
and Sethuraman N. Rao

**Abstract** Mobile and web apps are playing an essential role in people's daily life. ICT enhances internet connectivity to anyone and anything. In this case, the importance of mobile apps is critical in extreme environmental conditions such as ocean scenarios. Several mobile phone applications have been launched to encourage and assist fishermen through knowledge distribution, facilitating communication, monitoring geolocation, and assisting in boat-to-shore activities. However, these apps need internet connectivity to provide services. They are limited to working only on the shore using cellular connectivity. In this context, OceanNet, a marine communication infrastructure that our research center has developed and deployed in 2016. OceanNet provides affordable internet access up to 60 km from the shore. We have interviewed around hundred fishermen in the neighborhood to understand their requirements. We have also identified a few mobile and web applications developed for the fisherman. In this paper, we analyze and summarize the features provided by these web and mobile applications. We also discuss the requirements of the fisherman when they are sailing. In the future, we plan to develop suitable applications to address the gaps identified in this study.

**Keywords** OceanNet · MarineCommunication apps · Fishermen · Literature survey · App store · Play store

## 1 Introduction

Today, the massive technological revolution in mobile devices has characterized and driven ICT, which has completely transformed how people communicate and engage with applications, resources, and the environment. The ultimate objective of ICT is

---

M. R. Jivthesh · M. R. Gaushik  
Department of Computer Science & Engineering (CSE), Amrita Vishwa Vidyapeetham,  
Vallikavu, Amritapuri, India

N. B. S. Shibu (✉) · D. Raj · S. N. Rao  
Center for Wireless Networks & Applications (WNA), Amrita Vishwa Vidyapeetham, Vallikavu,  
Amritapuri, India  
e-mail: [saishibunb@am.amrita.edu](mailto:saishibunb@am.amrita.edu)



to connect people and devices and interact seamlessly. The proliferation of mobile devices and the Internet in society bridged the gap of the digital divide. There are under-connected communities in the world having no internet coverage due to a lack of communication technologies. There is a “digital divide” between the various demographic groups and geographical regions. For example, in rural areas, anglers are less able to access the internet or communication technologies while they are at sea for fishing. Fishermen who lack the necessary skills and knowledge of digital technology’s potential use are under-exploited in terms of using it for communication and their business use case.

Fishing at sea is among the foremost dangerous occupations in the world. A lost vessel and a lost fisherman have a significant impact on the coastal community. The life of a fisherman can be made more comfortable with the help of fishing apps. The advantage of GPS mapping and fish-finding tasks with the assistance of the fishing apps has created folks, the feeling that speed, time, and money can be saved. Angling, target mapping, and location are the prime pinpoints for a rewarding catch. Another fact is that a little device that most of the anglers carry with them in their hands and pockets can do lots of productive tasks.

An angler’s life has taken a massive shift considering that they will get pleasure from their work and stay connected with the earth out there with the assistance of the best fishing apps. This not only helps them to remain connected with the shore but together helps them to be productive within the ocean. Thus, fishing apps became a vital addition to every modern-day fisher. The options of current rising developments with technological progressions are close to fishing as a game. Fishing technologies have become a valuable aid for contemporary anglers and fishing thrill-seekers. This paper delivers an insightful and detailed literature survey on widely used mobile and web applications in the market for fisherman’s assistance at the ocean. We highlight the importance and utility of these apps to anglers in the OceanNet network context.

## ***1.1 Motivation***

Fishing is one of the major occupations worldwide providing a livelihood for millions of people. Fishermen venture into the ocean for 5–7 days, get totally isolated from the shore, and come across a variety of risky circumstances at the ocean. They travel more than a hundred kilometers into the ocean for fishing, but the coverage of the cellular network across the oceans is limited to around 15 km from the shore in most areas [1]. Satellite-based marine internet connectivity isn’t a suitable option for poor fishermen because of the excessive price. Much of the time marine VHF is quite loud and quite unusable. It becomes difficult to transfer emergency information concerning awful weather signals. This can put the lives of fishermen in danger at the sea.

Currently, there are no economically possible solutions to live connected among themselves in addition to the shore [2]. To deal with this situation, a project named ‘OceanNet’ was launched. OceanNet is a cost-effective communication system for

maritime network access that ensures a guaranteed 45 + km coverage from the shore, which can be expanded as far as a range of 20 km via multilevel point to multipoint (P2MP) networks [3]. Figure 1 shows the block diagram of the OceanNet network. OceanNet system consists of a long-range transceiver mounted on a rotating platform. Long-range base stations are set up across the coastal regions of Kerala. The rotary platform will align the transceiver to the base station with respect to the direction of the boat [4]. This network provides connectivity up to 60 km. Wi-Fi access points enable fishermen to connect their mobile phones to the Internet when they are in the ocean. The system also collects the live location, speed, and direction of the boat and relays it to the servers on the shores. A Raspberry Pi zero device is employed to collect the data and relay to shore [5].

Mobile applications support functionality such as VoIP and cross-messaging and have additional functionality, including position-based services and information searching. Smartphone applications allow quick access to information with a satisfactory network quality anywhere and at any time. As an outcome, they have an enormous potential to help anglers in any way.

Fishermen using smartphones can obtain a deep understanding of the market rate, business, potential fishing zone information, live weather, and distress alerts to improve their business and safety. In recent years, several fishing applications have been introduced into the market to educate and enhance the understanding of the marine environment while they are at sea [6]. By doing so, mobile apps improve the overall socioeconomic thrift of fishermen and, hence, greatly benefit community empowerment.

Mobile and web app developers also find it challenging to determine target users, their needs, and potential feedback to improve apps. Thus, a comprehensive survey of different apps helps them build a resilient app serving the user’s needs. We have already interviewed more than a hundred anglers to understand their requirements. We plan to develop suitable applications in the future to address the gaps identified in this survey. The rest of this paper is organized as follows. Section 2 identifies the web apps and mobile apps for different maritime applications and their requirements based

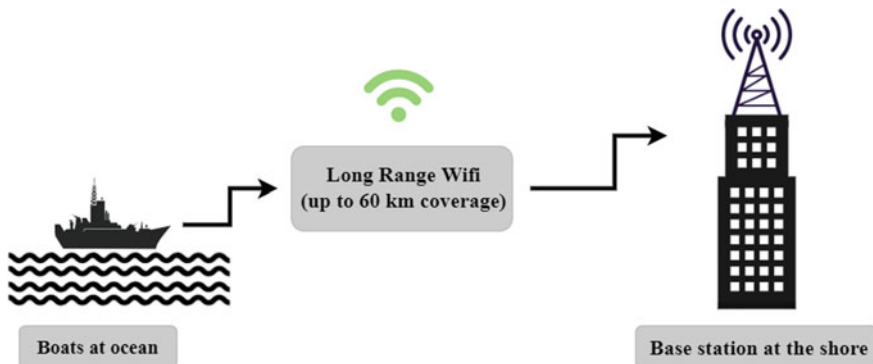


Fig. 1 Block diagram of the existing OceanNet Infrastructure

on their features. Section 3 discusses various web and mobile applications based on the cost, service, and overviews the key components associated with maritime web and mobile apps to meet the related maritime application requirements and features. Finally, Sect. 4 concludes the paper by highlighting the possible approach to developing resilient and reliable apps for fishing and information dissemination using maritime communication technology.

## 2 Literature Survey

There are many web and mobile applications to support fishermen. In this paper, we selected the web and mobile apps based on UI design, speed, navigability, and content readability, as explained by Peng et al. [7]. There are numerous fishermen-related applications available today. However, to build a reliable, effective, and secure app, you need to have in-depth knowledge about the state of the art of fishermen applications. In this paper, we chose only a few selective mobile and web apps to review. Spriestersbach et al. [8] performed the quality measurement of mobile and web apps based on functionality, cross compatibility, security, user interface, usability, and performance. In this paper, we aim to consolidate different apps by identifying the relevant features to existing fishing and fishermen-related applications, which are categorized in terms of user experience of different fishing community stakeholders such as fishermen, fishing vessel owners, managers, and fishing community and their functionalities required by each of them. This paper categorizes the survey of mobile and web apps separately and explains them in two sections as below.

### 2.1 Survey of Web Apps

In this section, we cover a set of important web apps focusing on the different maritime applications. We have reviewed more than 30 web apps and 50 mobile apps. We have listed a few best apps based on the features provided and customer reviews. The primary aim of this systematic review is to get updated information on web-related marine applications for fishermen. The research seeks to determine the extent to which existing applications provide functionalities and work under different communication technology, and user interface and to study the benefits and challenges of these applications.

- Tides for fishing: Weather temperature, water temperature, wave conditions, current standing of water, tidal coefficient, tide table, solunar info, fishing location for the complete world. NAUTIDE mobile app to see all this information [9].
- Indian Meteorological Department: Weather warnings, weather forecast, rainfall info, monsoon info, rainfall maps, marine forecast, solunar time and available separately for each district and station-wise [10].

- ESSO-Indian National Centre for Ocean Information Services: ESSO-INCOIS offers data like potential fishing area, ocean state prediction, early tsunami alerts, warnings on storm surge, coral bleaching warnings, Indian seismic and GNSS network, IN SITU data, remote sensing, and live access server [11].
- FishWeather: Free forecasts everywhere in the world, live wind, wind list, full wind chart and prediction page, wind warning via email, text, or mobile, onsite data from members around the world, wind archives, wind statics of all weather stations, tides and currents, radar and satellite, national weather service prediction, alerts, and wind widgets. These are publicly accessible features on the website or blog, many more features are available in the paid edition [12].
- Fishing Tackle: This is an associate degree Indian website that sells rods, reels, lines, lures, terminal tackle, fishing tools, accessories, boats and kayaks, and camping gears [13].
- National oceanic and atmospheric administration: This website provides weather, ocean, national environmental satellite, data, information and marine fisheries services. Provides warnings for hazardous weather, guides the utilization and protection of ocean and coastal resources, and conducts research to supply the understanding and improve environmental stewardship. However, this is not available for India and shows data for specific countries. [14].
- Fishing reminder: This website provides information on major and minor fishing hours, tide information, information about the sun and moon as well as weather information to plan fishing trips [15].
- Baywind: This website provides the marine forecast, weather forecast, weather radar maps, water temperature, humidity forecast, rainfall forecast, temperature forecast, UV index forecast, wind forecast, wave forecast, and information about storms, snow, fog, frost information for Australia [16].
- Marine Fish Sales: On this website, the fisherfolk can register as vendors and submit their wares as products to be sold, and customers can order these products subject to availability. Facilities exist for cash on delivery or net banking and email and SMS notification of orders and delivery. Online fish sale at Cochin within 10 km distance of CMFRI HOs is enabled now. This is a pilot research project envisioning scaling up at a later stage [17].
- Maxcatch fishing: This website offers its own fly reels, rods, lines, and clothing apparel that a rising fly angler would ever wish for [18].
- Fishermans World: This site offers accessories categorized under different categories like saltwater, freshwater, offshore, surf, and clothing [19].

## ***2.2 Survey of Mobile Applications***

This section summarizes mobile applications available that could be used by the fishermen through the OceanNet Network.

- Fishing Points GPS Navigation: The best fishing time, tide forecast, marine forecast, weather forecast, solunar information, compass, mapping facility, help to

locate fish points with GPS navigation, worldwide nautical charts, fishing logbook to save catches and create fishing logs [20].

- Fisher Friend Mobile Application (FFMA): Potential fishing zone and species-specific TUNA forecast, GPS facilities for direct navigation to PFZ and traditional fishing routes, ocean condition forecast data such as wave height, wind speed, and direction, sea current and surface temperature, disaster warnings, mark dangerous areas in the water, including sunken vessels, rock substrate, dead coral reefs, market prices for different species of fish, international borderline alert, SOS rescue option in the critical sea situation, when in an emergency, navigate harbor locations, calling facilities along with important contact information [21].
- Nautide: Daily tidal maps and tidal coefficient, tide table monthly, height and swell direction, wave quantity, surf table hourly, hourly activity charts, solunar periods, wind table, fishing barometer, pressure graph, and pressure table hourly with pattern indication show coastal weather details such as visibility, temperature, rainfall, humidity and hourly weather table, open water prediction for marine/sailing. It includes all weather indicators, water temperature, and hourly marine table [22].
- Boat US: Track your local waterways and forecasts for your favorite destinations up to 5 days before your journey for current tides, weather, and marine conditions, towing aid 24/7, hurricane warnings, manage insurance policy, and share position via text or email [23].
- Netfish: Solunar prediction based on barometric pressure and phases of moon, bait suggestions, fishing log, many map choices including topographic, detailed waterway information, including available species, fees, waterway capacity, depth [24].
- Odaku: Built-in GPS capabilities to store useful data in the cloud. These borders are cloud based and fishermen can download borders from their area, solving the problem of international borders and even local borders, ability to share location or fishing tracks between groups, backup data, online marketplace, buy or sell used boats, backup your catch data [25].
- Fishing Spots: Explore fishing charts showing sites, pictures, and details of fishing, find the right fishing hours with our robust prediction, create your own logbook and record each trip's information, share your area with anglers, post your catches, GPS fishing maps, fishing and solar forecasts [26].
- Fish Angler: Explore millions of data points on GPS charts, find the right fishing hours with a robust fishing prediction, real-time 7-day predictions of maritime conditions with temperatures of wind, wave, tide, and ocean, record and trace catches comprising weather conditions in real time, connect to other anglers, get fishing tips, track your catches and post them. [27].
- Navionics: Chart overlays like sonar chart and relief shading, satellite images, live-location sharing, advanced map options to customize chart views, and real-time weather, and tide information [28].
- Thoondil: Compass displaying the cardinal directions, longitude, speed, and direction, high tides, cyclone warning, strong wind directions, etc. It shows the PFZ

area on a regular basis, regular travels to register their journeys, my crew characteristics to record sailing crews on vessels, the boat owner can see the shore's live place. The positioning of the boat is very useful with hurricanes or cyclones in contact with the boat as well as in the process of rescue and search, the feature of the rescue plan uses offline maps to display the closest port locations, report incident monitoring feature to track suspicious activities such as oil spill/intrusion by foreign ship/illegal activities [29].

- WeFish: Create your fishing diary of all catches, best fishing forecasts and meteorological information, buy and sell fishing equipment [30].
- Fishing and Hunting Solunar Time: Save favorite places for future reference, significant and insignificant feeding/activity time interval, rating for the day, moonrise, moonset, sunrise and sunset times, present 5-day weather prediction, calendar for solunar data verification in advance [31].
- Deeper Smart Sonar (Hardware device with mobile application): Deeper smart sonar app together with a deeper sonar transforms your mobile into a high-grade sonar display, fishing log, download offline maps, save your favorite spots, solunar calendar, weather forecasts, data like depth, temperature, bottom structure, vegetation, fish location and more are displayed live to improve your angling, unlimited data storage to store data collected from deeper smart sonar [32].
- Fishing Calendar: The best fishing hours are predicted, infinite calendar forecast, fishing productivity view of day and month, moonrise, overhead moon and moonset times, lunar phase, sunrise times, intervals small and major, worldwide calendar for every place [33].
- Fishing Times Free: The feature provided by this app makes it easy to read the solunar fishing schedule, tide times, and sun/moon data to plan your fishing journey [34].
- Fish Weather: Wind and tide reports in real time, radar and map precipitation, satellite /cloud/ forecast map, and nautical charts, Also access to fishweather.com [35].
- mKRISHI Fisheries: Wind speed/direction, wave height, PFZ map, PFZ data, tuna PFZ map, weather information like rain, cloud, and temperature. This service is available to only those registered to use the service and belong to organizations such as TCS, ICAR, or CMFRI [36].
- Infish: Provides details on more than 150 inland waterways including directions, recent alerts, fishing reports, fishing tips, fishing regulations, and practices [37].
- Sagara: This application serves as a contact to track fishermen and vessels between anglers and government agencies, supporting various languages [38].
- PFZ Advisory: Potential Fishing Zone(PFZ) advisories, fishing folk often get regular tips on the availability of chlorophyll, sea temperature, water visibility which help them quickly find locations of abundant fish in the ocean while saving fuel and time spent looking for it [39].
- INCOIS: This app offers facilities for the advantage of the society of fishermen in India, such as Potential Fishing Zone (PFZ), and TUNA Potential Fishing Zone (TUNA PFZ) [40].

- Angler: Record fishing work, track and measure, track your fishing and catch fish. GPS fishing path with both catches and location, prediction based on your fishing results on a cell phone and net, fishing video report and trip analytics [41].
- Fishidy: Local spots for fishing, log catches, and places are 100 percent private, researched waterway information such as fishery details, boat ramps, access point, and stocking information, you get even more features with a paid membership, such as fishing insights, predictions, expert advice, access to maps offline in remote areas [42].
- ProAngler: Every major inshore, nearshore, and offshore fishing areas, guide your vessel to GPS hotspots and artificial reefs, provides updates on marine weather, tides, and solunar forecast, total species guide and methods for fishing, can access without internet connectivity [43].

### 3 Discussion on the Identified Web and Mobile Apps

This section discusses the requirements of fishermen, web and mobile app features available in existing apps, and features lacking in existing web and mobile apps.

#### 3.1 Requirements of the Fishermen

To understand the requirements of fishermen, we spoke to three fishermen in Kollam harbor about their requirements.

- i. Logbook to keep track of their catch and to generate a fishing log, save and share catches.
- ii. Save fishing areas, hot spots, waypoints; Record trolling paths and trotlines; Search saved GPS locations.
- iii. Maps including satellite view, radar weather, and sea surface temperature charts Navigation compass.
- iv. Distance to waypoints.
- v. Track mileage, presence of traffic, and other important stats.
- vi. Potential Fishing Zone and TUNA species-specific forecast downloadable maps or text data.
- vii. Instant display of warnings such as cyclones, tsunamis, and high waves.
- viii. Downloadable updates and offline accessibility.
- ix. Day-to-day fish activity forecast for any spot in the world.
- x. Severe weather alerts.
- xi. Real-Time weather, surf, wind, marine, tide, and solunar information hourly and weekly with archived info and 10-day forecast.
- xii. Support of multiple languages to select their preferred languages.
- xiii. Real-time marine traffic maps.

- xiv. GPS facilities for direct navigation to the PFZ and conventional fishing paths and to plot our position in real-time.
- xv. Maps illustrate dangerous areas in the ocean, such as sunken vessels, rocky substrates, and dead coral reefs.
- xvi. Border alert.
- xvii. SOS option and navigation to nearby harbor locations for rescue during emergency situations at sea.
- xviii. Government schemes and daily news.
- xix. Backup and restore feature to backup data if accidentally lost.
- xx. Calling facility along with crucial contact details during a crisis.
- xxi. Because of space utilization in mobile phones, a single customized mobile app with all the important features will be helpful for fishermen in productive angling.

We found many web/mobile applications like weather apps and disaster alert apps available during the survey. Based on their utility, offline and online modes, memory and application size, UI design, network support, and easy navigation feature are considered. The best apps are selected and discussed in this study. We believe that the above table's web apps are selected among other apps as they provide essential information such as weather prediction that includes visibility, temperature, precipitation, and weather table hourly. Surf information includes swell height and direction, wave period, and surf table per hour. Wind information such as wind speed, wind gust, wind force, land and sea conditions, and wind table per hour. Marine information includes weather indications, water temperatures, and marine table per hour. The fishing barometer includes a barometer for fishing, a pressure chart, and a pressure table per hour. Tide details such as regular tide graph, tidal coefficient, low and high tides, tide height, and tide table monthly. The solunar prediction provides the best times to catch fish based on sun and moon forecast, including sunrise, sunset, moonrise, moonset, moon phases, eclipses, transits, and other astronomical data. Fish activity information such as hourly activity graphs and similar periods of the day for the finest fishing moments are beneficial before venturing into the sea. Monthly tables for activity, tide, and solunar forecasts. Potential fishing zones (PFZ) information using interactive maps are sites in the ocean where there will be an excess of fish depending on ocean surface temperature and availability of chlorophyll; these PFZ maps and text are provided in the native languages of each sector and tuna advisory map showing demarcated areas for tuna fishing. Tsunami early warnings to convey information about hazards/risks at sea, storm surge warnings to take precautions before a disaster and coral bleaching alerts, easy to use interface. These features support anglers to save their time and maximize catch by understanding the right time for fishing. Some websites offer equipment and accessories needed for onshore and offshore fishing. Selected web apps also provide browser compatibility, security, and data visualization.

The mobile apps listed above lead because they provide useful fishing information like weather conditions that include visibility, temperature, precipitation, humidity, and weather table per hour. Surf information that includes swell height



and direction, wave period, and surf table per hour. Wind information provides wind speed, gust, force, wind table hourly, land and sea conditions, and coastal information including weather indications, water temperatures, and marine tables hourly. The fishing barometer offers general conditions for fishing and changes due to pressure patterns. Tide information like regular tide graph, tidal coefficient, low and high tides, tide height, and tide table monthly. Solunar forecast information includes sunrise, sunset, moonrise, moonset, moon phases, eclipses, transits, and other astronomical data. Fish activity information such as hourly activity graphs and solunar periods of the day for the finest fishing moments are beneficial before venturing into the sea. Monthly tables for fishing activity, tides, and solunar data these features are provided by web apps. Apart from this, mobile apps provide many unique features, such as compass options. Therefore, the boat owner can view the high-precision compass showing the cardinal directions, latitude/longitude, and speed. Fishing logbook to save fishing regions, hotspots, and waypoints to mark fishing spots, save day-by-day catches, dates and times of fishing, condition of that day, and favorite locations use this data to classify the best fishing areas and time to increase the chances of catching more. These records are stored in a cloud server and cannot be lost unexpectedly. Facility to save and access offline PFZ and traditional fishing areas for future use. GPS facility to determine and record coordinates, locate a vessel in the water, navigate potential fishing areas, and plot positions in real-time. Mark dangerous maritime locations such as sunken vessels, rock bodies, and coral reefs. Disaster alerts, including cyclones, tsunamis, and high waves, display immediate warnings that protect fishermen and their assets. International border line alerts notify the fishermen whenever the border is crossed by unauthorized means and safeguard them from international conflict and distress. Call facilities with essential contact information in emergencies, SOS (Save our Soul) feature works by transmitting a distress signal, which helps fishermen during search and rescue. Exchange of information with other anglers around the world. Catch details on what, where, and how the fishing net is being caught; this helps fishermen to have detailed reports of their catch data. Back up your data such as position and waypoints so that when necessary, you can restore them. Export and import data in KML or GPX format to manage paths, tracks, and conveniently share, store and view them on various applications, plotters, and compatible devices. User-friendly UI design, and multiple language support allows the boat owner to pick their preferred language, downloadable updates, easy usability, store data, ability to work in most weather conditions, low memory consumption, minor app size, and ability to run in different android versions as all the important features are combined into one distinctive mobile app which will be useful for fishermen in productive angling. These features stand-alone and integrated features are missing.

The characteristics we discover lacking from the web/mobile application are as follows:

- 10-day weather forecast.
- Archived weather information.

- Fishing prediction to identify the best species targeted during specific fishing times.
- Providing maritime traffic visibility in real-time, and connecting the app to a compatible AIS receiver without a connection to the internet.
- Travel distance tracker to calculate the distance traveled by vessels using GPS.
- Compass feature to get accurate direction.
- Measure distance option to measure how far the places are between each other.
- Marking of danger zones at sea like a sunken boat and dead coral reefs.
- Detailed map overlays give information about the underwater structure and bottom composition, submerged vegetation, submerged structures, and GPS coordinates with fishing hotspots.
- Fish-net locator that makes it possible for anglers in order to prevent cast nets in the sea.
- Advance chart plotter to plot the location in real-time using the device's built-in GPS.
- Advanced map tools to modify chart views to adjust chart-overlay combination, highlight shallow areas, seek several fishing ranges, and other pertinent details.
- Baits, lures, and rigs option to illustrate how different fishing knots are tied and where specific lures and baits are being used.
- Offline accessibility feature.
- Backup and restore option to back up their data such as locations and waypoints to be restored when needed.
- Notification during high fishing activity.
- Get real-time notifications about changes to regulations, emergency rules, etc.
- Report incident feature to report emergency action like Oil Spill/foreign vessel intrusion or prohibited actions.
- Calendar for pre-checking solar details.
- Downloadable PFZ and TUNA PFZ maps that display the positions of the PFZ regularly are useful for the fisherman to get a decent catch, and mark favorite fishing spots offline.
- The emergency Plan feature displays the closest harbor positions using offline charts. The fisherman can select any port nearby, and it shows the shortest distance from his current location.
- Import files like kmz or gpx from GPS devices or other applications.
- Multiple OS compatibility.

The above app's restriction is that they are only accessible for a particular territory and not open internationally. We found that these application limitations mostly lack multiple language support, are free for a short period, and pro-edition costs are high, not fully fledged, and have a messy UI interface. Several programs are regularly jammed and do not work correctly, ads are shown, which drains the battery faster. Some apps work with hardware devices, and the device's price is high.

## 4 Conclusion

In this paper, we discuss a few web and mobile applications developed for fishermen. These applications solve some of the problems faced by a fisherman when they fish in deep seas. OceanNet is an internet service provider for marine applications. We also tested these applications on our OceanNet Network. We met a few fishermen who use these applications to know the user experiences. We also discussed how these applications help them to improve the fishing experiences. We summarized the requirements put down by the fishermen. The insights we gathered from this survey will help us in developing an application for fishermen based on the requirements listed. We will optimize the application for the OceanNet infrastructure.

**Acknowledgements** We express our sincere gratitude to our beloved Chancellor and world-renowned humanitarian leader Dr. Mata Amritanandamayi Devi (AMMA), for her inspiration and motivation in all our endeavors.

## References

1. Rao SN, Raj D, Parthasarathy V, Aiswarya S, Ramesh MV, Rangan V (2018) A novel solution for high speed internet over the oceans. In: IEEE INFOCOM 2018—IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp 906–912
2. Gaushik MR, Jivthesh MR, Arjun D, Raj D, Rao SN (2020) Automatic virtual switching of adaptive backhaul equipment for marine internet applications. In: 2020 international conference on wireless communications signal processing and networking (WiSPNET), pp 15–19
3. Nadella H, Nagamalla C, Sai Shibu NB, Arjun D, Rao SN (2020) Low cost voip service for marine fishermen development and performance evaluation. In: 2020 international conference on wireless communications signal processing and networking (WiSPNET), pp 88–91
4. Guntha R, Sai Shibu NB, Rao SN (2019) Resource constrained onboard controller system for real-time fishing boat tracking in high seas. In: 2019 9th international symposium on embedded computing and system design (ISED), pp 1–5
5. Sai Shibu NB, Arjun D, Rao SN, Satish A, Navaneeth M (2018) Automatic antenna reorientation system for affordable marine internet service. In: 2018 IEEE international conference on computational intelligence and computing research (ICIC), pp 1–4
6. Do H-N, Shih W, Ha Q-A (2020) Effects of mobile augmented reality apps on impulse buying behavior: an investigation in the tourism field. 6(8):e04667. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2405844020315115>
7. Peng Q, Matterns J-B (2016) Enhancing user experience design with an integrated story-telling method. In: Marcus A (ed) Design, user experience, and usability: design thinking and methods. Springer International Publishing, Cham, pp 114–123
8. Priestersbach A, Springer T (2004) Quality attributes in mobile web application development. In: Bomarius F, Iida H (eds) Product focused software process improvement Berlin. Springer, Berlin, pp 120–130
9. Tides4fishing. [Online]. Available: <https://tides4fishing.com/>
10. India meteorological department. [Online]. Available: <https://mausam.imd.gov.in/>
11. Indian national centre for ocean information services. [Online]. Available: <https://incois.gov.in/>
12. Fish weather. [Online]. Available: <https://fishweather.com/>
13. Fishing tackle store. [Online]. Available: <http://www.fishingtackles.in/>

14. National oceanic and atmospheric administration. [Online]. Available: <https://www.noaa.gov/>
15. Fishing Reminder. [Online]. Available: <https://www.fishingreminder.com/>
16. Baywind.com.au. [Online]. Available: <http://www.baywind.com.au/>
17. Marine fish sales. [Online]. Available: <https://www.marinefishsales.com/>
18. Maxcatch. [Online]. Available: <https://www.maxcatchfishing.com/>
19. Fishermans world. [Online]. Available: <https://www.fishermansworld.net/>
20. Gps fishing app. [Online]. Available: <https://fishingpoints.app/>
21. Ffma. [Online]. Available: <https://tinyurl.com/y69ay3qd>
22. Nautide. [Online]. Available: <https://tinyurl.com/y6ydl4nw>
23. Boatus. [Online]. Available: <https://tinyurl.com/y23ryka5>
24. Netfish. [Online]. Available: <https://tinyurl.com/ybrx8gl5>
25. Odaku. [Online]. Available: <https://tinyurl.com/yxl7zwwn>
26. Fishing spots. [Online]. Available: <https://tinyurl.com/yynlfhup>
27. Fishangler. [Online]. Available: <https://tinyurl.com/y6b64t39>
28. Boating marine & lakes. [Online]. Available: <https://tinyurl.com/y2leydhw>
29. Thoondil. [Online]. Available: <https://tinyurl.com/y5py4sod>
30. Wefish. [Online]. Available: <https://tinyurl.com/y39xlfvl>
31. Fishing and hunting solunar time. [Online]. Available: <https://tinyurl.com/k2zbv2q>
32. Deeper smart sonar. [Online]. Available: <https://tinyurl.com/y51g7ytI>
33. Fishing calendar. [Online]. Available: <https://tinyurl.com/y2ytqfah>
34. Fishing times free. [Online]. Available: <https://tinyurl.com/y5r6zx93>
35. Fishweather. [Online]. Available: <https://tinyurl.com/y65r9b28>
36. Mkrishi® fisheries. [Online]. Available: <https://tinyurl.com/yyy6uu2u>
37. Infish. [Online]. Available: <https://tinyurl.com/y2kkddd6>
38. Sagara. [Online]. Available: <https://tinyurl.com/y5rt3ync>
39. Pfz advisory. [Online]. Available: <https://tinyurl.com/yyuoa8ha>
40. Incois. [Online]. Available: <https://tinyurl.com/yxnh4r2r>
41. Anglr fishing app. [Online]. Available: <https://tinyurl.com/y5lykkcl>
42. Fishidy. [Online]. Available: <https://www.tinyurl.com/y4tnprk4>
43. Pro angler fishing app. [Online]. Available: <https://www.tinyurl.com>

# Study of Machine Learning Classifiers for Intrusion Detection System



Akshita Mishra and Archana Thakur

**Abstract** Intrusion detection systems are an important part of our network security system. Furthermore, with the development in technologies, new attacks have been developed and attack types have also changed, hereby affecting more people and organizations. Cyber security research studies have been working on strengthening intrusion detection systems and intrusion detection methods. Machine learning has become a boon for this area also and is being actively applied to improving intrusion detection systems. Zero-day attacks or novel attacks have become a major problem. Therefore, machine learning methods need to be trained with newer attacks for detecting novel attacks quickly. Hence, we have used the CICIDS 2017, dataset which consists of many cutting-edge intrusions for testing and training purposes. In this work, we have concentrated on Supervised Machine Learning Classifiers and have selected 8 classifiers based on various works of literature available. These classification models have been evaluated on the basis of training time, testing time, train score, accuracy, and area under curve. As the result of this work, it has been found that decision trees and random forest have achieved an accuracy of almost 99% and within a considerable time limit. This work of ours is mainly concentrated on the use of classifiers for the intrusion detection system.

**Keywords** Intrusion detection system · Machine learning · CICIDS2017 · Ada boost · Decision tree · Random forest

## 1 Introduction

With the advent of new networking technologies and techniques, for increasing connectivity in whole of the world and making our lives easier, Internet has become an essential part of our everyday life whether it's a business or an individual. However,

---

A. Mishra (✉)

Acropolis Institute of Technology and Research, Indore, Madhya Pradesh, India  
e-mail: [akshitamishra1224@gmail.com](mailto:akshitamishra1224@gmail.com)

A. Thakur

School of Computer Science & IT, Devi Ahilya University, Indore, Madhya Pradesh, India

it has also increased the risk of compromising our network's and system's security. Although, there are lots of tools available to us for our rescue, like firewalls that prevent unauthorized traffic to enter our network, spam filters that filter out unwanted email messages, and anti-malware tools that protect end-points from malware. These systems are still not enough and hence, we need another security tool, which is an intrusion detection system. An intrusion detection system (IDS) is an important constituent of Network Security, monitors network traffic and aids in discovering, determining, and identifying unauthorized use, duplication, alteration, and destruction of information and information systems [1]. The whole concept of an Intrusion Detection System lies on the premise that the system or network under attack shows different and distinguishable patterns in comparison to the normal ones. Hence an IDS, monitors the network for any such distinguishable and suspected patterns or activities and issues an alert to either the Network administrator or Central Security system, whenever any suspicious activity occurs. Intrusions are basically of two types: Active attacks and Passive attacks. These attacks compromise availability, confidentiality, and integrity of information and network. In active attacks, the attacker tries to modify the content of messages. It is important to detect active attacks as they affect the integrity and availability of the network. They interrupt normal behavior of the network, like Denial of Service attack, Wormhole attack, Sinkhole, and Spoofing attacks. In passive attacks, the content of messages remains unchanged and is usually based on eavesdropping. The attacker observes the network traffic, copies them, and uses them for malicious purpose. It is important to prevent passive attacks as they attack the confidentiality of information system.

Intrusion Detection Systems can be classified on the basis of their usage and method of detecting intrusion [2]. On the basis of usage, IDS can be host-based and network-based. Network-based IDS keeps tab on network and analyze its packet traffic. It observes traffic through network devices. Networks based intrusion detection system (NIDS) works for real-time traffic and uses both signatures-based and anomaly-based intrusion detection methods. It generates an alarm whenever an intrusion is detected and stores all information in logs. Host-based intrusion detection system (HIDS) monitors intrusion on the basis of the system's configuration and application's activity. It is used to detect intrusion related to the software environment of a specific host. HIDS looks for abnormal behavior logs on the host system, therefore individual IDS is required for each host or system.

There are two types of IDS based on their Intrusion detection approach. Signature-based IDS and Anomaly-based IDS. Signature-based IDS, also known as Mis-use based IDS or Rule-based IDS, has been designed to detect known attacks based on the already available signatures of attacks in the form of rules defined in their database. Their advantage is that they are very effective in detecting known attacks and generate very few false alarms. And their disadvantage is that the stored signature of attacks needs frequent manual updates of rules and signatures of attacks in the database. Therefore, Signature-based IDS cannot be used for detecting novel attacks or zero-day attacks. Anomaly-based IDS is more prompt these days because of their quality to detect zero-day or novel attacks. This is because anomaly-based IDS detects intrusion on the basis of deviations from normal behavior, in this profile it is

defined for a group of users or each user individually. These profiles can either be created manually or dynamically and serves as a base for distinguishing normal and abnormal behavior. These profiles also aid in increasing the complexity for attackers to assess undetected actions for any particular user, system, or network, since these profiles can be particular to users. The disadvantage of anomaly-based IDS is that it has a high rate of false alarms. Both anomaly- and signature-based IDS have their advantages and disadvantages and a hybrid of both of these can be more efficient as data or log from anomaly-based IDS can be used for updating the rules and signature in signature-based IDS, thereby decreasing the false alarm rate for unknown attacks.

Machine learning is the branch of knowledge that aims to give computers or machines the potential to learn and improve from experience without being explicitly programmed. The main aim of Machine Learning is to develop programs or models which can use data to discover useful patterns and can make useful predictions based on examples provided on the basis of past data. That means computers should be able to learn without human assistance and intervention and can adapt to changing situations.

Machine Learning has the capability to identify hidden patterns in large amounts of data and this capability has become a boon for IDS. With the growing technologies, there are new attacks frequently and traditional methods of detecting intrusions are not enough for detecting novel and real-time attacks. Hence, we need to apply the power of machine learning to explore new patterns and identify novel and real-time attacks. Thus, adapting to machine learning for Intrusion detection is a revolutionary step in ensuring a more secure network. Using machine learning for intrusion detection has resulted in higher detection and lower false alarm rates and has also increased efficiency [3].

Machine learning models' performance largely depends on the data, which is being given as an example, i.e., on training data and on data used for checking its validity and performance, i.e., testing data [4]. Currently, most of the research done in the area of using machine learning for intrusion detection uses DARPA 98/99 and KDD Cup 99 datasets for training, which is not advisable now. The issue with these datasets is that they lack examples of newer attacks and only consist of traditional attacks. Hence, Intrusion detection models trained using these datasets are not able to perform well in today's scenario. CICIDS2017 dataset created by the Canadian Institute of Cyber security (CIC) has become the new favorite of researchers for training machine-learning-based IDS [5]. CICIDS 2017 consists of both benign and cutting-edge common attacks, aiming to match-real world data. The duration of the data collection process was from Monday, July 3, 2017 to Friday, July 7, 2017, i.e., for 5 days. And during this period, a total of 2,25,746 records were generated with 80 features such as Protocol, Flow ID, Source IP, Destination IP, Flow Duration, Total FWD Packets, etc. The attacks applied were Brute Force FTP, Brute Force SSH, DoS, Heartbleed, Web Attack, Infiltration, Botnet, and DDoS.

The paper is organized as follows. Section 2 consists of a literature review that uses machine learning for Intrusion detection. The paper has been selected from a vast repository of Google Scholar using keywords 'Machine learning' and 'Intrusion Detection' published in the years 2018, 2019, and 2020. The focus has been also

given to research works with CICIDS 2017 dataset. Section 3 provides a detailed description of the Dataset used and its preprocessing. In Sect. 4, a brief description of machine learning classifiers and the methodology used for this study has been given. In Sect. 5, the discussion of the performance results of methods used has been done on the basis of evaluation metrics used. And finally, Sect. 6 summarizes the paper.

## 2 Literature Review

In this section, we have presented literature since 2018 and they have been gathered from IEEE explore and science direct, using keywords machine learning classifiers and intrusion detection system”.

KDD-Cup 99 has been a very popular dataset for training machine learning models for the intrusion detection system. Krishna et al. in their work [7] has aimed at reducing false positive alarms in IDS. For this, they have used an artificial neural network with different optimizers. They also used particle swarm optimization and genetic algorithm for feature selection. The model proposed by them showed improved individual class detection and reduced false-positive alarm rate. They used accuracy, precision, recall, and f-measure for evaluating the performance of their proposed model. The authors in [8] have also used KDD-Cup 99 dataset for proposing an optimally approached anomaly NIDS. They used backpropagation neural network and optimization techniques for feature selection. On the basis of performance evaluation done using detection rate, false-positive rate, and f-score, they found that the proposed method showed improved performance. Anwer et al. in [6] have concentrated on yet another direction, they have presented a feature selection framework that aims to use minimum features for training the model. They have used J48 and naïve Bayes for showing the effect of using the GR ranking method for feature selection. They identified that J48 has achieved an accuracy of 88% using 18 features. Zhang et al. in their work [9] have addressed the issue of anomaly detection in wide area network meshes using two machine learning algorithms, Boosted Decision Tree and Simple feed-forward neural network. They used data gathered by the Open Science Grid using perfSONAR servers along with these algorithms. After observing the AUC score, they identified that boosted decision tree showed better performance. Another study has been done by Sahani et al. in [10], using C4.5, improved C4.5, CART, and ID3 on KDDCup 99 dataset for classifying intrusion detections. They used performance time, accuracy rate, and error rate for evaluating the performance of each above-mentioned classifier on the KDD Cup 99 dataset and concluded that improved C4.5 showed the best performance amongst selected classifiers. According to the study done by Mishra et al. in [11], the authors have proposed a bfs-nb hybrid model for intrusion detection, which uses the best-first search method for feature selection and naïve Bayes as machine learning classifier. They compared the accuracy, sensitivity, and time delay with and without feature selection and recorded improved results. Aksu et al. in [12] have done a comparative analysis of IDS with support vector machine, decision tree, and k nearest neighbor along with fisher score



for feature selection. They used the CICIDS2017 dataset for training their classifiers, which have many new attacks and measured their performance on the basis of accuracy, recall, precision, and f-score. They identified that the decision tree and support vector machine performed well among all. Sahil et al. in their work [13] proposed an SVM-based ensemble approach to develop an intrusion detection model aimed at distinguishing attacks. They used the ensemble approach of supervised (SVM) and unsupervised (K-Means) over three popular datasets KDDCup99, NSL-KDD, and GureKDD. The implementation has been done on Weka and performance metrics considered are recall, sensitivity, specificity, and accuracy. The authors in [14] have proposed a lightweight IDS using a multi-layer perceptron neural network and gain ratio for feature selection. On the basis of performance metrics precision, accuracy, test time, and confusion matrix, the authors showed that the technique adopted is promising and could be used for real-time intrusion detection. The technique of detecting attacks by using a combination of supervised learning and unsupervised learning has also been used by the authors in [15]. They have used k-medoids and decision trees and evaluated the performance of the proposed model on the basis of accuracy, training time, and testing time. The authors concluded that the proposed method maintained the detection rate and reduced detection time.

### 3 Methodology

In this work, we aim to make a comparative study of supervised machine learning techniques on the CICIDS 2017 dataset. Implementation has been done in Python and the dataset has been preprocessed for missing values. Classifiers to be studied are Ada boost, decision tree, random forest, naïve Bayes, logistic regression, k nearest neighbor, and kernel support vector machine.

#### 3.1 *Machine Learning Classifiers*

Following machine learning classifiers explained below have been used in this work. These classifiers have been selected for our study on the basis of different literature available. Since all these are traditional algorithms, a short description of each is given.

##### 3.1.1 **Ada Boost**

Boosting algorithms are being used recently for predictive analysis projects a lot. Boosting uses a series of weak learner algorithms. Ada boost is the first boosting algorithm, which combines multiple weak classifiers to form a strong classifier [15]. It uses a decision tree with a single split, called a decision stump, as weak classifiers.

It applies more effort to those instances which are difficult to classify instead of those that are already classified.

### **3.1.2 Decision Tree**

The decision tree is one of the most popular classifiers used in the real world. Its improved versions have also been used in real-life examples. It can be applied for classification and regression both. As the name implies, the decision tree uses a tree-like structure for demonstrating its decision and decision-making process [16]. Leaves of decision trees represent decision and branches represent how that decision has been reached. The main advantage of the decision tree is its easy implementation and high accuracy. But sometimes it can lead to overfitting and may not work well with imbalanced data.

### **3.1.3 K-Nearest Neighbor**

K-nearest neighbor is a supervised learning approach that is easy to implement and can be used for solving both classification and regression problems. In this, the learning is based on how similar is a data to another data. It uses the Euclidean distance formula to evaluate the distance between new instances of data and already available instances. After sorting, neighbors close to each other are selected based on minimum Euclidean distance. After that, the first K-sorted entries are selected and categories are formed. K-NN is easy to implement, versatile and simple but lacks a bit in the case of multi-dimensional datasets.

### **3.1.4 Kernel SVM**

SVM is a powerful machine learning technique but it cannot work well with nonlinear data. To handle nonlinear data, the kernel trick is being used. The idea is to map data in a high dimensional space such that it becomes linear and then SVM is applied. This method can be used for both classification and regression.

### **3.1.5 Logistic Regression**

Logistic regression is another machine learning technique that is being used for classification and predictive analysis. It uses the concept of probability. It can also be used for feature extraction. It is mainly used when the target variable is categorical in nature. Logistic regression classifier is easy to implement and takes less training time but is not advisable for high dimensional data as it can lead to overfitting.

### 3.1.6 Multi-layer Perceptron

Multi-layer perceptron is an add-on to the feed-forward network. It consists of input, output, and hidden layers. And can be used for predictive analysis and classification problems. It is very simple and gives high accuracy when the numbers of hidden layers are less for linear data.

### 3.1.7 Naïve Bayes

Naïve Bayes is one of the most powerful approaches used for predictive modeling. It is an application of the Bayes theorem and is highly advisable for binary or multi-class classification. It calculates the probability of every input feature and the feature with the highest probability is selected. Its modeling time is less, which means it learns fast and also gives high accuracy when used with textual data. Its disadvantage is that it treats each feature individually as if they are not related, which actually doesn't happen in real-world scenarios.

### 3.1.8 Random Forest

Random forest is a supervised machine learning method, which is an ensemble of the decision tree. It is a tree-based learning algorithm and can be used for both regression and classification. It combines the results from different decision trees to give the final result. It can also be applied for feature selection. It is easy to understand and can work with large datasets. Usually, it produces high accuracy and maintains performance in case of missing data also. Its disadvantage is that it can overfit some datasets.

## 4 Results and Discussion

The results evaluated for the above-mentioned machine learning classifiers are explained in this section with the help of plots. Based on the values of training time, testing time, train score, test score, and area under curve, it has been identified that the decision tree and random forest have the best performance overall which is indicated by the accuracy parameter. Then multi-layer perceptron also shows very good accuracy, which is lesser than decision tree and random forest but better than others. Multi-layer perceptron was also found out to be outstanding in terms of testing time and training time but lacks in terms of area under curve. Whereas, decision tree and random forest show better performance with respect to area under curve, which indicates that they are able to distinguish between benign and attack examples most of the time.

The graph shown above represents the time taken by selected machine learning classifiers to model on the training data provided. According to the graph shown above, we can observe that time taken by the naïve Bayes classifier is 0.371 s which is the least of all. Other better algorithms are random forest and decision tree in terms of time taken for training data. On the other hand, all other algorithms have taken considerable time for training the model and among them, kernel SVM has taken the largest time for training, which significantly affects the efficiency of the training model. It can be concluded from Fig. 1 that naïve Bayes, decision tree, and random forest can be the best classifiers when we don't want the system to spend the major portion of time on training the model.

The above-shown graph is representing the time taken by the classifiers, in the scope of this work, to test the performance of the trained model on test data. As it can be seen in Fig. 2, decision tree, random forest and logistic regression classifiers have taken the least amount of time to perform on test data, 0.01 s, 0.02, and 0.07 s, respectively. Among the remaining classifiers, a multi-layer perceptron can also be considered with a testing time of 0.13 s. It is also observed that k-NN and Kernel SVM took the highest amount of time 45.60 s and 25.09 s, respectively. According to Fig. 2, we can conclude that decision tree, logistic regression, and random forest are the best classifiers when we want to take testing time into account.

Train scores and test scores can help us in understanding overfitting and underfitting in the model. A higher value of training score and low value of test score indicates overfitting in the model, whereas a low training score and low test score indicate underfitting in the model. So, analyzing both train score and test score we can say that decision tree, random forest, multi-layer perceptron, and k nearest neighbor are the best choices among the available ones. They have shown high test scores and high train scores. Naïve Bayes have shown low test scores and low train scores and thereby suffering from underfitting (Figs. 3 and 4).

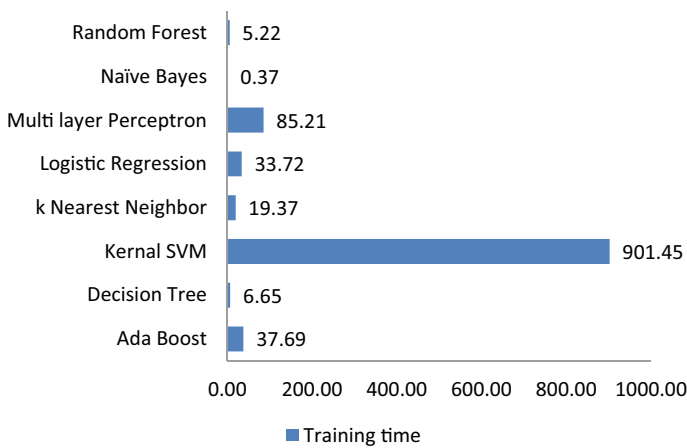


Fig. 1 Training time

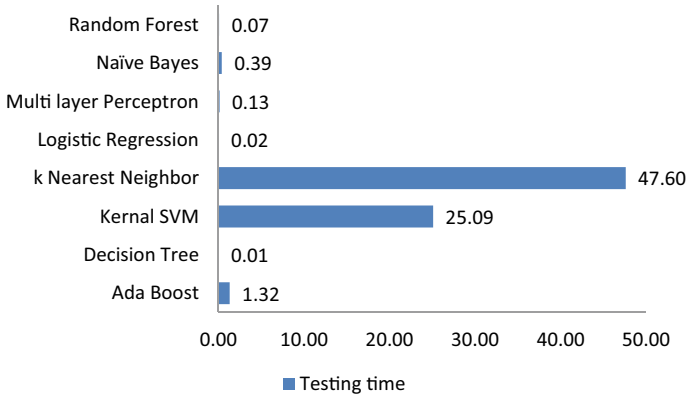


Fig. 2 Testing time

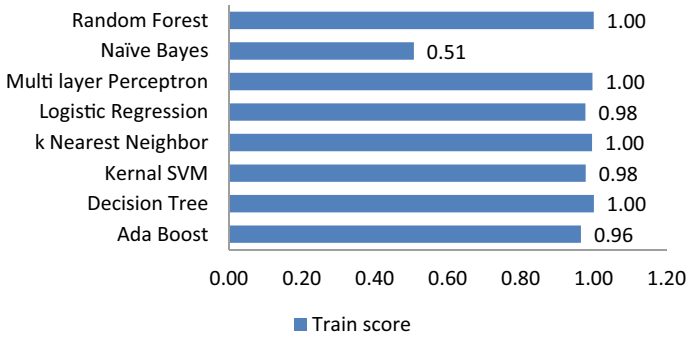


Fig. 3 Train score

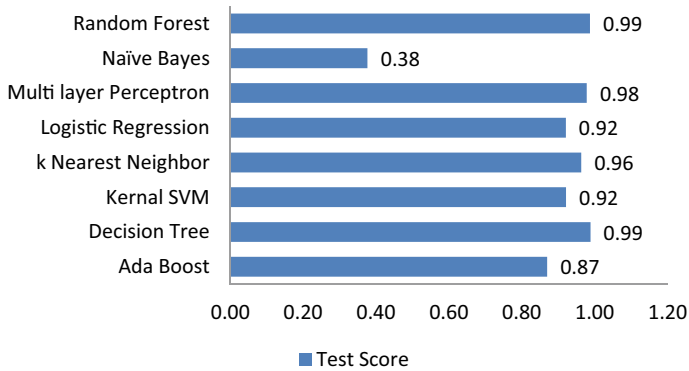


Fig. 4 Test score

A higher value of area under curve is desirable because the better the AUC value the better the performance of a model in terms of distinguishing between benign and malicious data. Looking at the plot in Fig. 5, it can be observed that for decision tree, random forest and Ada boost, the area under curve is more than 80% (0.88, 0.85, and 0.83, respectively), which has quite a good performance as they are able to distinguish between the classes most of the time.

Accuracy is not the only factor for evaluating a model but it is an important factor. As it can be observed from the plot mentioned in Fig. 6, decision tree, random forest, and multi-layer perceptron have accuracy above 95% (0.99, 0.99, and 0.98, respectively). Out of the remaining classifiers, k-NN has also performed well.

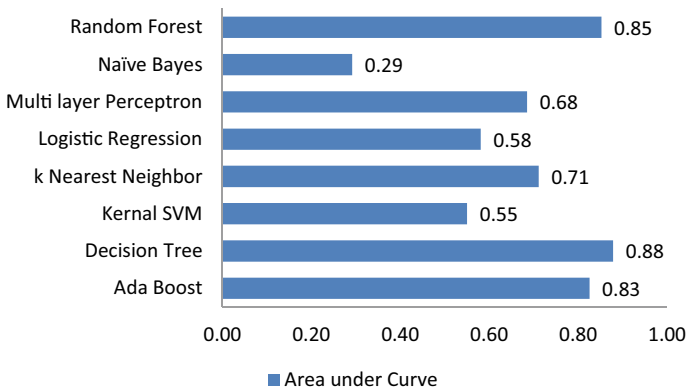


Fig. 5 Area under curve

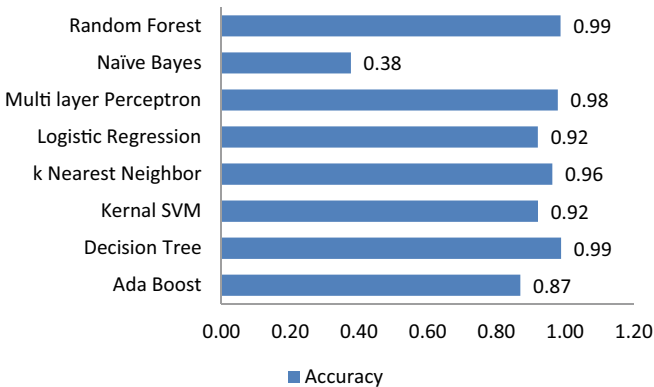


Fig. 6 Accuracy

## 5 Conclusion

Based on the analysis made till now, using individual evaluation plots, we have identified that decision tree and random forest have performed well. In each of the performance evaluation criteria, these two classifiers have outperformed other classifiers. The classifier which took minimum training time is random forest and the maximum time is kernel SVM. The classifier which took minimum testing time is the decision tree and the maximum time is k-NN. The classifier with the best train score and test score are decision tree, random forest and with the worst performance is naïve Bayes. The classifier for which area under curve is high is the decision tree and for low is naïve Bayes. Accuracy has been maximum for decision tree and minimum for naïve Bayes. So, in totality we can say that decision tree, random forest, and multi-layer perceptron have performed well on CICIDS 2017 dataset and naïve Bayes cannot perform well for this. Other classifiers have managed to perform well.

Our goal is to identify the classifiers which can be further explored for working in a real-time intrusion detection system. For this, we have picked the CICIDS 2017 dataset, consisting of cutting-edge intrusion attacks and benign traffic. We have selected this dataset so that our classifiers could be trained with new evolving attacks.

## References

1. Mukkamala S, Sung A, Abraham A, Vemuri Rao V (2005) Enhancing computer security with smart technology. *IEEE Comput Intell Mag* 3(2):70–71
2. Buczak A, Guven E (2015) A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Commun Surv Tutor* 1:99
3. Zamani M, Movahedi M (2013) Machine learning techniques for intrusion detection. Cornell University
4. Robin S, Paxson V (2010) Outside the closed world: on using machine learning for network intrusion detection. In: *IEEE symposium on security and privacy*, pp 305–316
5. Sharafaldin I, Lashkari AH, Ghorbani AA (2018) Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: *4th international conference on information systems security and privacy (ICISSP)*. Portugal
6. Anwer HMH, Farouk M, Hamid AA (2018) A framework for efficient network anomaly intrusion detection with features selection. In: *9th international conference on information and communication systems (ICICS)*
7. Krishna CR (2019) A hybrid approach to mitigate false positive alarms in intrusion detection system. In: *International conference on computer networks and communication technologies*. Springer, Berlin
8. Chiba Z, Abghour N, Moussaid K, Omri EL, Rida M (2018) A novel architecture combined with optimal parameters for back propagation neural networks applied to anomaly network intrusion detection. *Comput Secur* 75:36–58
9. James Z, Gardner R, Vukotic I (2019) Anomaly detection in wide area network meshes using two machine learning algorithms. *Futur Gener Comput Syst* 93:418–426
10. Sahani R, Shatabdinalini RC, Chandrakanta BJ, Jena AK, Das H (2018) Classification of intrusion detection using data mining techniques. In: *Progress in computing, analytics and networking*, Springer, Singapore, pp 753–764

11. Mishra S et al (2019) Implementation of bfs-nb hybrid model in intrusion detection system. *Recent developments in machine learning and data analytics*, Springer, Singapore, pp 167–175
12. Aksu D et al (2018) Intrusion detection with comparative analysis of supervised learning techniques and fisher score feature selection algorithm. In: *International symposium on computer and information sciences*, Springer, Cham
13. Sahu K et al (2019) An SVM based ensemble approach for intrusion detection. *Int J Inf Technol Web Eng* 14(1)
14. Mebawondu JO et al (2020) Network intrusion detection system using supervised learning paradigm. *Sci African* 9.
15. Schapire RE (2013) *Explaining adaboost*. Empirical inference, Springer, Berlin, pp 37–52
16. Safavian SR, David L (1991) A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern* 21:660–674



# The Analysis of Time Series Data



**Charu Kathuria, Deepti Mehrotra, Shalini Bhartiya,  
and Navnit Kumar Misra**

**Abstract** Patents contribute to technological development by providing insights into valuable information. In order to use the potential of patent documents and obtain significant content from them, it is a prerequisite to analyze structured as well as unstructured data. Patent mining is an add-on to data mining as here the tools are designed to handle unorganized and semi-organized data sets although data mining is mainly engrossed in analyzing the organized data. Owing to the enormous size of datasets of patents along with the incorporation of data found in the patent from diverse sources, it turns out to be nearly impossible and time-consuming to review relevant data from a researcher’s point of view. The discussion in this paper is based on a cancer data set that is used to categorize and invest in the true and growable technologies. By finding trends from the data set, this information helps in understanding prospects in the field of technology that existed former to it and in current use.

**Keywords** Trend analysis · Patent · Technology · Cancer domain · Time series analysis

## 1 Introduction

In the arena of patents, the term “trend analysis (time series analysis)” is a vital analysis that provides an insight into the research and innovation trends in a particular domain. Finding a trend in a set of patents is accomplished to get a full view of all this information or data available to us. A patent signifies the originality of work in any

---

C. Kathuria · D. Mehrotra (✉)  
Amity School of Engineering and Technology, Amity University Uttar Pradesh, Uttar Pradesh,  
Noida 201313, India  
e-mail: [mehdeepti@gmail.com](mailto:mehdeepti@gmail.com)

S. Bhartiya  
School of IT, Vivekanand Institute of Professional Studies, Pitam Pura, Delhi 110034, India

N. K. Misra  
Department of Physics, Brahmanand College, The Mall, Kanpur 208004, India

technical domain that gives comprehensive facts and data about the novel invention, including assignees' information, inventors, or owners, a technical area to which it belongs, publication details like dates of filing and granting IPC codes, etc. These numbers of patents are warehoused in numerous databases [1] across the globe like USPTO (the United States Patent and Trademark Office), WIPO (World Intellectual Property Organization), EPO (the European Patent Office), etc.

The trend analysis would provide a commercial perspective to patent analysis. Business units, researchers, and educational institutions can utilize this study to find new areas and trends in the patent. It would certainly enable a better understanding of technology. The technology roadmap analysis can be retrospective and prospective, i.e., analysis of patents filed in previous decades, present, and in the future direction like time series analysis, patent trend analysis, and citation analysis [2].

In time series analysis, the growth of technology in the past and its market implementation needs to be studied. Trends visible in the current market help in forecasting future innovations for a particular technology. It helps in understanding the growth/decline rate of various technologies over time.

The objective of this paper is to achieve and discover trends in the set of patents that (deal with cancer) to better understand on rich workload [3] dataset. The purpose of this research is to provide an instantaneous summary of a selection of approximately a few of the high-quality published patents using R with a computational time series research.

## 2 Literature Review

The information is gained from various research papers from recognized organizations. Analysis of Patent data extracts knowledge from the information stored in it. Literature has been searched in different domains [4] to identify the trend analysis and patent's contributions in a varied field of research.

### 2.1 *Biopharmaceutical Community*

Research has been conducted for analyzing cancer patents, especially that are related to immunotherapy, which is believed to make cancer treatment easy [5]. USPTO-granted patents of cancer are considered. 13 subfields are identified in these patents. The development of patent number growth in this domain has surpassed the background rate, with cytokine-related therapies, natural killer cell therapies, and immune checkpoint inhibitors growing at a rapid rate. The top 15 assignees have 27.6% (616) of the patents.

The experimental results show an ecosystem where the big industries and small-scale player industries each hold a niche. Succession and selection are expected to work in the long run in this young ecosystem.

## ***2.2 Statistical Computing and Programming Environment***

The core features of the R language for analyzing the basic time series are outlined [6]. Some intermediate level and advanced topics in time series analysis supported in the R language are discussed, such as state-space models, structural change, and generalized linear models. The result of this work is to understand the working of the R and R-studio software for the analysis of a dataset.

## ***2.3 Roadmap***

Most of the existing methodologies focus on market-driven road mapping. This [7] is a relatively robust method of quantity analysis and takes into account technology life cycles. Initially, patent data are collected on relevant technologies and citation information. The author talks about a patent quotation network comprising nodes (single patents) and arcs between nodes (quotations between patents) developed using these data. Secondly, consideration of the year of the patent application and the life cycle of technology is the phase of a selected technology during the life cycle. Finally, a roadmap for technology is drawn by linking these nodes into a layer of technology and estimating the time of development.

## ***2.4 Popularity of the Internet***

While general searches, such as those provided by Bing/Google/Yahoo, etc., are powerful, metadata could be a useful tool for searching for specific purposes. A tool [8] called the search guide is provided to help user to find relevant information on the basis of individual requirements in Learning Objects. The author proposes the concept of the “Reusability Tree” as a contribution.

## ***2.5 Spatial Associations and Trends in Mobility***

The era of large information can be readily accessed through massive streams representing the different levels of motions and communications. Existing stream clustering surveys, however, ignore the geometric characteristics of fluid information, for example, path and duration that show significant patterns in moving. In the present article, the author presents fresh readings of spatial and temporal resemblance [9] between flows and suggests a fresh, gradual-strategy, clustering method of stream information. This technique identifies clusters from different stream ranges and detects major spatiotemporal patterns from high stream information.

## ***2.6 Proxy for Inventive Product***

The research [10] has been based on the long-standing theory that the quote obtained by the applicant represents a proxy for the flow of information or the impact of the copyright. By designing the predictive system of a point-based patent quotation (self-quotation and non-self-quotation) that includes multiple patent data, we are opening up the option of carrying out a predictive evaluation of the patent, which is particularly helpful for applications with evolving technology recently awarded.

## ***2.7 Understanding of Technology***

It is very demanding to have automatic instruments to help innovators and patent technicians to get helpful data from patent papers. Text mining could be used to rapidly and automatically evaluate these text files and extract helpful data from big volumes of papers. A computerized method [11] is suggested to extract helpful data from patent papers in accordance with the TRIZ Inventive Principles.

Although the term trend analysis is frequently used to predict future events (trends), this analysis could also be used to estimate uncertain events in the past, such as how many patents were probably used or granted in ancient times regarding cancer, based on data provided such as the average years which other known patents reigned over the years.

With various analytical tools, trend analysis can be performed. R is the software used for this project. The R programming language is widely used by data miners and statisticians to develop statistical software and data analysis. Polls, data miner surveys, and scholarly literature database studies show that in recent years R's popularity has increased significantly.

## **3 Methodology**

Trend analysis comprises the main term trend that determines progressive changes in a system. The system is characterized by a sequence of measurements distributed over time. It emphasizes the current or future situation based on past details. Thus, the historical data when analyzed determines the trend direction. It's a statistical tool that emphasizes the future movements for a variable driven by historical trends. Trend analysis faces some difficulties due to inconsistent past data. So as if there is no limit for data to define the trend however the larger the data the better it covers the significant elements. The direction of the trend can be up or down showing the favorable or not favorable events. The overall objective of this analysis is to make a decision with past experience.

Different types of patents are being analyzed and have fascinated the research scholars and technocrats in the past two decades and ample work and content are presented in them, for which the patent data warehouse is considered to become the anticipated point for analysis.

Figure 1 shows a glimpse of a very large sample dataset that comprises more than 25 attributes and the patents that represent their domain are analyzed with respect to time (when these patents are granted that is the Grant Date) to find the trend in the unambiguous domain (such as Drugs and Chemistry, Diagnostic and Surgical Devices, Radiation Measurement, Data Science, Cells and Enzymes, Food and Nutrition, Model Systems and Animals, Other and Pre-classification, DNA RNA or Protein Sequence.) This is a sample dataset and this approach is used to find trends that can be applied to live datasets and also for analyzing them. A patent contains many information that is little difficult to extract.

## 4 Result and Observations

The trend analysis of the cancer data set comprising approx. quarter of a million patents in the domain of cancer and its treatment is performed. The patents were subdivided into nine attributes that majorly play role in this domain. Figures 2, 3, 4, 5, 6, 7, 8, 9 and 10. Depict the different trends of each particular technological attribute over the centuries from where the birth of patents started to how the growth in a particular domain of technology beamed at an increasing rate over the years. Mostly all technological attributes depict a sudden rise in drugs and diagnostics of cancer from the year 2000. New techniques and technologies were emerging at the start of the second millennium (2000 onwards).

Figures 2, 3, 4, 5, 6, 7, 8, 9 and 10 are the trend plots that are formed using the time series (ts) function of R programming. The time series function plots the time variable on the x-axis and the observed variable on the y-axis. The two indexes related to ts() function are frequency and start. Here frequency represents in number, the cycles per time unit, and start index includes the beginning of time variable which is presented on the x-axis of the plot. The plot.ts() function is specifically designed for the time series object which directly takes the x value from the ts() function. The dataset contains the research done in different areas of cancer research like patents in particular areas of drugs as shown in Fig. 2 and that of surgical and diagnostics equipment is shown in Fig. 3.

Similarly, how many patents are awarded for radiant measurement, and data science are given in Figs. 4 and 5, respectively.

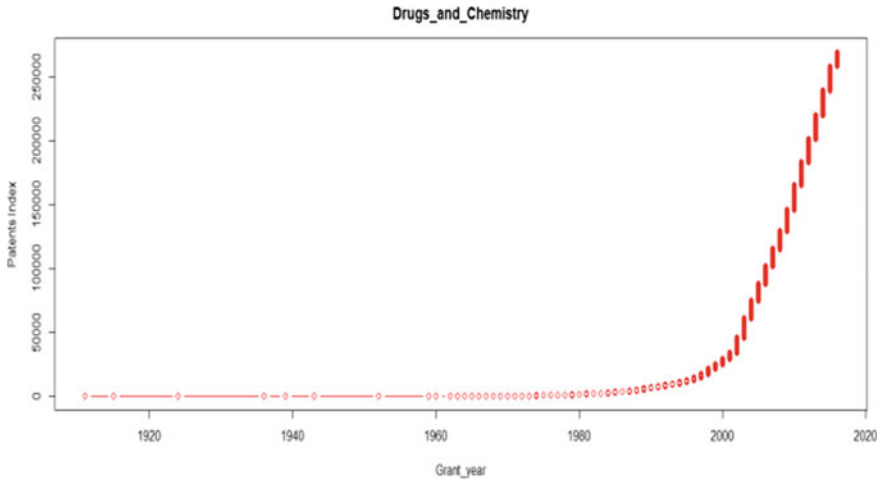
In Figs. 6 and 7, patent trends in the area of food and nutrition and model system and animals are shown.

Lastly, research trends in DNA and RNA reflected through patents awarded after 2000 are shown in Fig. 10.

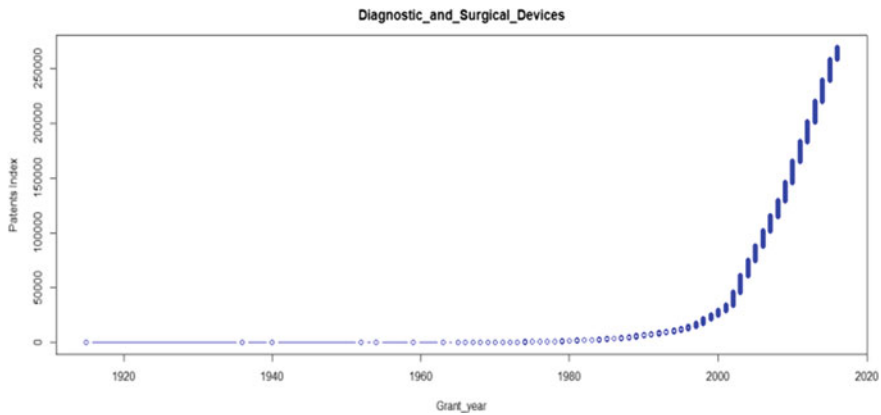
Some areas of the cancer domain have patents that are less in number (i.e., are still at the beginning stages like the DNA and RNA or protein sequence because less

Index	Family	Patent	Patent	Applic	ion	De	at	Grant	Publ	at	CPC	In	CPC	Ac	IPC	Pr	IPC	Se	USPC	USPC	Patent	Drugs	Diagn	Radiat	Data	S	Food	Model	Cells	a	Other	DNA	R
1	306593	US 0994660 A				1.9E+07	1911				A61K39/00	Y105435/931							424/234.1;424/269.1		TEXT NOT	1	0	0	0	0	0	0	0	0	0		
2	3219608	US 1151536 A				1.9E+07	1915				C07K16/18	G01N33/57492							424/174.1;424/130.1		TEXT NOT	1	1	0	0	0	0	0	0	0	0		
3	2.5E+07	US 1500985 A				1.9E+07	1924				C07K16/30	A61K2039/505							C07K16/1;424/172.1;424/559.1		Method f	1	0	0	0	0	0	0	0	0	0		
4	2.4E+07	US 1795580 A				1.9E+07	1931				A61D1/00								606/209		Test-tumc	0	0	0	0	0	0	0	0	0	1		
5	2.4E+07	US 1837503 A				1.9E+07	1931				A61D1/02								A61D1/02;606/159		Test mem	0	0	0	0	0	0	0	0	0	1		
6	2.4E+07	US 2036649 A				1.9E+07	1936				C07K16/00	G01N33/53							C07K16/0;424/9.81		435/29.4; Improvem	1	1	0	0	0	0	0	0	0	0		
7	2.4E+07	US 2165370 A				1.9E+07	1939				A61K39/0005	A61K39/00							424/9.81		436/543	1	0	0	0	0	0	0	0	0	0		
8	2.2E+07	US 2165371 A				1.9E+07	1939				A61K39/0005	A61K39/00							424/9.81		436/543	1	0	0	0	0	0	0	0	0	0		
9	2.2E+07	US 2194131 A				1.9E+07	1940				G01N33/53	G01N33/53							435/7.35		436/536; System for	0	1	0	0	0	0	0	0	0	0		
10	3895789	US 2310337 A				1.9E+07	1943				A61K35/742								A61K35/7;424/573		Pain-reliev	1	0	0	0	0	0	0	0	0	0		
11	2.2E+07	US 2591927 A				2E+07	1952				A61B10/02	A61B10/00							600/572		604/289	1	0	0	0	0	0	0	0	0	0		
12	4.1E+07	US 2606601 A				2E+07	1952				A47C3/12	A61K2039/5158; A61A47C3/12							424/7.3; 297/411.4		297/2; 29 Chair havi	1	0	0	0	0	0	0	0	0	0		
13	3690365	US 2697431 A				2E+07	1954				A61B1/30	A61B1/0669							A61B1/30		600/102	359/387; Microscop	0	1	0	0	0	0	0	0	0	0	
14	2.3E+07	US 2862867 A				2E+07	1958				C10G37/04								C10G37/0		208/3	208/177; Reduction	0	0	0	0	0	0	0	0	0	1	
15	2.4E+07	US 2917432 A				2E+07	1959				C07C309/65								C07C309/514/517		424/663	Leukemia	1	0	0	0	0	0	0	0	0	0	
16	2.4E+07	US 2905169 A				2E+07	1959				A61B10/0	Y105604; A61B10/00							600/572		604/15; 6 Device for	1	0	0	0	0	0	0	0	0	0		
17	2.4E+07	US 2953495 A				2E+07	1960				C07K7/06	Y105435/886							530/323		435/120; Antitance	1	0	0	0	0	0	0	0	0	0		
18	1.8E+07	US 3067100 A				2E+07	1962				A61K31/7	Y105435/886							424/119		435/71.3; Antibiotic	1	0	0	0	0	0	0	0	0	0		
19	2.2E+07	US 3060924 A				2E+07	1962				A61M31/00; A61M5/1016								A61M31/6		250/492.1	Apparatus	0	0	0	0	0	0	0	0	1	0	
20	2.5E+07	US 3019164 A				2E+07	1962				A61K31/5; A61K2300	A61K31/53							514/242		Urethan a	1	0	0	0	0	0	0	0	0	0		
21	2.5E+07	US 3041241 A				2E+07	1962				C07C309/66								C07C309/514/517		558/46	Method of	1	0	0	0	0	0	0	0	0	0	
22	2.5E+07	US 3057850 A				2E+07	1962				C07H15/203								C07H15/2; 536/17.6		564/446; Anti-tumo	1	0	0	0	0	0	0	0	0	0		

Fig. 1 Cancer dataset with few attributes



**Fig. 2** Trend for drug and chemistry in the dataset



**Fig. 3** Trend for diagnostic and surgical devices in the dataset

number of inventions have been discovered till now), while other domain (such as drugs and chemistry, diagnostic and surgical devices) stages have a large number of patent datasets that have grown over the years and are still growing. Figures 4, 5, and 6 have a less steep curve than Figs. 2, 3, and 9. So expecting a large number of technological advancements in these domains that can help us divert our focus onto the trend and where more work is required.

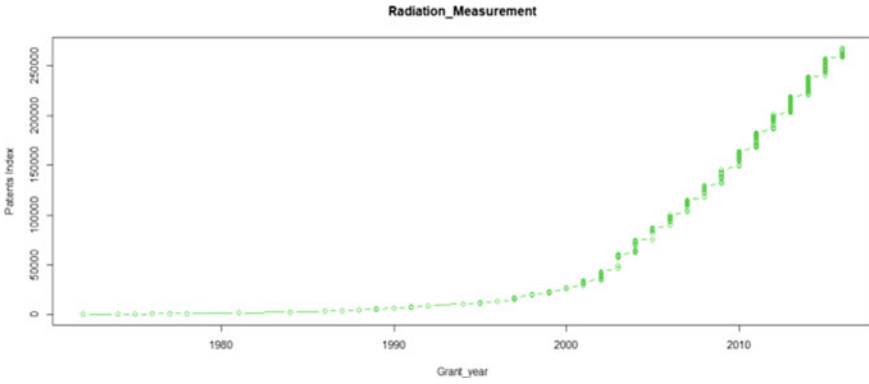


Fig. 4 Trend for radiation measurement in the dataset

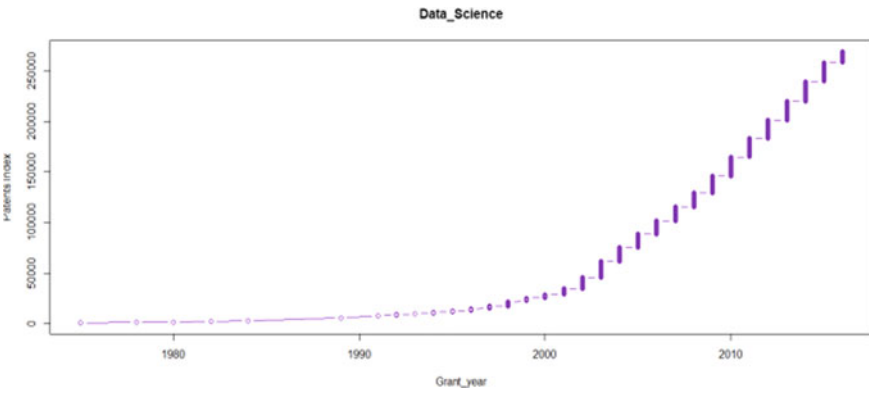


Fig. 5 Trend for data science in the data set

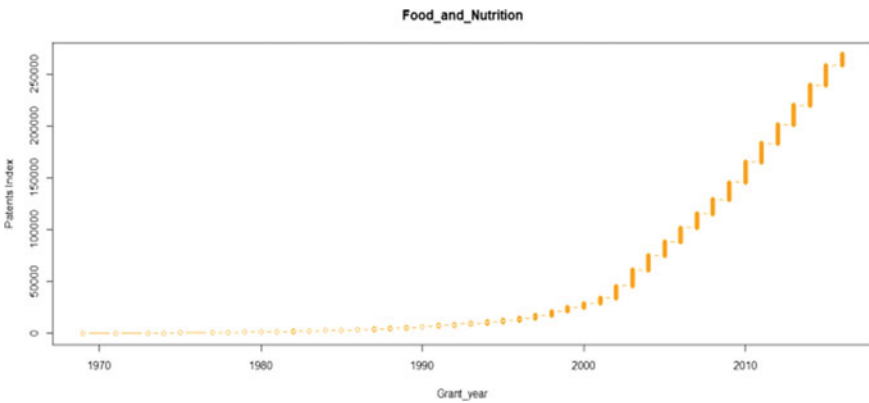


Fig. 6 Trend for food and nutrition in the dataset



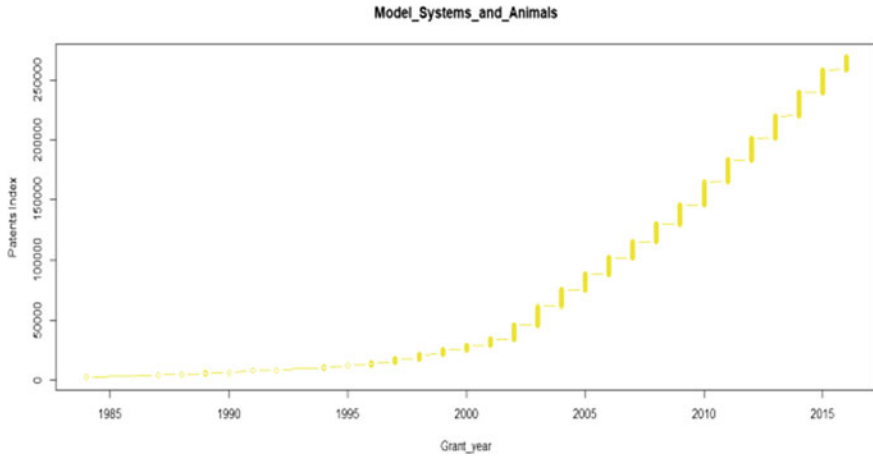


Fig. 7 Trend for model systems and animals in the dataset

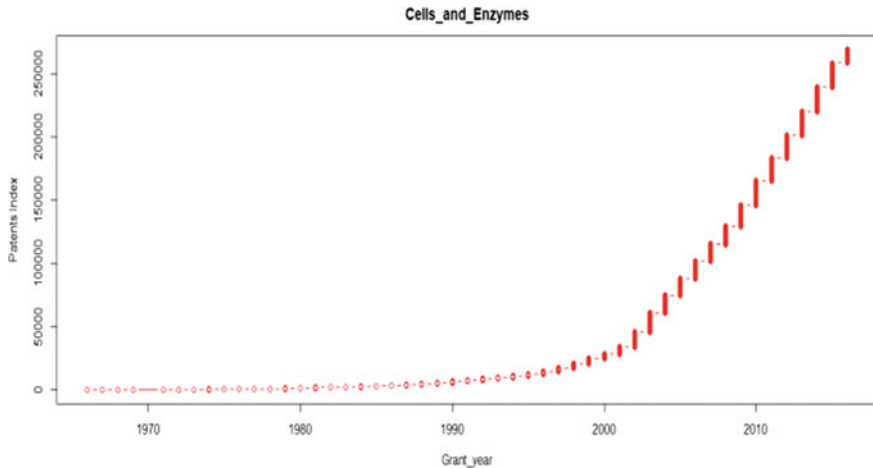


Fig. 8 Trend for cells and enzymes in the data set

## 5 Conclusion and Future Scope

In conclusion of the research, it has been seen that how trend analysis using time series data analysis can be successfully used to understand the technological trend in the medical sector in the cancer domain. The main benefit of the present-day research analysis is the fact that it analyzes different attributes of the cancer dataset with respect to time by performing time series analyses and deriving some conclusions for the prediction of key technological resources so that the amount of invention (patents) done till now should help the cancer domain grow and increase its technological

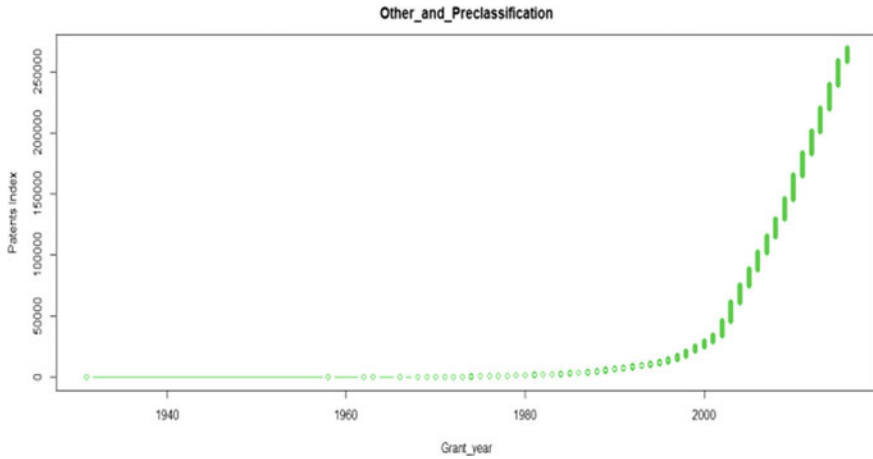


Fig. 9 Trend for other and pre-classification in the dataset

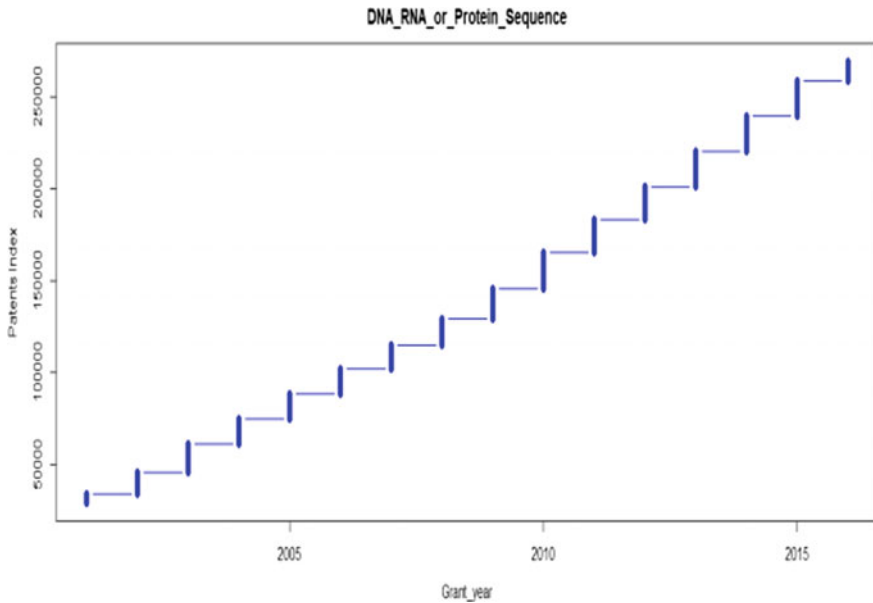


Fig. 10 Trend for DNA RNA or protein sequence in the dataset

territory of knowledge in curing it completely all over the world and help in focusing the new trend as well as using the older trend to analyze the market and help in developing and funding the major and trending technologies for cancer.

This paper talks about the various ways to analyze the literature on patent data set. The various analyzes were that the trend of the cancer dataset comprising different

attributes. The slope of the curves shows that every attribute has a positive upside growth around the 2000 millennium. Figures 2, 3, and 9 have a steeper curve as compared to others so Drugs and Chemistry, Diagnostic and Surgical Devices, and Other and Pre-classification have a large number of patents under this criterion that depicts the growth in these areas starting with the onset of the year 2000 and went on growing at an increasing rate. While the other attributes have a smaller number of patents in their domain and depict a very steady growth over the year since 2000 (this shows that there is the scope of improvement in this domain in the future) except for DNA RNA or Protein Sequence. As shown in Fig. 10, the number of patents is very few and that shows consistent growth over the years. This domain is less explored till now but over the years it has anticipated growth because of the advancement of technology.

## References

1. Singh V, Chakraborty K, Vincent L (2016) Patent database: their importance in prior art documentation and patent search.
2. Liu X, Yan J, Xiao S, Wang X, Zha H, Chu S (2017) On predictive patent valuation: forecasting patent citations and their types. In: Proceedings of the AAAI conference on artificial intelligence, vol 31, No 1
3. Lee S, Yoon B, Lee C, Park J (2009) Business planning based on technological capabilities: Patent analysis for technology-driven roadmapping. *Technol Forecast Soc Chang* 76:769–786
4. Abbas A, Zhang L, Khan SU (2014) A literature review on the state-of-the-art in patent analysis. *World Patent Inf* 37:3–13
5. Pan CL, Chen FC (2017) Patent trend and competitive analysis of cancer immunotherapy in the United States. *Hum Vaccin Immunother* 13:2583–2593
6. McLeod AI, Yu H, Mahdi E (2012) Time series analysis with R. In: *Handbook of statistics*, vol 30, pp 661–712. Elsevier
7. Jeong Y, Yoon B (2011) Technology roadmapping based On patent citation network considering technology life cycle. In: *First international technology management conference*, pp 731–738. IEEE
8. Yen NY, Shih TK, Chao LR, Jin Q (2010) Ranking metrics and search guidance for learning object repository. *IEEE Trans Learn Technol* 3:250–264
9. Yao X, Zhu D, Gao Y, Wu L, Zhang P, Liu Y (2018) A stepwise spatio-temporal flow clustering method for discovering mobility trends. *IEEE Access* 6:44666–44675
10. Shalaby W, Zadrozny W (2019) Patent retrieval: a literature review. *Knowl Inf Syst* 1–30
11. Tseng YH, Lin CJ, Lin YI (2007) Text mining techniques for patent analysis. *Inf Process Manage* 43:1216–1247

# Computed Tomography Image Processing Methods for Lung Nodule Detection and Classification: A Review



Ebtasam Ahmad Siddiqui, Vijayshri Chourasia, Madhu Shandilya, and Vivek Patel

**Abstract** The pulmonary nodules relate to a variety of lung irregularities that can be found in the early diagnosis of pulmonary patients. Radiologists can diagnose lung nodules by analyzing pulmonary images. The radiologists may be assisted by automatic sensing devices that identify nodules of various sizes within lung images. The identification and categorization of lung nodules on various images, such as CT images, electron microscopy, and Histopathology images, have attracted substantial mathematical, statistical, and observational study work during the past 50 years. Various methods appear to be promising suitable decision analysis systems to adequately handle the core problems in lung nodules diagnosis, such as extraction of features, nodule identification, true-negative reduction, and cancerous-noncancerous differentiation, as noted in a current remarkable and considerable improvements in machine learning for pulmonary nodules they attained across both research and commercial. The major goal of this study is to offer a complete state-of-the-art analysis of various methodologies for finding a better method for the classification of lung CT images in malignant and benign classes. An analysis of the methods of automatic detection of lung nodules is given in this paper. It implements a common framework that can be used to define and reflect current methods for the identification of lung nodules.

**Keywords** Deep learning · Linear discriminate analysis · Artificial neural network · Deep convolutional neural network · Support vector machine · Computer-aided design

---

E. A. Siddiqui (✉) · V. Chourasia · M. Shandilya · V. Patel  
Department of Electronics and Communication Engineering, MANIT, Bhopal, India  
e-mail: [eas.193114001@manit.ac.in](mailto:eas.193114001@manit.ac.in)

V. Chourasia  
e-mail: [vchourasia@manit.ac.in](mailto:vchourasia@manit.ac.in)

M. Shandilya  
e-mail: [madhushandilya@manit.ac.in](mailto:madhushandilya@manit.ac.in)

## 1 Introduction

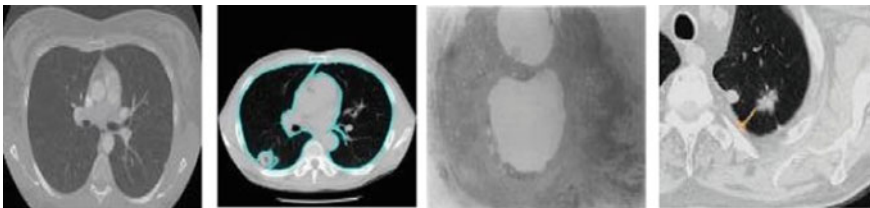
Uncontrollable abnormal cell development of lung tissue is the source of lung cancer. Early detection of irregularities in the lung tissue can assist pulmonary patients in early care. Abnormalities in lung tissue, which are approximately circular, are classified as lung nodules, with a red appearance and up to around 30 mm in diameter [1–3]. This may be categorized into a variety of groups including juxta-vascular, circumscribed, pleural tail, and juxta-pleural. Juxta-vascular has vital ties to the neighboring vessels. Taken together, the adjacent vessels and structures have little relation. The pulmonary tail is a quite narrow contact with its neighboring. The posterior tail is a part of the nodule's pleural membrane [4–6]. Juxta-pleural is connected to the top of the adjacent pleural. Figure 1 shows each of the nodule categories indicated, a sample picture. Advances in CT have been important to enable pulmonary nodules to seek premature care for this disease, as CT allows it is possible to image granules that are little or have a lower resolution that cannot be scrutinized by traditional radiograms. This review aims to classify and explain the current nodule detection techniques [7–10].

A summary of recent lung identification of nodules literature shows that a sufficient effort has been made to categorize known nodule detection methods according to their standards of operation. This paper provides a basic framework for the identification of lung nodules that can be used to identify most existing approaches and define them [11].

It contains multiple components: collection, pre-processing, pulmonary segmentation, nodule identification, and a false positive decrease. During each part of various systems, many algorithms have been employed. In this paper, such algorithms are explained. The paper, therefore, allows it easy for readers to gauge the effectiveness of the methods analyzed through a review of the results of current methodologies [12–14].

### A. Computed Tomography Scan

CT scanning or computed tomography scanning is an imaging diagnostic technique that utilizes multiple X-ray measurement device processing at various angles to generate cross-sectional photographs of particular areas of the body



**Fig. 1** For the four nodule groups, Juxta-vascular, juxta- pleural, circumscribed, and pleural tail nodule from left to right

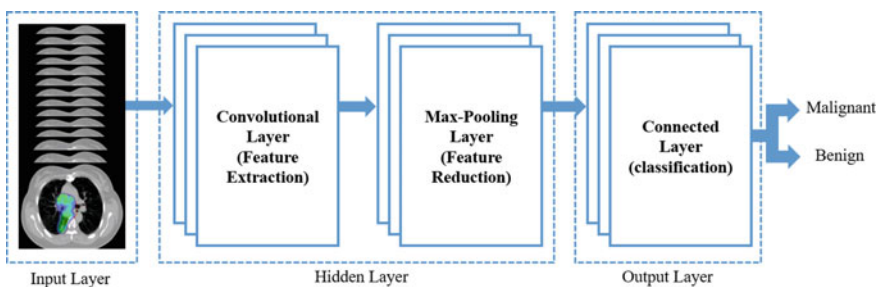
being studied to allow patients to see without cutting the specimens [15]. The CT searches for pixel values with the Hounsfield (HE) size. Comparisons [11] explain the definition of this scale. The Hounsfield Scale (HE) is clearly reported to be distance-dependent. The 0 Hounsfield units (HUs) are used as a reference for water and air and the surface is coated with  $-1000$  HU. Denser products such as bone are also around  $1000$  HU above [16–19].

**B. Deep Learning**

Deep learning, a complex and sophisticated algorithm, is a subfield of artificial intelligence distinguished by its architecture’s multiple layers [11, 20–22]. Such algorithms can overcome specific problems by learning from interactions such as pictures, time series, and images. More information on this subject can be found in the summary [8, 11, 23, 24]. Deep learning has been tremendously established for studying medical images in solving many automatic image processing tasks, including identification, position, detection, segmentation, and registration [13, 16, 25, 26].

**C. Convolutional Neural Networks**

CNNs are an ANN delineate for the analysis of images [5, 27–29]. In convolutional neural networks, there are mainly three layers, input layers, hidden layers, and output layers. In the input layer, we provide images from standard datasets; those images are operated in a variety of ways. operations like morphological and thresholding operations. After the input layer in the hidden layer, multiple procedures are performed like feature extraction and feature reduction. For feature extraction, mainly the convolutional layer is used; this layer uses ReLU activation functions. In the feature reduction max-pooling layer used, this layer uses the sigmoid activation function. In the last layer, which is called the connected layers, the classification process occurs by using the softmax activation function [9, 30–32] (Fig. 2).



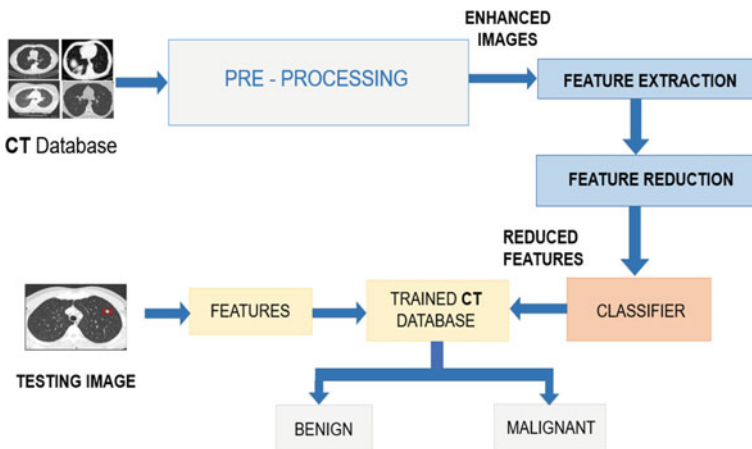
**Fig. 2** Basic Structure of Convolutional Neural Network

## 2 Existing Techniques

There are several papers explaining pulmonary nodule identification methods. An analysis of the methods produced reveals that various mechanisms are utilized for the methods. There are many algorithmic elements and their unique interrelationships in increasing structure. This paper includes a common framework for the identification of lung nodules, and can be used to categorize most of the existing approaches [33–36]. It consists of several components including input CT images, data acquisition and pre-processing, nodule detection, and categorization in Fig. 3. Any modern method for detecting nodules provide mentioned elements, whereas the remaining include batch of them. If a program does not have a predetermined element, the default framework offered may avoid the element, and the cost can be halved by constructing a simpler system module. A description of current structures focused on the built configuration allows readers to fully grasp program operational concepts and to equate the approaches with related frameworks [15, 37, 38].

It consists of several components including: input CT images, data acquisition, pre-processing, nodule detection, as well as categorization, as shown in Fig. 3. Any modern method for detecting nodules provides several elements, whereas old methods include a group of them. If a program does not have a specific element, the default framework offered avoids the element, and the cost can be lowered by constructing a simpler system module. A description of current structures focused on the built configuration allows readers to fully grasp program operational concepts and to equate the approaches with related frameworks [12, 39, 40].

- A. **Kostis et al.** [25] describe 3D models in high-resolution CT pictures for the volumetric doubling-time calculation of tiny pulmonary nodules. The CT volume factorization approaches the models used in this article are prototypes that use



**Fig. 3** Basic architecture of detection and classification process

the statistical morphological method. This describes four separate groups such as well-circumcised, vascularized, pleural tail and pleural pulmonary nodules. Examination of our open datasets shows that there are 200 nodules that showed that almost nearly half of the lesions were vascularized, about one-third were well limited and circumscribed, and almost a quarter was juxta-pleural and a very small (about 1%) were the pleural tail. Initially, low-dose screening CT tests detected the nodules studied here. The tomographic image samples were then collected using a regular dosage procedure utilizing GE High-Speed and Low-Speed CT scanners. The distinction between the nodules and the much lower attenuation was used in this segmentation of tissue that surrounds. This would imply a criterion for strength, area growth, or likely successful contour approaches.

- B. **Way et al.** [28] develop the CAD method to distinguish abnormal image identification in computed tomography images. A completely preprogrammed method has been developed for segmenting the nodule in a local volume of interest (VOI) from its surrounding organized context and for the extraction of classification pictures. Picture segmentation was conducted using an active contour (AC) process. A collection of data was used in this analysis of 96 pulmonary scans. A basic optimization approach was used to look for the right energy weights in the 3D AC model. The segmented nodule has extracted morphological features and gray-level features. The voxel shell containing the Nodule rubber band straightening transform (RBST) has been added. Texture attributes from the RBST picture have been extracted based on run-length statistics. Using a second simplex optimization, a linear discriminative analysis classification with step-by-step selection was developed to select the most effective features.
- C. **Al-Kadi et al.** [8] present the potential to distinguish between malignant and abnormal CT that is presented in the examination of fractals a series of CE-EC CT scans. As the splitting of a blood vessel can be taken into account as a fractus, the review investigates the vascularization of the tumor areas, which have strong fractal characterizations. The study is carried out after 15 patients with a contrast agent are administered and the CE CT pictures are converted into the fractal element from each patient at least 11 times series and related limitations are calculated.

A normal CT test for clinical tumor-staging was conducted in this paper for 15 patients (10 males and 5 females  $63 \pm 8$  yrs. and possessing lung cancers of more than  $10 \text{ mm}^2$ ). At the anatomical level, a 12-bit dynamic sequence of pixels DICOM thorax pictures representing the largest cross-sectional aspect of the lung tumor has been produced. The CE CT photographic images obtained are transformed into FD photographs by means of a differential box counting (DBC) algorithm for nodule region of interest recognition, after that, texture comparison. In contrast to advanced tumors (malignant) and early-stage (benign) malignant tumors, the fractal structure has spread across the tumor area and the quantitative distinction has shown a precision of 83.3%.

- D. **Sharma et al.** [41] The automatic computer-aided diagnostic (CAD) method is suggested for examining computer-generated tomography images in this



paper to diagnose lung cancer. This paper contains a cancer-sensitive framework focused on texture characterizations derived from the DICOM Lung CT slice to distinguish cancer nodules. We have transmitted the lung CT images and their database in this device in three essential steps, such that our experimental findings are accurate: The first aspect to fix is a pre-processing phase involving certain techniques for improving images. After scanning Lung CT Images, the images collected are analyzed (by increasing contrast, thresholding, filtering, and blob analyses) and the suspected nodule areas (SNA) are then removed from the picture by a segmentation procedure mechanism for threshold segmentation by Otsu algorithm of thresholds and techniques for area growth. Lastly, we relied on texture characteristics which help us compare cancer with noncancer images. A neural network to differentiate them is established in this article. We have educated the NIH / NCI Lung Image Databases Consortium (LIDC) network by using a backpropagation algorithm and checked it with various images from the collection of DICOM CT Lung pictures.

- E. **Choi et al.** [23] We are suggesting a modern genetic programming (GP) classification pulmonary nodule identification method. There are three phases in the suggested method. The first step consists of segmenting the lung volume by way of threshold and 3D part labeling. In the second stage, the identification of candidates for the segmented nodule is optimized by multiple thresholds and rule-based sizing. At this point, the identified nodule candidates extract a set of features and pick important 3D and 2D features. The final stage is to train and identify nodules and non-nodules using a GP-based classification (GPC). The proposal would then assess the results with the Lung Image Database Consortium (LIDC). The suggested approach showed that the amount of false positives in the candidate nodule may be greatly decreased and eventually, a 94.1% sensitivity at 5.45 false positive per scan was obtained.
- F. **Kuruvilla et al.** [30] present a computer-aided classification method in the use of an artificial neural network in computed tomographic (CT) images of the lung. The computer-aided diagnostic device is really useful for radiologists in faster and quicker identification and diagnosis. The complete nodule is split into computed tomography pictures, and the elements that are separated picture are determined. Classification is based on statistical criteria such as average and other statistical parameters, the 5th key level, and the 6th pivotal point. The categorization method is carried out by feeding and feeding back neural networks. The feedback propagation network has greater sorting relative to feed for forward networks. Images are taken from the LIDC collection and even from a well-known hospital. 155 doctors, both males and females, also received CT pictures. The patients averaged 64.2 years (lowest 18 years of age and highest 85 years of age of patient). Different neural systems perform morphologic operations for segmentation and classification. The accuracy of the classification of skewing is increased by 5–8%. The Traingdx feature provides overall classification accuracy of 91.1% among the already usable 13 training operations of the ANN. This article advances two different teaching features. The tests indicate that the suggested teaching position 1 has a 93.3% precision,

100% reliability, and 91.4% flexibility with an average 0.998 mean square error. A categorization of 0.933% and a minimum average fault is 94.20% are given in the proposed training function.

- G. **Gunavathi et al. [42]** A computerized classification method for measuring in this paper images of the lungs uses FIS and ANFIS. This is planned to implement an integrated neuro-fuzzy inference method. The complete pulm uses morphological operations, and CT scans are used to segment the lobe. GLCM are statistical and divided photo estimated four parameters are chosen for categorization through main element review from there are 14 GLCM variables and 3 attribute values in the model. The chosen variables are cluster hue, variance, and skew. FIS and ANFIS complete the classification process. ANFIS offers a clearer categorization relative to FIS. In ANFIS, the RNN is employed, and a new training technique is suggested. The system suggested offers 94% classification accuracy with 100% specific characteristics and 93% accuracy.
- H. **Dhara et al. [43]** the aim of this work is on classifying benign and malignant pulmonary nodules with support vector machine. A semiautomatic technique is implemented for segmentation of pulmonary nodules, which only needs a seed from the end consumer. Different forms, margins and Pulmonary nodules are measured for texture-based characteristics. For the efficient representation of nodules in the feature space a collection of related characteristics is described. A sample consisting in 891 nodules of the shared data base LIDC Development Initiative should offer a confirmation of this classification scheme. The classification efficiency is measured in region (az) below the practical attribute curve of the receptor. The Az of 0.9505, 0.8822, and 0.8488 respectively were achieved with the suggested configuration- 1, configuration-2, and configure-3 process. The suggested procedure succeeds the new methodology, which relies on a qualified radiologist manually segmenting the pulmonary nodules.
- I. **Yutong et al. [44]** In this paper, we offer an algorithm for the classification of lung nodules which combines structure, form, and color profound Fuse-TSD information at the threshold of selection. Its method uses a GLCM, a Fourier form descriptor for the variability of the nodule, and a DCNN to learn how to view nodules slowly. It uses each function category to build an Ada Enhanced Neural Propagation Network (BPNN) and mixes decisions made with 3 classifiers in order to classify nodules. In the LIDC-IDRI dataset, we tested this method with three strategies. When lesions with several malignancies score 3, referred to as mild or cancerous, were discarded proposed Fuse-TSD method, which was significantly higher than the AUC of 96.65%, 94.45%, and 81.24%, respectively.
- J. **Lakshmanprabu et al. [20]** A CT scan was used to detect the tumor's location and to determine the cancer extent in the body. In the current research, a new approach for automatically classifying computer tomography (CT) pulmonary imaging is introduced. In this article, an ODN and a Dynamic Differentiate Analysis (LDA) analyzer of the lung scan are conducted. The strong characteristics of lung nodules are marked as malignant or benign with CT and the dimensionality of function by utilizing LDR. The ODN is introduced

in CT images and then modified to classify the lung cancer type using the Modify-Gravitational Search Algorithm (MGSA). Comparative findings show that 96.20% of the classifier suggested is sensitive, 94.2% precise and 94.56% accurate.

- K. **Ying et al.** [17] The suggested technique helps identify whether a pulmonary tumor is cancerous or non-cancerous based on a 3D image reconstruction area of the nodules to aid in the diagnostic procedure. We built a unique multimodal regularized classification DBN (MRC-DNN) to conduct classification based directly on the multidimensional interpretation of tumor segmentation pictures, prompted by the idea that actual pattern across data is typically contained on a low-dimensional band. The classification is anticipated to benefit from the compact multidimensional model, which reveals significant data architecture. In contrast, the multidimensional normalization imposes substantial but reasonable limitations on the training phase, avoiding over-fitting. By evaluating training and testing results from that of the traditional DL-based SVM classifier, the efficacy of the learning phase was confirmed. It's worth noting that the variable is regulating the intensity of multidimensional regularization impacts MRC-DNN performances. The training sample results of MRC-DNN and C-DNN are analogous. At the same time, both models have immense ability and, therefore, can match the machine learning model to a high degree of accuracy. And from the other end, the MRC-DNN beats the C-DNN considerably from both the training and testing sets, with accuracy gains of 10% and 7%, respectively. More crucially, MRC-DNN achieves a 20% increase in susceptibility in both evaluations.
- L. **Hong et al.** [45] Moreover, because of the diversity of chest radiography and the morphological resemblance of computed tomography to various lung tissue, traditional DL-based detection approaches make it challenging to build a strong classifier. He presents a modified artificial neural structure based on the 3d deep neural networks to solve this challenge (MMEL-3DCNN). Three essential concepts are incorporated into this strategy: (1) The uniqueness of chest radiography may be successfully applied using a developed nonlinear and linear network architecture. (2) The data consists of a concatenation of the image pixels according to the CT image mask, the source picture, and the emerging trend related to may aid the learning approach in extracting innovative functions with greater detection accuracy. (3) Adaptively choose the represent substantial for varied lesion sizes for forecast, which may significantly raise the model's generalization capacity. Moreover particular, the learning algorithm is used to increase the reliability of the pulmonary CT scan classification framework for this study. Mostly on open dataset LIDC-IDRI, the suggested technique has been validated. Experiments demonstrate that the suggested MMEL-3DCNN design may provide acceptable prediction performance.
- M. **Rachel et al.** [18] Extensive information collections featuring increased abnormality labeling assist neural network models in imaging. A lung CT image collection of 36,316 slices from 19,993 distinct individuals was selected

and analyzed. That's the world's most significant dimensional computed tomography data set with many annotations. We created a mandate technique for retrieving abnormalities diagnoses and available medical image analysis with only a median F-score of 0.976 to label such existing data (min 0.941, max 1.0). We used a deep learning model to construct a nonlinear and linear, multi-disease categorization of lung CT data. The system has an aggregate AUROC of 0.773 for all 83 anomalies, indicating the possibility of training from raw complete volumetric CT images. Researchers demonstrate that learning on much more symbols makes a significant difference noticeably: because when the set of test identifiers was enhanced from 9 to all 83, the model's mean AUROC improved by 10% for a subcategory of 9 labels – tumor, occlusion, distention, pulmonary edema, centralization, density, pneumothorax, atelectasis, and lung injury. The volumetric irregularity predictive algorithm and the algorithms for volumetric processing, automating labeling retrieval, and the volumetric irregularity linear regression model are all open source.

- N. **Liyun et al.** [12] Computed Tomographic image detection classification and division algorithms are part of the ALIAS methodology. Researchers can extract the features and do analytical studies after tumor classification and localization. The 3D ReLU sequence FPN is explicitly used for lesion identification. Then, using the edges supplied by tumor classification, pictures are resized and analyzed through the lesion separated to get lesion overlays. Image characteristics such as the spectrum of HUs or closely and continuously parameters are retrieved for pulmonary image analysis. The variations among normal and abnormal lesions of different sizes may be analyzed. The ALIAS technology also makes it easier to create tumor mapping depending on density. Each Image data lung area will be first divided, then mapped into a preset pattern input data. We may convert the tumor positions within every Lung region to the equivalent locations in the pattern region, where its content-based contour maps are generated, using the deformed areas obtained by recognition.
- O. **Prasad et al.** [26] Initial diagnosis of lung disease is achieved by detecting early phase tumors (3-30mm), which can also significantly improve pulmonary patients' multiyear survival rates. Cells are confined tumors in the thorax that are exceedingly tiny and hard to find due to their small size. According to identically looking entities such as non-lesions that possess characteristics that make it recognizable as a lesion, identifying a tumor is also more difficult. As a result, we presented a unique technique for identification, segmentation, and identification of computed tomography from CT images to address these complicated challenges. We use a higher intensity prediction methodology for the portion of the picture processing approach in our suggested method. On public information LIDC-IDRI, challenge set of data, and primarily unbiased ILCID medical set of data, we indicated our research and testing. We suggested SquExUNet classification and 3D-NodNet classification algorithm. For tumor detection, we presented a 2D-3D feedforward CNN approach that produces properly divided and categorized nodules. In comparison to current lesion classification and localization techniques, the findings acquired with the suggested technique show that

we will have detected accurately and divided the lesion. We acquired a Dice-Coefficient of 0.80 for lesion classification and a Precision of 90.01% for tumor identification.

### 3 Result and Discussion

In this review, it becomes apparent that developments in these areas are still at an early stage when the methods for classifying lung nodules and lung cancer detection and the algorithms for characterizing lung nodules are examined. Although commercial approaches appear to be more advanced, experiments that have historically utilized very limited sample sets have not yet been tested, and their success in the general patient population. The commercial methods like CAD and neural techniques were focused on solid lung nodules for the detection of lung nodules. Because non-solid nodules are more likely to be malignant and are overlooked by radiologists, it is, therefore, necessary to improve CAD and neural techniques for such nodules. For lung nodule diagnosis, the growth rate calculated in regular CT exams is the most significant piece of information radiologists use to predict the risk of malignancy of a nodule.

Imaging methods of intelligence-based prediction strategies were found to be effective in forecasting and deciding on lung cancer. For their efficient evaluation for the analysis of lung nodules. Table 1 outlines the strategies of image detection and classification approaches. Table 1 gives a description of the identification and classification of lung nodules (Fig. 4).

#### Performance Parameters

Quality of most of the techniques suggested for the Lung CT image classification for detection of cancer is analyzed by following parameters:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

where TP, FN, FP, and TN represent the number of true positives, False negatives, false positives, and true negatives, respectively (Fig. 5).

#### Findings and Future Scope

- In the preprocessing stage, morphological operations are employed which causes loss of fine details and leads to lower sensitivity [12, 44].

**Table 1** Different methodology with datasets along with parameters

Methodology	Data set	Dimension	Performance
Volumetric doubling-time calculation	Clinical Database Testing: 75 scans, 210 nodule	2D and 3D	Accuracy: 95% Detection: 80% Sensitivity: 83%
Computer-aided diagnosis (CAD) with a three-dimensional (3D) active contour (AC) method	Lung Image Database Consortium (LIDC): Lung Nodule: 96 Malignant:44 Benign: 52	3D	Training data sensitivity: 91% Testing data sensitivity: 87%
Fractal analysis of time sequence contrast-enhanced (CE) computed tomography	Clinical Dataset 44 Scans 177 Nodules	3D	Accuracy: 83.3% segmented 14% requires alternative solution
Automated Computer-Aided Diagnosing (CAD) system	NIH/NCI - (LIDC) dataset: 1000 images	3D	Specificity: 99% Sensitivity: 90% Accuracy: 85%
Genetic programming (GP)-based feature transform	NCI – LIDC Dataset 165 Scans	2D and 3D	Training data accuracy: 95.5% Testing data accuracy: 89.0% 94.1% sensitivity at 5.45 FPs/scan
Computer aided classification method using artificial neural network(ANN)	Lung Image Database Consortium (LIDC): 23 Scans 110 Nodules	3D	Accuracy: 93.3% Specificity: 100% Sensitivity: 91.4% Mean square error:0.998
Classification using FIS and ANFIS method	Lung Image Database Consortium (LIDC): 23 Scans 110 Nodules	3D	Sensitivity: 94% Accuracy: 93% Specificity: 100%
Classification of benign and malignant pulmonary nodules using support vector machine	Image Database Resource Initiative (LIDC/IDRI): 142 Scans 891 Nodules	2D and 3D	Overlap: 0.935 Accuracy: 89.94% Sensitivity:92.45%
Fuses texture, shape, and deep-model-learned information (Fuse-TSD)	(LIDC-IDRI) dataset: Scans: 545 Nodules: 2171	3D	Sensitivity: 94.45% Accuracy: 96.65%
Optimal Deep Neural Network (ODNN) and Linear Discriminate Analysis (LDA)	LIDC: 70 scans for train, 30 scans for test	3D	Sensitivity: 96.2% Accuracy: 94.56% Specificity: 94.2%

(continued)

**Table 1** (continued)

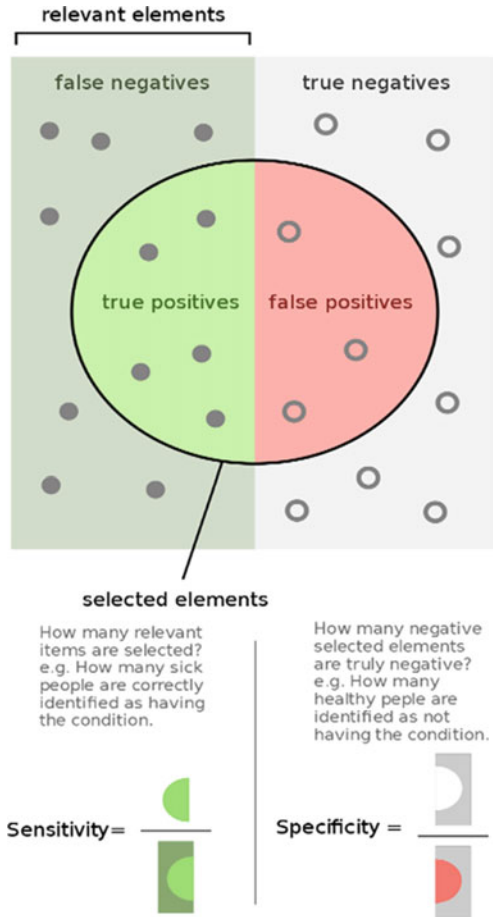
Methodology	Data set	Dimension	Performance
Manifold learning regularization approach to enhance 3D CT image-based lung nodule classification	Database Consortium (LIDC) of Image Database Resource Initiative (IDRI) 1018 CT scans 1226 nodules	3D	Accuracy:90% Sensitivity:81% Specificity:95%
Multi-model Ensemble Learning Architecture Based on 3D CNN for Lung Nodule Malignancy Suspiciousness Classification	LIDC-IDRI Dataset 754 Scans 1273 Nodules	3D	Sensitivity:87% Specificity:84% Accuracy:81% AUC:88.90%
ML-Based Various deformity Prediction with LS Chest CT 3D images	Clinical dataset 36,316 CT images 19,993 scans	3D	Accuracy:90.90% AUROC:90% F-Score:97.60% Sensitivity: 93.42% Specificity:91.67%
An artificial-intelligence lung imaging analysis system (ALIAS) for population-based nodule computing in CT scans	Clinical Dataset 8540 pulmonary CT images 7716 patients	3D	Accuracy:94.21% Specificity:73.60% Sensitivity:90.60%
LNCDS A 2D-3D cascaded Convolutional Neural Network	LIDC_IDRI Dataset 889 patients 2361 Scans	2D and 3D	Accuracy: 95.46% Sensitivity:90.01% Specificity:96.28%

- Most of the lung cancer classification techniques are employing single-layered DCNN, which offers low accuracy in feature extraction [17, 26, 27].
- Modified Gravitational Search Algorithm (MGSA) based DCNN models require large classification training time due to the complex weight update process [18, 26].
- In lung CT image, classification methods, inclusion of Deep Belief Network in MGSA-based DCNN reduces the complexity of weight updation on the cost of sensitivity [8].
- Lung cancer CT image classification can be done by using a Deep Belief Network with significant optimization techniques [3, 20, 30, 46].
- In machine learning, Spiking Neural Network can be used for lung cancer CT image classification in their respective classes [2, 5, 21, 36].

## 4 Conclusion

In order to synthesize emerging developments and recognize more problems, we presented an analysis of the 10 widely common methodologies. We discuss features such as data acquisition, pre-processing, lung segmentation, nodule recognition, and

**Fig. 4** Performance parameters



reduction. In contrast to others, we found that LIDC-IDRI is commonly used for data sets. In addition, we have outlined the strategies and classifications for improved pulmonary nodule identification dependent on sensitivity, accuracy, and specificity. The techniques for image analysis are often used in lung cancer diagnosis, as well as for early identification and care for lung disease prevention. There is also a valuable pattern recognition technique to predict lung cancer, utilizing specific features from the images. A systematic analysis of previous studies using image recognition methods for the diagnosis of lung cancer has been published. The predictive summary of lung cancer CT image processing techniques was also presented by previous researchers.



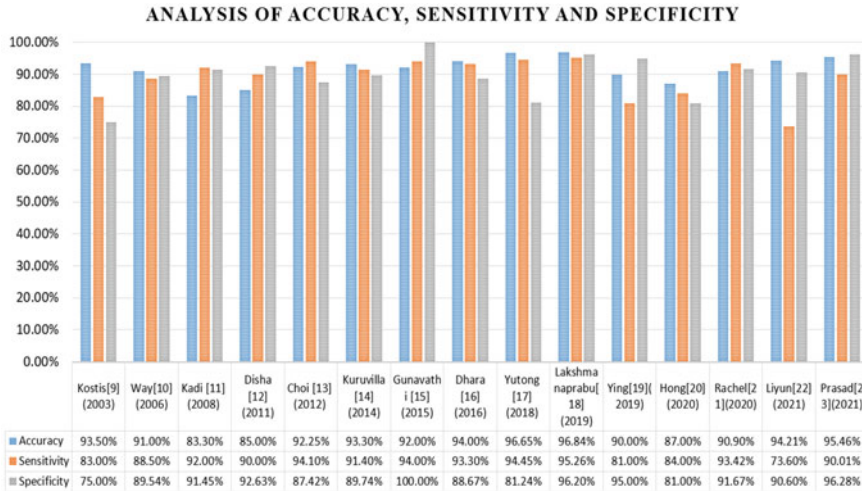


Fig. 5 Comparison of classification methods based on Accuracy, Sensitivity, and Specificity

## References

- Chao W-J, Choi T-S (2014) Pulmonary nodule detection based on three-dimensional shape-based feature descriptor. *Comput Methods Programs Biomed* 113:37–54
- Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M (2015) Imagenet large scale visual recognition challenge. *Int J Comput Vis* 115:211–252
- Li X, Kao Y, Shen W, Li X, Xie G (2017) Lung nodule malignancy prediction using multi-task convolutional neural network. In: *Proceedings of SPIE*
- Armato SG III, Altman MB, Wilkie J, Sone S, Li F, Doi K, Roy AS (2003) Automated lung nodule classification following automated nodule detection on CT: a serial approach. *Med Phys* 30:1188–1197
- Al-Shabi M, Lan BL, Chan WY, Ng K-H, Tan M (2019) Lung nodule classification using deep local-global networks. *Int J Comput Assist Radiol Surgery* 14:1815
- Armato SG, McLennan G, Bidaut L, McNitt-Gray MF, Meyer CR, Reeves AP, Zhao B, Aberle DR, Henschke CI, Hofman EA (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 38:915–931
- Chaudhary A, Singh SS (2012) Lung cancer detection on ct images using image processing. In: *International conference computing sciences*. IEEE
- Al-Kadi OS, Watson D (2008) Texture analysis of aggressive and nonaggressive lung tumor CE CT images. *IEEE Trans Biomed Eng* 55(7)
- Ma J, Wang Q, Ren Y, Hu H, Zhao J (2016) Automatic lung nodule classification with radiomics approach. In: *Medical imaging 2016: PACS and imaging informatics: next generation and innovations*. International Society for Optics and Photonics, p 978906
- Ardila D, Kiraly AP, Bhardwaj S, Choi B, Reacher JJ, Peng L, Tse D, Etemadi M, Ye W, Corrado G, Naidich DP, Shetty S (2019) End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat Med* 25:954–961
- Silveira M, Nascimento J, Marques J (2007) Automatic segmentation of the lungs using robust level sets. In: *Engineering in medicine and biology society, 29th annual international conference of the IEEE*, vol 1, pp 4414–4417

12. Chen L, Gu D, Chen Y, Shao Y, Cao X, Liu G, Gao Y, Wang Q, Shen D (2021) An artificial-intelligence lung imaging analysis system (ALIAS) for population-based nodule computing in CT scans. In *Computerized medical imaging and graphics*, vol 89
13. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) *Advances in neural information processing systems* 25. Curran Associates Inc., pp 1097–1105
14. Kalpathy-Cramer J, Mamomov A, Zhao B, Lu L, Cherezov D, Napel S, Echegaray S, Rubin D, McNitt-Gray M, Lo P (2016) Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features tomography
15. Niranjana G, Ponnavaikko M (2017) A review on image processing methods in detecting lung cancer using CT images. In: *International conference on technical advancements in computers and communications*. IEEE
16. Katre PR, Thakre A (2017) Detection of lung cancer stages using image processing and data classification techniques. In: *2nd international conference for convergence in technology*
17. Ren Y, Tsai M-Y, Chen L, Wang J, Li S, Liu Y, Jia X (2020) A manifold learning regularization approach to enhance 3D CT image-based lung nodule classification. *Int J Comput Assist Radiol Surg* 15:287–295
18. Draelos RL, Dov D, Mazurowski MA, Lo JY, Henao R, Rubin GD, Carin L (2020) Machine-learning-based multiple abnormality prediction with large-scale chest computed tomography volumes. *Med Image Anal*
19. Aresta G, Araújo T, Kwok S, Chennamsetty SS, Safwan M, Alex V, Marami B, Prastawa M, Chan M, Donovan M, Fernandez G, Zeineh J, Kohl M, Walz C, Ludwig F, Braunewell S, Baust M, Vu QD, To MNN, Kim E, Kwak JT, Galal S, Sanchez-Freire V, Brancati N, Frucci M, Riccio D, Wang Y, Sun L, Ma K, Fang J, Kone I, Boulmane L, Campilho A, Eloy C, Polonia A, Aguiar P (2019) BACH: grand challenge on breast cancer histology images. *Med Image Anal* 56:122–139
20. Lakshmanaprabu SK, Mohanty SN, Shankar K, Arunkumar N, Ramirez G (2019) Optimal deep learning model for classification of lung cancer on CT images. *Futur Gener Comput Syst* 92:374–382
21. Manning DJ, Ethell SC, Donovan T (2004) Detection or decision errors? Missed lung cancer from the poster anterior chest radiograph. *Br J Radiol* 77(915):231–235
22. Caruana R, Lawrence S, Giles CL (2000) Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping. In: *BT—advances in neural information processing systems 13, Papers from neural information processing systems (NIPS)*. Denver, CO, USA, pp 402–408
23. Choi W-J, Choi T-S (2012) Genetic programming-based feature transform and classification for the automatic detection of pulmonary nodules on computed tomography images. *Inf Sci* 12:57–78
24. Han F, Zhang G, Wang H, Song B, Lu H, Zhao D, Zhao H, Liang (2013) A texture feature analysis for diagnosis of pulmonary nodules using LIDC-IDRI database. In: *2013 IEEE international conference on medical imaging physics and engineering*, pp 14–18
25. Kostis WJ, Reeves AP, Yankelevitz DF, Henschke CI (2003) Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images. *IEEE Trans Med Imaging* 22:10–26
26. Dutande P, Baid U, Talbar S (2021) LNCDS: A 2D-3D cascaded CNN approach for lung nodule classification, detection and segmentation. In: *Biomedical signal processing and control*, vol 67
27. Wang S, Zhou M, Gevaert O, Tang Z, Dong D, Liu Z, Tian J (2017) A multi-view deep convolutional neural networks for lung nodule segmentation. In: *2017 39th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, pp 1752–1755
28. Way TW, Hadjiiski LM, Sahiner B, Chan H-P, Cascade PN, Kazerooni EA, Bogot N, Zhou C (2006) Computer-aided diagnosis of pulmonary nodules on CT scans: segmentation and classification using 3D active contours. Department of Radiology, University of Michigan, Ann Arbor, Michigan, p 48109

29. Schneider LS et al (2015) Reduced lung cancer mortality with low dose computed tomographic screening. *New Engl J Med* 687–696. <https://doi.org/10.1056/NEJMoa1505949>
30. Kuruvilla J, Gunavathi K (2014) Lung cancer classification using neural networks for CT images. *Comput Methods Programs Biomed* 113:202–209
31. Jiang J, Hu YC, Liu CJ, Halpenny D, Hellmann MD, Deasy JO, Mageras G, Veeraraghavan H (2019) Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images. *IEEE Trans Med Imaging* 38(1):134–144
32. Amorim P, Moraes T, Silva J, Pedrini H (2018) 3D adaptive histogram equalization method for medical volumes. In: *BT—Proceedings of the 13th international joint conference on computer vision, imaging and computer graphics theory and applications (VISIGRAPP 2018)*, vol 4. VISAPP, Funchal, Madeira, Po, pp 363–337
33. Gopi K, Selva Kumar J (2017) Lung tumor area recognition and classification using EK-mean clustering and SVM. In: *International conference on nextgen electronic technologies*. IEEE
34. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT Press, 2016.
35. Armato SGEA (2011) The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Med Phys* 38(2):915–931
36. Buty M, Xu Z, Gao M, Bagci U, Wu A, Mollura DJ (2016) Characterization of lung nodule malignancy using hybrid shape and appearance features. In: *BT—Ourselin S, Moskowitz L, Sabuncu MR, Unal G, Wells W (eds) Medical image computing and computer-assisted intervention—MICCAI 2016*. Springer International Publishing, Cham, pp 662–670
37. He X, Niyogi P (2003) Locality preserving projections. In: *Thrun S, Saul LK, Schölkopf B (eds) Proceedings of the 16th international conference on neural information processing systems (NIPS'03)*. MIT Press, Cambridge, MA, USA, pp 153–160
38. Roy S, Ghosh P, Bandyopadhyay SK (2015) Contour extraction and segmentation of cerebral hemorrhage from MRI of brain by gamma transformation approach. In: *Satapathy SC, Biswal BN, Udgata SK, Mandal JK (eds) BT—Proceedings of the 3rd international conference on frontiers of intelligent computing: theory and applications (FICTA) 2014*. Springer International Publishing, Cham, pp 383–394
39. Shen W, Zhou M, Yang F, Yang C, Tian J (2015) Multi-scale convolutional neural networks for lung nodule classification. In: *International conference on information processing in medical imaging*. Springer, pp 588–599
40. Mzoughi H, Njeh I, Ben Slima M, Ben Hamida A (2018) Histogram equalization-based techniques for contrast enhancement of MRI brain Glioma tumor images: comparative study. In: *2018 4th international conference on advanced technologies for signal and image processing*, pp 1–6
41. Sharma D, Jindal G (2011) Computer aided diagnosis system for detection of lung cancer in CT scan images. *Int J Comput Electri Eng* 3(5)
42. Kuruvilla J, Gunavathi K (2015) Lung cancer classification using fuzzy logic for CT images. *Int J Med Eng Inf* 7:233–249
43. Dhara AK, Mukhopadhyay S, Dutta A, Garg M, Khandelwal N (2016) A combination of shape and texture features for classification of pulmonary nodules in lung CT images. *J Dig Imaging* 29:466–475
44. Yutong X, Jianpeng Z, Yong X, Fulham M, Yanning Z (2018) Fusing texture, shape and deep model-learned information at decision level for automated classification of lung nodules on chest CT. *Inf Fusion* 42:102–110
45. Liu H, Cao H, Song E, Ma G, Xu X, Jin R, Liu C, Hung C-C (2020) Multi-model ensemble learning architecture based on 3D CNN for lung nodule malignancy suspiciousness classification. *J Dig Imaging* 33:1242–1256
46. Cui W, Zhou Q, Zheng Z (2018) Application of a hybrid model based on a convolutional auto-encoder and convolutional neural network in object-oriented remote sensing classification. *Algorithms*
47. Bakr S, Gevaert O, Echegaray S, Ayers K, Zhou M, Shafiq M, Zheng H, Benson JA, Zhang W, Leung AN, Kadoch M, Hoang CD, Shrager J, Quon A, Rubin DL, Plevritis SK, Napel S

- (2018) Data descriptor: a radiogenomic dataset of non-small cell lung cancer. *Sci Data* 5:1–9. <https://doi.org/10.1038/sdata.2018.202>
48. Messay T, Hardie RC, Rogers SK (2010) A new computationally efficient CAD system for pulmonary nodule detection in CT imagery. *Med Image Anal* 14(3):390–406
  49. Kubota T, Jerebko AK, Dewan M, Salganicoff M, Krishnan A (2011) Segmentation of pulmonary nodules of various densities with morphological approaches and convexity models. *Med Image Anal* 15(1):133–154
  50. Zhu J-Y, Krahenbuhl P, Shechtman E, Efros AA (2016) Generative visual manipulation on the natural image manifold. In: *European conference on computer vision*. Springer, Berlin, pp 597–613
  51. Li S, Xu P, Li B, Chen L, Zhou Z, Hao H, Duan Y, Folkert MR, Ma J, Huang S (2019) Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. *Phys Med Biol* 64:175012
  52. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, Moore S, Phillips S, Maftit D, Pringle M (2013) The cancer imaging archive (tcia): maintaining and operating a public information repository. *J Digit Image* 26:1045–1057
  53. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. *Adv Neural Info Process Syst* 25:1097–1105
  54. Kang G, Liu K, Hou B, Zhang N (2017) 3D multi-view convolutional neural networks for lung nodule classification. *PLoS ONE* 12:e0188290
  55. Yan X, Pang J, Qi H, Zhu Y, Bai C, Geng X, Liu M, Terzopoulos D, Ding X (2017) Classification of lung nodule malignancy risk on computed tomography images using convolutional neural network: a comparison between 2D and 3D strategies. In: Chen C-S, Lu J, Ma KK (eds) *BT—computer vision—ACCV 2016 workshops*. Springer International Publishing, Cham, pp 91–101

# Implementation and Analyzing SURF Feature Detection and Extraction on WANG Images Using Custom Bag of Features Model



Roohi Ali and Manish Maheshwari

**Abstract** A novel technique of image classification using BOVW model also known as Bag of Visual Words is very popular for retrieval of images using features instead of text vocabulary. The entire process first involves feature detection of images by selecting key points or forming a Grid over images, the choice made in order to speed up the process of detection. Then comes the stage of feature extraction for which SURF, a binary feature descriptor is employed. K-means clustering is then applied in order to quantize and make the bag of visual words. Every image, expressed as a histogram of visual words is fed to a supervised learning model, SVM for training. SVM is then tested for classification of images into respective classes. Matlab is used for implementation using bag class with Extractor function over 1000 image dataset WANG with 10 different categories.

**Keywords** BOVW · K-means clustering · SURF · Extractor function · SVM

## 1 Introduction

An image is a set of signals sensed by the human eye and processed by the visual cortex in the brain creating a vivid experience of a scene that is instantly associated with concepts and objects previously perceived and recorded in one's memory.

One of the prominent applications of image processing is Image feature extraction that laid the base for many image processing algorithms prominently image classification. Images are classified between the predefined classes within the big datasets of images using compact vector representations of local neighborhood property. There are many techniques available to categorize the images. All they need some feature extraction technique that helps in categorizing them on the basis of the extracted features. A diverse range of extractors and detectors are available vary upon kinds of detection over interest points, repeatability, complexities over time and

---

R. Ali (✉) · M. Maheshwari

Department of Computer Science and Applications, Makhnallal Chaturvedi National University of Journalism and Communication, Bhopal 462001, MP, India

e-mail: [roohi.ali2006@gmail.com](mailto:roohi.ali2006@gmail.com)

space. Features such as Blob, Corner, and Grid etc are available along with certain descriptors. The Feature Detection and Extraction tab under The Computer Vision Toolbox TM delivers Corner detection features methods such as Harris-Stephens algorithm, Features from Accelerated Segment Test (FAST) [1], Methods by Shi and Tomasi, Oriented FAST and rotated BRIEF (ORB)[4]. Blob detection features such as Maximally Stable External Region Tracking (MSER), Speeded Up Robust Features (SURF) [2] and KAZE. In Addition to this certain detectors are also available like FREAK [6], BRISK [5], ORB [4], HOG, SURF and KAZE descriptors. A good mix of a detector along with the corresponding descriptors will be used according to the need of the application.

In this chapter retrieval of images is performed from one of the ten categories present in the WANG Dataset. Despite of the various techniques available in this genre, the methodology used here is unique as it uses a feature extraction style technique and implementation with creating a bag of local features. Detection Methods like SURF and GRID are put to use along with descriptors deployed in the Bag of Feature model accelerate the efficiency and fastness of the algorithm. K-means clustering provides a base to form the dictionary of the visual words known as the bag of visual words, which is the most important framework used in the entire process. Support vector machine a supervised learning technique is deployed as a classifier.

## 2 Detection and Extraction of Local Features

Various patterns and different structures that comprise an image are nothing but the local features of an image. It may be the small patches, edges of an image etc. Basically significance of a feature is the difference it possesses from its corresponding surroundings i.e. local neighborhood like texture, intensity or color of one patch will be different from other. It laid the foundation of most of the image processing algorithms like image tracking, object detection and classification etc. Vector of numeric values (i.e. nothing but a raw pixel value) which describes a piece of image is a descriptor. Key point is again a driving force for a descriptor; it is actually the interest point (intensity of certain patch or corner) of an image. Descriptors along with the key points will define the local features. A local feature is a good mix of gradient-based and intensity variation approaches includes edges, blobs and regions. Figure 1 shows different techniques for detection and extraction of features according to interest points.

One of the main advantages of using the local features is that it represents contents of image precisely without the use of any segmenting technique for the tasks of classification, extraction and Detection.

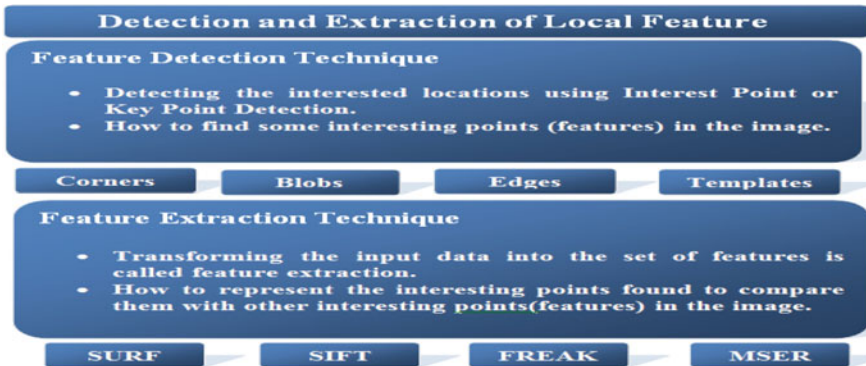


Fig. 1 Depicting various detection and extraction techniques

### 2.1 Feature Detection and Extraction Using SURF

While processing an image the initial step is to perform Feature Detection. It will detect the low level entities by the selection of unique interest/key points (IP/KP). These selected distinct key points ensure the competency of the Procedure. IP/KP are the features like a Corner, edge or Blob identified and detected by a detector from an image that seem to be useful for further processing. Feature Detector can be selected according to the Region to be abstracted such as Edge, Corner, and Blob etc.

Feature extraction, is a form of reduced dimensionality, the extraction of particular set of features from full size input to form a feature vector. Extracted features contain only relevant information about the image and describe the input data accurately.

SURF is inspired by Scale-Invariant Feature Transform (SIFT).SURF was first introduced by Herbert Bay et al. at the 2006 European Conference on Computer Vision. In Computer Vision, Speeded up Robust Features (SURF) is a technique used for object recognition, classification, and registration. SURF is inspired by Scale-Invariant Feature Transform (SIFT) [9]. Interest points detected by the SURF are done using the Hessian blob detection with integer approximation method. The Haar Wavelet response around the interested features retrieves its feature vector [12].

## 3 Bag of Visual Words

The dictionary of visual words which can be used to define each image in terms of the frequency of each word present in an image is known as the bag of visual words.

Figure 2 above demonstrates the bag of visual words model showing the text images as histogram of visual words. Now, this gives a fixed length vector of descriptors irrespective of the number of key points detected. This is one of the advantages

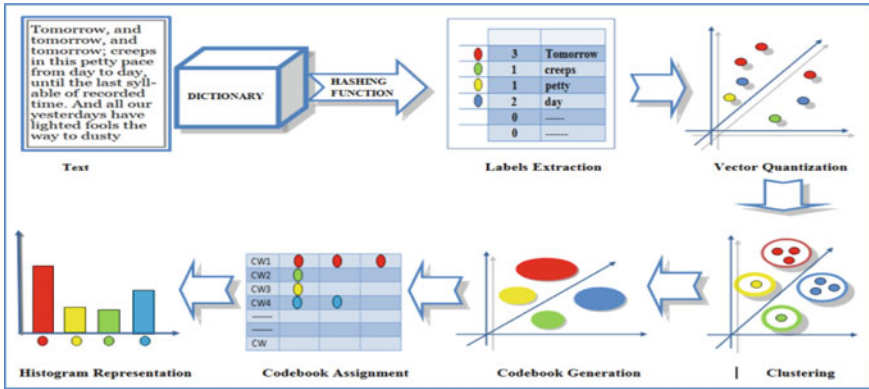


Fig. 2 Bag of visual words model

of using a bag of visual words model. The fixed length feature vector thus obtained is fed to Support Vector Machine or any other supervised learning model for training.

#### 4 Classifier: SVM (Supervised Vector Machine)

In machine learning, support vector machines (SVMs) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis [9]. SVM can be defined as a discriminative classifier, formed of an optimal hyper plane. The hyper plane separates the various classes of examples which help in classification. SVM runs an algorithm which helps in finding the hyper plane which is optimal based on the training data that is fed to it. The optimal hyper plane is chosen such that the distance of the hyper plane from the nearest data point on either side is highest. However, SVM can also be used for performing non-linear classification. SVM is thus an important model used in machine learning and finds its application in object classification.

### 5 Dataset, Implementation, and Results

#### 5.1 Dataset

Wang database with 10 different categories as listed in Fig. 3 is used. Each group consists of 100 images in JPEG format, from Wang database, downloaded from the website <http://wang.ist.psu.edu/iwang/test1.tar>. All these images in the database are natural images. Each image is of size 384 \* 256 or 256 \* 384 pixels. The dataset images are in the RGB format.





Fig. 3 WANG dataset with 10 image categories

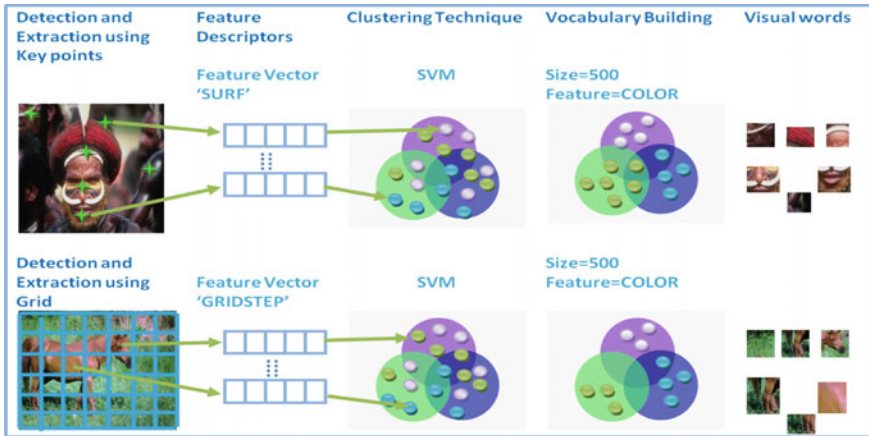


Fig. 4 BoVW model for representation of images

### 5.2 BOVW Bag of Feature Model

As Shown in Fig. 4. Below shows the stepwise visual representation of how the process of Bag of feature works. How to create a histogram or clusters of visual words that represent an image and then these histogram/clusters are used to train classifiers.

In the Fig. 4 above two techniques are depicted on two different images of African and Horses. The first one is using Point selection dense SURF for feature extraction and the next is using Grid step method. These are two different strategies for feature extraction only while all the other steps are same like vector formation, quantizing and training phase are all same.

### 5.3 Algorithm BOVW for Bag of Feature Model

Following are the steps involved in extracting BoVW features model implemented on WANG images:

Algorithm		
Step 1	Preprocessing the images	RGB to Grayscale
Step 2	Images are divided into test and training sets	Test (30%) Training (70%)
Step 3	Feature detection and extraction	SURF DetectSURFFeatures
Step 4	Training classifier	Bag of features metric MultiscaleSURFPoints.Metric
Step 5	Generate feature histograms for each image using clustering	SVM

Prediction method classify the query images

## 5.4 Implementation

The Feature Detection and Extraction tab under The Computer Vision Toolbox TM of MATLAB TM delivers methods for creating bag of visual features using `bagOfFeatures` class [11].

There are two ways for implementation:

- i. Without Extractor function

```
bag = bagOfFeatures (... , Name, Value)
```

- ii. Using Custom Extractor function

```
bag=bagOfFeatures(imds,'CustomExtractor', extractorFcn)
```

It will return a bag of features that uses a custom feature extractor function to generate the vocabulary features. Given below shows all the arguments of bag class along with their descriptions.

- i. `CustomExtractor`: A function handles to a custom feature extraction function.
- ii. `VocabularySize`: Number of visual words held by the bag.
- iii. `StrongestFeatures`: Fraction of strongest features to use from each label.
- iv. `PointSelection`: Method used to define point locations for feature extraction.
- v. `GridStep`: Step in X and Y directions defining the grid spacing.
- vi. `BlockWidth`: Patch sizes from which SURF descriptor is extracted.
- vii. `Upright`: Whether or not to extract upright SURF descriptors.

While implementing there are two files one is an ExtractorMain file named 'Bag1withExtractorsMain.m' which is the main file and the other one is BagOne-Extractor2SURF.m which is a SURF extractor function file. By executing the 'Bag1withExtractorsMain.m' on the various categories of WANG images, with vocabulary size of 500. The SURF features are detected and extracted numerically, after training and testing Results are displayed in Result Section (Figs. 5 and 6).

extractor = @BagOneExtractor2SURF;

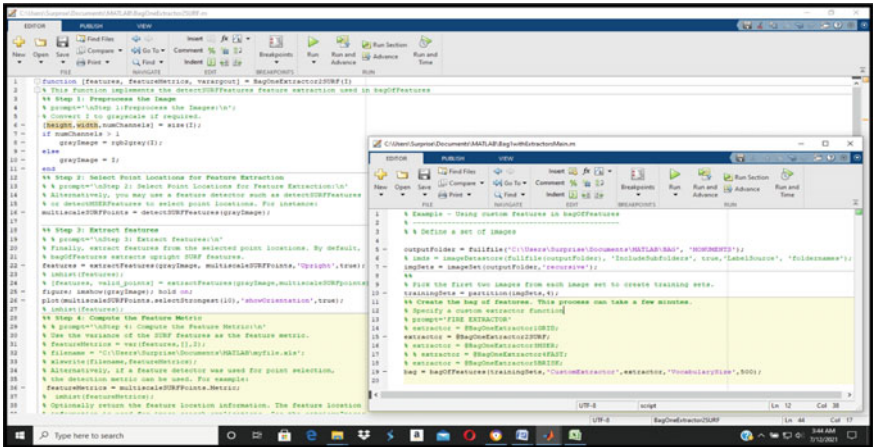


Fig. 5 Shows Editor/Implementation of Extractor and main file for SURF in MATLAB

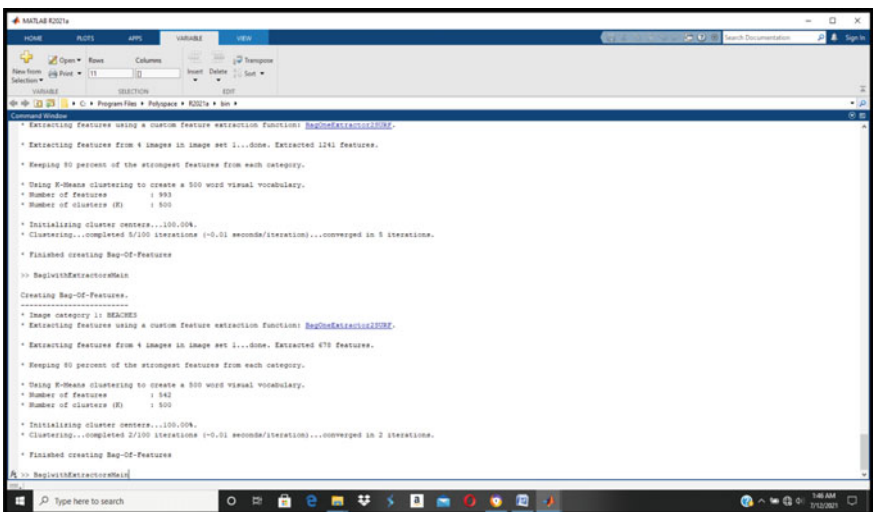


Fig. 6 Shows the command line view of results in MATLAB

WITHOUT EXTRACTOR							
Total images=1000 WANG DATASET							
Testing set=30%							
Training set=70%							
			Categories (No. of images per category after preprocessing)	No. of strongest features from each category	Number of features	Number of clusters (K)	Clustering completed
1	Selecting feature point locations	Detector method.	102 (*31)	546	55692	500	20/100 iterations (~0.64 seconds/iteration)
	Extracting Features	SURF					
	GridStep	[8* 8]					
	BlockWidth	[32 64 96 128].					
	918 images Extracted features.	202674					
	least number of strongest features	Image category 6 (546)					
WITH EXTRACTOR FUNCTION							
Testing set=30%							
Training set=70%							
			Categories (No. of images per category after preprocessing)	Number of strongest features from each category	Number of features	Number of clusters (K)	Clustering completed
2	Selecting feature point locations	BagOneExtractor2SURF	102 (*31)	11059	1128018	500	24/100 iterations (~15.26 seconds/iteration)
	Extracting Features	SURF					
	GridStep	[8* 8]					
	BlockWidth	[16 32 48 64].					
	918 images Extracted features.	4139784					
	least number of strongest features	Image category 4 (11059)					

Fig. 7 Results of WANG images when Bag of features Implemented without Extractor and with Extractor Function

bag = bagOfFeatures.

(trainingSets,'CustomExtractor', extractor,'VocabularySize', 500);

Meanwhile the implementation could be done with or without the use of Extractor function. The results show some of the wide differences in the number of features extracted. The extracted features and strongest features are more promising with the Extractor function. Therefore, it will be continued in implementation part and corresponding results are displayed in Sect. 5.5 (Fig. 7).

### 5.5 Results

The graphical representation of detected SURF features in the images of all the 10 categories when implemented using feature extractor function are presented in the Fig. 8 below. The features are marked in green circles taking radial distance from targeted objects. It can be inference from the results that 'flower' category images have least number of feature detected.

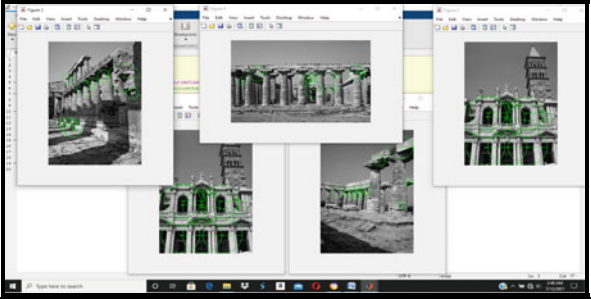
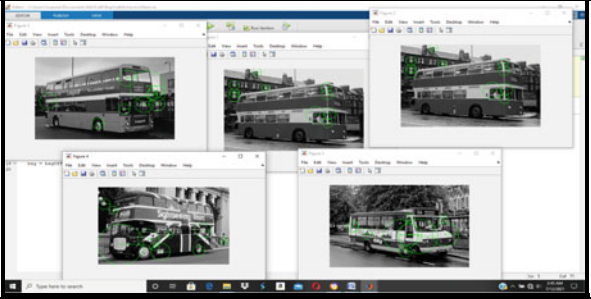
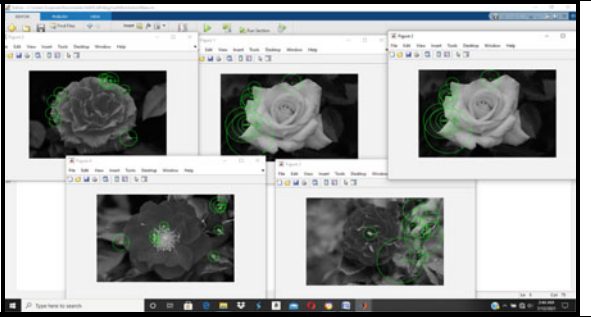
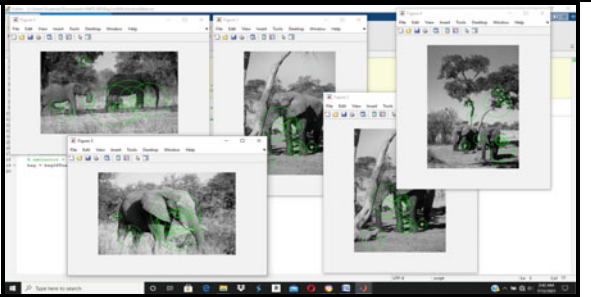
Category		
Monuments	 A collage of five screenshots from a software application showing SURF feature detection on various monument images. The detected features are highlighted with green circles. The images include classical buildings with columns and arches.	
Bus	 A collage of five screenshots from a software application showing SURF feature detection on various bus images. The detected features are highlighted with green circles. The images show buses from different angles and in different environments.	
Flowers	 A collage of five screenshots from a software application showing SURF feature detection on various flower images. The detected features are highlighted with green circles. The images include roses and other floral patterns.	
Elephants	 A collage of five screenshots from a software application showing SURF feature detection on various elephant images. The detected features are highlighted with green circles. The images show elephants in natural habitats.	

Fig. 8 Display the detected features in green circles in all the 10 categories of images

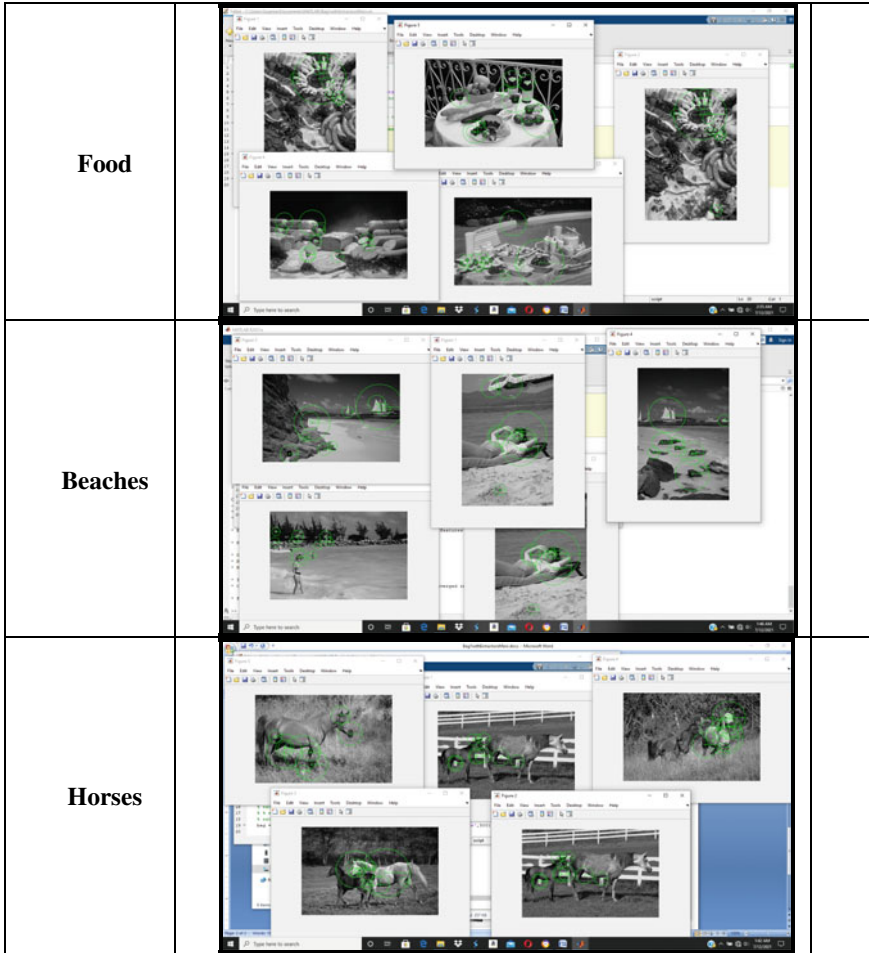


Fig. 8 (continued)

Images from all the 10 categories are trained and tested for detecting SURF features. It uses grayscale images therefore in preprocessing stage all the images are converted from RGB to grayscale.

### 5.6 Analysis

Table 1 depicts the analytical view of Extracted features using SURF detector. The size of clusters (K) is fixed to 500 for all Categories. Number of Iterations and time taken per second for iterations is also specified.

It is cleared from the above figures that FLOWER category performance is quite weak, as it has lowest number of features detected and extracted, while BUS has maximum number of features extracted as the local feature vectors extracted are based on color features, followed by MONUMENT and FOOD. It is also evident that iterations for clustering is directly proportional to the number of features. But it is not the case with time taken for completing iterations (Figs. 9, 10 and 11).

**Table 1** Analysis of extracted feature using SURF

Keeping 80% of the strongest features from each category  
 Extracting features using a custom feature extraction function: BagOneExtractor2SURF

S.NO	Category	Number of feature extracted	Number of Cluster(K)	Number of features Using K-Means clustering to create a 500 word visual vocabulary	Clustering completed/100 iterations	Seconds/iteration
1	African	1216	500	973	10	~0.01
2	Monument	1869	500	1492	22	~0.02
3	Bus	2053	500	1642	16	~0.02
4	Flower	306	245	245	1	~0.00
5	Elephant	1282	500	1026	6	~0.04
6	Food	1726	500	1381	11	~0.02
7	Beach	678	500	542	2	~0.01
8	Hourse	1121	500	897	10	~0.02
9	Mountain	1241	500	993	5	~0.01
10	Dianasour	734	500	587	1	~0.01

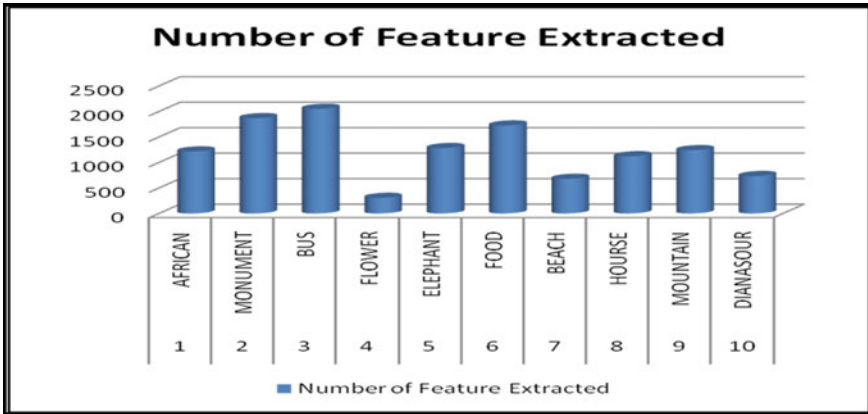


Fig. 9 Chart showing Number of Features Extracted from images per category

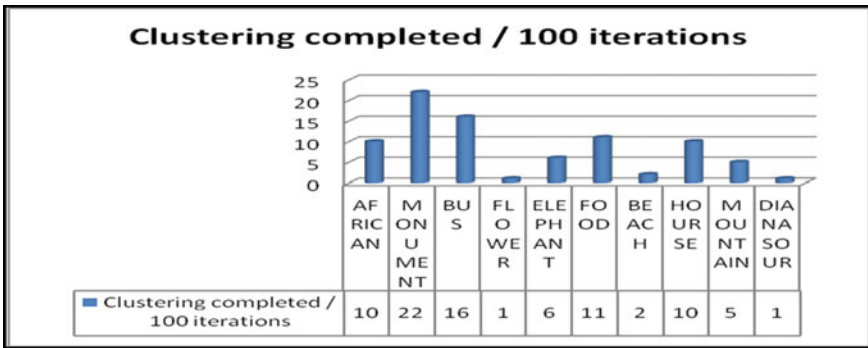


Fig. 10 Chart showing No. of cluster formed from Features Extracted of images/category

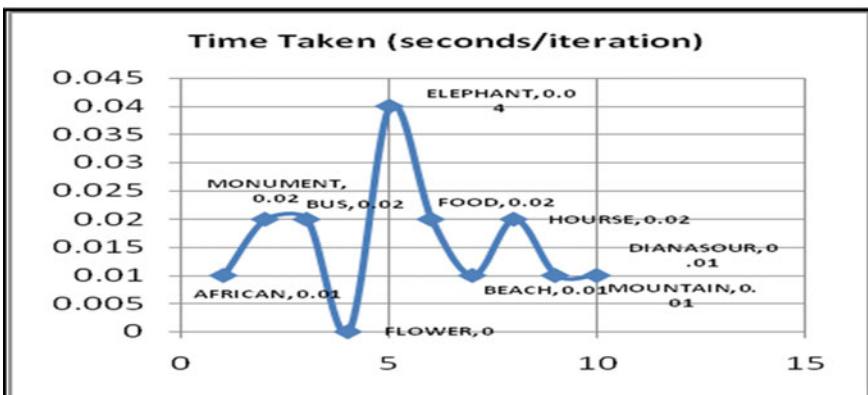


Fig. 11 Chart showing Time Taken in seconds for completing iterations for Features Extraction of images per category



## 6 Conclusions and Future Work

In this paper, we have discussed a new approach Bag of Visual Words for Image retrieval based on Bag of Word model that is successful in text mining. There are many positive points that drive researcher's attention like fast retrieval and bulk amount of features extracted and used. Therefore the retrieval performance goes quite well. In future it could be clubbed with different detectors other than SURF and with/without Custom Extractor function. There is a wide approach available to club it with traditional approaches like color Moments, Histogram or any new color, texture approaches for better classification results. It will be one of our next strategies to club the Bag feature vectors with Generative Adversarial Networks or pre trained networks.

## References

1. Bay H, Ess A, Tuytelaars T, Gool LV (June 2008) SURF: speeded up robust features. *Comput Vis Image Underst* 110(3):346–359
2. Calonder M, Lepetit V, Strecha C, Fua P (2010) BRIEF: binary robust independent elementary features. In: *European conference on computer vision*, vol 6314, pp 778–792
3. Rublee E, Rabaud V, Konolige K, Bradski G (2011) ORB: an efficient alternative to SIFT or SURF. In: *IEEE international conference on computer vision*, pp 2564–2571
4. Stephan Leutenegger MC, Siegwart R (2011) BRISK: binary robust invariant scalable keypoints. In: *IEEE international conference on computer vision*, pp 2548–2555
5. Alahi A, Ortiz R, Vanderghenst P (2012) FREAK: fast retina keypoint. In: *IEEE conference on computer vision and pattern recognition*, pp 510–517
6. Wohrer A (2008) Model and large-scale simulator of a biological retina with contrast gain control. Ph.D. thesis, University of Nice Sophia-Antipolis
7. Alsabti K, Ranka S, Singh V (1998) An efficient k-means clustering algorithm. In: *Proceedings of first workshop high performance data mining*
8. Vapnik V (1995) *The nature of statistical learning theory*. Springer, NY
9. [http://ai.stanford.edu/~jkrause/cars/car\\_dataset.html](http://ai.stanford.edu/~jkrause/cars/car_dataset.html)
10. [http://www.emt.tugraz.at/~pinz/data/GRAZ\\_02/](http://www.emt.tugraz.at/~pinz/data/GRAZ_02/)
11. [https://en.wikipedia.org/wiki/Speeded\\_up\\_robust\\_features/detectinterestpointsSURF,around\\_pointinterest](https://en.wikipedia.org/wiki/Speeded_up_robust_features/detectinterestpointsSURF,around_pointinterest)
12. <https://ww2.mathworks.cn/help/vision/ug/image-classification-with-bag-of-visual-words.html>

# Machine Learning Approaches for Image-Based Screening of Cervical Cancer



Priyanka Rastogi, Kavita Khanna, and Vijendra Singh

**Abstract** Cervical cancer can be prevented if it is diagnosed at an early stage using screening methods. The need for automated cancer screening methods is emphasized due to the tedious task of cytotechnicians and medical experts, who spend hours analyzing the cellular abnormalities under the microscope. The objective of this review paper is to gain awareness from the recent researches on automated screening methods using image processing techniques, classical machine learning, and more recent deep learning approaches. The paper primarily focuses on the classification of cervical cell types, both binary and multiclass. It also provides a comprehensive understanding of the datasets and the performance metrics used in these studies. It targets cancer studies promote more focused researches to be undertaken in the future. The recent literature shows that deep-learning-based methods are quite popular in delivering promising results in image-based classification and segmentation tasks in diagnosing cervical cancer. The development of such automated systems can assist doctors or pathologists in a much quicker error-free decision-making process.

**Keywords** Decision support system · Cervical cancer · Deep learning · Traditional machine learning · Pap smear · Cervical cancer screening

## 1 Introduction

The latest epidemiological study of cervical cancer states that [1] India accounts for about 33% of the deaths from cervical cancer worldwide. After breast cancer, it is the most frequent type of cancer among women. This cancer ascends from the cervix and is caused due to the irregular growth of cells that can invade or spread to other

---

P. Rastogi (✉)  
CSE &IT Department, The NorthCap University, Gurugram, India  
e-mail: [livepriyanka09@gmail.com](mailto:livepriyanka09@gmail.com)

K. Khanna  
Delhi Skill Entrepreneurship University, Delhi, India

V. Singh  
Department of Computer Science, UPES, Dehradun, India

areas of the body. More than 90 percent of cases are caused by human papillomavirus (HPV). The lack of detection in the early stages is one of the causes of high mortality. Screening methods can help early diagnosis and treatment of pre-cancer and early cancer thus helping to reduce cases of cervical cancer, thereby reducing the mortality rate. Conventional cytology screening programs have helped immensely in reducing the cases of cervical cancer in other developed countries.

Cervical cancer includes two types of cellular screening tests namely conventional pap test and Liquid-based cytology (LBC). The Pap smear is a screening procedure that detects pre-cancerous and cancerous processes, by collecting a sample of cervix cells that are smeared on a glass slide stained by the Papanicolau and analyzed under a microscope. In the more recent liquid-based pap cytology examination, cervical cells extracted with a brush or other tool are put in a vial of liquid preservative and sent for analysis to a laboratory for the presence of high-risk variants of HPV along with cellular abnormalities. The ability to detect cellular anomalies tends to be similar in both traditional and liquid-based pap tests. If any of the screening tests are found to be positive, further testing is done to determine whether the changes are cancerous. A colposcopy may be carried out on a small sample of tissue collected from the cervix and/or a biopsy may be carried out. Figure 1 shows a sample image of pap smear image with carcinoma in situ and a sample cervigram image with CIN2 grade cervical cancer.

There have been constant efforts for the development of computer-assisted diagnosis (CAD) systems that would reduce the tedious workload of the pathologists screening the pap smear or LBC. The cervical cells under the microscope are screened for cellular abnormalities and cervigram images are analyzed for physician interpretation. Repetitive work can cause fatigue and be prone to human error while reporting incidences. Also, the development of such computerized systems would reduce the lag in test results and can contribute to timely follow-up and treatment planning. These automated systems can assist doctors or pathologists in a much quicker error-free decision-making process.

There have been multiple research areas that have contributed to the development of CAD systems for the diagnosis of cervical cancer. Several kinds of research are



**Fig. 1** Sample pap smear image with cell-type carcinoma in situ (abnormal cell) on the left side. Sample cervigram image with CIN2 grade (cancerous) on the right side

attributed to the automated screening of cancer from cervigram images obtained from colposcopy [2, 3], and at the cellular level [4–7]. In the past few decades, researchers were mainly focused on the computer-assisted diagnosis of cervical cancer based on pap smears images and diagnosis workflow followed segmentation and designing a novel handcrafted feature for cervical nuclei as it is an important indicator in the classification of cancerous cells using traditional machine learning methods [5, 8, 9]. With the technological improvements, deep-learning-based methods are being employed for providing end-to-end automated solutions for cancer diagnostics due to their self-learning ability[10].

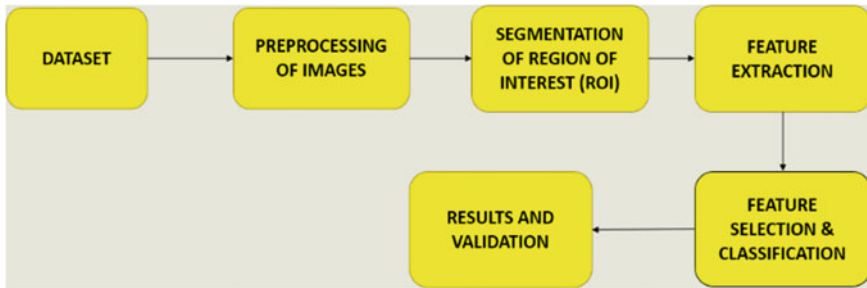
The objective of this review paper is to help new researchers gain insight into the various studies for computer-aided screening of cervical cancer. Such kind of targeted cancer studies gives a total round-up of the researches being carried out in this direction. This paper is laid out in five sections. Section 2 reviews the work done in the past by various researchers toward automating the diagnosis of cervical cancer, the basic workflow followed by them, the challenges faced by using traditional machine learning techniques in the diagnosis of cervical cancer cell image data, and the currently existing methods used. Section 3 describes the different publicly available benchmarked cervical cancer datasets. Section 4 discusses the performance measures used for assessing the different models. Finally, Sect. 5 concludes the paper by providing a discussion on the reviewed researches and giving pointers for further research opportunities in this direction.

## 2 Approaches Used for Image-Based Applications

### 2.1 *Traditional Machine Learning-Based Methods*

The focus of the classical machine learning approach was on input data representation using handcrafted features for regions of interest (ROI) from the pre-processed biomedical image and later training the classifier based on these features. The general steps followed for the analysis of cytopathological images (like pap smear cells) are shown in Fig. 2. The first step in the pipeline is pre-processing of the cell images by using various image processing methods, later segmenting the region of interest (ROI), e.g., cell nuclei, thereby extracting several features from the segmented regions using various feature extraction techniques. Among the extracted features, the relevant features are selected using feature selection algorithms (or feature vectors are reduced using dimensionality reduction methods). Lastly, the selected subset of features is used for training the classifier.

Many kinds of research have been focused on using traditional machine learning methods for the classification of pap smear images in the Herlev dataset. The dataset is described in brief in Sect. 3 of this paper. The initial set of methods applied for cell classification on an earlier version of the Herlev dataset [11] used inductive methods FCM, GK, ANFIS, and transductive methods like NNH. Jantzen [11] from



**Fig. 2** General steps followed for medical image analysis using classical machine learning approaches

a group of segmented and unsegmented cell images, extracted 20 numerical feature descriptors. A lot of researches have been focused on classifying pap images in the Herlev dataset based on this initial set of extracted features. Raina et al. [12] used the ensemble learning methods technique for pap smear cell classification using the initial 20 features. They trained the Naïve Bayes model and various other models by using combinations of bagging ensemble techniques and the Greedy Forward Selection (GFS) feature selection technique. In another study involving the use of an initial 20 feature set, Dewi et al. [13] compared a simple Naïve Bayes classifier with a combination of Naïve Bayes (NB) classifier and weighted PCA for the classification of single-cell pap smears. Simple Naïve Bayes performed far better for binary classification compared to other methods.

Apart from that, many researchers proposed a binary class and multi-class classification model by training the classifier directly on pap smear images from the Herlev dataset. Mariarputham et al. [7] developed a cervical cancer multi-class classification system for classifying PAP smears into one of the seven classes using nominal texture-based features. They trained SVM and Neural Network classifiers using 24 textural features for classification. Mbaga et al. [6] used median filter and linear contrast enhancement on greyscale pap smear images to remove unwanted noise and increase contrast respectively as a pre-processing step. They used Otsu thresholding for cell segmentation thereby extracting 20 salient features from the segmented cells. They used a feature selection wrapper method called Recursive feature elimination (RFE) with SVM (SVM-RFE) for classification. Sharma et al. [5] used Gaussian Filter and Histogram equalization to remove unwanted noise and further improve image quality respectively as pre-processing steps and used Sobel gradient filter for segmentation and extracted 7 morphological descriptors. They experimented on a cervical pap smear dataset collected from Fortis Hospital Mohali in 2016. S. Athinayanan et al. [8] extracted features using three newly proposed nuclei feature extraction techniques from segmented nuclei from an unpublished pap smear dataset collected at Muthamil Hospital in Tirunelveli (TN) in 2013. The dataset contained a total of 952 images out of which 440 were normal and 512 were abnormal cell images. The nuclei-based descriptors were then trained on 2 classifiers SVM and

Kernel-SVM for binary classification problems. The results showed that EEETCM + KSVM performed best. The important details of the semi-automated approaches used in cervical cell classification are tabulated in Table 1.

Studies related to cytology segmentation using image processing have also been taken up by researchers. Nisar et al. [14] used various image processing algorithms for segmentation of cell clump using Otsu thresholding method with Iterative Region (IRO), MSER method, and level set evolution using distance regularization (DRLSE) for nuclei segmentation and cell cytoplasm segmentation respectively. They used overlapping cervical cytology images released as a part of the ISBI-2014 and ISBI 2015 segmentation challenge. A similar study to segment overlapping cells was taken by Kumar et al. [4]. Cell clumps were segmented using adaptive histogram equalization and H-maxima transform followed by the newly proposed Otsu method with prior class weights for precise nuclei segmentation from the cell clumps and

**Table 1** Semi-automated approaches used in cervical cancer classification reviewed in this paper

References	Preprocessing techniques	Segmentation techniques	No. of extracted features	Classifier
Mariarputham et al. [7]	Image inversion, binarization	Morphological operations on image binarization	24	SVM, Neural network
Mbaga et al. [6]	Grayscale conversion, Median filter, Linear contrast enhancement	Cell Segmentation using Otsu thresholding	20	SVM-RFE
Sharma et al. [5]	Gaussian Filter, Histogram equalization	Sobel gradient filter	7	KNN
Athinarayanan et al. [8]	Anisotropic filter	Nucleus Segmentation using thresholding and feature extraction using EEETCM, EEETCM, CFE	3	K-SVM, SVM
Raina et al. [12]	Numerical descriptors of Herlev benchmark dataset	–	20	Ensemble methods using NB with Greedy Forward Selection (GFS)
Dewi et al. [13]	Numerical descriptors of Herlev benchmark dataset	–	20	Simple Naïve Bayes (NB), weighted PCA with NB

the distance regularized level set model for cytoplasm segmentation on ISBI 2015 challenge dataset.

## ***2.2 Challenges with the Traditional Approaches***

The major challenges faced by the research community for analysis of cervical pap images are the occurrence of overlapping of cytology images, smear thickness due to cell clumps, un-wanted particles in the smear, presence of cell bands, and pathologies. All of these factors can lead to erroneous detections and impede the development of automated methods using traditional machine learning methods. Also, identification of the areas of interest and extracting handcrafted features for training the classifier have greatly influenced the performance of the machine learning model. This plays a key role in limiting the performance of ML methods. However, with the advent of GPU and technological advancements, deep-learning-based approaches are gaining popularity due to their self-learning ability to overcome the problem of feature extraction and designing segmentation-free detection architectures to make quicker diagnosis pipelines.

## ***2.3 Deep-Learning-Based Methods***

Deep learning is one of the approaches being investigated in recent times due to the success of tasks involving computer vision in other domains. Deep learning algorithms are a subcategory of machine learning algorithms that can learn significant low-level features (like edges/line or textural features) and high-level features and integrate them in subsequent layers to form discriminative nodes capable of classifying. They are implemented using a deep neural network architecture having numerous hidden layers. Convolutional Neural networks (ConvNets or simply CNNs) are one of the most ubiquitous supervised deep learning methods in computer vision tasks because of their ability to preserve spatial relationships in images. These relationships are critically important, especially in biomedical image data and it is because of this reason that they have been exploited most in medical image analysis tasks as well in recent times [15–17].

Xiang et al. [10] presented a fully segmentation-free method for directly detecting cervical cells or clumps using the latest variant of the object detection algorithm, YOLOv3. They cascaded an additional classifier for taking care of the task-specific requirements thereby improving the classification results for hard examples. They experimented on their own established dataset containing 12,909 cervical images with 10 categories of objects from cervical cell images along with 58,995 ground truth boxes. Zhang et al. [18] proposed a “DeepPap”, a convolutional neural network with transfer learning to directly classify cells based on deep features. The pre-trained network (with ImageNet) was fine-tuned on augmented coarsely nuclei-centered

image patches. The proposed model was assessed on two kinds of cervical cytology images—the Herlev dataset and the HEMLBC dataset. The model obtained superior performances for both datasets as compared to the top promising traditional machine learning methods. Bora et al. [9] automated feature extraction method using deep CNN Alex model thereby selecting features using unsupervised feature selection method based on similarity measure. The features were trained on two classifiers LSSVM and Softmax. The model was trained on a database of 1611 PAP smear images collected from two diagnostic centers in Assam and validated on generated and the Herlev dataset. Nguyen et al. [19] proposed a methodology in which they concatenated the features extracted from 3 pre-trained deep CNNs. They tested their work on the Herlev dataset and 2D-Hela dataset. Plissiti et al. [20] conducted various experiments using a novel SIPaKMeD dataset and achieved the best accuracy with raw image pixels trained over the CNN model. Bhatt et al. [21] harnessed the power of transfer learning with a scaled CNN network, EfficientNetB3. They carried out experiments for both binary and multiclass classification over multiple datasets. Rahaman et al. [22] proposed DL based framework using hybrid deep feature fusion for classification called “DeepCervix”.

Apart from the cellular-level diagnosis, research towards automated diagnosis/screening based on images from colonoscopy are also being carried out. Xu et al. [3] carried out a study for cervical dysplasia classification on a subset of cervigram images from the National Cancer Institute, United States. Several features were extracted to capture color, edge/gradient, and texture information, respectively. They benchmarked the cervigram image dataset with the ground truth annotations. A comprehensive comparison among seven popular classifiers was conducted to differentiate images of high-risk patient cases from those of low-risk patient cases. The results illustrated that the ensemble-tree-based models performed better than other classifiers. Saini et al. [2] proposed “ColpoNet”, a deep CNN-based model, inspired by DenseNet for the early screening of cervical cancer using cervigram images from the National cancer institute. The proposed model was compared with other CNN-based state-of-the-art models and achieved an accuracy of 81.353%. Table 2 provides the summary of various application areas for cervical cancer studies based on reviewed papers along with the datasets and their results.

## 3 Dataset

### 3.1 Herlev Dataset

Herlev dataset<sup>1</sup> is a benchmark Pap smear dataset developed with the motive to recognize the different categories of cervical cancer cells using computer vision. The data were collected at the Herlev University Hospital. Cytology experts and

---

<sup>1</sup> <http://labs.fme.aegean.gr/decision/downloads>.



**Table 2** Important application areas in cervical cancer research

Applications	References	Datasets	Results
Cellular-level image classification (Binary Classification)	Mbaga et al. [6]	Herlev Dataset	Accuracy = 97.02%
	Athinarayanan et al. [8]	Unpublished pap smear dataset	EEETCM + KSVM with Accuracy = 94.5%, Sensitivity = 97%, Specificity = 92%
	Raina et al. [12]	Herlev numerical descriptor dataset	NvB + GFS model with accuracy = 92.15%
	Dewi et al. [13]	Herlev numerical descriptor dataset	For Binary class, Naïve Bayes with accuracy = 90.42%
	Bora et al. [9]	Herlev dataset	2 level LSSVM-Accuracy = 94.61%, Precision = 92.16%, Recall = 97.06%, Specificity = 88.35%, F-score = 90.21%
	Zhang et al. [18]	Herlev Dataset, HEMLBC dataset (Unpublished dataset)	For Herlev Dataset, Accuracy = 98.3%; For HEMLBC dataset, Accuracy = 98.6%
	Nguyen et al. [19]	Herlev Dataset	Accuracy = 92.63%
	Xiang et al. [10]	Unpublished LBC dataset	Accuracy = 89.3%, Sensitivity = 97.5%, Specificity = 67.8%
	Bhatt et al. [21]	SIPaKMeD dataset	Accuracy = 99.01%, Precision = 99.157%, Recall = 98.896%, K-score = 98.88%, F-score = 99.026%
Rahaman et al. [22]	Herlev dataset, SIPaKMeD dataset	Accuracy (Herlev) = 98.91%; Accuracy (SIPaKMeD) = 99.85%	
Cellular-level Image classification (Multiclass classification)	Mariarputham et al. [7]	Herlev Dataset	Linear kernel SVM with AUC > = 0.75 for all classes, Precision > = 0.92 for all classes
	Sharma et al. [5]	Unpublished Pap smear dataset	Accuracy = 82.9%

(continued)

**Table 2** (continued)

Applications	References	Datasets	Results
	Raina et al. [12]	Herlev numerical descriptor dataset	NvB + GFS + BG model with accuracy = 63.25%
	Dewi et al. [13]	Herlev numerical descriptor dataset	Naïve Bayes + Weighted-PCA with accuracy = 87.24%
	Plissiti et al. [20]	SIPaKMeD dataset	Accuracy = 95.35%
	Diniz et al. [23]	CRIC Dataset; Herlev Dataset	Accuracy (CRIC) = 0.9686; Accuracy (Herlev) = 0.9543
	Bhatt et al. [21]	SIPaKMeD dataset, Herlev Dataset	Accuracy (SIPaKMeD) = 99.7%; Accuracy (Herlev) = 93.14%
	Rahaman et al. [22]	Herlev dataset, SIPaKMeD dataset	Accuracy (SIPaKMeD -3 class) = 99.38%; Accuracy (SIPaKMeD -5 class) = 99.14%; Accuracy (Herlev - 7 class) = 90.32%;
Cytology Segmentation	Nisar et al. [14]	ISBI-2014, 2015	Pixel-based evaluation, Nucleus segmentation in real images, Dice = 0.789 TPR = 0.754 Cytoplasm segmentation in real images, Dice = 0.729, TPR = 0.841
	Kumar et al. [4]	ISBI- 2015	On training set, Dice coefficient = 0.86, TPR = 0.88
Cervigram based image classification	Xu et al. [3]	Cervigram images from NCI	On balanced dataset, RF achieved AUC = 84.63%, Accuracy = 80.00%, Sensitivity = 84.06%, Specificity = 75.94%
	Saini et al. [2]	Cervigram images from NCI	Accuracy = 81.353%

trained cytotechnicians had manually classified every single cell image into one of the 7 cell type classes. Out of which 3 cell type classes are categorized as normal cells while 4 classes are of precancerous or cancerous cell types. There is a total of 917 cells in the database with two times more abnormal cells than normal cells. The image dataset is also provided with the initial measurements where each sample is described by 20 features extracted from images of single cells of the dataset. The

cell selection was done to make a virtuous assortment of the significant classes and not to mimic a typical distribution of classes in the real world.

### 3.2 *SIPaKMed Database*

The SIPaKMeD database<sup>2</sup> consists of 4049 images of isolated cervical cells, which have been cropped from 966 cluster cell images of pap smear slides manually. The database has cell images of 1600 normal cervical cells, 793 benign metaplastic cervical cells, and 1638 abnormal cervical cells. The cropped cell image database has a total of 5 cell type categories. All cells have been classified by expert cytopathologists depending on their cellular appearance and morphology.

### 3.3 *CRIC Searchable Image Database*

It is a cervical cell image database developed by the Centre for Recognition and Inspection of Cells (CRIC) to support pap smear analysis. The cervical cells from the conventional cytology have been standardized using The Bethesda System (TBS) of nomenclature, which is the most commonly used system of reporting pap smear results in the clinical setting. The database<sup>3</sup> is divided into two collections—for classification and segmentation. The group focused on classification containing markings of the cell's center while the other group contains annotations of the cell's nucleus and cytoplasm for segmentation. There are a total of 3233 images with 862 normal cervical cells and 2371 abnormal (or cancerous) cervical cells.

### 3.4 *Overlapping Cells Dataset of ISBI 2014 and ISBI 2015 Challenge*

The cervical overlapping cells dataset<sup>4,5</sup> was released as part of the ISBI-2014 and ISBI-2015 segmentation challenge. The ISBI-2014 dataset consisted of 16 EDF actual cervical cytology images and 945 synthetic images. The dataset was released in two phases with 135 synthetic images for training and testing and the rest images for qualitative and quantitative assessments. The objective of the challenge was to extract the exact borders of the nucleus and cytoplasm of individual cells from the overlapping cervical cytology images. Analysis of overlapping cells is challenging

---

<sup>2</sup> [www.cse.uoi.gr/\\_marina](http://www.cse.uoi.gr/_marina).

<sup>3</sup> <https://database.cric.com.br/>.

<sup>4</sup> [https://cs.adelaide.edu.au/~carneiro/isbi14\\_challenge/index.html](https://cs.adelaide.edu.au/~carneiro/isbi14_challenge/index.html).

<sup>5</sup> [https://cs.adelaide.edu.au/~zhi/isbi15\\_challenge/description.html](https://cs.adelaide.edu.au/~zhi/isbi15_challenge/description.html).

since in the manual examination under the microscope, focus adjustments are made to help in the reading and understanding of overlapped cells, and automating this focus adjustment is a difficult task. Therefore, an extended depth of field (EDF) cytology image is used that puts all objects in focus into a single image.

The input data consisted of a multi-layer cytology volume in the ISBI-2015 overlapping cervical cell segmentation challenge, which meant that the input data was now a volume consisting of a series of multi-focal images obtained from the same specimen. This dataset was potentially more informative for nucleus detection tasks and effective cervical segmentation. It consisted of a collection of 17 multi-layer cervical cell volumes, from which 8 were used for training and 9 for testing.

### 3.5 *National Cancer Institute (NCI)*

Cervigram images are obtained by conducting cervicography. The cervigram dataset for study in [3] was obtained from the Guanacaste project of the National Cancer Institute (NCI). The collection consisted of 10,000 anonymized women's records. The collected dataset<sup>6</sup> composed of 345 positive (CIN2/3/4) patient cases and 767 negative (CIN1/Normal) cases. The ground truth assessment for each patient included the CIN grade of cancer. Another study [2] used one more subset of Cervigram images data from NCI, including various types of images reflecting the different cancer stages. The collected dataset had 2 sets—Type 1 images (includes Normal/CIN1) and Type 2 images (includes CIN2/CIN3/CIN4), 400 in each set, chosen based on image clarity.

## 4 Performance Measures

### 4.1 *Evaluation Metrics for Recognition Tasks at the Cellular-Level and Cervigram Images*

For the various classification and detection tasks at the cellular-level and cervigram images, the classical measures for assessing the classifier are Accuracy, Specificity, Sensitivity, Precision, Recall, F-score, and AUC. Accuracy is the overall performance of the classifier and is given by Eq. (1). Specificity is the measure of the ability of the model to correctly identify normal cells while Sensitivity (or Recall) is the measure of the ability of the model to correctly identify abnormal cells. Precision is the ratio of correctly identified abnormal cells by model among all retrieved instances as abnormal by the model.

---

<sup>6</sup> <http://www.cse.lehigh.edu/idealab/cervitor>.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

where TP indicates the number of cancerous cells that are correctly classified as cancerous (True Positives). TN is the number of normal cells that are correctly classified as normal (True Negatives). FP is the number of normal cells that are wrongly classified as cancerous (False Positives). FN is the number of cancerous cells that are wrongly classified as normal (False Negatives).

The F-score (or F1-score or F-measure) tells the model's accuracy on a dataset by combining the model's precision and recall. It ranges between 0 and 1. F-score indicates perfect precision and recall. In simple terms, a "good" F-score is when there are low false positives and low false negatives. Another important metric to evaluate the model is ROC AUC (AUROC), which is a probability curve to display the performance of a classification model at various thresholds of TPR and FPR. The higher AUC indicates the model's efficacy at distinguishing between the classes.

#### ***4.2 Evaluation Metrics for Segmentation Task in ISBI 2014 and ISBI 2015 Challenge***

The quantitative performance measures used in ISBI 2014 and 2015 challenge for assessing segmentation of individual cell cytoplasm was the average Dice Coefficient. DC is twice the area of overlap divided by the union of pixels in both images. It is given by Eq. (2).

$$\text{DC}(A, B) = 2(|A \cap B|) / (|A| + |B|) \quad (2)$$

The acceptable cell segmentation was expected to have DC greater than 0.7, any value less than it was considered as false negative as per challenge evaluation criteria. For the acceptable cell segmentation, pixel-based TPR and FPR were calculated for both training and test sets.

## **5 Discussion and Conclusion**

The importance of an automated cervical screening system is extremely vital for developing nations like India to aid in reducing the workload of lab technicians and also help doctors in decision-making. It has been observed from the image classification research that those high accuracies have been reported for binary classification tasks while multiclass classification of cancer grades/cell types from images still needs a lot of improvement. Consistent efforts are being made to design complex algorithms to solve such challenging tasks where the images look visually very

similar but belong to different class labels. Rapid technological advancements have also empowered more extensive research in this field. A lot of deep-learning-based methods are being used for developing fully automated screening systems using LBC, colposcopy, and cervigram images. DL-based research has shown promising results; however, these methods are data-hungry and require ample amounts of data for training. Also, there is a dearth of benchmarked datasets because of which the performance of models built by various researchers cannot be compared on common grounds. The absence of availability of clinical public standard cervical multi-cells dataset is one of the reasons which impede the growth of robust automated systems. Herlev dataset was the oldest and most studied dataset for cervical cell classification, it is because of this reason a major portion of research literature is based on it. The addition of new classification datasets in recent times offering comprehensive coverage of different cervical cancer cell types have the potential to boost research in this area. This review has tried to provide the current state of research in the automated image-based screening of cervical cancer which is necessary for targeted cancer studies and gives a total round-up of the researches being carried out in this direction.

**Declaration of Originality** I certify that this is my work, and it has not previously been submitted for any other conference or journal publication. I also certify that the use of material from other sources has been properly and fully acknowledged in the text.

## References

1. Monica, Mishra R (2020) An epidemiological study of cervical and breast screening in India: district-level analysis. *BMC Women's Health* 20(1):225. <https://doi.org/10.1186/s12905-020-01083-6>.
2. Saini SK, Bansal V, Kaur R, Juneja M (2020) ColpoNet for automated cervical cancer screening using colposcopy images. *Mach Vis Appl* 31(3):15. <https://doi.org/10.1007/s00138-020-01063-8>
3. Xu T et al (2015) A new image data set and benchmark for cervical dysplasia classification evaluation. In: *Machine learning in medical imaging*, pp 26–35
4. Kumar P, Happy SL, Chatterjee S, Sheet D, Routray A (2016) An unsupervised approach for overlapping cervical cell cytoplasm segmentation. In: *2016 IEEE EMBS conference on biomedical engineering and sciences (IECBES)*, pp 106–109. <https://doi.org/10.1109/IECBES.2016.7843424>
5. Sharma M, Kumar Singh S, Agrawal P, Madaan V (2016) Classification of Clinical Dataset of Cervical Cancer using KNN. *Indian J Sci Technol* 9(28). <https://doi.org/10.17485/ijst/2016/v9i28/98380>
6. Mbagi AH, Zhijun P (2015) Pap smear images classification for early detection of cervical cancer. *Int J Comput Appl* 118(7):975–988
7. Mariarpatham EJ, Stephen A (2015) Nominated texture based cervical cancer classification. *Comput Math Methods Med* 2015:586928. <https://doi.org/10.1155/2015/586928>
8. Athinarayanan S, Srinath MV, Kavitha R (2016) Detection and classification of cervical cancer in pap smear images using EETCM, EEETCM and CFE methods based texture features and various classification techniques. 2(5):533–549. <https://doi.org/10.18535/ijecs/v5i7.32>

9. Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK (2016) Pap smear image classification using convolutional neural network. In: ACM international conference proceeding series. <https://doi.org/10.1145/3009977.3010068>
10. Xiang Y, Sun W, Pan C, Yan M, Yin Z, Liang Y (2020) A novel automation-assisted cervical cancer reading method based on convolutional neural network. *Biocybern Biomed Eng* 40(2):611?623. <https://doi.org/10.1016/j.bbe.2020.01.016>
11. Jantzen J, Norup J, Dounias G, Bjerregaard B (2005) Pap-smear benchmark data for pattern classification. In: Proceedings of NiSIS 2005. Albufeira, Portugal, pp 1–9
12. Riana D, Hidayanto AN, Fitriyani (2017) Integration of Bagging and greedy forward selection on image pap smear classification using Naïve Bayes. In: 2017 5th international conference on cyber and IT service management, CITSM 2017. <https://doi.org/10.1109/CITSM.2017.8089320>
13. Dewi YN, Riana D, Mantoro T (2018) Improving Naïve Bayes performance in single image pap smear using weighted principal component analysis (WPCA). In: 3rd international conference on computing, engineering, and design, ICCED 2017, vol. 2018, pp 1–5. <https://doi.org/10.1109/CED.2017.8308130>
14. Nisar H, Wai LY, Hong LS (2018) Segmentation of overlapping cells obtained from pap smear test. In: 2017 IEEE life sciences conference, LSC 2017, vol 2018. Janua, pp 254–257. <https://doi.org/10.1109/LSC.2017.8268191>
15. Ker J, Wang L, Rao J, Lim T (2017) Deep learning applications in medical image analysis. *IEEE Access* 6:9375?9379. <https://doi.org/10.1109/ACCESS.2017.2788044>
16. Hu Z, Tang J, Wang Z, Zhang K, Zhang L, Sun Q (2018) Deep learning for image-based cancer detection and diagnosis—a survey. *Pattern Recogn* 83:134?149. <https://doi.org/10.1016/j.pat.cog.2018.05.014>
17. Rastogi P, Singh V, Yadav M (2018) Deep learning and big datatechnologies in medical image analysis. In: PDGC 2018–2018 5th international conference on parallel, distributed and grid computing, pp 60–63. <https://doi.org/10.1109/PDGC.2018.8745750>
18. Zhang L, Lu L, Nogue I, Summers RM, Liu S, Yao J (2017) DeepPap: ddeep convolutional networks for cervical cell classification. *IEEE J Biomed Health Inform* 21(6):1633?1643. <https://doi.org/10.1109/JBHI.2017.2705583>
19. Nguyen LD, Lin D, Lin Z, Cao J (2018) Deep CNNs for microscopic image classification by exploiting transfer learning and feature concatenation. In: 2018 IEEE international symposium on circuits and systems (ISCAS), pp 1–5. <https://doi.org/10.1109/ISCAS.2018.8351550>
20. Plissiti ME, Dimitrakopoulos P, Sfikas G, Nikou C, Krikoni O, Charchanti A (2018) Sipakmed: a new dataset for feature and image based classification of normal and pathological cervical cells in pap smear images. In: 2018 25th IEEE international conference on image processing (ICIP), pp 3144–3148. <https://doi.org/10.1109/ICIP.2018.8451588>
21. Bhatt AR, Ganatra A, Kotecha K (2021) Cervical cancer detection in pap smear whole slide images using convNet with transfer learning and progressive resizing. *PeerJ. Comput Sci* 7:e348?e348. <https://doi.org/10.7717/peerj-cs.348>
22. Rahaman MM et al (2021) DeepCervix: A deep learning-based framework for the classification of cervical cells using hybrid deep feature fusion techniques. *Comput Biol Med* 136:104649. <https://doi.org/10.1016/j.combiomed.2021.104649>
23. Diniz DN et al (2021) A hierarchical feature-based methodology to perform cervical cancer classification. *Appl Sci* 11(9). <https://doi.org/10.3390/app11094091>

# A Comparative Analysis of Image Steganography Based on Discrete Wavelet Transform (DWT) and Exploiting Modification Direction (EMD)



Abhishek choubey and Shruti Bhargava Choubey

**Abstract** The art of hiding the existing data over a covering medium which may be an image or any other media file is called steganography. In this paper, we will try to review major steganographic methods that are existing until now for the various image formats. We will majorly do a comparative analysis between discrete wavelet transform (DWT)-based steganography and exploiting the modification direction (EMD) algorithm. In DWT-based steganography, we will try to conceal data in lower frequency bands of the image data which is less perceivable by the human eye, and in EMD-based steganography we will try to find an extraction function based on an algorithm for embedding data. For practical analysis, we will be using Matlab as a software tool that is robust while dealing with image processing.

**Keywords** Discrete Wavelet Transforms (DWT) · Exploiting Modification Direction (EMD) · Steganography

## 1 Introduction

In this world of the digital era, privacy became the main concern for the people who need a lot more efficient secured signal transmission techniques. From early human life itself, this secured communication began to develop in the form of various sign languages for various tribes of people. After that, civilized man used to convey their messages using different sign languages on cave walls, thus different languages evolved; with the increase in the knowledge base and greediness of humans, some sort of techniques needed to be developed to communicate securely with a particular set of communities [1]. Based on the way they hide data, these data hiding techniques are mainly classified into three types: one is steganography and the other two are

---

A. choubey (✉) · S. B. Choubey

Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, Hyderabad, India

e-mail: [abhishek@sreenidhi.edu.in](mailto:abhishek@sreenidhi.edu.in)

S. B. Choubey

e-mail: [shrutibhargava@sreenidhi.edu.in](mailto:shrutibhargava@sreenidhi.edu.in)



cryptography and watermarking, while the first two techniques are for hiding data and the watermarking is for protecting the owner's writing of particular content. Watermarking is not a technique to hide data but helps us to protect the ownership rights of our content. Earlier, they use to implement this watermark using transparent text or logo-based embedding in our content, but nowadays with the advancement of programming digital watermarking came into wide existence.

The word cryptography is derived from the Greek words 'crypt' which means encrypted/hidden and 'graphy' means writing. So, it involves the conversion of message or data into a format using some key so that unauthorized users won't be able to read it. There are various cryptographic techniques available today which will enable users to authenticate each other using key pairs; some of them are public-key cryptography, secret-key cryptography, and hash function-based cryptography. But the main disadvantage here is the intruder will know that some secret message is being communicated so if he tries his level best, he can crack the code. Steganography is also derived from the Greek words which mean covered writing. In this technique, the intruder won't be able to that data is hidden so he won't even try to crack the message itself [3–5]. Steganography is dated long back and is used by the Chinese people in the form of punch hole writing; in this method, they have a mask of holes by keeping it on a paper; they write a secret message text and after removing the mask, they complete the text without a gap so that the intruder will have a view of nothing but randomized text. The message text can be viewed only by the receiver with the mask. During the world times, they use several techniques related to this steganography some of them including the microdots technique, invisible ink, and cipher-based technique.

The Block diagram of the steganography method is shown in Fig. 1. The various steganographic techniques have evolved till now from the least significant bit models to the frequency domain models. Whatever may be the techniques, the main aim is to reduce the noise ratio while retrieving back and increased security using different algorithms. In this paper, we will study different methods that are available up to now and we will mainly focus to do a comparative analysis between discrete wavelet transform and exploit modification direction-based techniques.

Arrhythmia monitors for ambulatory patients that analyze the ECG in real time are currently in development [5–8]. For software in most QRS detectors, one or more of the following processing steps are used: nonlinear transformation, linear digital filtering, and decision rule algorithms [4].

## 2 Review

In this section, we will have a summary of the most significant steganographic techniques that are used in the case of various image formats Portable Network Graphics (PNG), Joint Photographic Experts Group (JPEG), Bitmap format (BMP), and Graphics Interchange Format (GIF). The main idea of steganographic techniques involves embedding a message in the above-mentioned survey of papers [5–12]; each

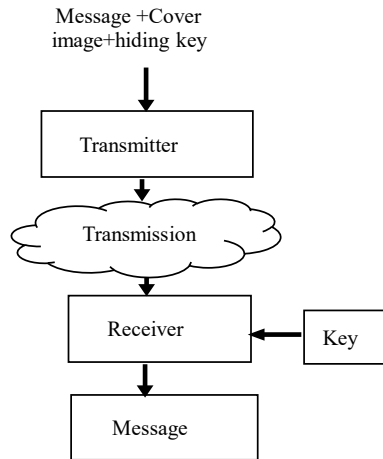


Fig. 1 Block diagram of a steganographic method

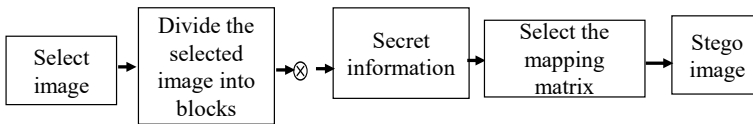


Fig. 2 Block diagram of the steganographic encoding

one induced a different method to advance the steganographic techniques available at the time of publishing the papers. The block diagram of steganographic encoding is shown in Fig. 2. We will try to study some spatial domain approaches which are widely used and some transform domain approaches (Table 1).

### 2.1 Least Significant Bit (LSB)-Based Methods

These are the basic algorithms that are developed at the beginning of the era of steganography. In the least significant bit-based algorithm, the message data that should be hidden is converted to binary format and is hidden in the least significant bits of RGB pixels from the cover image. The pixels in which we will hide data are selected based on the stego key. So the receiver can view the message only if he had the stego key in general. But there are so many brute force algorithms available today using which these LSB-based techniques can be cracked easily even without the use of a stego key.

In this paper, the author [1] introduced a technique of cropping and doing LSB-based steganography to improve the security features that we have as of now. In this approach, we will have a predefined set of coordinates in which cropping and storage

**Table 1** Summary of steganography proposed in the literature

Year	Authors	Method	Major Focus
2008	Abbas Cheddad (et al.)	Irreversible Fast Fourier Transform and advancing security hash algorithms	Pre-processing level encryption
2009	Abbas Cheddad (et al.)	Changing hue saturation values so that data can be hidden efficiently	Reducing unnecessary data
2010	Anjali A. Sejul (et al.)	Implementing steganography in spatial domain by the use of Discrete Wavelet Transform	Reducing the Peak signal-to-noise ratio is the main concern here
2013	Shabnam Samima (et al.)	Key secured mapping info algorithm	Increasing the encryption and security for available methods
2014	K. Sathish Shet (et al.)	Integer Wavelet Transform-based spatial domain method	Cost-effective and more secured
2014	Ratnakirti Roy (et al.)	Block entropy-based segmentation	Highly efficient embedding
2016	Khalid A. Al-Afandy (et al.)	Image cropping and LSB based	Increasing security
2018	Aditya Kumar Sahu. Gandharba Swain	Bit replacement algorithm	Improving PSNR and embedding capacity
2018	S. Saha (et al.)	Direction-based using dynamic weightage array	Reduced Quality distortion
2019	K. Sathish Shet (et al.)	FPGA usage for EMD-based algorithm	Increasing speed and variable image resolutions
2020	Mukhopadhyay, S (et al.)	Integer sequence named Catalan Transform (CT)	Improved the PSNR parameter

of main message bits are done using LSB techniques that we have as of now thereby the security of the hidden message is increased furthermore. But if we increase the cropping more to increase the security, then there is a disadvantage in the form of increased PSNR.

## 2.2 Least Significant Bit (LSB)-Based Methods

Pixel pair matching is another prominent steganographic method that utilizes a pixel pair as a coordinate of reference to search other coordinates within the neighbor set of  $\Phi$ . An adaptive approach for secured key-based steganography based on mapping information was proposed in [5]. In this method rather than sending embedded data directly, the mapping information is sent over the transmission medium. Some

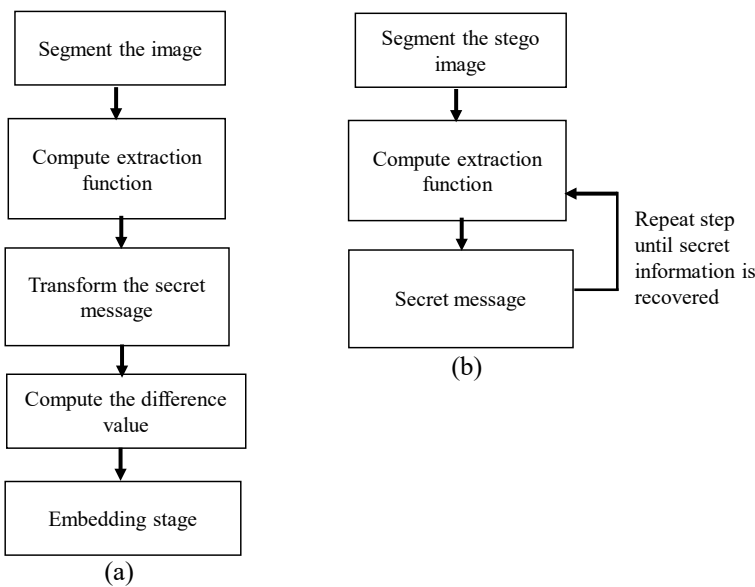
mapping is also embedded inside the stego image so the realization of actual information can be done through multiple security layers available.

It's a popular technique with high fidelity stego image outputs. During the embedding process, the message is transformed in  $2j + 1$  array notation by applying the algorithm where  $j$  is the number of pixels in the covering image. The EMD will use a specified kind of base for varying pixel intensity of the cover image so that more textured parts of the image can be embedded with more secret data. As a result, EMD has more visually satisfying outputs compared to other spatial domain techniques discussed earlier.

$$f_{emd}(a_1, a_2, \dots, a_n) = \left[ \sum_{b=1}^j a_b, b \right] \text{ mod } (2j + 1) \tag{1}$$

where  $a_1, a_2, \dots$  are the gray values of the cover image pixels, and  $j$  is the number of pixels chosen to embed the secret data.

The encoding as per the EMD algorithm is as follows in which the cover image will be divided into  $j$  pixel groups and the extraction function is calculated based on that  $j$  pixel groups. By transforming the message into  $2j + 1$  array notation, we will calculate the difference value. Based on this difference value, we will change the pixel values of the cover image. Then we will perform the exact algorithm in reverse fashion for the extraction purpose as per the author [4]. The block diagram of DWT encoding of image and decoding of the image is shown in Fig. 3.



**Fig. 3** Block diagram of DWT **a** Encoding of image **b** Decoding of image

### 3 Discussion

In this DWT algorithm, the cover image is split into RGB bit-planes and each bit-plane is transformed by DWT into four subbands. The data is embedded into all three bands except the lower frequency one which is more perceivable through the human eye. For the extraction process, the algorithm will perform the reverse, i.e., the flowchart of steganographic implementation using DWT is shown in Fig. 4. In this, we will compare the practical results that we obtained by the implementation of DWT-based steganography in Matlab and EMD algorithm-based steganography. First, we will discuss the DWT-based steganography that has been implemented practically and will try to analyze the results that are obtained.

In comparison to other image processing algorithms, the experimental results show that this algorithm maintains good image quality. The cover image, embedded images, and extracted images obtained in the practical operation are shown in Fig. 4.

The peak signal-to-noise ratio is crucial in determining how efficiently the algorithm is implemented. If the algorithm has a higher PSNR, we can say it's more efficient, i.e., with the increase in the value of PSNR the efficiency increases (Fig. 5).

The following are the steps to using DWT to implement steganographic

- Step 1: Segment the color stego image into R-plane, G-plane, and B-plane.
- Step 2: For R-plane, G-plane, and B-plane,  $N$ -level DWT is applied.
- Step 3: Used extraction algorithm.

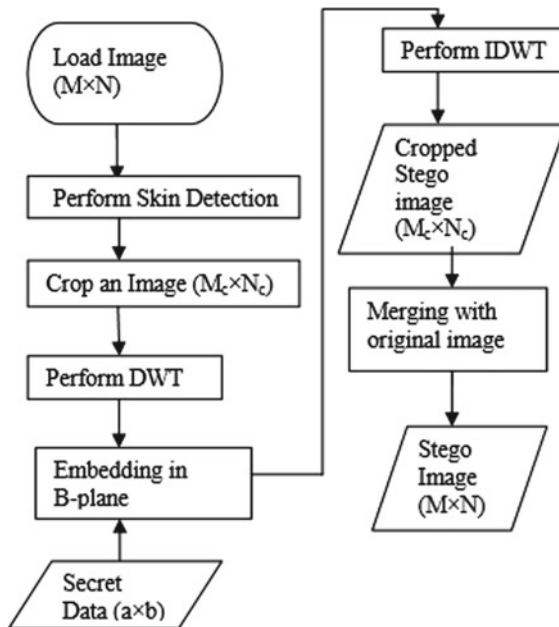


Fig. 4 Flowchart of steganographic implementation using DWT

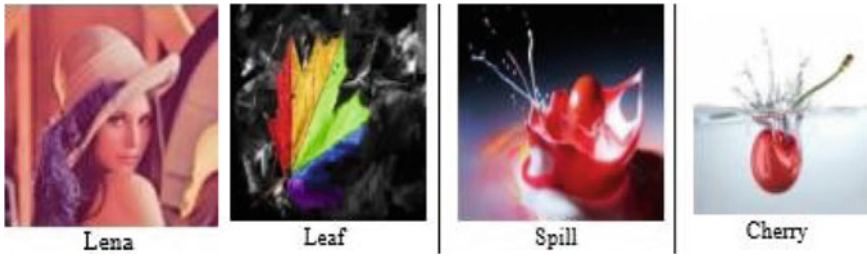


Fig. 5 Sample images

Table 2 Comparison of MSE and PSNR for the sample input image

	Means square value (MSE)	Peak signal-to-noise ratio (PSNR)	Security	FPGA implementation
Discrete wavelet transform	0.49	51.17	High	Difficult
Exploiting modification direction (EMD)	0.79	45.12	Low	Easy

Step 4: Extract the image into the secret image.

This method of analysis, namely the structural Similarity Index Matrix is good for the efficiency calculation than conventional PSNR and other evaluating parameters. The experimental results showed that the average SSIM value is 0.46. Steganographic weight is the factor that decides how many details should be embedded. It is the key factor to decide the embedding rate and SSIM values. The following table will give SSIM values for the given input images and stego images.

The experimental results in Matlab show that this algorithm has a highly appealing output, but the peak signal–noise ratio is nowhere nearer to the DWT algorithm. But what makes this algorithm unique is that its FPGA implementation is quite easy due to a step-by-step algorithm that it has and its less complex implementation (Table 2).

## 4 Conclusion

In this paper, we tried to study briefly all the available techniques for steganography in the spatial and transform domains. From the above experimental results, we can conclude that transform-based domain steganographic techniques are a bit dominating in the terms of PSNR values and security based on the key used for increasing their complexities. But EMD has its standing-out advantages when it comes to FPGA implementation as discussed in the paper. Transform domain approaches has been good since they deal with embedding in the frequency domain which provides a lot

of bit storing space. Our suggestion for getting better steganographic results is to go through all available basic techniques in the domain in which one wants to develop an algorithm and try to improve the hybrid mix of algorithms to yield better results in terms of PSNR, and security factors.

## References

1. Sathish Shet K, Aswath AR, Hanumantharaju MC, Gao X-Z (2019) Novel high-speed reconfigurable FPGA architectures for EMD-based image steganography. In: Proceedings of Springer Science+Business Media, LLC, part of Springer Nature
2. Ahirwar R, Choubey A (2011) A novel wavelet-based denoising method of SAR image using interscale dependency. In: International conference on computational intelligence and communication networks, pp 52–57. [https://doi.org/10.1109/\(2011\)CICN.2011.11](https://doi.org/10.1109/(2011)CICN.2011.11)
3. Ahuja B, Doriya R (2021) Visual chaos steganography with fractional transform. In: Reddy VS, Prasad VK, Wang J, Reddy KTV (eds) Soft computing and signal processing. advances in intelligent systems and computing, vol 1325. Springer, Singapore. [https://doi.org/10.1007/978-981-33-6912-2\\_27](https://doi.org/10.1007/978-981-33-6912-2_27)
4. Mukhopadhyay S, Hossain S, Ghosal SK et al (2021) Secured image steganography based on Catalan transform. *Multimed Tools Appl* 80:14495–14520. <https://doi.org/10.1007/s11042-020-10424-4>
5. Solak S (2020) High embedding capacity data hiding technique based on EMSD and LSB substitution algorithms. *IEEE Access* 8:166513–166524. <https://doi.org/10.1109/ACCESS.2020.3023197>
6. Baby D, Thomas J, Augustine G, George E, Rosia Michael N (2015) A Novel DWT based image securing method using steganography. *Procedia Comput Sci* 46:612–618. ISSN 1877–0509
7. Hussain M, Wahab AW, Idris YI, Ho AT, Jung KH (2018) Image steganography in spatial domain: a survey. *Signal Process: Image Commun* 65:46–66
8. Mohanty BK, Choubey A (2017) Efficient design for Radix-8 Booth multiplier and its application in lifting 2-D DWT. *Circuits Syst Signal Process* 36:1129–1149. <https://doi.org/10.1007/s00034-016-0349-9>
9. Konyar MZ, Akbulut O, Öztürk S (2020) Matrix encoding-based high-capacity and high-fidelity reversible data hiding in HEVC. *Signal Image Video Process* 14(5):897–905
10. Kim P-H, Yoon E-J, Ryu K-W, Jung K-H (2019) Data-hiding scheme using multidirectional pixel-value differencing on colour images. *Secur Commun Netw* 2019:1–11
11. Goel A, Bhujade R (2019) A functional review of image encryption techniques. *Int J Sci Technol Res* 8(9):1203–1205
12. Flesch BF, Tedeschi I, De Figueiredo RM, Prade LR, Da Silva MR (2020) A functional safety methodology based on IEC 61508 for critical reliability FPGA-based designs. *Int J Emerg Technol Adv Eng* 10(7):12–19

# Web Attack Detection Using Machine Learning



Ruturaj Malavade, Harshali Upadhye, Heena Jamadar, Deepak Kshirsagar, and Jagannath Aghav

**Abstract** Network security is critical in the new age of the Internet. Web attacks tend to target sites such as e-commerce sites, social media, and email websites. Even if web attacks are detected, some attacks may be able to get through, and hence we desire better performance on attack detection. This paper proposes a feature selection-based IDS to detect web attacks. Filter-based feature selection methods with top percent ranked feature selection strategies achieve relevant feature subsets. The proposed system achieves a higher accuracy of 97.9826% with 17 features, using Random Forest 10-part cross-validation. Finally, we compare our system against traditional systems.

**Keywords** Web attacks · Intrusion detection · Random forest · ReliefF

## 1 Introduction

Web security is an important aspect of our everyday lives. Web security is important to prevent hackers and cyber-thieves from being able to access sensitive and/or private information. A web attack can also cause a company or business to suffer substantial financial losses, due to theft of financial and/or corporate information. Without a good

---

R. Malavade (✉) · H. Upadhye · H. Jamadar · D. Kshirsagar · J. Aghav  
Department of Computer Engineering and Information Technology, College of Engineering Pune (COEP), Pune, Maharashtra, India  
e-mail: [malavaders17.comp@coep.ac.in](mailto:malavaders17.comp@coep.ac.in); [ruturaj.malavade@gmail.com](mailto:ruturaj.malavade@gmail.com)

H. Upadhye  
e-mail: [upadhyehm18.comp@coep.ac.in](mailto:upadhyehm18.comp@coep.ac.in); [hmupadhye66@gmail.com](mailto:hmupadhye66@gmail.com)

H. Jamadar  
e-mail: [jamadarhm17.comp@coep.ac.in](mailto:jamadarhm17.comp@coep.ac.in); [jamheena03101999@gmail.com](mailto:jamheena03101999@gmail.com)

D. Kshirsagar  
e-mail: [ddk.comp@coep.ac.in](mailto:ddk.comp@coep.ac.in); [kdeepak83@gmail.com](mailto:kdeepak83@gmail.com)

J. Aghav  
e-mail: [jva.comp@coep.ac.in](mailto:jva.comp@coep.ac.in)



web attack detection strategy, a business or website owner runs a risk of spreading malware and attacks on other businesses, websites, or even networks.

Web attacks are attacks on network security; cyber-criminals utilize weaknesses or vulnerabilities in a network to get access to confidential data. There are many types of web attacks that exist, such as malware, which is any kind of malicious software (such as viruses, spyware, or ransomware, fuzzers, in which an attacker tries to exploit flaws in system security by crashing it using large amounts of randomized data, analysis, an attack that penetrates web applications generally via email or web scripts, backdooring, which is bypassing of standard authentication methods, exploits, which take advantage of glitches or bugs in a system, Denial of Service (DoS), which prevents authorized requests [1] from accessing a device, worms, which is an attack that replicates itself over and over to spread to other devices, spoofing, in which an unknown source disguises itself as a trusted one, and so on.

Web attacks generally affect websites such as e-commerce websites and social media, networks such as a private networks of some corporate businesses, and other forms of IT infrastructures. In 2020, it was found that over 68% of business leaders think that cybersecurity risks are increasing. Security breaches have been found to increase by 11% since 2018, and by 67% since 2014. 34% of data breaches in the world have involved internal action, about 92% of all malware has been shared over email, and the average cost of a ransomware attack on a business has averaged over \$133,000. The banking industry was hit the hardest, and has incurred nearly \$18.3 million in cybercrime costs in 2018.

According to OWASP, the top web application attacks were caused due to exposure of private and sensitive data, remote code execution caused by insecure deserialization, bad enforcement of access control rules, external entities using XML, insufficient monitoring and logging, injection of malicious scripts into a valid website, usage of frameworks and libraries that are known to be risky, injection of untrusted data, failure of proper security implementations, and failure of proper implementation of authentication functions.

Studies in 2021 have shown that personal data was involved in over 58% of security breaches in 2020, 64% of Americans have never even checked if they were affected by a data breach, and 56% of Americans do not even know what to do if a data breach occurs. In 2020, the average ransomware payment rose by 33% from 2019. Phishing attacks have been found to account for more than 80% of reported security incidents. From the data, it is clear that web attacks can wreak havoc on the world if not detected early and taken care of. Hence, it is essential to detect web attacks as soon as possible, and prevent them from happening in the future.

A popular method of web attack detection is by using an Intrusion Detection System (IDS). An IDS [2] monitors a network or system for any malicious activity and raises alerts to detect any such activity. An IDS works by looking for signatures of known attacks, or activity that is different from normal activity, and sends alerts to a system administrator if it detects any such kind of abnormal activity.

IDSs are classified into 2 main categories, based on where detection takes place: host-based IDSs and network-based IDSs. A network-based IDS monitors traffic on all devices on a network, and detects malicious traffic, if any. It does so by scanning

network packets at the host level, and logging any suspicious packets that might be present, whereas a host-based IDS runs on individual hosts in the network, and analyzes traffic and logs malicious behavior on the system on which it is installed. It uses a database of important objects of the system such as environment variables that it should monitor, and for each object, it creates a checksum that is stored in a database for comparison later.

IDSs can also be classified into 2 categories based on the detection method they use: signature-based IDSs, and anomaly-based IDSs. A signature-based IDS detects attacks by looking for specific patterns such as byte sequences in network traffic or known malicious instruction sequences in malware. It relies on a list of known indicators that indicate a system is compromised, such as malicious network behavior, email subject lines, malicious domain names, and so on, whereas an anomaly-based IDS monitors system activity and tries to classify it as normal or anomalous. It does so by creating a model of normal (valid or trustworthy) activity using machine learning (ML), and then compares new behavior in a system against this model.

A signature-based IDS suffers from a major drawback, namely that it is difficult for it to detect new attacks, as no prior information is available on said new attacks as compared to attacks that may have already had information on them. Since an anomaly-based IDS uses ML to create a model to compare against, and these models can be trained better, an anomaly-based IDS does not suffer from this issue.

A signature-based IDS also requires reprogramming for every new pattern that is to be detected in a system. However, it has a low false positive rate. While an anomaly-based IDS can suffer from false positives more, it does not need to be reprogrammed, because the ML model can learn the pattern of the new attack occurring. Hence, anomaly-based IDSs are used more compared to signature-based IDSs nowadays.

ML plays a major role in intrusion detection in modern times. This is because, with recent developments, the amount of data that is handled becomes too large for classic methods to be usable. Further, classic methods are not able to actively learn data patterns and easily find web attacks as compared to ML-based intrusion detection methods, which are much more easily able to detect web attacks. ML finds use in intrusion detection as part of anomaly-based IDS methods, which use ML to build a model to compare potentially malicious behavior against.

This paper proposes a filter-based feature selection method to detect web attacks. Feature reduction is performed because all features are not essential; some features are noisy, and keeping these causes the system to degrade, giving us less accuracy and higher build times. Feature selection and removal of noisy features improve the system's accuracy and give us shorter build times as well.

The contributions of this work are as follows:

1. It proposes a technique of filter-based feature selection method by applying top percentage selection to form subsets, which we use for web attack detection.
2. It tests on the UNSW-NB15 dataset with reduced features and compares our results with traditional systems.

Section 2 describes the Literature Review of the paper. Section 3 describes the system we have proposed. Section 4 describes the implementation of the system and

the results obtained from the analysis. Section 5 describes the conclusion of the paper and the future scope of our research.

## 2 Literature Review

In the paper [3], an intelligent system is proposed which first performs feature ranking based on Information Gain (IG) and Correlation (COR). Then, feature reduction is performed by combining ranks obtained from both IG and COR, using a novel approach to identify features to be removed. The model is then trained and tested on the KDD99 dataset and it gave outstanding results.

In the paper [4], an ensemble-based multi-filter feature selection method is proposed that takes four filter methods, namely IG, Gain Ratio (GR), ReliefF, and Chi-Squared (CS), and puts together their analyzed outputs to achieve optimal selection. This method is then evaluated using the NSL-KDD dataset, and the evaluation produces a higher attack classification accuracy and attack detection rate in comparison with other classification methods.

The paper [5] proposes a network intrusion detection system using Extreme Gradient Boosting (XGBoost) on the UNSW-NB15 dataset. For analysis purposes, a subset of 23 features was taken, and the Random Forest classifier was applied to this subset. This gave an accuracy of 97.9060%.

The paper [6] provides an overview of the KDD99 and UNSW-NB15 datasets' significant features for network IDSs. For analysis, it uses an ARM (Association Rule Mining) algorithm to generate the strongest features from both datasets. Analysis shows that the KDD99 dataset gives more accuracy as compared to the UNSW-NB15 one.

The paper [7] presented an attack procedure and its impact for each attack, execution of attacks wherever appropriate, as well as countermeasures that can be taken to prevent said attacks. To cope with threats like these, a Web developer must know about the risks in the system, as well as the impact they can have. Furthermore, the existing Web standards should be examined even more for potentially undiscovered vulnerabilities, so as to develop more advanced countermeasures against them.

The paper [8] has proposed an integrated classification-based IDS and evaluates its performance on an existing dataset and a newly generated one. This paper evaluates the proposed model's performance on the UNSW-NB15 dataset, which covers more recent attacks (such as Denial-of-Service, Probe, Fuzzers, Generic, Exploits, and so on) compared to another dataset, the KDD99 dataset. It is observed that compared to older models that utilize Decision Trees, this model gives better values in terms of certain evaluation metrics. Further, this paper also generates a real-time dataset at the NIT Patna CSE lab (RTNITP18), and it acts as the testing dataset to evaluate the performance of the proposed model, which gives an accuracy of 83.8% on this dataset.

In the paper [9], a novel Hierarchical Intrusion Detection System (IDS) is proposed, combining three different classifiers: JRip algorithm, REP Tree, and Forest

PA, on decision tree and rules-based concepts. The proposed model uses three classifiers, where the first two methods take the dataset and classify its features as being either Normal or Attack, while the third method uses this classification along with the original dataset to classify even further.

The paper [10] explores the XGBoost algorithm's application for feature selection in conjunction with multiple ML techniques, including Decision Trees, Support Vector Machines, k-Nearest Neighbors, Logistic Regression, and Artificial Neural Networks, in order to implement accurate IDSs. This paper uses multiclass and binary classification techniques on the UNSW-NB15 dataset. Further, applying an XGBoost-based feature selection algorithm to the dataset gave us 19 optimal features.

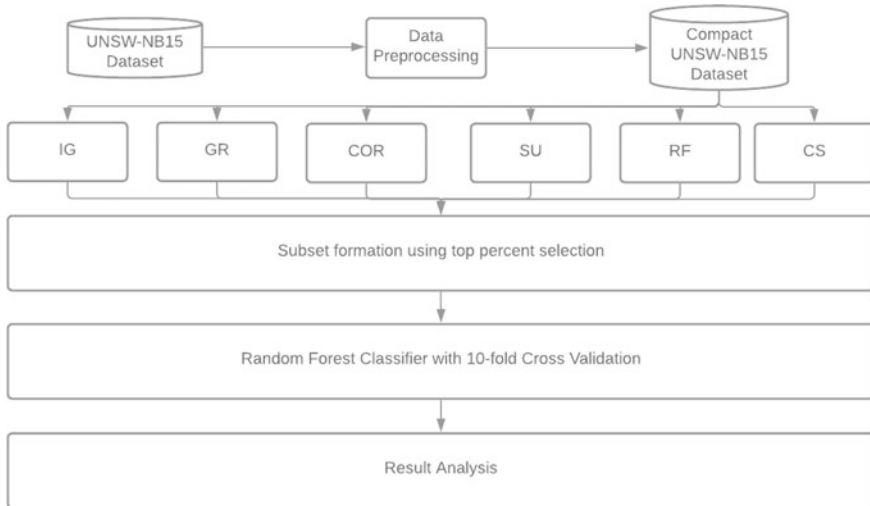
The paper [11] presents an intrusion detection approach that uses a genetic algorithm. This paper uses a Support Vector Machine along with genetic algorithms, and introduced feature selection and attack detection methods that improve the performance of IDSs. The feature selection method that uses the new fitness function measures TPR and FPR (true positive rate and false positive rate, respectively), and selects features of the training data better than traditional methods. The experimentation and result analysis by using the UNSW-NB15 and KDD99 datasets are measured using metrics like classification accuracy, Receiver Operating Characteristic (ROC) curve, false positive rate, and true positive rate.

The paper [12] aims to use ML techniques to identify rare cyber-attacks using the UNSW-NB15 dataset. It uses a hybrid feature selection method to identify the most important features for each type of attack in the dataset, of which there are nine: Exploits, Denial-Of-Service (DoS), Worms, Reconnaissance, Fuzzers, Generic, Analysis, Shellcode, and Backdoor. Overall, the most important features found were as follows: `ct_srv_src` (feature number 41), which occurred in 5 of the 9 attack types, `service` (feature number 14), which was found in 7 of the attack types (all except Exploit and Denial-Of-Service), `sttl` and `dttl` (features 10 and 11, respectively), which were found in 6 of the attack types, and `dur` (feature number 2), which was found to be in 8 of the attack types, all except worms. They used the Naïve Bayes classifier, and using the feature subsets found in the hybrid feature selection method, they were able to achieve a lower false acceptance rate as well as a higher classification rate compared to other methods.

### 3 Proposed System

As shown in Fig. 1, the proposed web attack detection consists of data preprocessing, feature selection, and a tree-based classifier called the Random Forest classifier.

The original dataset contains noisy data; hence, we apply to preprocess [13] to the dataset. We use a LabelEncoder to encode the alphanumeric entries in the dataset into purely numeric entries in preprocessing. The LabelEncoder works by assigning a unique positive number value to each unique alphanumeric entry in the dataset, and then replacing the corresponding alphanumeric entry with the corresponding numeric entry. Data preprocessing gives us consistent data.



**Fig. 1** Proposed IDS using feature selection methods

The feature reduction method uses this consistent data to find relevant reduced features. It uses various filter-based feature selection algorithms, namely Information Gain (IG), Gain Ratio (GR), Correlation (COR), Symmetrical Uncertainty (SU), ReliefF, and Chi-Squared (CS). We apply top percent selection on these algorithms by selecting the top 10, 20%, and up to 90%; this gives us subsets. The subsets generated are passed on to the Random Forest classifier; we choose the feature subset that improves performance.

The system uses Random Forest with 10-fold cross-validation, and performance is measured using parameters like Accuracy, Incorrectly classified instances, and Time is taken to build the model.

## 4 System Implementation and Result Analysis

The proposed system is implemented using Python 3 and Weka 3.8.4. A Python script was written and used for data preprocessing, while the Weka tool is used for feature selection techniques and classifiers. The system is implemented and tested on a machine configured with an Intel(R) Core(TM) i7-7700HQ CPU @ 2.81 GHz and an Nvidia GeForce GTX 1050 Ti GPU with 8 GB of RAM.

The system is tested on the UNSW-NB15 dataset [1]. This dataset consists of 45 attributes, including labels. We initially apply to preprocess using Python to the training and testing datasets to convert the alphanumeric entries into purely numeric entries, or in other words, label encoding the dataset, giving us a compact dataset. Out of these 45 features, we remove “attack\_cat” because we perform binary

**Table 1** Number of records per attack type in training dataset

Attack type	Number of records per attack type
Analysis	2000
Backdoor	1746
DoS	12,264
Exploits	33,393
Fuzzers	18,184
Generic	40,000
Reconnaissance	10,491
Shellcode	1133
Worms	130

classification and “id” since this causes system bias that gives us a dataset with 42 features, excluding labels.

The training dataset contains 175,341 records, out of which 56,000 are Normal records, and 119,341 are Attack records, while the testing dataset contains 82,332 records, out of which 37,000 are Normal and 45,332 are attack records. The attack records consist of 9 different types of attacks, namely Exploits, Denial-Of-Service (DoS), Worms, Reconnaissance, Fuzzers, Generic, Analysis, Shellcode, and Backdoor. The number of records of each attack type for the training and testing dataset are described in Tables 1 and 2, respectively.

We apply various classifiers such as IG, GR, COR, SU, ReliefF, and CS [14] to the training dataset to calculate each feature’s score. IG gives us scores ranging from 0 to 0.67048, GR from 0 to 0.40367, COR from 0 to 0.692741, SU from 0 to 0.481425, ReliefF from 0 to 0.15148558, and CS from 0 to 134,615.2876.

The Top percent feature selection strategy is applied to all of these feature selection techniques. We have formed ten subsets for each feature selection technique by selecting 10–90% of all feature indifference of 10%. The subsets obtained from this are applied to the Random Forest tree-based classifier. The tree-based classifier consists of various classifiers such as Random Forest, Random Tree, LMT, J48,

**Table 2** Number of records per attack type in testing dataset

Attack type	Number of records per attack type
Analysis	677
Backdoor	583
DoS	4089
Exploits	11,132
Fuzzers	6062
Generic	18,871
Reconnaissance	3496
Shellcode	378
Worms	44

**Table 3** Performance analysis of our proposed system on UNSW-NB15 dataset using random forest

Method	Selected features (%)	Number of features	Accuracy (%)	Incorrectly classified instances (%)	Time taken to build model (s)
ALL	100	42	97.9255	2.0745	59.06
IG	70	29	97.4481	2.5519	42.59
GR	90	38	97.9182	2.0818	50.77
COR	50	21	97.1044	2.8956	39.5
SU	80	34	97.5064	2.4936	52.2
ReliefF	40	17	97.9826	2.0174	25.27
CS	70	29	97.4481	2.5519	34.78

Hoeffding Tree, REP Tree, and Decision Stump; all of these classifiers are applied datasets. However, out of all of these, Random Forest gave us a higher accuracy than other tree-based classifiers on the dataset. Therefore, our system selects Random Forest for experimental purposes. Therefore, further experimentation is performed using the Random Forest classifier, and performance is measured. The best feature subset obtained for each feature selection method using a top percent selection strategy that provided higher performance when applied to the testing dataset is described in Table 3.

Table 3 shows the experimentation on the UNSW-NB15 dataset using the top percent feature selection. IG takes 70% of all attributes and gives us less build time than all attributes, but also gives us less accuracy, as does GR, which takes 90%, COR, which takes 50%, SU, which takes 80%, and CS, which takes 70%. However, ReliefF takes 40% of all attributes and gives us a better accuracy of 97.9826%, with build time nearly half of the base case. Therefore, our system selects the feature subset obtained using ReliefF that gives higher accuracy than other feature selection methods.

Further, the proposed system is compared with traditional systems. The analysis is performed using various traditional system approaches on the testing dataset. Table 4 shows the comparison of the proposed system with traditional feature selection-based IDSs.

The approaches [3–6] give feature selection methods for feature selection-based IDSs, and we applied the presented feature selection methods and obtained accuracy, as shown in Table 4.

The approach [3] gives an accuracy of 97.5064% by taking 28 features, the approach [4] gives an accuracy of 93.7461% by taking ten features, the approach [5] gives an accuracy of 97.9060% by taking 23 features, and the approach [6] gives an accuracy of 97.9461% by taking 29 features.

However, our approach gives a higher accuracy of 97.9826% by taking 17 features. Therefore, we select the ReliefF approach with the Random Forest classifier.

**Table 4** Comparison of performance analysis on UNSW-NB15 dataset using random forest with other papers

Work	Number of features	Accuracy (%)	Incorrectly classified instances (%)	Time taken to build model (s)
All F	42	97.9255	2.0745	59.06
Akashdeep et al. [3]	28	97.5064	2.4936	32.88
Osanaiye et al. [4]	10	93.7461	6.2539	24.49
Husain et al. [5]	23	97.9060	2.0940	34.5
Moustafa and Slay [6]	29	97.9461	2.0539	37.47
Proposed	17	97.9826	2.0174	25.27

## 5 Conclusion

This paper proposes a feature selection-based IDS to detect web attacks. Initially, a filter-based feature selection method with the top percent feature selection is applied for feature selection. Finally, we have compared our system against traditional systems. We have obtained a higher accuracy of 97.9826% by taking 17 features compared to traditional systems.

In the future, we will use other kinds of algorithms such as genetic algorithms for feature selection in web attack detection.

## References

1. Moustafa N, Turnbull B, Raymond CK (2019) An ensemble intrusion detection technique based on proposed statistical flow features for protecting network traffic of internet of things. *IEEE Internet Things J* 6(3):4815–4830
2. Kshirsagar D, Kumar S (2020) Identifying reduced features based on IG-threshold for DoS attack detection using PART. In: *Lecture notes in computer science*, vol 11969. Springer, Cham. [https://doi.org/10.1007/978-3-030-36987-3\\_2](https://doi.org/10.1007/978-3-030-36987-3_2)
3. Akashdeep, Manzoor I, Kumar N (2017) A feature reduced intrusion detection system using ANN classifier. *Expert Syst Appl* 88:249–257
4. Osanaiye O, Cai H, Choo KK, Dehghantanha A, Xu Z, Dlodlo M (2016) Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing
5. Husain A, Salem A, Jim C, Dimitoglou G (2019) Development of an efficient network intrusion detection model using extreme gradient boosting (XGBoost) on the UNSW-NB15 dataset. In: *IEEE international symposium on signal processing and information technology (ISSPIT)*, pp 1–7
6. Moustafa N, Slay J (2015) The significant features of the UNSW-NB15 and the KDD99 data sets for network intrusion detection systems. In: *4th international workshop on building analysis datasets and gathering experience returns for security (BADGERS)*, pp 25–31
7. Jensen M, Gruschka N, Herkenhöner R (2009) A survey of attacks on web services. *Comput Sci Res Dev* 24(4):185–197. <https://doi.org/10.1007/s00450-009-0092-6>



8. Kumar V, Sinha D, Das AK, Pandey SC, Goswami RT (2020) An integrated rule based intrusion detection system: analysis on UNSW-NB15 data set and the real time online dataset. *Cluster Comput* 23:1397–1418
9. Ahmim A, Maglaras L, Ferrag MA, Derdour M, Janicke H (2019) A novel hierarchical intrusion detection system based on decision tree and rules based models. In: 15th international conference on distributed computing in sensor systems (DCOSS), pp 228–233
10. Kasongo SM, Sun Y (2020) Performance analysis of intrusion detection systems using a feature selection method on the UNSW-NB15 dataset. *J Big Data* 7:105
11. Gharaee H, Hosseinvand H (2016) A new feature selection IDS based on genetic algorithm and SVM. In: 8th international symposium on telecommunications (IST), pp 139–144
12. Bagui S, Kalaimannan E, Bagui S, Nandi D, Pinto A (2019) Using machine learning techniques to identify rare cyber-attacks on the UNSW-NB15 dataset. *Secur and Priv* 2:e91
13. Kshirsagar D, Kumar S (2021) A feature reduction based reflected and exploited DDoS attacks detection system. *J Ambient Intell Hum Comput* 1–13
14. Kshirsagar D, Kumar S (2020) An ensemble feature reduction method for web-attack detection. *J Discret Math Sci Cryptogr* 23(1):283–291

# Pragmatic Analysis of Web Service Discovery Models and Architectures from a Qualitative Perspective



Moumita Majumder Sarkar and Manish Dhananjay Sawale

**Abstract** Web services and application programmer interfaces (APIs) have become an integral part of modern-day software development. In today's connected world, these services and APIs must possess characteristics like the ease of availability, low delay, high security, and minimum memory utilization. Moreover, these models must be easily discoverable, and highly usable from a development standpoint. In order to improve the discoverability of these models, various approaches like domain ontology approach, public ontology approach, syntax and semantic approach, context-aware approach, etc. have different performance in terms of metrics like the delay of discovery, computational complexity, accuracy of discovery, etc. Due to such a wide variation in the performance, it becomes difficult for web administrators to select the best possible combination of service discovery mechanisms suited for their application. Web administrators are therefore required to either study the internal mechanisms of each discovery model, or deploy the recommended models individually and evaluate their performance. Both approaches are time-consuming and increase the overall cost of system deployment. In order to reduce the probability of this drawback, the underlying text evaluates the performance of some of the recent web service discovery models; and quantifies them in terms of performance parameters like the delay of discovery, accuracy of discovery, and fuzzy computational complexity. Using this review, both researchers and website administrators will be able to evaluate the best possible solution for their specific application, thereby reducing both time and cost needed for web deployment. Moreover, this text also recommends various methods to improve the performance of these models, thereby assisting in further cost and deployment time optimization.

**Keywords** Web services · Discovery · Machine learning · Cost · Deployment time · Accuracy

---

M. M. Sarkar (✉) · M. D. Sawale

Department of Electronics and Communication Engineering, Oriental University, Indore, India  
e-mail: [mukunsarkar1985@gmail.com](mailto:mukunsarkar1985@gmail.com)

## 1 Introduction

The application of web services discovery belongs to the field of web crawling wherein a multitude of cloud-enabled services are used. A combination of these services in tandem assists in the development of a highly efficient service discovery platform that possesses low delay of discovery, high throughput, high accuracy of discovery, and low computational complexity. In order to design such a system, the following sub-system components must be modelled and optimized:

- Web Services Search Engine.
- Service clustering model.
- Web Service Description Language (WSDL) parser.
- Feature extraction and evaluation model.
- Web Crawler model.

Modelling of each of these components requires efficient design practices, and optimized architectures. For instance, to model an efficient web crawler; it is required that a database of the most relevant web engines must be prepared, and this database must have linkages with context-sensitive web documents. Similarly, the designing of web services search engine must be done in such a manner that it can efficiently use crawler results for similarity matching; and then use these similarity results for efficient search result reranking [1]. A generalized architecture for this model can be observed in Fig. 1, wherein service discovery is divided into pre-process and discovery phases.

Based on this architecture, it can be observed that the majority of computational work is performed by the pre-processing block. Here, operations like content extraction, type identification, message passing evaluation, tag identification, etc. are performed. Results of these operations are given to a clustering mechanism, wherein web data is categorized into multiple small fragments, and each of these fragments is used by web services search engine for efficient retrieval.

A survey of different available implementations for these micro-components can be observed in the next section. This is followed by the performance evaluation of these models, and their comparison in terms of discovery delay, accuracy of discovery, and fuzzy computational complexity. The comparison will assist readers of this text to select the best possible combinations of these models for their given application, which will eventually assist them in increasing system deployment speed, and reducing overall deployment costs. Finally, this text concludes with some interesting observations about the proposed model, and recommends methods to improve its performance.

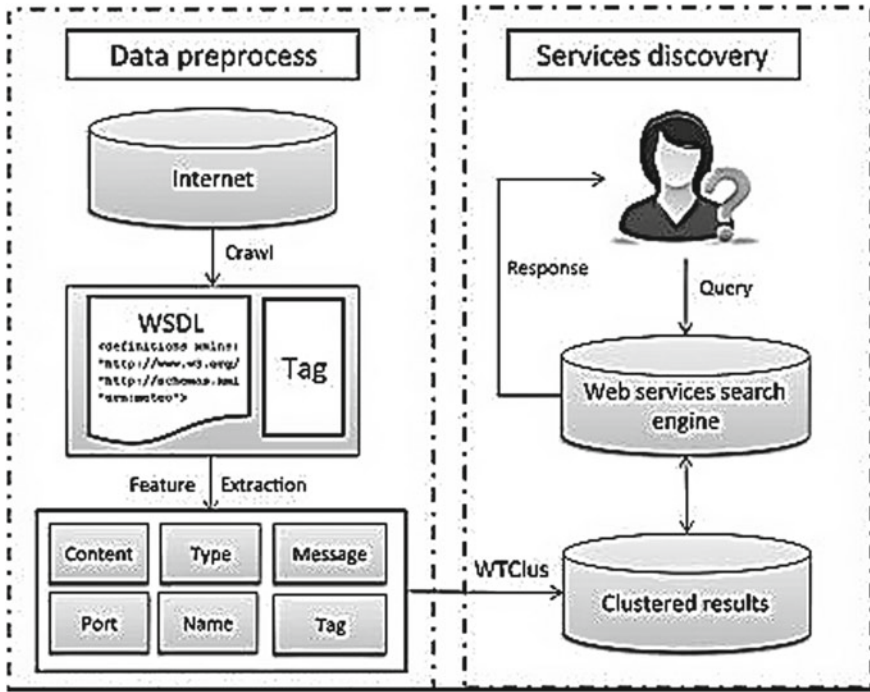


Fig. 1 Web services discovery architecture

## 2 Literature Review

Web service discovery (WSD) is modelled via a combination of intelligent crawling, clustering, and semantic-based matching algorithms. The crawling mechanism is responsible for the extraction of query-sensitive document extraction, while clustering assists in categorizing this data into groups of similar content. Finally, the semantic-based matching model compares the input query with the clustered documents in order to extract the service discovery results. A model that uses topic modelling using latent Dirichlet allocation (LDA) for extraction of query-dependent topics using web crawlers can be observed in [2]. In this model, k-Means is combined with agglomerative clustering in order to decide the optimum cluster size; and group the data into 1 of 'N' different buckets. Each of these buckets is then compared with the input query, and relevant web service results are evaluated. This model replaces word-based vectors with topic-based vectors in order to improve result relevancy and reduce delay needed for clustering and feature extraction. Due to the use of k-Means clustering with topic modelling, an accuracy of 59% is achieved which is higher than simple agglomerative clustering that provides an accuracy of 39%, and word-vector-based k-Means clustering that provides an accuracy of 47% on the same

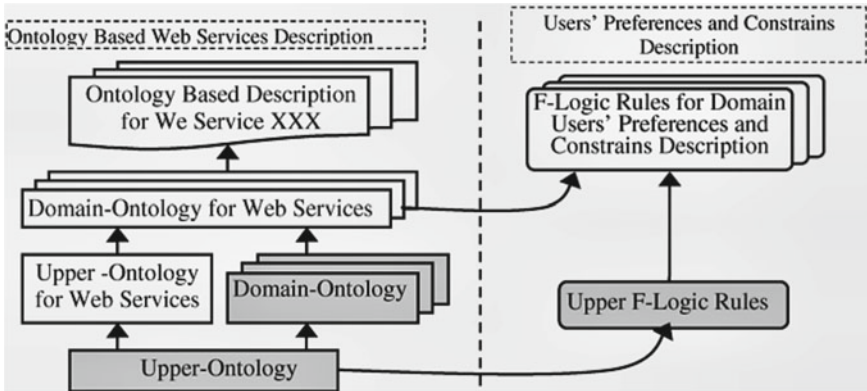


Fig. 2 Semantic-based web services discovery with fuzzy rules [3]

set of query inputs. This accuracy can be improved via the addition of semantics-aware web service discovery models like the one described in [3]. In this model, a combination of semantic data with user preferences and constraints is defined in order to improve the relevance of obtained results. The architecture for this model can be observed in Fig. 2, wherein fuzzy rules (F-rules) are used in order to combine the domain ontology results with user preferences and application-level constraints. These models are combined in order to obtain better relevancy of service discovery results, thereby resulting in accurate service discovery.

This model is able to obtain data from multiple data sources that include, but are not limited to, upper ontology, upper ontology for web services, domain ontology, domain ontology for web services, and service-based ontology description. All these data sources are combined using fuzzy-logic rules, and the final web service results are obtained. The efficiency of this model can be improved using shared service discovery, wherein data retrieved using one WSD model is used by other models using query-to-query matching via collaborative filtering as proposed in [4]. In this architecture, the Latent Factor Model (LFM) is used as observed from Eq. 1, wherein the service consumer feature matrix ( $k$ ) and service provider feature matrix ( $x$ ) are used.

$$\text{LFM} = k_{\text{consumer}} * x_{\text{provider}} \quad (1)$$

Based on this metric, an evaluation of the retrieved result is done, and its relevancy is decided. A result with a higher value of LFM indicates that it is relevant to both consumer and provider, a moderate value indicates that it might not be relevant to one of them, while a low value indicates that both the parties mildly agree on the discovered result. The model is able to achieve an accuracy from 46 to 55% depending upon the selection of regularization coefficients and learning rate values. This accuracy can be improved via the use of deep learning models that provide

better insights into user’s behaviour by long-term analysis of data patterns available at the input. Such a deep learning model can be observed from [5], wherein ServeNet or Service description-based neural network is defined. The model uses context-aware embedding using Bidirectional Encoder Representations from Transformers (BERT) architecture. It takes service name and service description as input parameters, and provides relevant discovered services at the output. An accuracy of 91.13% is achieved using this model, which is improved when compared to a similar model that uses only service description as input and uses Global Vectors for word representation (GloVe) for feature extraction that provides an accuracy of 88.4% on the same dataset. Architecture of this model can be observed in Fig. 3, wherein multiple convolutional neural networks (CNNs) are combined with long short-term memory (LSTM) models and fully connected (FC) layers in order to obtain the final service category. The model is evaluated in terms of accuracy of discovery and accuracy variance across multiple categories that include Advertising, Tools, Financial, Messaging, eCommerce, Payments, Social, Enterprise, Application Development, Analytics, etc.

It is observed that the proposed model outperforms other models in terms of accuracy and accuracy variance, wherein the proposed model has an accuracy of 91.58%. This accuracy is higher than CNN (58.46%), AdaBoost (64.92%), LDA with linear support vector machine (LSVM) (71.91%), LDA with radial basis function (RBF) SVM (73.84%), Naïve Bayes (NB) (78.94%), Random Forest (RF) (80.24%), LSTM (80.1%), recurrent CNN (84.29%), convolutional LSTM (84.32%), bi-directional LSTM (86.7%), and ServeNet with GloVe (88.4%). Categories like eCommerce, Tools, Financial, Music, Advertising, Application development, etc. are observed to have over 95% accuracy; while fields like entertainment, video matching, transportation, etc. have a lot of improvement scope which can be worked upon via application-specific models or by feedback-based optimization. Such a model can be observed from [6], wherein incremental correction of results is performed using

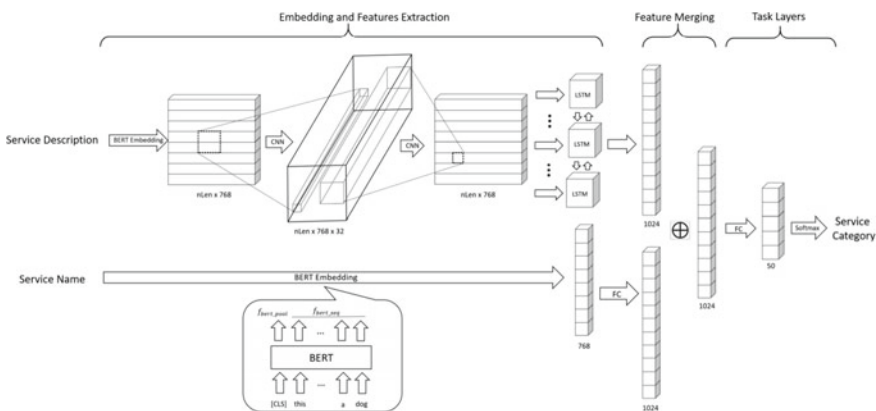
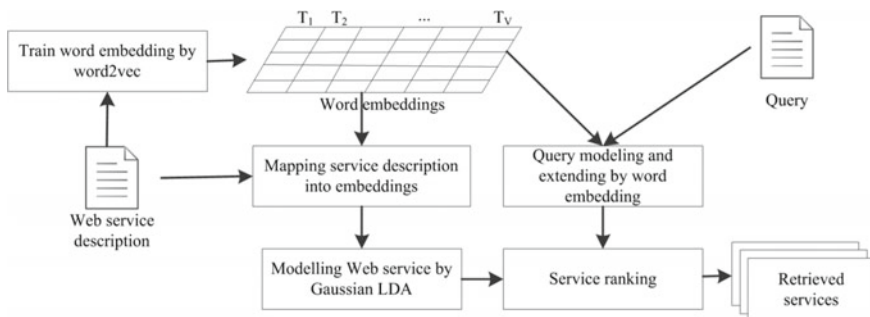


Fig. 3 BERT with CNN and LSTM for improved accuracy of web service discovery

local and global feedbacks along with global propagations. The model is able to achieve an accuracy of 68.73%, which is higher than Annotation with Optimized Concept (AOC) that has an accuracy of 62.66% on the same dataset.

A service composition framework for efficient web service discovery can be observed from [7], wherein a consistent framework is designed for the global discovery of service composition. This model is able to improve trust levels during service discovery, and thereby must be used as a model for highly secure service discovery applications. The model can also be used with high-quality of service (QoS) discovery applications like the one mentioned in [8], wherein web services discovery Harvesting as-a-Service (HaaS) is defined. The HaaS model is able to Generate different varieties of crawled data by adapting to the underlying application, thereby improving result usability. The model is able to achieve an accuracy of 95% for the discovery of relevant web services, with precision and recall values in the range of 99–99.5% depending upon the given dataset. A similar high accuracy model that uses Gaussian latent Dirichlet allocation (GLDA) with word embedding can be observed in [9]. This model initiates by examining query context, and based on this context GLDA is initialized for the acquisition of service descriptors. These service descriptors are ranked according to their similarities in order to obtain the final search results. The framework for this model can be observed in Fig. 4, wherein Word2vec is used as a feature vector for extraction of word embeddings and using them for final comparison.

The GLDA-QE model is able to achieve an accuracy of 81% which is higher than GLDA (76%), LDA with word embedding (LDA-WE) (70%), Doc2vec (68%), LDA (62%), and Probabilistic Latent Semantic Analysis (PLSA) (67%), thereby making it useful for real-time deployments. This accuracy can be improved using the BiLSTM model as described in [10], wherein features like N-grams, word order, context information, and linear regression are combined in order to obtain a highly accurate web service discovery model. The wide model is also used here, which allows for fully connected layer deployment, and linear regression-based training. This combination of Wide and BiLSTM (WBILSTM) produces an accuracy of 92.5%, which is higher



**Fig. 4** GLDA and query embedding (QE) with context-aware service ranking and discovery [9]

than BiLSTM (91%), Wide and Deep model (77%), LSTM (71%), and word embedding LDA (WE-LDA) (75%) when compared on the same dataset. Thus, it can be observed that WBiLSTM outperforms other models and must be used for real-time web service discovery. This performance can be improved via the use of a knowledge graph as suggested in [11], wherein matching templates are used for the conversion of discovery results into knowledge graphs that can be queried using knowledge query languages. An example knowledge graph can be observed in Fig. 5, wherein operations like corpus processing, user intent classification, service input identification, service knowledge graph generation, and similarity calculations can be seen. It uses Word2vec for feature extraction, and LSTM for classification of the crawled results. The proposed Word2vec graph LSTM (WGLSTM) model is able to achieve an accuracy of 95.5% on different datasets, which is higher than Word2vec graph SVM (WGSVM) model that possesses an accuracy of 94.4% on similar datasets.

Event-driven models for web service discovery can be used in order to find the best services in case of event triggers. Such a model for Internet of Things (IoT)

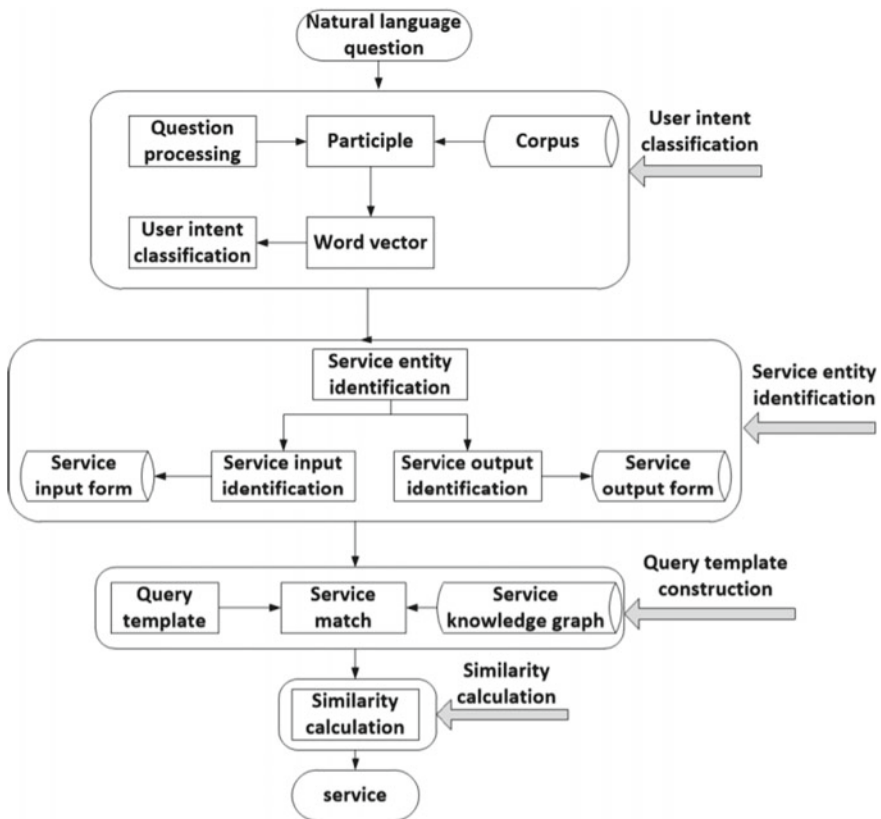


Fig. 5 Graph based model for service discovery [11]



can be observed from [12], wherein different models for event recognition and event handling are used. These models are able to evaluate semantic similarity via word frequency evaluation, which is calculated using word-level contexts and semantic relationships. This evaluation can be observed from Eq. 2, wherein target word ( $w$ ) and its context ( $c$ ) are used for the evaluation of continuous bag of words (CBOW) similarity in a set of ' $T$ ' words.

$$\text{CBOW} = \frac{1}{T} * \sum_{i=1}^T \log_p(w_i | w_{i-c}^{i+c}) \quad (2)$$

Depending upon similar metrics, a hybrid learning algorithm is devised that uses word embeddings to evaluate learned word embedding results.

These results are then used by classification models for efficient web service discovery. As a result of this, an accuracy of 83.1% is achieved using a hybrid model that uses cosine-retrofitting (Hybrid CSR), which is higher than hybrid-retrofitting (Hybrid R) (81.3%), multiple semantic fusion (MSF) (81.8%), and only CBOW (79.3%), thereby making Hybrid-CSR an ideal candidate for real-time deployments. This efficiency can be improved by a combination of these models with Embeddings from Language Models (ELMo) as suggested in [13]. The ELMo process combines language model pre-training with representation generation in order to learn from word embeddings. Due to a combination of ELMo with CNN, an accuracy of 96% is achieved, which is higher than Word2vec-based CNN (67.5%), Doc2vec (70%), and LDA (91%) across multiple datasets. Thereby, it is recommended that this model must be used for any kind of real-time web service discovery system. A similar model is proposed in [14, 15] wherein deep learning is combined with semantic analysis to improve the accuracy of service discovery as observed in Fig. 6, wherein probabilistic maximization is used for real-time semantic feature classification.

The proposed Deep word semantic CNN with LDA and heuristics (DWSCLH) model is able to achieve an accuracy of 63.79%, while a similar model without heuristics (DWSCL) is able to achieve an accuracy of 57.08%, due to this it is recommended that to include context-sensitive heuristics while designing such models. Other models like DCL, LDA, term frequency-inverse document frequency (TF-IDF), and LDA with k-Means (LDA-K) are able to achieve an accuracy of 54.92%, 52.85%, 46.73%, and 52%, respectively. This accuracy can be further improved via the use of semantic mining as proposed in [16], wherein along with semantic mining, result indexing is done in order to improve overall system speed. The model utilizes start operation sets (SOps) and end operation sets (EOps) in order to evaluate the initial and final indexing values as observed in Fig. 7, wherein web service language (WSL) is used for data query.

Due to the use of multiple semantic mining and data indexing (MSMDI), the model is able to achieve an accuracy of 91% for different types of query sets, which is higher than single semantic mining and data indexing (SSMDI) (90%), service pooling (80%), Woogle (79%), Schema matching (61%), and keyword-based matching (54%). The model is evaluated on general-purpose web datasets, and can

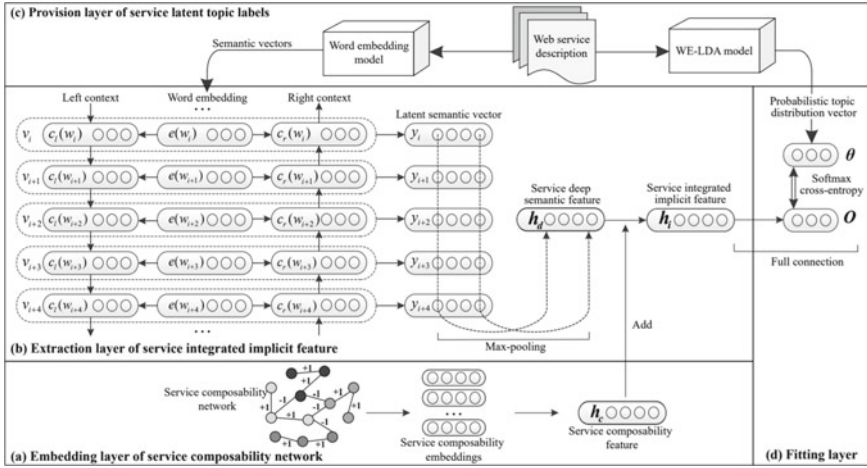


Fig. 6 Deep learning model for efficient semantic analysis [14]

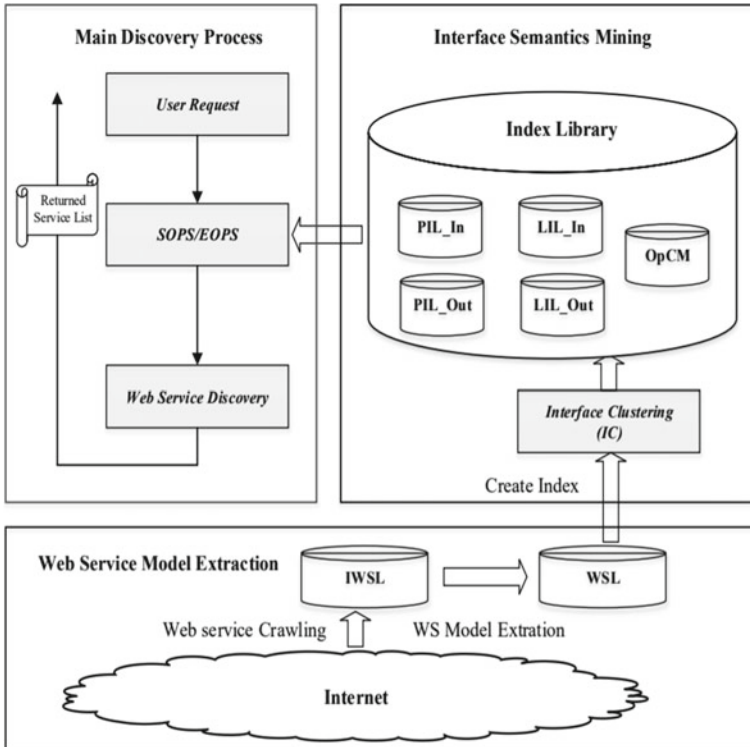
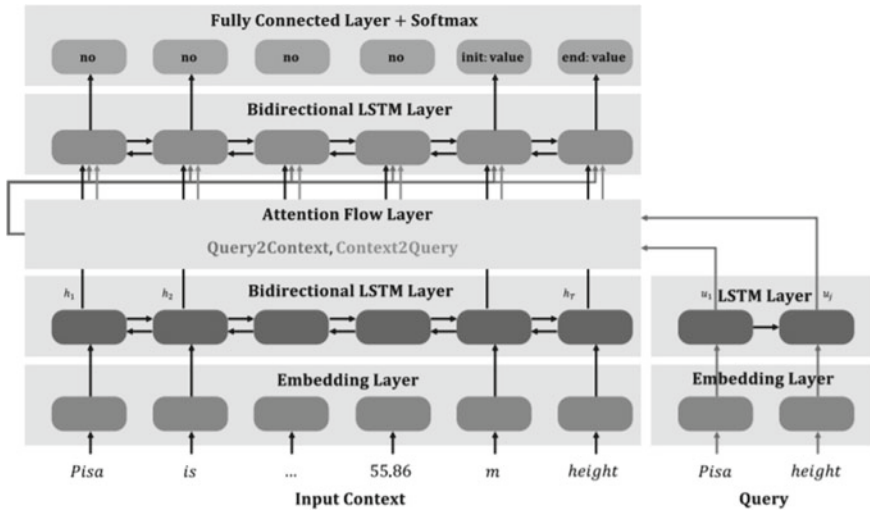


Fig. 7 Indexing with semantic mining for improved web service discovery [16]

be extended to social media datasets as discussed in [17], wherein Latent Semantic Indexing (LSI) is used for discovering Twitter feeds. The model takes into consideration social media-related features that include a number of lists, number of tweets, number of followers, etc. These features are used for the evaluation of popularity, social activity, and other social features. Finally, a weight-based algorithm is used in order to find ranked web search results via LSI matching. The LSI model is able to achieve an accuracy of 65% on Twitter data, which is higher than the list-only model that provides an accuracy of 48% on the same Twitter dataset. Another application-specific model can be observed from [18], wherein Internet of Smart City (IoSC) is used for service composition and discovery. The framework is able to perform distributed processing on IoSC devices, thereby improving the speed and reducing the computational complexity needed for service discovery. This model is able to achieve over 85% accuracy using adaptation and monitoring feedback loops. The performance of these models can be improved via the use of test-oriented service discovery, wherein unwanted services are pre-filtered from the system via semantic matching as observed in [19]. Here, the underlying system model is able to remove unsuitable services using semantic similarity, interface compatibility, functional availability evaluation, and testing candidate responses with actual responses. Due to this 4-level checking, the Test-Oriented API Search with Semantic Interface Compatibility (TASSIC) model is able to achieve an accuracy of 85%, which is higher than existing models like input-output covering (IoC) (83%), similarity semantic scores (SimSS) (55%), and full text information retrieval (FTIR) (54%). The system is highly efficient in terms of removing unwanted tasks and services, but lacks in terms of accurate service discovery and delay of processing. The delay can be reduced via modular semantic processing as observed from [20], wherein web API design is improved via adding modularity to procedural interfaces, while the accuracy can be improved via deep learning models that use a combination of CNNs, LSTMs, and other high-efficiency semantic tools. Such a model is described in [21], wherein Bi-directional Attention Flow (BiDAF) framework is combined with a neural engine. This implementation can be observed in Fig. 8, and showcases the use of bi-directional LSTM, attention flow model, fully connected softmax layers, and word embedding layers to form the Neural Ensemble Knowledge Extraction Model (NEM) architecture. A combination of these layers allows the system to reduce redundancies, and select the most efficient feature vectors for highly accurate web service discovery.

It is observed that the proposed NEM-BiDAF model is highly complex, but has an accuracy of 94.3%, which is higher than the original BiDAF model (76.9%), and original NEM model (76.5%), thereby making the proposed model highly applicable for real-time web service discovery. Further evaluation of these models can be observed in [22–25] wherein discussion about model performance, its applications to medical appointments, neural topic modelling, and service mapping can be observed. These models utilize the same underlying deep learning technique for improving the efficiency of semantic web discovery. These improved efficiency models can be used for lower computational power devices like smartphones via the addition of data augmentation for improved accuracy. The work in [26] proposes such a model,



**Fig. 8** NEM-BiDAF model for highly accurate web service discovery [21]

wherein mobile resource augmentation is performed in order to improve the overall efficiency of web service discovery. Similar models and their standards are described in [27, 28], wherein simple object access protocol (SOAP), web services description language (WSDL), and Universal Description, Discovery, and Integration (UDDI) are introduced for regulating the service discovery capabilities of standard models. Optimizations to these models can be done via methods like distributed computing [29], context-aware text mining [30], service data network [31], service clustering mechanisms [32], privacy awareness [33], and self-organizing maps (SOM)-based models. These models are able to incrementally improve the overall efficiency of the underlying system, and thus must be used for coarse-grained improvement in the system. A statistical comparison of these models can be observed in the next section, which will assist researchers and web service discovery system designers to select the best possible models for their application.

### 3 Quantitative Analysis

In order to quantitatively analyse the performance of the reviewed models, each of these models is evaluated based on their absolute accuracy (AA), delay of computation (D), and computational complexity (CC). While accuracy has an absolute value, the values for the delay and computational complexity are present in relative terms. Thus, they are converted into fuzzy levels of low (L), moderate (M), high (H), and very high (VH). Based on this, performance evaluation for each of the models can be observed from the following Table 1, wherein the given parameters and their

**Table 1** Quantitative evaluation of web service discovery models

Method	AA (%)	D	CC
LDA, k-means and topic modelling [2]	59	M	H
LDA with simple clustering [2]	39	M	M
Word-vector k-means [2]	47	M	H
Fuzzy rules [3]	62	M	M
LFM [4]	55	M	M
ServeNet BERT [5]	91.13	VH	VH
ServeNet GloVe [5]	88.4	VH	VH
AdaBoost [5]	64.92	H	M
CNN [5]	58.46	H	H
LSVM [5]	71.91	H	M
LDA RBF SVM [5]	73.84	H	M
NB [5]	78.94	M	M
RF [5]	80.24	M	H
LSTM [5]	80.1	H	H
Recurrent CNN [5]	84.29	H	H
Convolutional LSTM [5]	84.32	H	VH
Bi-directional LSTM [5]	86.7	VH	VH
Global feedback [6]	68.73	M	H
AOC [6]	62.66	M	L
HaaS [8]	95	H	H
GLDA-QE [9]	81	H	M
GLDA [9]	76	H	M
LDA-WE [9]	70	H	M
Doc2vec [9]	68	M	M
LDA [9]	62	M	M
PLSA [9]	67	M	M
WBiLSTM [10]	92.5	VH	H
BiLSTM [10]	91	H	H
Wide and Deep [10]	77	H	VH
LSTM [10]	71	H	H
WE-LDA [10]	75	H	H
WGLSTM [11]	95.5	H	H
WGSVM [11]	94.4	H	H
Hybrid CSR [12]	83.1	M	L
Hybrid R [12]	81.3	M	L
CBOW [12]	79.3	M	L

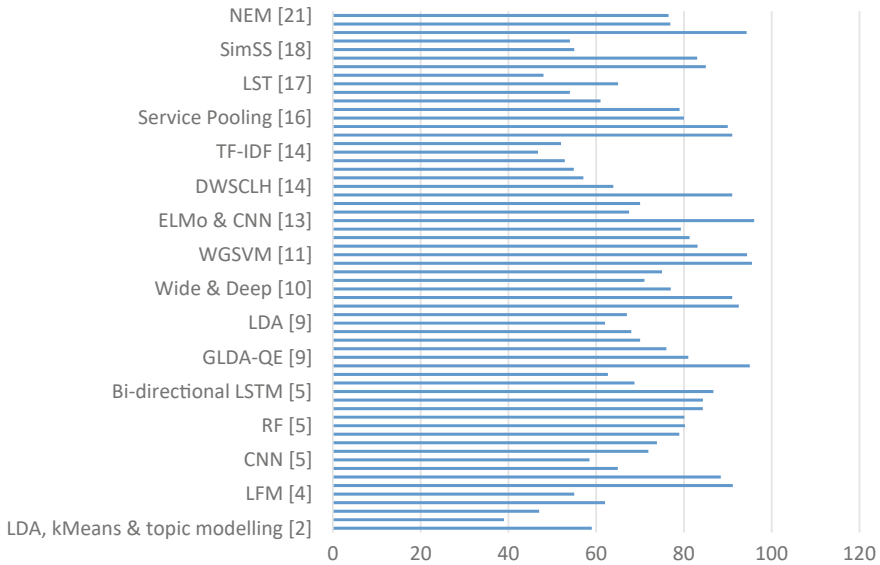
(continued)

**Table 1** (continued)

Method	AA (%)	D	CC
ELMo and CNN [13]	96	H	VH
Word2vec CNN [13]	67.5	H	VH
Doc2vec [13]	70	M	H
LDA and CNN [13]	91	H	VH
DWSCLH [14]	63.9	H	M
DWSCL [14]	57.08	H	H
DCL [14]	54.92	H	M
LDA [14]	52.85	H	M
TF-IDF [14]	46.73	L	L
LDA-K [14]	52	M	L
MSMDI [16]	91	H	M
SSMDI [16]	90	M	M
Service Pooling [16]	80	H	M
Woogle [16]	79	M	M
Schema Matching [16]	61	M	L
Keyword-based matching [16]	54	L	L
LST [17]	65	M	L
List only [17]	48	L	L
TASSIC [18]	85	L	L
IoC [18]	83	M	L
SimSS [18]	55	M	M
FTIR [18]	54	M	M
NEM-BiDAF [21]	94.3	VH	H
BiDAF [21]	76.9	H	H
NEM [21]	76.5	H	H

observed estimates are showcased. Based on this evaluation, researchers should be able to select the most accurate web service discovery model that has a moderate delay and moderate computational complexity, so that it can be suited for a large number of applications.

The result evaluation can be visualized using Fig. 9, wherein absolute accuracy values are compared along with the given method. It can be observed that deep learning models like ELMo and CNN [13], WGLSTM [11], NEM-BiDAF [21], WBiLSTM [10], ServeNet BERT [5], and BiLSTM [10] outperform other models in terms of accuracy. But these models require a large computational delay, and have high computational complexity. For systems that require high speed and low complexity calculations, it is recommended that the TASSIC [18] model, or Hybrid CSR [12] model, must be used.



**Fig. 9** Accuracy for different algorithms (%)

## 4 Conclusion and Future Scope

Deep learning models have always proven to have better performance than linear classification models. This can also be observed in the case of web service discovery applications, wherein models like ELMo and CNN [13], WGLSTM [11], NEM-BiDAF [21], WBiLSTM [10], ServeNet BERT [5], and BiLSTM [10] have proven to have over 91% accuracy in terms of service recommendation. Though these models have high accuracy, the delay needed to train them is very high, thereby they must be used with cloud-based deployments for utmost efficacy. Lower complexity models like TASSIC [18], Hybrid CSR [12], CBOW [12], LDA [9], and LDA-K [14] can also be used for moderate accuracy applications like intranet web search; wherein computational resources are limited. Moreover, lower delay models like TASSIC [18], and Keyword-based matching [16] can also be used for delay-sensitive applications, where accuracy is a secondary factor of evaluation. It is recommended that LSTM, Gated Recurrent Units (GRUs), and recurrent neural networks (RNNs) must be incorporated with these models to improve their internal performance.

## References

1. Mukhopadhyay D, Chougule A (2012) A survey on web service discovery approaches. In: Wyld D, Zizka J, Nagamalai D (eds) *Advances in computer science, engineering and applications. Advances in intelligent and soft computing*, vol 166. Springer, Berlin, Heidelberg. [https://doi.org/10.1007/978-3-642-30157-5\\_99](https://doi.org/10.1007/978-3-642-30157-5_99)
2. Bukhari A, Liu X (2018) A Web administration internet searcher for enormous scope Web administration disclosure dependent on the probabilistic point displaying and grouping. *SOCA* 12:169–182. <https://doi.org/10.1007/s11761-018-0232-6>
3. Ben Seghir N, Kazar O, Rezeg K, Bourekkache S (2017) A semantic web administrations revelation approach dependent on a versatile specialist utilizing metadata. *Int J Intell Comput Cybern* 10(1):12–29. <https://doi.org/10.1108/IJICC-02-2015-0006>
4. Wen Z, Liang Y, Li G (2020) Exploration on shared service discovery model based on collaborative filtering algorithm. In: 2020 IEEE fourth information technology, networking, electronic and automation control conference (ITNEC), pp 2589–2592. <https://doi.org/10.1109/ITNEC48623.2020.9084742>
5. Yang Y et al (2020) ServeNet: a deep neural network for web services classification. In: 2020 IEEE international conference on web services (ICWS), pp 168–175. <https://doi.org/10.1109/ICWS49710.2020.00029>
6. Huang K, Zhang J, Tan W, Feng Z, Chen S (2019) Improving semantic annotations for web service invocation. *IEEE Trans Serv Comput* 12(4):590–603. <https://doi.org/10.1109/TSC.2016.2612632>
7. Ache S, Gao Q, Liu T, He H, Xu G, Liang K (2019) A behavior based trustworthy service composition discovery approach in cloud environment. *IEEE Access* 7:56492–56503. <https://doi.org/10.1109/ACCESS.2019.2913432>
8. Alkalbani A, Hussain W, Kim J (2019) A unified cloud administrations vault (CCSR) system for ideal cloud administration ad disclosure from heterogenous online interfaces
9. Tian G, Zhao S, Wang J, Zhao Z, Liu J, Guo L (2019) Semantic sparse service discovery using word embedding and Gaussian LDA. *IEEE Access* 7:88231–88242. <https://doi.org/10.1109/ACCESS.2019.2926559>
10. Ye H, Cao B, Peng Z, Chen T, Wen Y, Liu J (2019) Web services classification based on wide and Bi-LSTM model. *IEEE Access* 7:43697–43706. <https://doi.org/10.1109/ACCESS.2019.2907546>
11. Guodong L, Zhang Q, Ding Y, Zhe W (2020) Exploration on service discovery methods based on knowledge graph. *IEEE Access* 8:138934–138943. <https://doi.org/10.1109/ACCESS.2020.3012670>
12. Liu F, Deng D, Jiang J, Tang Q (2018) Occasion driven semantic service discovery based on word embeddings. *IEEE Access* 6:61030–61038. <https://doi.org/10.1109/ACCESS.2018.2876029>
13. Huang Z, Zhao W (2020) Blend of ELMo representation and CNN approaches to enhance service discovery. *IEEE Access* 8:130782–130796. <https://doi.org/10.1109/ACCESS.2020.3009393>
14. Zou G, Qin Z, He Q, Wang P, Zhang B, Gan Y, DeepWSC: clustering web services by means of integrating service composability into deep semantic features. *IEEE Trans Serv Comput*. <https://doi.org/10.1109/TSC.2020.3026188>
15. Zhang F, Zeng Q, Duan H, Liu C (2019) Piece context-based web services similarity measure. *IEEE Access* 7:65195–65206. <https://doi.org/10.1109/ACCESS.2019.2915371>
16. Cheng B, Li C, Chen J (2021) Semantics mining & indexing-based rapid web services discovery framework. *IEEE Trans Serv Comput* 14(3):864–875. <https://doi.org/10.1109/TSC.2018.2831678>
17. Sivrikaya F, Ben-Sassi N, Darn X, Görür OC, Kuster C (2019) Web of smart city objects: a distributed framework for service discovery and composition. *IEEE Access* 7:14434–14454. <https://doi.org/10.1109/ACCESS.2019.2893340>



18. Mama S, Chen Y, Syu Y, Lin H, FanJiang Y, Test-oriented RESTful service discovery with semantic interface compatibility. *IEEE Trans Serv Comput.* <https://doi.org/10.1109/TSC.2018.2871133>
19. Barros A, Ouyang C, Wei F (2020) Static analysis for improved modularity of procedural web application programming interfaces. *IEEE Access* 8:128182–128199. <https://doi.org/10.1109/ACCESS.2020.3008904>
20. Kim M, Lee S, Oh Y, Choi H, Kim W (2020) A near-real-time answer discovery for open-domain with unanswerable questions from the web. *IEEE Access* 1–1. <https://doi.org/10.1109/ACCESS.2020.3020245>
21. Pourghebleh B, Hayyolalam V, Aghaei A (2020) Administration disclosure in the Internet of Things: audit of latest things and examination challenges. *Remote Netw* 26:5371–5391. <https://doi.org/10.1007/s11276-020-02405-0>
22. Sharifi O, Ataee SM, Bayram Z (2020) Casing Logic-based determination and disclosure of semantic web administrations with application to clinical arrangements. *Master Syst* 37:e12387. <https://doi.org/10.1111/exsy.12387>
23. Tian G, Liu P, Peng Y, Sun C (2018) Labeling increased neural theme model for semantic scanty web administration disclosure. *Simultaneousness Comput Pract Exper* 30:e4448. <https://doi.org/10.1002/cpe.4448>
24. Dong H, Hussain F, Chang E (2013) Semantic web service intermediaries: state of the craftsmanship and difficulties. *Simultaneousness Comput: Pract Exp* 25. <https://doi.org/10.1002/cpe.2886>
25. Anitha S, Padma, T (2020) A web administration level web of things system for versatile asset expansion. *Int J Commun Syst* 33:e4475. <https://doi.org/10.1002/dac.4475>
26. Curbera F, Duftler M, Khalaf R, Nagy W, Mukhi N, Weerawarana S (2002) Unwinding the web administrations web: a prologue to SOAP, WSDL, and UDDI. *Web Comput IEEE* 6:86–93. <https://doi.org/10.1109/4236.991449>
27. Pakari S, Kheirkhah E, Jalali M (2014) Web service discovery methods and techniques: a review. *Glob J Comput Sci Eng Inf Technol* 4. <https://doi.org/10.5121/ijcseit.2014.4101>
28. Seghir NB, Kazar O (2017) A new framework for web service discovery in distributed environments. In: 2017 first international conference on embedded and distributed systems (EDiS), pp 1–6. <https://doi.org/10.1109/EDIS.2017.8284046>
29. Shafi S, Qamar U (2018) [WiP] web services classification using an improved text mining technique. In: 2018 IEEE eleventh conference on service-oriented computing and applications (SOCA), pp 210–215. <https://doi.org/10.1109/SOCA.2018.00037>
30. Gu Q, Cao J, Yang X (2018) A web services composition discovery approach based on service data network. In: 2018 IEEE international conference on progress in informatics and computing (PIC), pp 344–350. <https://doi.org/10.1109/PIC.2018.8706331>
31. Yang B, Sailer A, Jain S, Tomala-Reyes AE, Singh M, Ramnath A (2018) Administration discovery based blue-green deployment technique in cloud native environments. In: 2018 IEEE international conference on services computing (SCC), pp 185–192. <https://doi.org/10.1109/SCC.2018.00031>
32. Lee H, Chow R, Haghghat MR, Patterson HM, Kobsa A (2018) IoT service store: a web-based system for privacy-minded IoT service discovery and interaction. In: 2018 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), pp 107–112. <https://doi.org/10.1109/PERCOMW.2018.8480260>
33. Xiao Q, Cao B, Zhang X, Liu J, Hu R, Li B (2018) Web services clustering based on HDP and SOM neural network. In: 2018 IEEE SmartWorld, ubiquitous intelligence and computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people and smart city innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCOM/IOP/SCI), pp 397–404. <https://doi.org/10.1109/SmartWorld.2018.00097>

# An Integrated Semi-Supervised Learning Framework for Image Compression Using DCT, Huffman Encoding, and LZW Coding



Jamandlamudi Sravya Sruthi, Rekh Ram Janghel, and Lokesh Singh

**Abstract** The application of compression of the data to digital files is called image compression. The encoding of images tackles the question of reducing the volume of data needed for digital image representation. The storage space for image loading is also limited. The compression of the picture can be lossy or lossless. This text aims to use different strategies including Discrete Cosine Transformation, LZW Coding, and Huffman Encoding to implement simple JPEG compression. Digital picture data is transformed from spatial domain to frequency domain by the Digested Cosine Transformations (DCT). The compression methods used in this article do not impact data loss in the field of picture transparency. We use JPEG for this reason. JPEG is a format for still compression frames based on the Discrete Cosine Transform and is also appropriate for most compression applications. In a python framework, the project has been tested and a compressed picture was finally created. We were able to achieve 96.11%, 46.61%, and 67.91% compression in DCT, LZW, and Huffman, respectively.

**Keywords** DCT · LZW · Image compression · Huffman encoding

## 1 Introduction

Compression of photographs is a key aspect of digital computation. By compressing pictures to a fraction of their original dimension, you will conserve precious (and costly) storage space [1]. Further, it is simpler and less time-consuming to transport pictures from one device to the next (that is why compression is essential in

---

J. S. Sruthi · R. R. Janghel  
National Institute of Technology Raipur, G.E. Road Raipur, Raipur, C.G, India  
e-mail: [sravyasruthij222@gmail.com](mailto:sravyasruthij222@gmail.com)

R. R. Janghel  
e-mail: [rrjanghel.it@nitrr.ac.in](mailto:rrjanghel.it@nitrr.ac.in)

L. Singh (✉)  
Alliance College of Engineering & Design, Alliance University, Bangalore, Karnataka, India  
e-mail: [lokesh.singh@alliance.edu.in](mailto:lokesh.singh@alliance.edu.in)

the growth of the internet) [2]. In the last 15 years, image compression has been considered a significant topic of study. This topic is as relevant as image processing, pattern recognition, and more specifically biometrics [2, 3] since it has become a tool for recognizing applicants who have participated in numerous national and foreign academic exams for e-passports and for identifying them [4].

For successful image transmission and storage, image compression is quite necessary. Exploding demand for the communication of media data across the network and the access of multimedia data across the Internet [5]. The demands for collection, editing, and transfer of digital images have risen explosively as digital devices are utilized. This demand has increased. The photo files may be really huge and have a lot of storage space. A  $256 \times 256$  pixel gray image has 65,536 storage objects and almost one million of the standard  $640 \times 480$  color images [6]. It can be a very time-consuming process to retrieve these files from the internet [7]. Picture data constitute an important part of multimedia data and comprise an important part of multimedia connectivity bandwidth. Thus, it has become absolutely important to establish effective techniques for image compression [8].

Image compression plays an important part in the effective transfer and storing of photographs and has multiple uses [9]. The compression of the picture is intended to minimize image data redundancy by storing or transferring a limited number of samples and we can recreate from this an outstanding accession of the original image according to human visual perception [10].

It is a known that the Huffman algorithm is used to reduce the codes for compression. The Huffman code-works used for the compression of documents like image, video, MPEG-2, MPEG-4 and H.263 framework systems. The methodology of Huffman coding is used for special source image symbols and its estimation direction meaning is used for compression of the image. The JPEG method is a very common Lossy medium. Picture, compression, the discrete core Transforming Cosine (DCT). The DCT is separated by Pictures of various frequencies in sections. In a little bit Quantization is named. Where is the compression? The least relevant frequencies are discarded, So the expression “Lossy” is used. The most, then, Significant remaining frequencies are used to recover the image in the process of decompression. As a consequence, The photos that have been restored include some distortion; this distortion degree can be changed soon during the process of compression. The JPEG protocol is implemented for black and white pictures as well as color [5].

A number of ways are accessible to compact picture details. Certain formats of files include TIFF, JPEG, GIF, PNG, and wavelets. Massive file measurement is provided by TIFF photos. This photographic file is the most common form [11]. The JPEG and GIF formats are the two most common web application compressed image formats. For pictures, the JPEG technique is sometimes used; for linear art and other photographs with very fundamental geometric shapes, the GIF type is used. In order to replace GIF in an open format, PNG was created [12, 13].

The illustration below demonstrates the need for digital image compression. a. To store a moderately large color image, for example, 512 Pixels, 0.75 MB of disk space is needed. b. For internet image with 35 mm diaphragm with  $12 \mu\text{m}$  resolution, 18 MB space needed. c. A PAL digital (Phase Alternation Line) video, 27

MB memory space required. Learn more. It's important that to preserve and include these images, compression methods over network (e.g., internet) are required. It's important. Compression of photographs solves the issue of reducing the amount of digital data needed picture. The basis for the reduction method is redundant data deletion. As mentioned, This amounts to transformation in quantitative terms: A two-dimensional pixel sequence in a mathematical data set. The shift is introduced beforehand for image preservation or switch.

The JPEG cycle is a broadly utilized type of lossy picture, compression that is based on the Discrete Cosine Transform (DCT) [3]. The DCT works by isolating pictures into parts of varying frequencies during a stage called quantization. Where part of compression really happens, the less significant frequencies are disposed of, thus the utilization of the term "lossy" [4]. At that point, just the most significant frequencies that remain are utilized to recover the picture in the decompression cycle. Thus, recreated pictures contain some contortion; however, as we will observe, this degree of bending can be changed during the compression stage. The JPEG strategy is utilized for both shading and high-contrast images.

## 2 Literature Review

In this study, the literature reviews the advances achieved with Discrete Cosine Transform (DCT) in the area of image compression effects. The DCT algorithms are being tested and only the JPEG compression is applicable. In this analysis, progress is made along with several other algorithms on the use of face reconnaissance in the DCT algorithms of an image, namely, 2D DCT. Raid et al. [2] include the titanic automated data for image compression. A framework that is expected to store, send, and recuperate epic picture data. DCT changed picture pressure system diminishes the abundance pieces for capable limit, transmission, and recuperation of the image by stuffed picture keeping up the equivalence with the main picture subject to human wisdom. Straight change, quantization, and entropy coding are the methods drawn in with the lossy pressure technique [3]. In facial recognition, the resolution or scale of the photo plays a significant part. The easier it's, the higher the resolution. However, the impact of image compression on the face recognition device does not take into consideration its significance in recent years [4]. Photos may be compressed for numerous uses, including storage of images in memory such as handheld devices or small devices, transfer of massive files across the network, or storage of large quantities of images for scientific or testing purposes in databases. This is significant since compressed images consume less memory or can be transferred more easily due to their limited size. This has contributed to an important rise in the impact of image compression on facial detection and to becoming one of the important study fields for other biometric methods as well as for the identification of iris and fingerprints. More recent contributions have been made in the field of iris recognition [6, 14] and fingerprint recognition [6]. Hudson et al. [7] depict that picture pressure dependent on lossless pressure incorporates numerous calculations, however, even the best of these

couldn't give a compression ratio (CR) not more than 2–3. This wasteful estimation of the pressure proportion brings about unacceptable outcomes. Accordingly, the use of the lossy pressure strategy for the picture pressure becomes necessary [8]. The photo should be in decompressed mode, which is one of the most significant disadvantages of face recognition using compressed photos. However, it is computationally costly to decompress a compressed picture to recognize the face, and facial recognition technologies would benefit if total decompression could somehow be avoided. This ensures that facial detection is done when the photographs are compressed and that the computer speed and the average output of a face recognition device are, therefore, improved. JPEG [9] and its associated transformations are the most common compression methods and discrete transformations of the cosine. It is regarded as the highest implementations of image compression such as JPEG and JPEG2000 provide since the image must still be decompressed and displayed at any stage to the human being. This review examines in depth the advancement in the implementation of face recognition of the DCT and other algorithms of a single image, namely, 2D DCT [11]. The endless favorable support of Discrete Wavelet Transform don't legitimize wide substitution of DCT [13]. Discrete Cosine Transform is a method for changing over a sign into rudimentary recurrence parts, broadly utilized in picture compression. The quick development of computerized imaging applications, including work area distributing, mixed media, remotely coordinating, and top-quality TV (HDTV) has expanded the requirement for successful and normalized picture pressure methods [15]. Regardless of whether various investigations have been created in this field of interest, the pressure of biomedical pictures [10] stays a significant issue. Since the rise of computerized procurement in clinical imaging, information creation is consistently developing. As of late, it has been dependent upon a semi-dramatic increment, specifically due to broad utilization of MRI and CT (computer tomography) images. These are both volume modalities that can be seen as a grouping of 2D pictures (cuts). The progressive upgrades of procurement hardware tend to intensify the goal of those pictures, which strengthens the mass of information to chronicle [11]. This makes them truly much bulkier than other imaging modalities. This is the reason we zeroed in on CT and MRI [12]. Huffman Coding depends on a number rendition of a discrete cosine change and a novel, low intricacy yet proficient, entropy encoder utilizing a versatile Golomb–Rice calculation rather than Huffman tables [16]. Quantization, as a center module of wavelet transform-based on lossy picture code, adequately diminishes the visual repetition. Thus, it is basic to far off detecting picture pressure [12]. For the pressure of a RGB picture, it is first changed into shading space utilizing a RCT. After the shading change, the luminance channel Y is compacted by a customary lossless picture coder. Pixels in chrominance channels are anticipated by the various leveled disintegration and directional expectation [17]. Gonzalez et al. [18] depict that the typically used picture plan is JPEG considering the way that it gives raised degree of weight and moreover maintenance is given to it by most of the web programs. Still, picture pressure is used in JPEG. The monster extent of picture quality is maintained by JPEG and most of the mechanized pictures apply this [5]. JPEG uses a benchmark system as a weighting procedure also, and it is DCT based. In spite of the way that the idea of the conveyed picture is lesser than

that of the principal picture, it gets difficult for the characteristic eye to recognize [13].

### 3 Methodology

#### 3.1 Compression Techniques

Lossless versus Lossy compression: the restored image is numerically similar to the initial picture in the lossless compression method after compression. However, the compression without loss can only be minimal. Lossless encoding is superior to medical imagery, scientific sketches, clip art, or comics for archival purposes. This is because compression methods introduce compression artifacts, especially when used at low bit rates. A recreated picture is deteriorated compared to the initial compression. This is mostly because the compression device discards obsolete details entirely. Yet loss systems can achieve even higher compression. Loss strategies are particularly appropriate for natural images such as portraits, where a small (sometimes unnoticed) allegiance loss is acceptable in order to minimize the bit rate considerably. Visually lossless may be considered the lossful compression that occurs in imperceptible discrepancies.

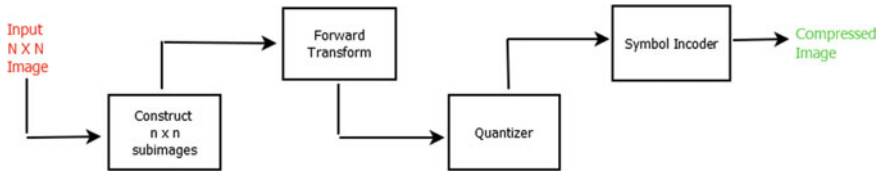
#### 3.2 Discrete Cosine Transform (DCT)

Due to its almost optimized efficiency compared to the statistically optimal Karhunen-Loeve transform [19], the Two-Dimensional Discrete Cosine Transform (2D DCT) has been one of the most common transforms for many image compression applications. The 2D DCT is computer-intensive, and therefore, demands high-speed, high-performance, and low latency computer architectures. The 2D DCT processor architecture was based on tiny nonoverlapping blocks because of the strong calculation requirements (typical  $8 \times 8$  or  $16 \times 16$ ). A variety of 2D DCT algorithms were suggested to minimize machine complexity and thus improve operating speed and output [10]. The 2D DCT may be processed as two 1D DCT compositions in each dimension. The concept is formal

$$X(l, m) = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} X[i, j], C_{li} C_{mj} \tag{1}$$

The multiplication of 2D DCT basis functions can be defined as

$$C_{li} C_{mj} = \cos \left[ \frac{\Pi}{N} \right] \left( i + \frac{1}{2} \right), \cos \left[ \frac{\Pi}{N} \right] \left( j + \frac{1}{2} \right) m \tag{2}$$



**Fig. 1** Flow chart for DCT

With an  $8 \times 8$  block in the diagram, a square of frequency components is the product of 2D DCT [5], which is treated as the position (0, 0) on the top left corner. There is an X (0, 0) DC coefficient in the zero coefficient. This DC coefficient is concerned with the most significant elements of the signal [9]. Vector quantization algorithms are one of the most common face recognition techniques. Various methods are already suggested for extracting image functional vectors [7] (Fig. 1).

The initial picture is broken down into  $8 \times 8$  blocks. Pixel value for a pixel in a black and white picture varies from 0 to 255, but DCT is intended for pixel values between  $-128$  and  $127$  DCT matrix estimation. For any row, DCT is added by multiplying the changed block with the left DCT matrix and transposing the correct DCT matrix. Then each block is quantitatively compact. The matrix is then encoded as Quantized Matrix. A reverse step is carried out to recreate the compressed file. Inverse DCT for decompression is used.

### 3.3 Lempel-Ziv-Welch (LZW) Coding

LZW is a fully dictionary code (Lempel-Ziv-Welch). LZW is often split into static and dynamic encoding. During the coding and decoding procedures, the dictionary is statically fixed. The dictionary is modified with dynamic dictionary coding [11]. Single codes are used to substitute character strings with LZW encoding. No review of the incoming text is carried out. It just adds any new character string from the string table. Although it may be of any arbitrary length, the code that the LZW algorithm produces should have more bits than one character. For files containing several repetitive details, LZW compression works well. The LZW dictionary retains the compression. Both stream and code are contained in this dictionary [12] (Fig. 2).

As an input, use the stream. Initialize the dictionary to include each stream character. Interpret the stream if the real byte is the stream's end, then leave. Read the next stream character from the dictionary. If the dictionary does not include the character, then add a new string to the dictionary and write the current string in to the entry code, then write out the string and exit encoding code from the dictionary. If the bunch is always around, send them a special code (according to the diagram). Read the next stream character in the dictionary. If the dictionary does not include the character, then: A. Add a new string to the dictionary, B. Write the current string entry code, and C. Take move 4; write out the string and exit encoding code.

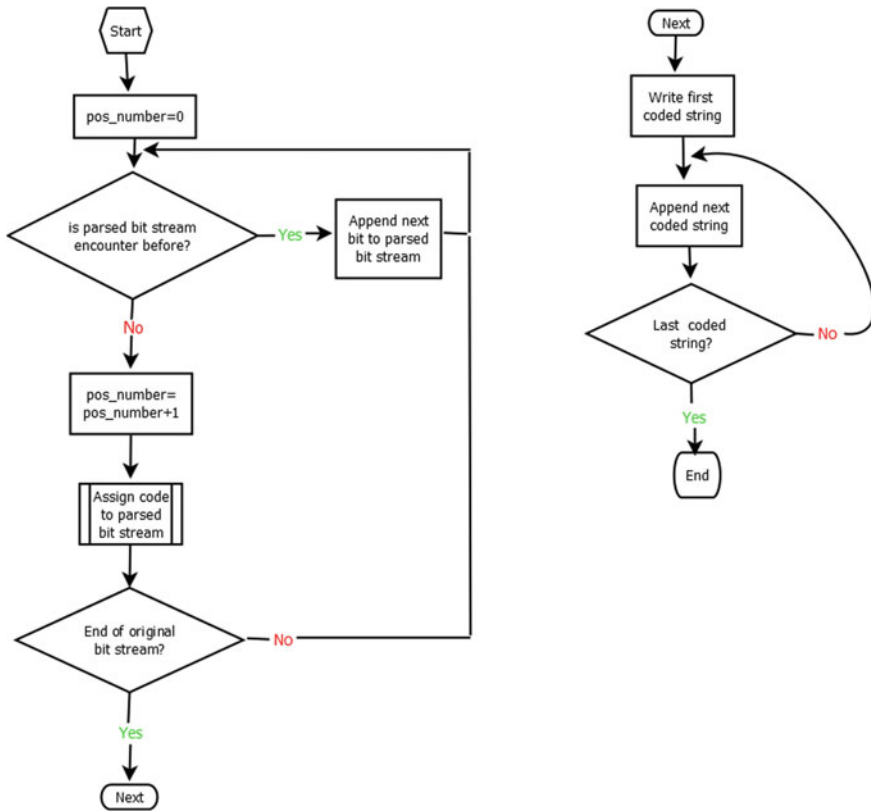
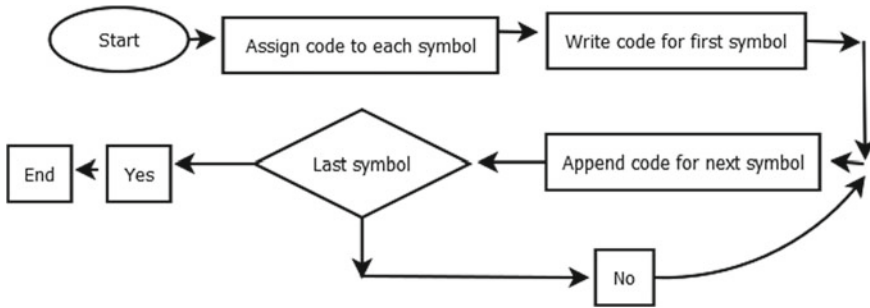


Fig. 2 Flow chart of LZW coding

### 3.4 Huffman Encoding

Huffman coding is a computation entropy encoding used for data compression without loss. The word applies to the usage of an encoding code table of a variable length for a picture of the source where a code table of the variable length is deduced from a particular purpose based on the event's calculated possibility. Coding of Huffman depends on the recurrence of the case, Bahrami A. Set et al. [16]. The pixels are treated like photos in the frame. Fewer items are given to the photographs that take place as often as possible while the images that take place less frequently are assigned a greater number of pieces. The code of Huffman is a prefix. Any picture's double coding is not the code prefix for any other picture. In the previous pressure phases, the majority of instructions for image coding use loss approaches and use change as the last move. Huffman coding [29]. The estimates from Huffman are static as scalable in two territories. The measurement of Static Huffman is a process that encrypts data in two parts. Firstly, the recurrence of any picture is determined, and secondly, the Huffman tree is formed. In the second pass, a scalable calculation





**Fig. 3** Flow chart for Huffman encoding

of Huffman was made, which constructs a Huffman tree in one pass, but requires longer space than the calculation of Static Huffman, Singh [13]. An et al. Overall, using Huffman's standard document coding could recover them anywhere between 10 and 30 percent, based on the dissemination of the character. As Huffman encodes the configuration of a tree, the string is encoded (Fig. 3).

Take image as input and apply the symbols-finding function (un-repeated pixel value). Call a functionality to measure any symbol which is likelihood. Now sign odds are structured and the likelihoods are get reduced. Name the symbols-finding function (un-repeated pixel value). Name a functionality to measure any symbol's likelihood. Sign odds are structured and the likelihoods are reduced. This process remains same until only two probabilities left and codes are allocated according to the rule that a shorter-length code is available for the most likely symbol. More encoding of Huffman will be done, i.e., the code words will be mapped with the corresponding symbols. Restoration of the original file, i.e., decompression with Huffman decoding. To retrieve the picture reconstructed, fit the code terms with the code dictionary. Build the encoding tree's tree counterpart. Read the input feature-wise until it hits the last part. Leave the leaf character encoding and return to the core, obey phase 9 before all the symbols codes are identified.

### 3.5 Quantization

In the region of higher frequency coefficients, the human eye may eliminate many redundant details [20]. This can be accomplished if the frequency components are separated into a corresponding constant and rounded to the nearest integer [21, 22]. As a result, all of the higher frequency rates are zero.

In instances where  $X(l, m, n)$  are the pre-quantization frequency coefficients, and  $Q(l, m, n)$  is post-quantizing frequency coefficients [23]. This procedure generates lost information and thus, since the coefficients are rounded, certain components cannot be recovered in the decompression phase because of loss of information [24, 25]. The benefit of such an action is that it contributes to less data being saved. The

constant of which and frequency variable is used is to be determined, so the cube of quantization must be specified. Its segments specify the video sequence compression ratio and output rate [17].

The next step of JPEG compression is quantifying where lower-frequency components are expected. JPEG is separated into a quantizing factor for each DCT value and is then rounded to the nearest integer. Since DCT variables are  $8 \times 8$ , an  $8 \times 8$  table is used, referring to each DCT performance expression [26–28]. This table is then preserved by the JPEG file such that this table or a regular table can be used by the decoding process. Notice that several component files must have multiple tables, including one for each component Y, Cb, and Cr.

### 4 Results and Analysis

Examples showing Compression results for our dataset (Figs. 4, 5, 6, 7 and 8).

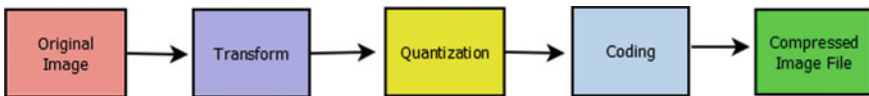


Fig. 4 A typical image compression system

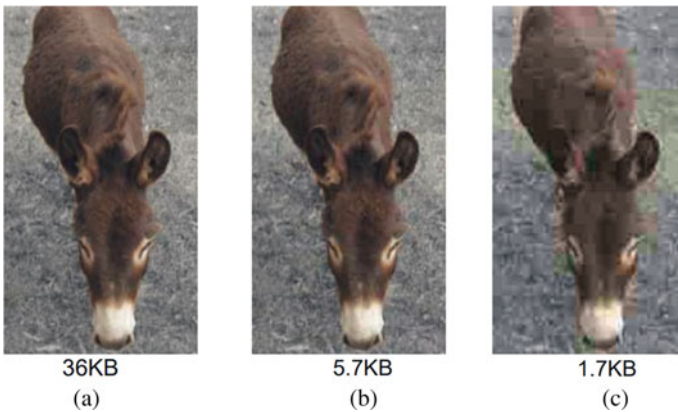
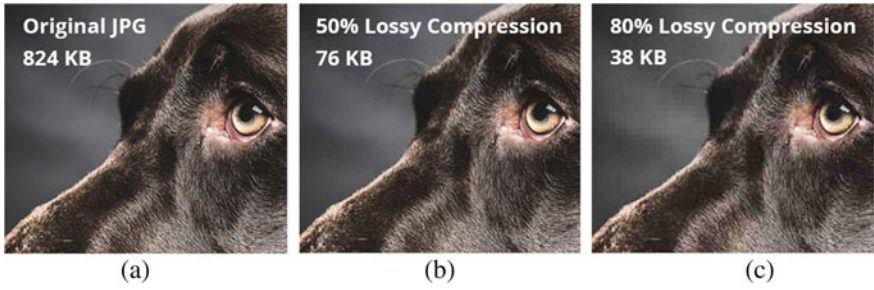
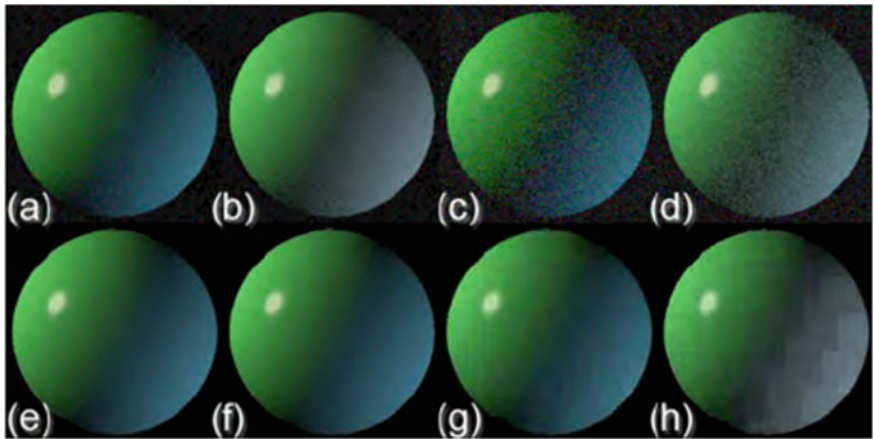


Fig. 5 a Original image. b Compressed image using Huffman encoding, and c Compressed image using DCT



**Fig. 6** a Original image, b 50% lossy compressed image, and c 80% lossy compressed image



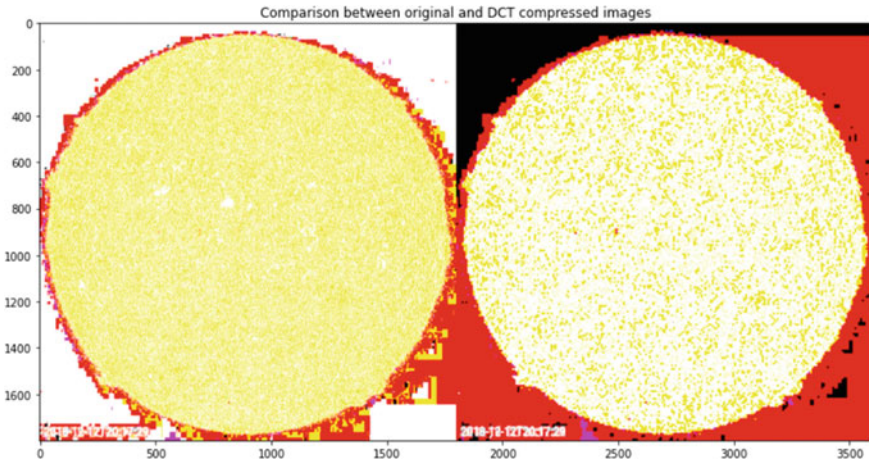
**Fig. 7** JPEG compression: a and c noise 4.0 and 9.0; b and d reconstructions; e and g 20 and 80% compressed; f, h reconstructions

### 4.1 DCT Technique

Compression Ratio (CR) and Compression Percentage (CP) obtained over 5 different images using DCT Technique (Table 1 and Fig. 9).

### 4.2 LZW Technique

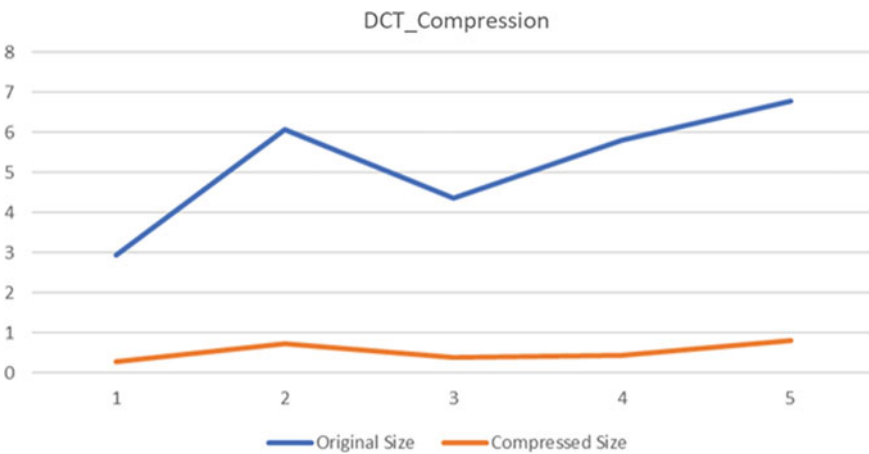
Compression Ratio (CR) and Compression Percentage (CP) obtained over 5 different images using LZW Technique (Table 2 and Fig. 10).



**Fig. 8** Image of Solar flares: [NASA Directory]. **a** Original image. **b** Compressed DCT. \*Test data used may or may not be the same as the images showcased here

**Table 1** Compression percentage obtained using DCT

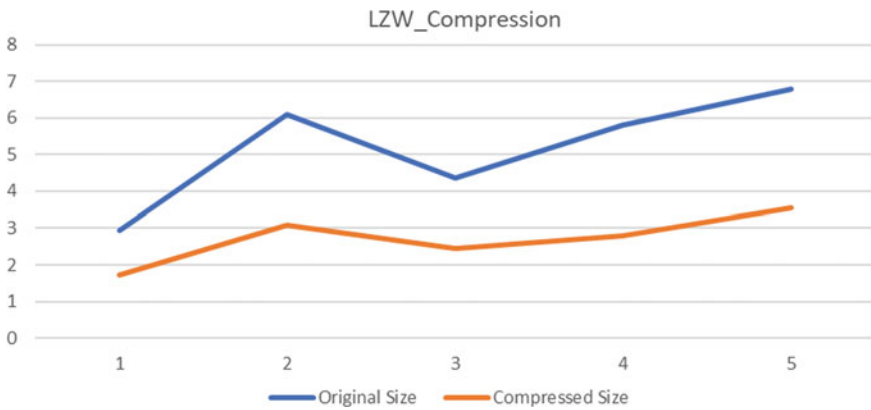
DCT technique of compression				
File number	Original size	Compressed size	Compression ratio	Compression percentage
1	2.94	0.13	0.044217687	95.57823129
2	6.08	0.26	0.042763158	95.72368421
3	4.37	0.12	0.027459954	97.25400458
4	5.8	0.19	0.032758621	96.72413793
5	6.78	0.32	0.04719764	95.28023599



**Fig. 9** Graphical representation of compression obtained using DCT

**Table 2** Compression percentage obtained using LZW

LZW technique of compression				
File number	Original size	Compressed size	Compression ratio	Compression percentage
1	2.94	1.74	0.591836735	40.81632653
2	6.08	3.08	0.506578947	49.34210526
3	4.37	2.46	0.562929062	43.70709382
4	5.8	2.81	0.484482759	51.55172414
5	6.78	3.55	0.52359882	47.64011799



**Fig. 10** Graphical representation of compression obtained using LZW compression

### 4.3 Huffman Encoding Technique

Compression Ratio (CR) and Compression Percentage (CP) obtained over 5 different images using Huffman Encoding (Tables 3 and 4, Figs. 11 and 12).

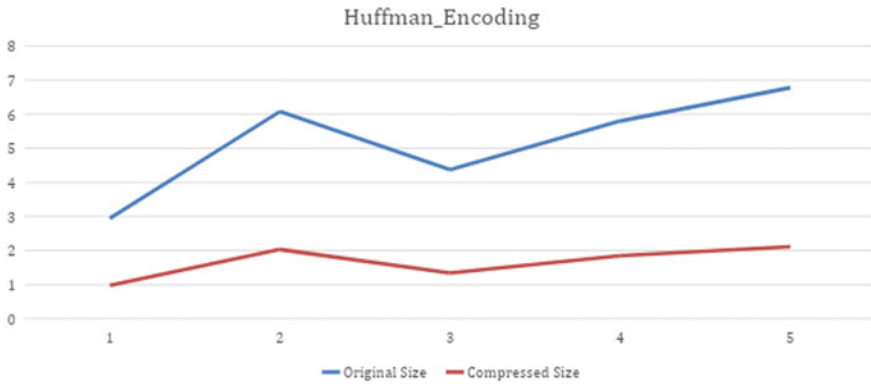
So we were able to achieve a mean compression of 89.94% using DCT for compressing JPEG images. We obtained 67.67% using LZW Encoding and 61.68%

**Table 3** Compression percentage obtained using Huffman encoding

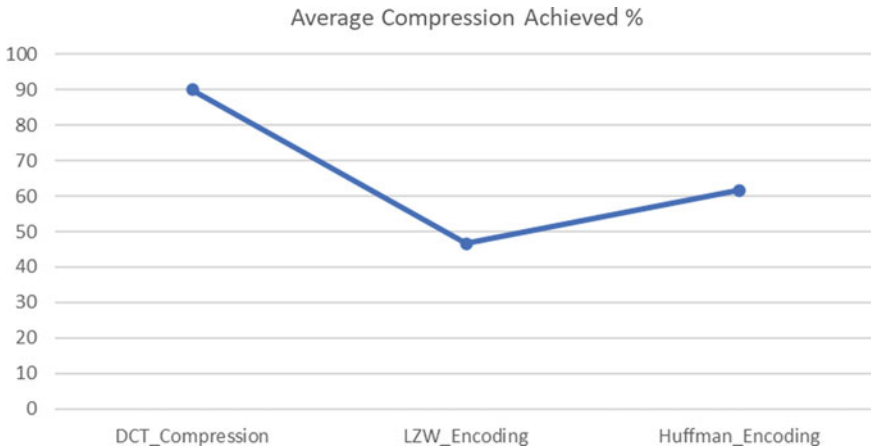
Huffman encoding technique of compression				
File number	Original size	Compressed size	Compression ratio	Compression percentage
1	2.94	0.98	3	66.66
2	6.08	2.03	2.99	66.61
3	4.37	1.34	3.26	69.33
4	5.8	1.85	3.13	68.10
5	6.78	2.11	3.21	68.87

**Table 4** Comparison of different techniques (DCT vs. LZW vs. Huffman)

Comparison (LZW vs. Huffman vs. DCT)		
S.no	Technique used	Average compression achieved (%)
1	DCT_compression	96.11
2	LZW_encoding	46.61
3	Huffman_encoding	67.91



**Fig. 11** Graphical representation of compression obtained using Huffman encoding



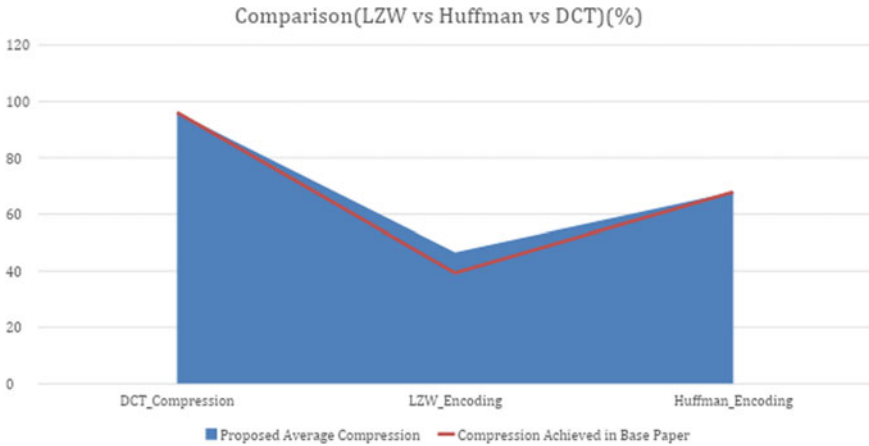
**Fig. 12** Graphical representation of average compression obtained

using Huffman Encoding. These are the mean of total compression obtained in 5 different images (Table 5, Fig. 13).

From the above graph, we can conclude that our performance was satisfactory for Huffman and LZW encoding, whereas we need to improve a bit in our DCT

**Table 5** Comparison of CP with paper of different techniques (DCT vs. LZW vs. Huffman)

Comparison (LZW vs. Huffman vs. DCT) (%)		
S.no	Compression methods	Proposed compression ratio
1	DCT_compression	96.1120588
2	LZW_encoding	46.61147355
3	Huffman_encoding	67.91947951



**Fig. 13** Graphical representation of comparison with base paper

implementation, and the difference is mostly due to the blocking effect which has been an issue throughout the implementation and is discussed in the research paper.

## 5 Conclusion

This paper is an undertaking that effectively actualized the JPEG picture compression. The framework is planned by utilizing Python and Jupyter programming available over the web. This task has been tried for all potential circumstances on the Jupyter cloud and on Google Cloud. Our major objective was to compare various compression techniques with the CR they provide. DCT, LZW, and Huffman coding were compared on the same set of images. The results show that there is a vast difference in the compression ratio achieved between these techniques. Huffman and LZW were close to each other in the compression percentage, whereas DCT provided a much better compression than the other two, although DCT provides lossy compression. In future, we can include more techniques and comparing them to find the most suitable technique for different images to get the best possible compression ratio. Also, we can use different compression techniques to improve our efficiency.

## References

1. Isac B, Santhi V (2011) A study on digital image and video watermarking schemes using neural networks. *Int J Comput Appl* 12(9): 0975–8887
2. Gonzalez RC, Woods RE, Eddins SL (2003) *Digital image processing using MATLAB*, 1st edn. Prentice Hall. ISBN-10:0130085197. ISBN-13:978-0130085191
3. Wallace GK (1991) The JPEG still picture compression standard. *Commun ACM* 34(4):30–44
4. Jain AK (1989) *Fundamentals of digital image processing*. Englewood Cliffs: Prentice Hall information and system sciences series. Prentice\_Hall International, London
5. Rao KR, Yip P (1990) *Discrete cosine transform: algorithms, advantages, applications*. Academic Press, San Diego, CA
6. ISO/IEC: Information technology-JPEG 2000 image coding system-Part 1: core coding system (2000) ISO/IEC 15444-1:2000(ISO/IEC JTC/SC 29/WG 1 N1646R)
7. Russ JC (2006) *The image processing handbook*, 5 th edn. CRC Press. ISBN-10:0849372542. ISBN-13:978-0849372544
8. Ames G (2002) *Image compression*
9. Boliek M, Gormish MJ, Schwartz EL, Keith A (1997) Next generation image compression and manipulation using CREW. In: *Proceedngs of the IEEE ICIP*
10. Smith BC, Rowe LA (1993) Algorithms for manipulating compressed images. *IEEE Trans Comput Graph Appl* 13
11. Kaimal AB, Manimurugan S, Devadass CSC (2013) Image compression techniques: asurvey. *Int J Eng Invent* 2(4)
12. Singh A, Gahlawa M (2013) Image compression and its various. *Int J Adv Res Comput Sci Softw Eng* 3(6)
13. Shahbahrami A, Bahrampour R, Rostami MS, Mobarhan MA (2003) Evaluation of humman and arithmetic algorithm for multimedia compression standards
14. Liu C-C, Hang H-M (2005) Acceleration and implementation of JPEG 2000 encoder on TI DSP platform. In: 2007. *ICIP 2007. IEEE international conference on image processing*, vol 3, pp III-329–339
15. Adiego J, Navarro G, Fuente PDL (2004) Lempel-Ziv compression of structured text. In: *Data compression conference, Spain, 2004*; CmojeviC V, Senk V, Trpovski Z (2003) Lossy Lempel-Ziv algorithm. *Telsiks*, pp 523–525
16. Nguyen H, Shwedyk E (2009) *Introduction to digital and data commuicaions*. Campbridge, New York
17. Weizheng R, Haobo W, Lianming X, Yansong C (2011) Research on a quasi-lossless compression algorithm based on huffman coding. In: *International conference on transportation, mechanical, and electrical engineering*, pp 1729–1732
18. Gonzalez RC, Woods RE (2007) *Digital image processing*, 3rd edn. Prentice Hall. ISBN-10:013168728X. ISBN-13:978-0131687288
19. Bonnie L (1996) Stephens, student thesis on “Image Compression algorithms.” California State University, Sacramento
20. Kirovski D, Landau Z (2004) Generalized Lempel-Ziv compression for audio. In: *IEEE 6th workshop on multimedia signal processing, USA*
21. Gonzales R, Woods R (2001) *Digital image processing*. Prentice Hall, New Jersey
22. Das R, Tuithung T (2012) A novel steganography method for image based on Huffman encoding. In: *3rd National conference on emerging trends and applications in computer science (NCETACS), Shillong*
23. CmojeviC V, Senk V, Trpovski Z (2003) Lossy Lempel-Ziv algorithm. *Telsiks*, pp 523–525
24. Săcăleanu DI, Stoian R, Ofrim DM (2011) An adaptive huffman algorithm for data compression in wireless sensor networks. *Bucharest*
25. Bedruz RA, Quiros AR (2015) Comparison of Huffman algorithm and Lempel-Ziv algorithm for audio, image and text compression 1–6. <https://doi.org/10.1109/HNICEM.2015.7393210>



26. Jagadish HP, Lohit MK (2002) A new lossless method of image compression and decompression using huffman coding techniques; Zeimer R, Peterson R (2001) Introduction to digital communication. Prentice Hall, New Jersey
27. Ashida S, Kakemizu H, Nagahara M, Yamamoto Y (2004) Sampled-data audio signal compression with huffman algorithm. In: SICE annual conference in Sapporo, Hokkaido
28. Maghari A (2019) A comparative study of DCT and DWT image compression techniques combined with Huffman coding. *Jordanian J Comput Inf Technol (JJCIT)* 5:73–87. <https://doi.org/10.5455/jcit.71-1554982934>

# Analyzing the Need for Video Summarization for Online Classes Conducted During Covid-19 Lockdown



Shikha Sharma and Madan Lal Saini

**Abstract** The evolution of wearable cameras has revolutionized the exponential growth of the industry due to the production of video content. This has ultimately generated the need for storage management, video management, video summarization, methods to reduce the cost of resources, etc. The present paper compares and evaluates the different techniques available for video summarization nowadays and to visualize the need of applying these methods to the content of online classes delivered during the COVID-19 pandemic. This pandemic has affected the lives of children as they all have got stuck totally onto the screen for their education and learning through their respective organizations. But this trend has raised many challenges among the parents, educational lists, and health experts. These challenges mainly include the availability of online gadgets with internet connectivity, especially in rural areas, and the health issues generated because of the overexposure to online gadgets by students all over the world. As per the educational experts, including UNESCO and UNICEF, this continuous exposure to online gadgets has alarmed the different potential dangers to their health including obesity, stubbornness, heart diseases, vision loss, etc. This paper mainly focuses on the need for summarizing the online class videos not only to reduce the burden of the overhead cost of Internet and resources, but also to reduce the harmful effects produced on the health of these students due to the continuous exposure of the unnatural light and waves generated through these gadgets.

**Keywords** Video summarization · Resource management · Online class · COVID-19

---

S. Sharma (✉)  
Research Scholar, Poornima University, Jaipur, India  
e-mail: [shikha.sharma@poornima.edu.in](mailto:shikha.sharma@poornima.edu.in)

M. L. Saini  
Poornima University, Jaipur, India  
e-mail: [madan.saini@poornima.edu.in](mailto:madan.saini@poornima.edu.in)

# 1 Introduction

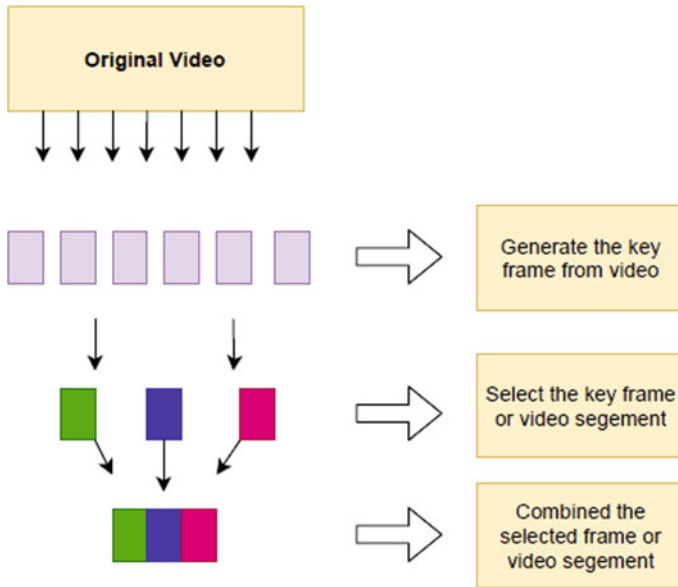
With the rapid growth of digitization and the development of the IT industry along with the proliferation of wearable cameras for capturing and watching videos, a big flood of users has come forward to get video hosting services both on the internet and cloud. This rapid increase in the popularity of videos has generated the demand for efficient video retrieval [1], which has given rise to the requirement of video management as well as video summarization. The major objective of video summarization is to get a short and precise picture of the content available in the video through the proper analysis of it including the removal of redundant data and extraction of relevant keyframes from it [2]. Therefore, the objective of video summarization is to handle massive and redundant data for storage management, resource management, bandwidth management, etc. [3].

The video summarization thus helps to extract the silent information from the lengthy videos by suppressing the redundant data perceptively which is used for surveillance purpose [4], action or activity recognition and retrieval, generating movie trailer, developing sports highlight, traffic monitoring, crime investigation [5], etc. Thus, the summarized video can be used for reviewing the crucial content and important aspects of a specific video along with indexing and faster browsing.

Conclusively we can say that summarization of the video is a task for creating and presenting a meaningful abstract view of the content within a short period of time [5] which includes the segmentation of the original video into small shots to extract the video frames constituting the precise and concise presentation. The steps involved in this procedure can be explained in Fig. 1.

This method of video summarization is of two types:

- Static video summary/key frame-based approach
  - Dynamic Video Skimming.
  -
1. The key frame-based approach: In this technique, the key frames can be evoked through the investigation of change point, low level-based clustering or clustering depending on objects [7]. These key frames are very useful in case of video indexing. This method can be further classified as event based, object based, indexed based, attention based or hierarchical code-book methods, etc.
    - (a) Event based: This method is used by the anomalous events within the recordings and following the changes by classifying them through some characterization features for watching the unusual highlights. At that point, a synopsis calculation joins all the chosen edges to develop a new video.
    - (b) Object based: In this technique, we select an article in the video as our key component. At that point, all casings having this featured article are viewed as used to make the new video or video outline. Here, the achievement of the strategy relies upon the substance of the video.



**Fig. 1** Basic procedure for video summarization [6]

- (c) Hierarchical Codebook: It is a powerful arrangement of pictures in the nearness of Divisive or Agglomerative grouping that recognizes the substance of a video. There are restrictions on spatial transient relationships and items spread over the entire video. The upside of this procedure is, it goes after steady and makes a montage that can be seen in zoomed mode as well. This structure utilizes the Region of Interest strategy to make the codebook.
  - (d) Index Based: It utilizes some key edges or highlights to make a file so as to suit the looking-through procedure in the video being perused. In content-based recovery frameworks, record monitors articles, casings, and portions identified with the extraction procedure of one of the kind key edges. In reconnaissance recordings generally face and shapes are the highlights extricated to file the recordings. This method utilizes the extraction of elements and the identification of the strong component through the extraction strategies.
  - (e) Attention Based: In this, the framework creates a whole number an incentive for all the shots or key-outlines. These number qualities are considered as significance scores or effect scores for the casing. The higher the score, higher will be the significance of the casing. These high-scored outlines are consolidated to produce the abstraction.
2. Dynamic Video Skimming approach: In case of video skimming, we select the short segments of the video for summarization. While selecting these segments the care must be taken that each segment is a complete video to attain

the objective of getting the useful content through summarization [7]. This approach can be further classified as Trajectory based and sparse dictionary based to get video skimming.

- (a) Trajectory based: It helps in recognizing the object of intrigue and following the path of areas where the article has visited. This strategy is generally utilized in reconnaissance recordings to expel vulnerability and find the articles in worldly information. This strategy is introduced by the hidden Markov [8] Model to recognize irregular content during the web test and classes. The conduct relies upon the head postures of the understudies and a grouping is considered as strange if its registered likelihood is more noteworthy than ordinary.
- (b) Sparse dictionary learning: It alludes to making an unaided system that can follow screen changes and make logs of information after some time. It would require some investment to make the word reference or log, however, it helps in the future to make the synopses dependent on a powerful rundown. Meng et al. [9] gave multi-see inadequate word reference determination with centroid co-regularization strategy to advance the agent choice of casings. Li et al. [10] proposed to develop a video videography word reference to speak to each video as arrangement of words for unconstrained proposed to develop a video videography word reference to speak to each video as arrangement of words for unconstrained recordings. The word reference is framed utilizing bunching and shot limit procedure.

## 2 Literature Survey

Ajmal et al. [11] summarized various video abstraction techniques and design a tree like structure which shown different techniques to store and processed large amount of data, produced due to recording of various events. The author has also explained application-wise categorization in order to design an optimum model to get the best result.

Khattabi et al. [12] applied Key frame approach; where the important frame was the representative frame about the shot. These have been extracted indiscriminately and evenly from unique video sequence. It concerned for choosing one or multiple frames as represented the content of the video for producing video summaries.

Jeong1 et al. [13] presented a routine by which the stationary summary will be presented. Considering that, the videos usually had a few doubtful boundaries that Low-quality video boundaries also have the problem of suitable frame rates, so no technological disturbance generated due to the lower bandwidth or unsuitable frame rate. They have used two step approach. In step 1 delete all duplicate frames and in the 2nd step, performed segmentation and afterward utilized nice clustering for each concerning the segments, the place every frame is represented by way of the rare coding of SIFT features. Author chronic UTE or ATL datasets.

Otani et al. [1] introduced a video briefness method to provide a rapid pathway in imitation of overview its content. They have used deep video attributes that encode a range of ranges on content semantics, which includes objects, actions & scenes, they have designed a flagrant neural community up to expectation map videos as well so descriptions in accordance with a common semantic area then mutually trained such including associated pairs over movies and descriptions. For generating summary, they have extracted the deep attributes out of every segment over the unique video and utilized a clustering-based briefness technique after them.

Zhu et al. [14] proposed a video short approach based upon machine learning techniques for automatic video gait prediction. Several functions had been extracted in imitation of characterized video boundary, such as cut, vanish in, vanish oversea then disappear because facilitating the understanding content structure then area guidelines of a video.

Bora et al. [6] explained a basic flow chart that will explain the step-by-step summarization process. The author has also explained the various demanding video summarization areas. Where one needs to get a summary of the full video. They have talked about types of videos and what is the most prominent video format, they have also discussed about types of summaries generated by video according to their application. The author has given a comparative study of the previous research, provided details of the data set, approaches, and frameworks on which research is already been done.

Workie et al. [2] designed a framework that will combine the benefits of supervised and unsupervised video summary techniques. It is a GAN-based model which consists of two different types of networks, for each network, they will calculate the loss to improve outcomes of the whole network. They have also discussed various applications, methods, and trends in that domain.

Srinivas et al. [7] proposed an algorithm to generate a summary of the video. These algorithms consist of three step first of all they gave score of every frame calculate at the beginning and on the basis of key frame score they combined them and at the end remove duplicate frames to generate appropriate summary of video. For that they worked on various features of video frame and used K-means clustering to choose key frame.

Otani et al. [9] compared the random generated summary with human generated summary for various video segments. They have worked on two different datasets, SumMe and TvSum, and also evaluated which data set gives proper results. The author has also calculated co-relation between human observation and the estimate scoring of the random segment.

Yuan et al. [15] constructed a deep semantic embedded network by using multiple hidden layers to minimize semantic loss and feature recreation loss. They have used encoders between layers and also trained their model by using Thumb1K and TVSum50 datasets. Author has generated summary by using captions in future.

Venkatesh et al. [16] explained the procedure of key frame extraction and video skim, and also described various attributes of both techniques. The author also discussed summarization techniques according to various application areas and also

described the performance benefits of abstraction techniques. They also presented the various gaps in this domain.

Rochan et al. [17] designed a deep learning framework consisting of two different networks. The first network was used for key frame selection and the other one was used to discriminate the summary. They have also designed a mapping function that will produce a mapping between original and abstract video. Authors have also compared hand-crafted data summary with summary generated by supervised learning. They have worked on unpaired data and also used different datasets to present which data set will produce a good F-score on their designed model.

Jadon et al. [18] designed a project in which they compared various methods for summarization on the SumMe data set and discussed which model gives good F-scores over others. According to the author, CNN Gaussian performed well. They worked on various clustering algorithms and compared their outcomes with human generated outcomes.

Ejaz et al. [19] introduced a group based design for extraction of important casings from recordings. The important virtue of this case is the center of backhanded importance input from clients within education mold boundaries for correctly demonstrating human consider of similitude. Exploratory outcomes demonstrate, so much the utilization on non-straight extended premise capacities in imitation of demonstrate the dissolution within highlights vectors diminishes the semantic gap among classy level semantics or low-level file highlights. Besides, the results exhibit up to expectation the lesson bases over RBFN lets in the frame after effectively be brought familiar including the fluffy concerning client's objectives. When the framework boundaries are discovered outdoors of getting ready stage, framework does not require any info boundaries from the client. Furthermore, creator format is properly fabulous because of non-stop internet video perusing by means of demonstrating on-the-fly video synopses. In future, our structure can test recordings of concerning distinctive sorts and operate crucial modifications.

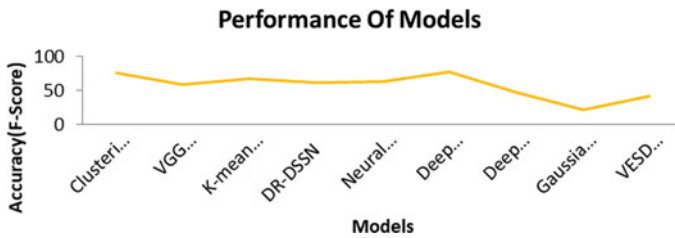
Cai et al. [20] discussed one trouble that creates a problem with eyesight due to the resolution of the device which we are using for attending the online class. To focus on this important issue, the author brought a multiplicative briefness fabric referred to as VESD in conformity with leveraging the net movies for higher latent semantic modeling or in imitation of decreasing the gloom on video summarization in a principled way. The author integrated flexible net previously outgiving into a variation mold and then presented simple coding and decoding techniques over data and collected information or data. The capacities over VESD frame because of big-measure along the ground-truth honor scores (cyan background) significantly (Table 1).

We have generated a table after reviewing multiple papers; here tables show F-score of an algorithm. We summarized all reviews according to the summarization techniques, proposed methods, and data sets used by various authors; the final column shows the accuracy of the above techniques in terms of F-score.

Based on the above analysis, we will draw a performance chart of various summarization techniques using different models (Fig. 2).

**Table 1** Summary report of various summarization techniques

Techniques	Video summary
Motion based	Dynamic summary
Color based	Static summary
Dynamic contents	Dynamic summary
Gesture based	Static summary
Audio-Visual based	Dynamic summary
clustering based	Both Static and dynamic summary
Trajectory based	Static summary
Shot selection based	Static summary
Event based	Static summary



**Fig. 2** Performance of various models

Now we will design one more table which will list out various types of videos, and based on that content, we must choose the summarization technique, i.e., whether it is a static or dynamic summarization (Table 2).

**Table 2** Suitable techniques for various types of summarizations [6]

S.no	Summarization techniques	Purposed method	Data-sets	Results (F-score)
1	Static video summarization [13]	clustering-based and the visual attention-based methods	UTE ADL	0.763 0.793
2	Dynamic video summarization [6]	VGG based video summary	SumMe dataset 25 VIDEOS	58.8

(continued)



**Table 2** (continued)

S.no	Summarization techniques	Purposed method	Data-sets	Results (F-score)
3	Key frame based summarization [7]	K-means clustering	open-video.org	66.6
4	Video segmentation [21]	DR-DSSN	SumMe and TVSum	61.2
5	Automatic video transition [14]	Neural network	Traffic video (V248)	63.3
6	Key frames approach [15]	Deep semantic embedding	Thumb1K and TVSum50	76.6
7	Key frame-based approach [16]	Deep learning framework	UnpairedVSN	47.6
8	Key frame-based approach [17]	Gaussian clustering along with CNN	SumMe	21.2
9	Variational autoencoder [20]	VESD Framework	TVSum	40.8

### 3 Conclusion

Through this paper, it has been observed that the video summarization is done by various advanced machine learning, Deep learning, Computer Vision, and Deep Fusion Learning approaches including CNN, ResNet, VGG, DenseNet, along with the combination of multiple encoders and decoders. Though these techniques are available in abundance, none of these methodologies is efficient enough to produce high-level accuracy which undoubtedly directs us to do more work in this area. Moreover, due to the flood of online classes during COVID-19, the demand for video summarization has also increased to a great extent leading to the demand for techniques providing more précised and accurate results which are not available through the present methodologies. Conclusively, we can say that the presently available methods of video summarization have limitations such as:

1. Summarization does not give appropriate results on noisy video.
2. The accuracy of the work is lower.

### 4 Future Work

Through this paper, we have created a focus on the need for a proper framework for video summarization. This requirement for video summarization has increased more in case of the online classes conducted during the COVID-19 pandemic. So, the future work of this paper is mainly focused on the need for a proper and efficient framework or a model for video summarization. This framework is actually needed in case of online classes which are being conducted due to the change in teaching

and learning methods during the COVID-19 pandemic. This need is emphasized here as the stress level among the students is increasing every day due to harmful effects on their health. So, one of the major challenge arisen during COVID-19 is to smoothen the delivery of valuable lecture content to the student for the sake of Teaching and Learning Methodologies is suggested to be solved here through the video summarization method. This idea would be helpful in the field of education and development reducing the overhead of the student from going through the entire lectures and expanding the continuous cost of the Internet including costly gadgets. Furthermore, we can say that this suggestion of summarizing video lectures would be very beneficial for the improvement of Education leading to the upgradation in proficiency and robustness of the efforts done in online learning.

## References

1. Otani M, Nakashima Y, Rahtu E, Heikkilä J, Yokoya N (2016) Video summarization using deep semantic features. In: Asian conference on computer vision. Springer, Cham, pp 361–377
2. Workie A, Sharma R, Chung YK, Digital video summarization techniques: a survey
3. Ghafoor HA, Javed A, Irtaza A, Dawood H, Dawood H, Banjar A (2018) Egocentric video summarization based on people interaction using deep learning. *Math Probl Eng*
4. Muhammad K, Hussain T, Del Ser J, Palade V, De Albuquerque VHC (2019) DeepReS: a deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. *IEEE Trans Industr Inf* 16(9):5938–5947
5. Sebastian T, Puthiyidam JJ (2015) A survey on video summarization techniques. *Int J Comput Appl* 132(13):30–32
6. Bora A, Sharma S (2018) A review on video summarization approaches: recent advances and directions. In: 2018 international conference on advances in computing, communication control and networking (ICACCCN). IEEE, pp 601–606
7. Srinivas M, Pai MM, Pai RM (2016) An improved algorithm for video summarization—A rank-based approach. *Procedia Comput Sci* 89:812–819
8. Cote M, Jean F, Albu AB, Capson D (2016) Video summarization for remote invigilation of online exams. In: 2016 IEEE winter conference on applications of computer vision (WACV). IEEE, pp 1–9
9. Meng J, Wang S, Wang H, Yuan J, Tan YP (2017) Video summarization via multi-view representative selection. In: Proceedings of the IEEE international conference on computer vision workshops, pp 1189–1198
10. Li K, Li S, Oh S, Fu Y (2017) Videography-based unconstrained video analysis. *IEEE Trans Image Process* 26(5):2261–2273
11. Ajmal M, Ashraf MH, Shakir M, Abbas Y, Shah FA (2012) Video summarization: techniques and classification. In: International conference on computer vision and graphics. Springer, Berlin, Heidelberg, pp 1–13
12. Elkhatabi Z, Tabii Y, Benkaddour A (2015) Video summarization: techniques and applications. *Int J Comput Inf Eng* 9(4):928–933
13. Jeong DJ, Yoo HJ, Cho NI (2016) A static video summarization method based on the sparse coding of features and representativeness of frames. *EURASIP J Image Video Process* 2017(1):1
14. Ren W, Zhu Y (2008) A video summarization approach based on machine learning. In: 2008 international conference on intelligent information hiding and multimedia signal processing. IEEE, pp 450–453

15. Yuan Y, Mei T, Cui P, Zhu W (2017) Video summarization by learning deep side semantic embedding. *IEEE Trans Circuits Syst Video Technol* 29(1):226–237
16. Truong BT, Venkatesh S (2007) Video abstraction: a systematic review and classification. *ACM Trans Multimed Comput Commun Appl (TOMM)* 3(1):3-es
17. Rochan M, Wang Y (2019) Video summarization by learning from unpaired data. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7902–7911
18. Jadon S, Jasim M (2019) Video summarization using keyframe extraction and video skimming. [arXiv:1910.04792](https://arxiv.org/abs/1910.04792)
19. Ejaz N, Baik SW (2012) Video summarization using a network of radial basis functions. *Multimed Syst* 18(6):483–497
20. Cai S, Zuo W, Davis LS, Zhang L (2018) Weakly-supervised video summarization using variational encoder-decoder and web prior. In: *Proceedings of the European conference on computer vision (ECCV)*, pp 184–200
21. Otani M, Nakashima Y, Rahtu E, Heikkila J (2019) Rethinking the evaluation of video summaries. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 7596–7604

# Local Roughness Binary Pattern for Texture Classification



Sumit Kumar Gupta, Susheel Yadav, Dharendra Pratap Singh,  
and Jaytrilok Choudhary

**Abstract** The texture is an essential property of an image that tells about the arrangements of pixels of different intensities in an image matrix. Researchers have used many descriptors for texture classification of the images in computer vision and image processing. Local Binary Pattern (LBP) is one of the descriptors, which is computationally efficient and straightforward for texture classification. In recent years, researchers have made many enhancements to the basic LBP method to improve the quality of extracted features for an image. This paper presents a Local Roughness Binary Pattern (LRBP) descriptor for feature extraction of an image to increase LBP discriminative capability. This new descriptor extracts features from the local region of an image called a partial feature vector and concatenates these local features to obtain a feature vector of an image. This feature vector is used as input to a classifier to classify the class of an image. We tested the proposed descriptor for the images available on CURET and KTH-TIPS2b databases with Support Vector Machine (SVM) and k-nearest Neighbors (KNN) classifiers. We have presented the experimental results obtained from the analysis process regarding classification accuracy and confusion matrix. Finally, we have presented a performance comparison between the proposed method, basic LBP descriptor, and its variants. Results show that the proposed descriptor gives much better results than other descriptors on the KTH-TIPS2b Database when we use the KNN classifier.

**Keywords** LBP · Local binary pattern · LRBP · Local roughness binary pattern · SVM · Support vector machine · KNN K-nearest neighbors

## 1 Introduction

The texture properties of an image are used for its classification. Several methods are proposed in the literature to extract the texture features of an image LBP [1, 2] and Census Transform (CT) [3]. The LBP is a very efficient and powerful descriptor in

---

S. K. Gupta (✉) · S. Yadav · D. P. Singh · J. Choudhary  
Department of Computer Science and Engineering, MANIT, Bhopal, MP, India  
e-mail: [sumitgupta888@gmail.com](mailto:sumitgupta888@gmail.com)

extracting the local features of an image. In the last few decades, the LBP has inspired methods that are deeply studied or used in image processing and computer vision. To apply the LBP in any image, generally the image is divided into multiple regions. LBP is first applied on each reach of the image. To evaluate the LBP code for a region, each pixel of the region is compared with its neighboring pixels. After evaluating the LBP code for every region, we evaluate the LBP histogram of the region. The Histogram of all regions is concatenated to create the Histogram of the image. This Histogram represents a spatial feature vector of the image. LBP features conation two important properties; tolerance regarding monotonic illumination change and computational simplicity.

In recent years, LBP and its variants have been used in many applications and help to solve many problems like image analysis [4], face image analysis, image and video retrieval [7, 8], motion analysis [9, 10], environment modeling [11, 12], and remote sensing [13].

There is a lot of information present in an image. Generally, the rough region of an image contains more information because corners and edges are present in this region. The corners and edges of an image provide more information. This information is extracted to identify the texture of an image. There are various descriptors present to extract information from a rough region, but the way to extract information is different. If we select more number of neighboring pixels around a pixel to calculate the LBP code, then we can tell in a better way the effect of neighboring pixels on a pixel. More precise and detailed information gives better texture classification of the images.

This paper proposes a new descriptor that increases the discriminative capability of an image and gives better texture classification results. Here LBP discriminative capability means how well a descriptor explains the binary number sequence formation process. If we choose more number of neighboring and next to neighboring pixels around a pixel, then we can tell more clearly the effect of neighboring pixels on a given pixel. This work also involves a comparison between the recent variants of LBP and the proposed descriptor.

The rest of the paper is organized into five sections. Section two reviews some recent research work related to texture classification. Section three describes the proposed method and explains it with the help of an example. Section four talks about the data set used in this research. Section five presents the results and its analysis. Finally, section six presents our conclusions and some future directions.

## 2 Literature Review

In recent years, the operator LBP has come with plenty of variations to improve performance in different applications. Extended LBP [14, 15] has worked on the discriminative capability of LBP and shows improved results in comparison to LBP. It uses the GD between the central and its neighboring pixels. Guo et al. [16] proposed a new operator Complete LBP [16], which is very much similar to the extended LBP. It

uses two things; sign and the GD between the central and its neighboring pixels. Local Smoothness Pattern [17] captures the local edge and curve patterns in an image. It uses nonlinear modeling. Soft LBP [18, 19] was introduced with a fuzzy membership function to make LBP sensitive to an image's noise. Adaptive LBP descriptor [20] proposed to minimize the direction difference along with different orientations. It minimized the variation of means and standard deviation of differences. Liao et al. proposed a dominated LBP descriptor [21] to overcome the inability of the basic LBP to capture the shapes having high curvature edges, crossing boundaries, or corners by finding the dominated features in an image. Median Binary Pattern [22] uses the median value of the selected neighbor set for deciding the threshold value. LBP Histogram Fourier descriptor [23] uses the combination of uniform LBP descriptor and Discrete Fourier Transform. Ahonen et al. [24] proposed monogenic LBP to enhance the classification performance of basic LBP by combining uniform local binary patterns, local phase information, and local surface information. LBP Variance [25] was used as an adaptive weight to adjust the LBP contribution in Histogram calculation. It generates the Histogram for different scales. Neighbor Intensity LBP [26] uses the basis of pixel intensity and pixel difference. They proposed four different descriptors and combined all four to form a joint histogram to represent texture.

LBP variants that are used for texture classification vary according to the following points:

1. How to select the threshold value?
2. How to choose neighboring pixels, whether it is lying on the perimeter of a circle or ellipse or sphere (for 3D images) or square or rectangle?
3. The number of neighboring pixels.

LBP variant calculates LBP code by comparing the intensity of each neighboring pixel with a threshold value (also the possible intensity of a given pixel) or by comparing two neighboring pixels. So by reviewing many texture classification-related papers, which tells how we will choose neighborhoods around a given pixel and how we will compare each neighboring pixel with a given pixel. By reviewing all these papers, we knew how to select a neighborhood around a pixel, compare each neighboring pixel with a given pixel, and choose threshold values.

So taking inspiration from the above, we proposed a new descriptor to extract feature vector from the local region of an image by comparing pixels of a region with their neighboring pixels and neighboring pixel with their next neighboring pixel in a particular direction.

### 3 Proposed Work and Methodology

The proposed descriptor is named Local Roughness Binary Pattern (LRBP). It extracts information from the local region of an image. It first divides the image into several regions. We apply the LRBP to every pixel of the region, to calculate the LBP code and then a new decimal value corresponding to pixels. Calculate Mean

and Variance or Histogram of the region after obtaining decimal values for every pixel of the region. We use the mean and Variance or Histogram as a partial feature vector for the regions. The partial feature vector of the regions is combined to form a feature vector of the image.

To calculate the LBP code by using the LRBP select every pixel of the region as a center value and consider eight neighboring pixels and 16 next neighboring pixels around the center value, as shown in Fig. 1.

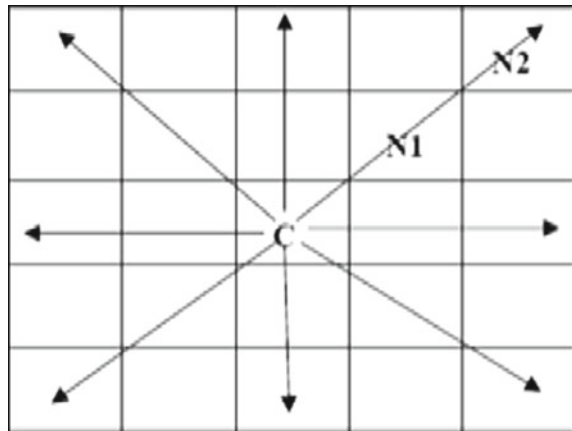
To calculate the LBP code by using the proposed LRBP descriptor, we use the following procedures. If there is a regular increment or decrements in pixel intensity (there is no variation) in a particular direction concerning center value, then put values zero instead of a neighboring pixel in that particular direction. If there is not a regular increment or decrements in a particular direction, then put one (because here edge detection can occur) instead of a neighboring pixel in that particular direction, as shown in the equation below.

$$s(x) \begin{cases} 1, & \text{if } \begin{cases} C > N1 \text{ AND } N1 < N2 \\ \text{OR} \\ C < N1 \text{ AND } N1 > N2 \end{cases} \\ 0, & \text{if } \begin{cases} C > N1 \text{ AND } N1 > N2 \\ \text{OR} \\ C < N1 \text{ AND } N1 < N2 \end{cases} \end{cases}$$

Where,  $s(x)$  is the binary number corresponding to the neighboring pixel in a particular direction.  $C$  is the intensity of the central pixel and,  $N1$  and  $N2$  are intensities of neighboring pixel and next to neighboring pixel in a particular direction, respectively.

After evaluating all neighboring values corresponding to the center pixel, we need to calculate the decimal value of a center pixel. In binary representation of center

**Fig. 1** Working procedure of LRBP descriptor



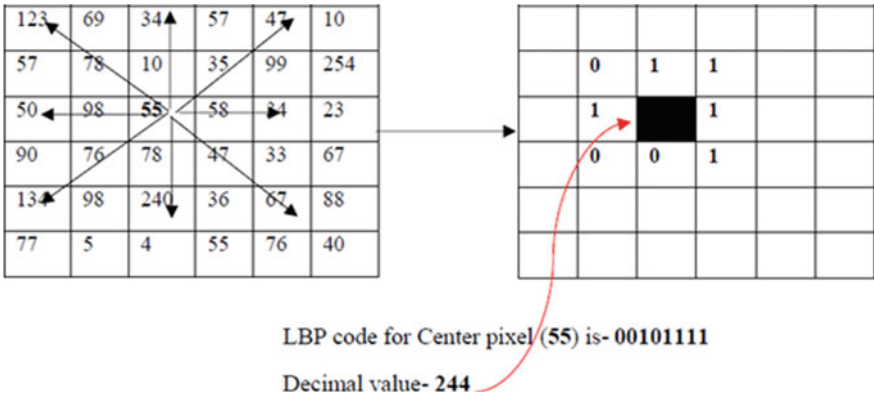


Fig. 2 Example of LBP code using the proposed descriptor

pixel value, pixel at below to the center pixel represents the MSB bit, and pixel at the right corner of below to the center pixel represents the LSB bit of a binary number of eight bits. All other neighboring pixel values are considered in clockwise from the pixel at below to the center pixel to pixel at the right corner of below to the Centre pixel. Figure 2. shows an example of calculation of the intensity of Centre pixel using the LRBP.

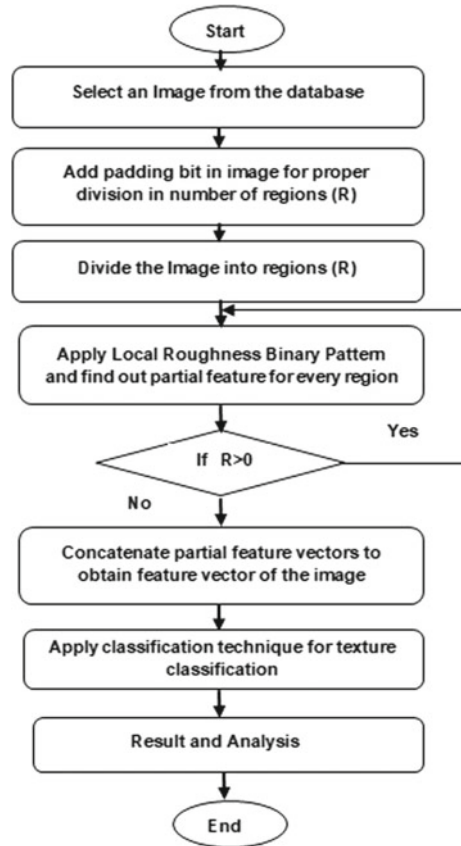
### 3.1 Proposed Algorithm

In the proposed algorithm of texture classification with some labeled classes, we train a model that tells us the class of the images. To evaluate the performance of our model, we split the data into training and testing set. The model learns using the training set. Their performance is measured using the testing set. Testing results allow us to measure how well the algorithm generalizes to unseen instances. The main measure of performance for a classifier is accuracy. It is the proportion of instances in the test set that are classified correctly. Steps of the algorithm with the proposed LRBP descriptor are described below:

- Step 1: Select an image from the database.
- Step 2: Add a padding bit in the image if the dimension of this image is not proper for the division.
- Step 3: Divide the image into a number of regions.
- Step 4: Apply Local Roughness Binary Pattern descriptor to calculate LBP code and new intensity for every pixel. Find the mean and Variance or Histogram for each region as a partial feature vector.
- Step 5: Prepare the feature vector by concatenating the partial feature vector obtained for each region.
- Step 6: Apply the feature vector obtained in step 5 to classification techniques.



**Fig. 3** The flow chart shows the working procedure of texture classification for an image by using the proposed descriptor



*Step 7: Repeat steps 2–5 for all training images.*

*Step 8: For the images not included in the training set, repeat steps 2–5 and test using the trained classification techniques.*

*Step 9: Analyze the results.*

Figure 3 shows the flow chart of the procedure of the texture classification for one image by using the LRBP descriptor.

## 4 Test Data Sets

To test the performance of the proposed descriptor for texture classification, we use CURET and KTHTIPS2b databases. There are 61 texture classes in the original CURET [27] database, and each class contains 205 images. We are considering only 92 images per class, similar to other CURET studies [28, 29]. These images are obtained by cropping images of original databases. Each image is an extraction of

200 × 200 pixels foreground region of texture from the original CURET image database. We are using the same subset of images that are used by [28, 29].

KTHTIPS2b database is divided into four samples (Sa, Sb, Sc, Sd). Every sample contains 11 texture classes, and in each class, 108 images are there. We perform experiments on KTHTIPS2b in two ways. One is by considering every sample as a database like Sa, Sb, Sc, and Sd so that there are 4 different databases. Another way is to consider KTHTIPS2b as a single database by merging every sample. CURET database is collected from [30] and KTHTIPS2b image database is collected from [31]. The proposed descriptor is named Local Roughness Binary Pattern (LRBP). It extracts information from the local region of an image. It first divides the image into several regions. We apply the LRBP to every pixel of the region to calculate the LBP code and then a new decimal value corresponding to pixels. Calculate the Mean and Variance or Histogram of the region after obtaining decimal values for every pixel of the region. We use the mean and Variance or Histogram as a partial feature vector for the regions. The partial feature vector of the regions is combined to form a feature vector of the image.

## 5 Experimental Results and Evaluations

This section represents the performance of our proposed descriptor in terms of accuracy (%) with the help of classification systems. Here we extract features from regions of an image by using either mean and Variance, or Histogram with the help of our proposed approach. Feature vectors are obtained after concatenation of partial feature vectors, these vectors are given as input to KNN or SVM classifier for texture classification. Outputs obtained are correctly classified class labels (%).

### 5.1 Evaluation Metric

The evaluation metric is used to evaluate the proposed descriptor and relate the theoretical and practical developments of the model. It consists of a set of measures that follow a common rule. If a system has high recall and low precision, then most of its predicted labels are incorrect when compared to the training labels. If high precision and low recall are there for a system, it returns very few results, but most of it is correctly labeled. In an ideal model, all results returned by it are accurately labelled if high precision and high recalls are present. Precision, Recall, Accuracy, and so on evaluation metrics are calculated with the help of True Positive (TP), False Positive (FP), False Negative (FN), and True Negative (TN) with the probability of classification of images. Evaluation matrices used in this research are given below.

Accuracy: The weighted percentage of correctly classified images among total tested images. Formally accuracy is defined as

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

Accuracy can also be calculated in terms of positives and negatives for binary classification from as

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

**Precision:** Precision is a fraction of correctly classified test images out of total test images for a class

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

**Recall:** Recall is a fraction of correctly classified test images for a class form total correctly classified test images

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

**F-Measure:** F-Measure is the harmonic mean of Precision and Recall

$$\text{F - Measure} = \frac{2 * (\text{Precision} * \text{Recall})}{\text{Precision} + \text{Recall}}$$

KNN and SVM classification techniques are used for texture classification in our experiments and result in analysis. We have divided the database into two parts (one is the training part and another is testing). We extracted the feature vector by using the proposed descriptor for each image. Feature vectors are given as input to classification systems and output correctly classified class labels in terms of classification accuracy (%). There is 61 texture class in the training database for the CURET database and the same classes for testing. For the KTHTIPS2b database, 11 classes for both training and testing.

## 5.2 *Experimental Results*

We have divided the database into two parts; one is training, and another is testing. Table 1 shows the results for the division of data into 80 and 20% ratios, and Table 2 shows the results for the division of data into 50 and 50% ratios. These tables show the classification accuracy (%) for both the CURET and the KTHTIPS2b databases. Results are shown for models in which image features are extracted through mean and Variance, and SVM and KNN classifiers are used.

**Table 1** Classification accuracy (%) when division of data into 80 and 20% ratio for both SVM and KNN classifier & feature extraction through mean and variance

Database	SVM classifier (%)	KNN Classifier (for K = 1 and K = 3)
CUReT	71.11	63.943% and 61.68%
KTHTIPS2b	62.35	53.78% and 54.21%

**Table 2** Classification accuracy (%) when division of data into 50 and 50% ratio for both SVM and KNN classifier & feature extraction through mean and variance

Database	SVM classifier (%)	KNN classifier (for K = 1 and K = 3)
CUReT	67.07	58.02% and 54.95%
KTHTIPS2b	59.09	51.87% and 52.68%

**Table 3** Classification accuracy (%) on KTHTIPS2b database sample wise for both SVM and KNN classifier and feature extraction through mean and Variance

Database	SVM classifier (%)	KNN classifier (for K = 1 and K = 3)
Sample A	64.86	60.09% and 56.88%
Sample B	71.52	67.87% and 68.33%
Sample C	63.47	56.89% and 56.47%
Sample D	55.14	44.50% and 41.28%

Table 3 shows the results for the KTHTIPS2b database sample-wise. For these results evolution, data is divided into 80% and 20% ratios for training and testing. Results are shown for both SVM and KNN classifier with feature extraction through mean and Variance.

### 5.3 Comparison of LRBP with LBP Variants

In this section, we compare the proposed descriptor with base paper [32] descriptors. Table 4 shows the classification Accuracy (%) of the proposed descriptor and LBP variants on the CUReT database for the KNN classifier (for k = 1) and feature extraction through Histogram. Figure 4 shows the comparison of the proposed approach with some LBP variants on the CUReT Database. The proposed descriptor is not giving the result at par with the other descriptors. When we are using Histogram to extract feature, then classification accuracy (%) for KNN classifier (for k = 1) on CUReT database is 88.1996% and 84.6881%, when split into 80–20% and 50–50%, respectively. We can see in Table 4 that when the number of bins is different during feature extraction from a region for different LBP variants, results may be different. Figure 4 represents the comparison between the proposed descriptor with some LBP variants on the CUReT Database, where the feature is extracted through Histogram.

**Table 4** Classification accuracy (%) of the proposed descriptor and LBP variants on CURET database for KNN classifier (for  $k = 1$ ) and feature extraction through Histogram

Descriptor	Neighborhood size	Bins	Accuracy (%)
LBP	$5 \times 5$	416	94.00
VZMR8 [28]	$49 \times 49$	2440	97.48
VZjoint [29]	$19 \times 19$	610	97.17
VZMRF [29]	$11 \times 11$	219, 600	98.03
CLBP [19]	$7 \times 7$	2200	97.39
NI/RD/CI (Multi-resolution scheme) [32]	$11 \times 11$	2200	97.29
Proposed descriptor	$68 \times 68$	256	88.25

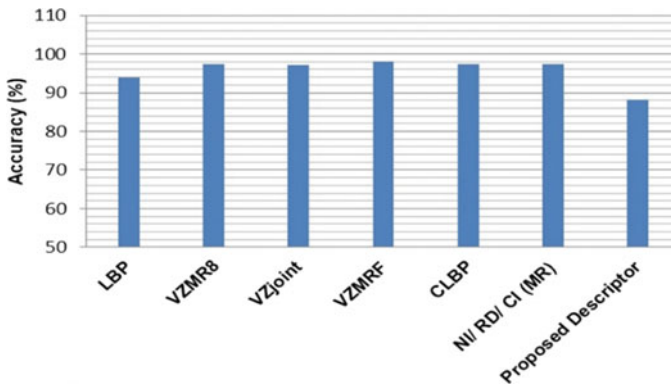
**Fig. 4** Comparing the proposed approach with some LBP variants on CURET database where the feature is extracted through histogram

Table 5 shows the classification Accuracy (%) of the proposed descriptor and LBP variants on the KTHTIPS2b database for the KNN classifier (for  $k = 1$ ) and feature extraction through Histogram. The proposed descriptor has got 80.90% and 82.53% classification accuracy when data is split into 50–50 ratio and 80–20 ratio, respectively. Figure 5 shows the comparison of the proposed approach with some LBP variants on the KTHTIPS2b Database. Results show that the proposed descriptor gives much better results than other descriptors on the KTHTIPS2b Database.

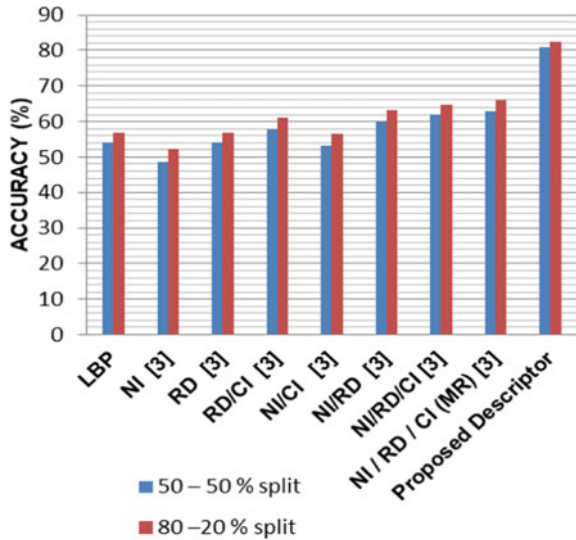
## 6 Conclusions

An image contains various types of information. Different descriptors are used to extract information from an image for applications like texture classification and face recognition. LBP is one of the most commonly used descriptors. The authors proposed many LBP variants to avoid the limitations of the basic LBP approach

**Table 5** Classification accuracy (%) of the proposed descriptor and LBP variants on KTHTIPS2b database for KNN classifier ( $k = 1$ ) and feature extraction through histogram

Descriptor	Accuracy (%) With 50–50% split	Accuracy (%) with 80–20% split
LBP	54.2	56.8
NI [32]	48.7	52.3
RD [32]	54.2	56.9
RD/CI [32]	57.8	61.2
NI/CI [32]	53.1	56.4
NI/RD [32]	60.0	63.1
NI/RD/CI [32]	61.9	64.8
NI/RD/CI (Multi-resolution scheme) [32]	62.9	66.0
Proposed descriptor	80.90	82.53

**Fig. 5** Comparing the proposed approach with some LBP variants on KTHTIPS2b Database where the feature is extracted through Histogram by using KNN classifier ( $k = 1$ )



by choosing different ways to select the threshold values and neighboring regions. Many researchers have combined the LBP approach with other approaches and use different types of image types (2D or 3D images). The proposed approach is also a variant of basic LBP, which gives more weightage to the roughness of the surface. The rough surface of the image conations more information, which helps to classify the image accurately. Experimental results show that the proposed approach with the KNN classifier has outperformed the other LBP variants on KTHTIPS2b Database.

## 7 Future Scope

In the proposed descriptor, we have used value one when there is not a regular increment or decrement in a particular direction, whether the difference is 1 or 100. In the future, we can introduce a threshold value ( $t$ ) to give value one for those whose difference is more than  $t$  when there is no regular increment or decrements. Here we are introducing  $t$  because when the difference is almost one or two, then that region behaves like a smooth region. In our proposed descriptor, neighboring pixels of a pixel are lying on the square. Many neighboring regions like a circle, ellipse, sphere (for 3D images), and rectangle are possible. If neighboring pixels lie on the perimeter of these regions, the performance of our proposed descriptor may improve. So in the future, we will select different types of regions to check the performance of the proposed descriptors. If we select different techniques to extract features, then the performance of the proposed method may vary. So in the future, we can try to check which feature extraction technique is better for our proposed descriptor in terms of accuracy.

## References

1. Ojala T, Pietikainen M, Maenpaa T (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Pattern Anal Mach Intell* 24(7):971–987
2. Ojala T, Pietikainen M, Harwood D (1996) A comparative study of texture measures with classification based on featured distributions. *Pattern Recogn* 29(1):51–59
3. Zabih R, Woodfill J (1994) Non-parametric local transforms for computing visual correspondence. In: *European conference on computer vision*. Springer, Heidelberg, pp 151–158
4. Pietikainen M (2005) Image analysis with local binary patterns scandinavian conference on image analysis: image analysis. Springer, Heidelberg, pp 115–118
5. Ahonen T, Hadid A, Pietikainen M (2004) Face recognition with local binary patterns. In: Pajdla T, Matas J (eds) *ECCV 2004, LNCS, vol 3021*. Springer, Heidelberg, pp 469–481
6. Hadid A (2008) The local binary pattern approach and its applications to face analysis. In: *IEEE workshop on image processing theory, tools and applications*. IEEE Press, New York, pp 1–9
7. Huijsmans DP, Sebe N (2003) Content-based indexing performance: size normalized precision, recall, generality evaluation. In: *IEEE international conference on image processing*. IEEE Press, New York, pp 733–736
8. Hadid A, Pietikainen M, Ahonen T (2004) A discriminative feature space for detecting and recognizing faces. In: *IEEE computer society conference on computer vision and pattern recognition*. IEEE Press, New York, pp 797–804
9. Marcel S, Rodriguez Y, Heusch G (2007) On the recent use of local binary patterns for face authentication. *Tech Rep*
10. Kellokumpu V, Zhao G, Pietikainen M (2008) Human activity recognition using a dynamic texture based method. In: *Proceedings of the British machine conference*. BMVA Press, Guildford, pp 1–10
11. Ali W, Georgsson F, Hellstrom T (2008) Visual tree detection for autonomous navigation in forest environment. In: *IEEE intelligent vehicles symposium*. IEEE Press, New York, pp 560–565

12. Nanni L, Lumini A (2008) Ensemble of multiple pedestrian representations. *IEEE Trans Intell Transp Syst* 9(2):365–369
13. Lucieer A, Stein A, Fisher P (2005) Multivariate texture-based segmentation of remotely sensed imagery for extraction of objects and their uncertainty. *Int J Remote Sens* 26(14):2917–2936
14. Huang D, Wang Y, Wang Y (2007) A robust method for near infrared face recognition based on extended local binary pattern. In: *International symposium on visual computing: advances in visual computing*. Springer, Heidelberg, pp 437–446
15. Huang Y, Wang Y, Tan T (2006) Combining statistics of geometrical and correlative features for 3d face recognition. In: *British machine vision association*, pp 879–888
16. Guo Z, Zhang L, Zhang D (2010) A completed modeling of local binary pattern operator for texture classification. *IEEE Trans Image Process* 19(6):1657–1663
17. Kumar TS, Nagarajan V (2015) Local smoothness pattern for content based image retrieval. In: *IEEE international conference on communications and signal processing*. IEEE Press, New York, pp 1190–1193
18. Ahonen T, Pietikäinen M (2007) Soft histograms for local binary patterns. In: *Finnish signal processing symposium*
19. Iakovidis DK, Keramidas EG, Maroulis D (2008) Fuzzy local binary patterns for ultrasound texture characterization. In: *International conference image analysis and recognition: image analysis and recognition*. Springer, Heidelberg, pp 750–759
20. Guo Z, Zhang L, Zhang D, Zhang S (2010) Rotation invariant texture classification using adaptive LBP with directional statistical features. In: *17th IEEE international conference on image processing*. IEEE Press, New York, pp 285–288
21. Liao S, Law MW, Chung AC (2009) Dominant local binary patterns for texture classification. *IEEE Trans Image Process* 18(5):1107–1118
22. Hafiane A, Seetharaman G, Zavidovique B (2007) Median binary pattern for textures classification. In: *International conference image analysis and recognition*. Springer, Heidelberg, pp. 387–398
23. Huang D, Zhang G, Ardabilian M, Wang Y, Chen L (2010) 3D face recognition using distinctiveness enhanced facial representations and local feature hybrid matching. In: *Fourth IEEE international conference on biometrics: theory applications and systems*. IEEE Press, New York, pp 1–7
24. Ahonen T, Matas J, He C, Pietikäinen M (2009) Rotation invariant image description with local binary pattern histogram fourier features. In: *Scandinavian conference on image analysis*. Springer, Heidelberg, pp 61–70
25. Zhang L, Zhang L, Guo Z, Zhang D (2010) Monogenic-LBP: a new approach for rotation invariant texture classification. In: *17th IEEE international conference on image processing*. IEEE Press, New York, pp 2677–2680
26. Guo Z, Zhang L, Zhang D (2010) Rotation invariant texture classification using LBP variance (LBPV) with global matching. *Pattern Recogn* 43(3):706–719
27. Dana KJ, Ginneken BV, Nayar SK, Koenderink JJ (1999) Reflectance and texture of real-world surfaces. *ACM Trans Grap* 18(1):1–34
28. Varma M, Zisserman A (2005) A statistical approach to texture classification from single images. *Int J Comput Vis* 62(1):61–81
29. Varma M, Zisserman A (2009) A statistical approach to material classification using image patch exemplars. *IEEE Trans Pattern Anal Mach Intell* 31(11):2032–2047
30. Liu L, Fieguth P (2010) Texture classification using compressed sensing. In: *IEEE Canadian conference on computer and robot vision*. IEEE Press, New York, pp 71–78
31. Mallikarjuna P, Fritz M, Targhi AT, Hayman E, Caputo B, Eklundh J (2006) The *K*th-tips and *K*th-Tips2 databases
32. Liu L, Zhao L, Long Y, Kuang G, Fieguth P (2012) Extended local binary patterns for texture classification. *Image Vis Comput* 30(2):86–99



# Speech Enhancement Using VAD for Noise Estimation in Compressive Sensing



Vasundhara Shukla and Preety D. Swami

**Abstract** This paper presents a novel approach to speech enhancement that is based on compressed sensing. Voice activity detection is performed to identify the speech and non-speech (pauses) frames. The non-speech frame is then used for noise estimation. Orthogonal matching pursuit is employed to obtain improved speech by sparse recovery. During pauses, the noise is measured and used to update the orthogonal matching pursuit termination threshold. The proposed technique is tested for white noise. The comparison with a few state of art algorithms is carried out in terms of segmental signal-to-noise ratio and perceptual evaluation of speech quality. The outcomes show the benefits of the proposed approach.

**Keywords** Compressed sensing · Orthogonal matching pursuit · Speech enhancement · Voice activity detection

## 1 Introduction

Compressive sensing (CS) is a new field of fast-growing research proposed in 2006 by Candes and Romberg [1], Donoho [2], and Tao [3]. It suggests that by discovering solutions to underdetermined linear systems, a signal can be reconstructed with less samples than typical Nyquist-based approaches need.

The Shannon theorem is acceptable for most traditional applications, however, in cases where the required sampling rate is too high, the use of CS offers a promising solution for measurements that could otherwise be very costly or sluggish. Compressive sampling has been recognized in different kinds of signal processing applications as one of the most effective, simple, and attractive methods. CS applications can generally be classified into five broad categories: compression imaging, applications

---

V. Shukla (✉)

Department of Electronics and Communication Engineering, Oriental College of Technology, RGPV, Bhopal, India

e-mail: [shuklav2015@gmail.com](mailto:shuklav2015@gmail.com); [vasundharashukla@rgtu.net](mailto:vasundharashukla@rgtu.net)

P. D. Swami

Department of Electronics and Communication Engineering, RGPV, Bhopal, India

in the field of biology [4, 5], compression RADAR [6], analog-to-digital converters [7], and communication and networks [8–11].

In applications such as cameras [12–14], Seismic Imaging [15], and Medical Imaging [16, 17], compression imaging itself can be defined as CS. In addition, compression sampling has recently become a growing topic in speech processing. In a variety of applications [18–21], compressive sampling has been developed in speech signal processing. In these implementations, instead of using compressive sensing directly for the speech signal, the residual excitation signal is used. In this way, the measurement matrix is generated by considering the impulse response (or random matrix) and then the recovery of the signal is done by using an optimization technique such as  $l_1$  minimization or the Basis Pursuit.

This paper is organized as follows. Section 2 provides theoretical background. Section 3 is dedicated to the description of the methods used in this paper. In Sect. 4, the evaluation measures followed by simulation results for the proposed technique are provided. Also, a comparison among different methods is tabulated in this section. Finally, the conclusion of the paper is presented in Sect. 5.

## 2 Theoretical Background

### 2.1 Amplitude Normalization and DC Component Elimination

Since there are many factors (such as microphones, amplifiers, digitizers, etc.) influencing the speech signal measurements. If there is no reference value that can be compared, it is difficult to describe the voltage obtained from such a measuring system. Hence, in the absence of a normalization process, the analysis of the raw speech signal amplitude is difficult. Normalization translates the signal to a known value relative to a scale. In the proposed approach, the following normalization procedure is utilized:

$$S_{\text{norm}} = 2 \times \left( \frac{S - \min(S)}{\max(S) - \min(S)} \right) - 1 \quad (1)$$

where  $S$  is the set of the sampled signal,  $S_{\text{norm}}$  is the normalized signal,  $\max$ , and  $\min$  are the functions that return the maximum and minimum sample amplitude in the set  $S$ . Normalization by Eq. (1) scales the signal in the interval  $[-1, 1]$ .

## 2.2 Discrete Cosine Transform (DCT) and Inverse DCT (IDCT)

As with Discrete Fourier Transform (DFT), DCT also transforms the signal into the frequency domain. However, the DCT uses only real-valued cosine functions in comparison with DFT and thus has reduced computational complexity. Another property that makes the DCT preferable over DFT is its high degree of “spectral compaction”. Because of this, the DCT transformation of a signal contains more concentrated energy in very few coefficients compared to other transformations such as the DFT [5].

This is desirable for sparse representation of a signal where it is required to represent a signal with few nonzero samples. If DCT is applied to a speech signal only a few DCT coefficients will hold significant amounts of energy, hence all other coefficients can be set to zero without loss of signal information. This will be a sparse representation of the speech signal.

Four versions of DCT transform, DCT-I to DCT-IV, are available. Variations in these versions are quite small. Among all the four versions, the DCT-II, also known as the even symmetric DCT, or simply as “the DCT”, is the most popular and hence is used in this work.

The general equation for calculating the DCT of a one-dimensional ( $N$  sample length) signal  $x$  is

$$F[k] = \begin{cases} \sum_{n=0}^{N-1} 2x[n] \cos\left[\frac{\pi}{2N}k(2n+1)\right], & 0 \leq k < N \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

The inverse of DCT can be calculated as

$$x[n] = \begin{cases} \frac{1}{N} \sum_{k=0}^{N-1} \Lambda[k] F[k] \cos\left[\frac{\pi}{2N}k(2n+1)\right], & 0 \leq n < N \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where  $\Lambda[k]$  is defined as

$$\Lambda[k] = \begin{cases} \frac{1}{2}, & k = 0 \\ 1 & 1 \leq k \leq N \end{cases} \quad (4)$$

### 2.3 Framing and De-Framing

Framing is a fundamental signal processing technique that consists of dividing the original signal into small blocks of samples often called frames. The frames are created in such a way that they overlap with the consecutive frames. Overlapping the frames is performed to avoid information loss in between adjacent frames. Three frames  $F_a, F_b, F_c$  and two overlapped frames  $OL_{ab}, OL_{bc}$  are shown in Fig. 1.

In the case of stationary signals (signals whose statistical characteristics do not vary with time), traditional spectral evaluation methods are accurate [22].

Since speech is a non-stationary signal, its statistical properties do not remain constant over time. As a result, spectral features and other distinguishing characteristics (such as short-time energy, Mel Frequency Cepstral Coefficients (MFCC), to name a few) should be extracted from small blocks of the signal. This assumes that the signal in this small frame is stationary (that is, its statistical properties are constant within this region). Furthermore, frame blocking is often used in real-time systems because it maximizes the system’s performance by spreading the fixed process overhead over several samples. Also, in the proposed work, compressive sensing is used which is a block-based processing algorithm, and processing a large block of the signal at once will require a very large amount of memory.

Dividing the signal into blocks results in spectral distortion. To minimize spectral distortion, multiplication of each frame with a Hamming window is done. Hamming window is expressed as

$$w(n) = 0.54 - 0.64\cos\left(\frac{2\pi n}{N - 1}\right), 0 \leq n \leq N - 1 \tag{5}$$

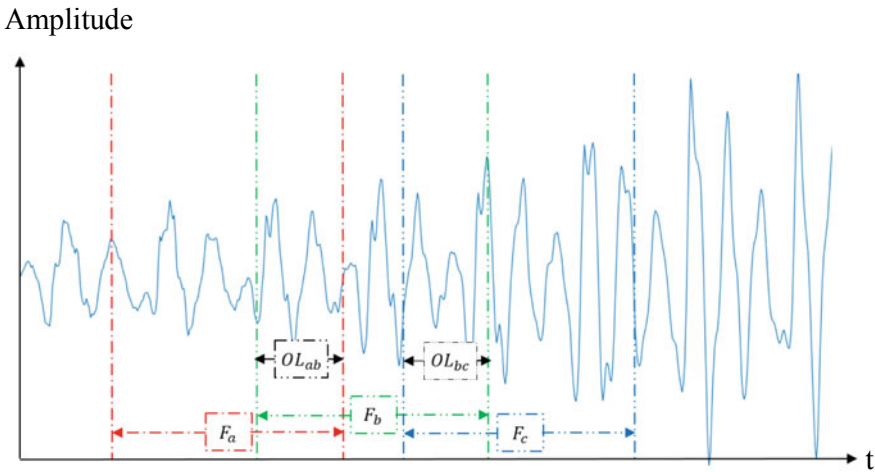


Fig. 1 Framing process and frame overlapping

Where the duration of the frame is given by  $N$ . After windowing, the output  $y(n)$  of the signal is

$$y(n) = x(n) \times w(n) \tag{6}$$

This windowing results in enhancement of the signal at the center, and because of the tapering nature of the window, smoothens the signal at the edges.

### 2.4 Voice Activity Detector

The problem of distinguishing a target speech sound from non-speech frames is known as voice activity detection (VAD). Robust detection of speech in noisy audio signals is an essential and fundamental pre-processing phase in various speech processing modules and applications, such as noise reduction algorithms, language identification, speaker recognition, speech coding, and automated speech recognition [23].

In the proposed work for performing VAD, the process used is shown in Fig. 2 and the various blocks are explained as follows [23].

**Spectral Shape:** Human speech is a pressure waveform produced by the tongue, velum, jaw, and lips, which control the shape of the vocal tract cavity and produce specific speech sound segments known as phones. The envelope of the short-time

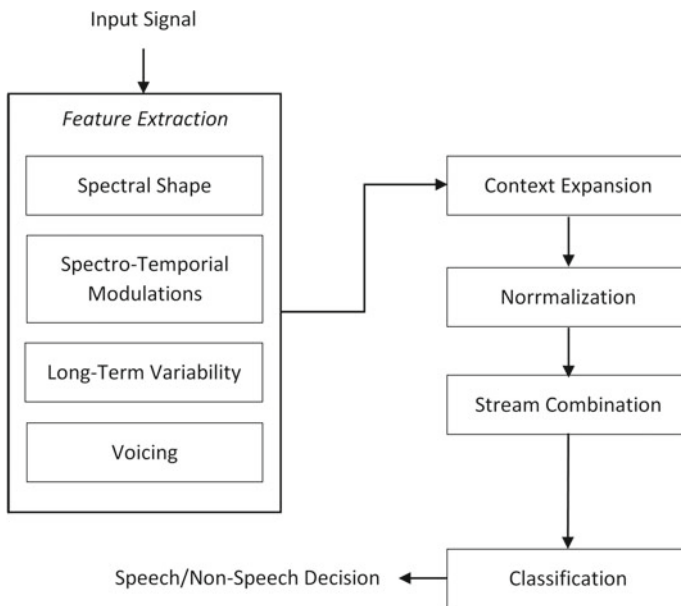


Fig. 2 Overview of steps used for VAD [23]

power spectrum of signals reflects the vocal tract shape at a given time. Because of their robustness [23], the Gammatone Frequency Cepstral Coefficients (GFCC) feature is used in the VAD.

**Spectro-temporal modulations:** Subsequent articulation of different phonetic sounds are caused by changes in the vocal tract shape, which are reflected by temporal variations in the spectral energy of speech. A wavelet-based approach with a time window size of 100 ms is used in VAD.

**Voicing:** The vocal folds display quasi-periodic vibrations of speech. Thus, speech is characterized by strong periodicity during voiced speech periods which is expressed in the spectrum by the presence of a fundamental frequency or pitch and its harmonics. VAD robustness may be improved by features that specifically measure the voicing state of speech.

**Long-term variability:** The fourth stream of data aims to model speech variability over a long period of time. Phones in various languages have durations ranging from 10 to 200 ms and are uttered at a rate of three to seven syllables per second during a normal conversation, with formant features that change over time.

**Context expansion, normalization, and stream combination:** The time period over which we integrate the information from the probability measures is referred to as the context. The stream vectors are then variance normalized using a global variance vector computed on the training data, which yields slightly better results than per-file normalization. Finally, using a hidden layer that has a size equal to the feature dimension, an (Multi-layer Perceptron) MLP classifier trains on these feature vectors. The ratio of all MLP outputs when thresholded detects the speech segments.

## 2.5 Compressive Sensing

Compressive sensing is also termed as compressed sampling, compressive sensing, or sparse recovery. The traditional way of the reconstruction of sampled signal imposes two basic limitations:

- The sampling rate must be twice or greater than the signal bandwidth which is also known as the Shannon theorem.
- According to linear algebra, the number of collected data samples should be at least as long as its length.

Compressive sensing overcomes these common limitations. It can recover certain signals from incomplete samples. In terms of a suitable basis, these certain types of signals can be approximated by a sparse expansion.

Compression is obtained by simply starting at only the largest basis coefficients. The CS strategy is good when full signal information is easily available. The concept of CS is to obtain the compressed signal more directly during sampling. CS provides a way to reconstruct a compressed version of the original signal by acquiring a small amount of non-adaptive and linear measurements. The basic equation for CS Sparse

signal recovery is as follows:

$$[y]_{M \times 1} = [\phi]_{M \times N} \cdot [x]_{N \times 1} + [v]_{M \times 1} \tag{7}$$

where  $y$  is the measurement matrix,  $\phi$  is the recovery matrix,  $x$  is the sparse signal, and  $v$  is the noise. Here the important considerations are:

1.  $M \ll N$ , hence infinite many solutions are possible.
2.  $x$  is a  $k$  sparse signal and  $k \ll N$ .
3. All rows in matrix  $\phi$  are linearly independent or it's a full rank matrix.

and the goal is to recover  $x$  from  $y$ .

The solution of Eq. (7) is not straightforward as the estimation of  $x$  requires the inverse of  $\phi$ ,

$$\phi^{-1} = \frac{1}{|\phi|} \cdot \text{adj}(\phi) \tag{8}$$

Which in this case is not possible as  $\phi$  is not a square matrix and also  $\phi$  must be non-singular.

To solve the above problem, we need to calculate the pseudo-inverse of the matrix  $\phi$ .

The pseudo-inverse can be calculated using singular value decomposition (SVD). Factorization of a real or complex matrix, in linear algebra, is defined as the SVD. Given a positive semi-definite normal matrix, SVD is the generalization of its Eigen decomposition.

The SVD decomposes a matrix into three other matrices as follows:

$$\phi = USV^T \tag{9}$$

where  $\phi$  is a  $M \times N$  matrix,  $U$  is an  $M \times M$  orthogonal matrix,  $S$  is a  $N \times N$  diagonal matrix, and  $V$  is a  $N \times N$  orthogonal matrix. For  $M \geq N$

$$a_{ij} = \sum_{k=1}^N u_{ik} \cdot s_k \cdot v_{jk} \tag{10}$$

The variable  $s_k$  is called singular values and are normally sorted in the form of largest to smallest  $s_{i+1} \leq s_i$ . The left and right singular vectors are the columns of  $U$  and  $V$ , respectively. Since  $U$  and  $V$  are orthogonal  $UU^T = VV^T = I$ . Since  $U$  is not square, hence we cannot say that  $UU^T = I$ , so it's orthogonal in one direction.

The pseudo-inverse from SVD of the non-square and singular matrix can be given as

$$\phi^{-1} = (\phi^T \phi)^{-1} \phi^T \tag{11}$$

The problem of  $\phi$ , non-being square is solved by  $\phi^T \phi$ . The problem of  $\phi$ , being singular is solved as  $\phi$  is not in the denominator.

Hence, in the original CS problem, the value of  $x$  from the matrix  $y$  can be obtained as follows:

$$x = \phi(\phi^T \phi)^{-1} y \quad (12)$$

Since Eq. (9) may have an infinite number of solutions, we have to find the  $\phi$ , with the minimum number of nonzero entries, this is called the  $l_0$  norm (counts simply how many nonzero entries in a vector).

Let we have known values of  $x$  and  $y$

$$\min \|x\|_0, \text{ subject to } y = \phi x \quad (13)$$

where  $\min \|x\|_0$  is the  $l_0$  norm. In partial reconstruction case

$$\min \|x\|_0, \text{ subject to } \|y - \phi x\|_2 \leq \beta \quad (14)$$

where  $\|y - \phi x\|_2$  is the  $l_2$  norm, minimization of error  $= \sqrt{z_1^2 + z_2^2 + \dots + z_n^2}$ , and  $\beta$  is the termination threshold.

## 2.6 Orthogonal Matching Pursuit

The CS Eqs. (13) and (14) can be solved by finding  $x$  with only a few nonzero elements that satisfy the given condition. This problem is NP-hard (nondeterministic polynomial time) and hence requires too much time. However, there is another way that can be used to find  $x$ , known as the Orthogonal Matching Pursuit (OMP) algorithm. To use this, the matrix  $\phi$  and vector  $x$  must satisfy the following conditions:

1. The  $x$  must be a  $k$ -sparse signal, which means the number of nonzero entries in  $x$  must be less than or equal to  $k$ . Mathematically, it can be given as  $x \in \mathbb{R}^m$ , then  $|\text{supp}(x)| \leq k$ , where  $\text{supp}(\cdot)$  defines the support of  $x$ , which contains all the indices  $i$  for which  $x_i \neq 0$ , or  $\text{supp}(x) = \{i : x_i \neq 0\}$ .
2. The mutual coherence of matrix  $\phi$  must satisfy the inequality  $M_\phi < \left(\frac{1}{2k-1}\right)$ , where  $M_\phi$  is the mutual coherence between the column vectors of  $\phi$  and  $k$  is the sparsity of  $x$ . The  $M_\phi$  is the largest absolute value of the normalized correlation between the column vectors of  $\phi$  and is calculated as follows:

$$M_\phi = \max_{i \neq j} \frac{|\phi_i \cdot \phi_j|}{\|\phi_i\|_2 \|\phi_j\|_2} \quad (15)$$



where  $\phi_i$  represents the  $i$ th column of  $\phi$  matrix.

OMP is an iterative greedy algorithm as it tries to find the solution for  $x$  element by element in a step-by-step manner and the problem is solved optimally at each stage. The OMP algorithm requires the following steps:

**Inputs:**

A  $k$  – sparse signal vector  $x \in \mathbb{R}^N$ ,

Mutual coherence of matrix  $\phi \in \mathbb{R}^{M \times N}$

Acceptable loss  $\beta = 0.1$ ;

**Initialization:**

Iteration counter  $\text{iter} = 0$ ;

Calculate  $y = \phi x$ ;

Residue  $\gamma_0 = y$ , presents the remaining (unrecovered) part of the  $x$ ;

Dominant Columns Indices  $C_{\text{iter}} = \emptyset$ , stores the column number having the highest correlation with  $\gamma_{\text{iter}}$ ;

Dominant Columns Contribution  $\lambda_{\text{iter}}$ , stores the correlation values of columns in  $C_{\text{iter}}$ ;

$\phi_{\text{iter}}^{\text{selected}} = \emptyset$ , contains all the columns whose indices are listed in  $C_{\text{iter}}$

**Normalization:**

Firstly the normalization of each column of matrix  $\phi$  is performed, this step is necessary to make sure that the dot product between any two columns remains within the interval  $[1, -1]$ .

Normalization is done as follows:

$$\phi_i^{\text{norm}} = \frac{\phi_i}{\|\phi_i\|}, \forall i \in \{1, 2, \dots, N\}$$

**Main Loop:**

While ( $\text{iter} < M \& \& \gamma_{\text{iter}} > \beta$ )

/\*finding the column having the highest correlation with residue.\*/

$$c = \underset{i}{\text{argmax}} |\gamma_{\text{iter}} \cdot \phi_i^{\text{norm}}|$$

/\*updating the dominating column indices matrix\*/

$$C_{\text{iter}} = C_{\text{iter}-1} \cup \{c\}$$

/\*Projection Magnitude Calculation\*/

/\*Superscript “+” indicates the pseudo-inverse.\*/

$$\gamma_{\text{tmp}} = \left( C_{\text{iter}} \cdot (C_{\text{iter}}^T \cdot C_{\text{iter}})^+ \cdot C_{\text{iter}}^T \right) \cdot y$$

/\*updating the residue.\*/

$$\gamma_{\text{iter}} = y - \gamma_{\text{tmp}}$$

/\*calculating the selected dominating column contribution\*/

$$\lambda_{\text{tmp}} = |\gamma_{\text{iter}} \cdot \phi_c|$$

/\*updating the dominant columns contribution matrix\*/

$$\lambda_{\text{iter}} = \lambda_{\text{iter}-1} \cup \{\lambda_{\text{tmp}}\}$$

/\*updating the reduced  $\phi$  matrix\*/

$$\phi_{\text{iter}}^{\text{selected}} = \phi_{\text{iter}-1}^{\text{selected}} \cup \{\phi_c\}$$

```

/*updating the iteration counter*/
iter = iter + 1
End While
 $\phi_{\text{iter}}^{\text{reduced}}(i, j) = (\phi_{\text{iter}}^{\text{selected}}(:, i))^T \cdot \phi_{\text{iter}}^{\text{selected}}(:, j), \forall i, j \in \{1, 2, \dots, N\}$ 
 $y_{\text{reduced}} = y^T \cdot \phi_{\text{iter}}^{\text{selected}}$ 
 $x_{\text{est}}^{\text{reduced}} = (\phi_{\text{iter}}^{\text{reduced}})^+ \cdot y_{\text{reduced}}$ 
 $x_{\text{est}}(C_{\text{iter}}(i)) = x_{\text{est}}^{\text{reduced}}(i), \forall i \in \{1, 2, \dots, \text{card}(C_{\text{iter}})\}$ 

```

**Output:**

The reduced  $\phi_{\text{iter}}^{\text{reduced}}$  matrix

Dominant Columns Contribution  $\lambda_{\text{iter}}$

Dominant Columns Indices  $C_{\text{iter}}$

### 3 Proposed Technique

Figure 3 introduces the strategy for enhancing speech by the proposed method. This method is divided into three stages:

1. Framing and transforming the frames into the frequency domain through DCT and zeroing the insignificant components to make the signal sparse. Speech framing is carried out using overlapping Hamming windows.
2. VAD is used to identify the speech and non-speech frames and estimate the OMP termination threshold.
3. Applying CS recovery to each frame, converting frames into the time domain through IDCT and de-framing.

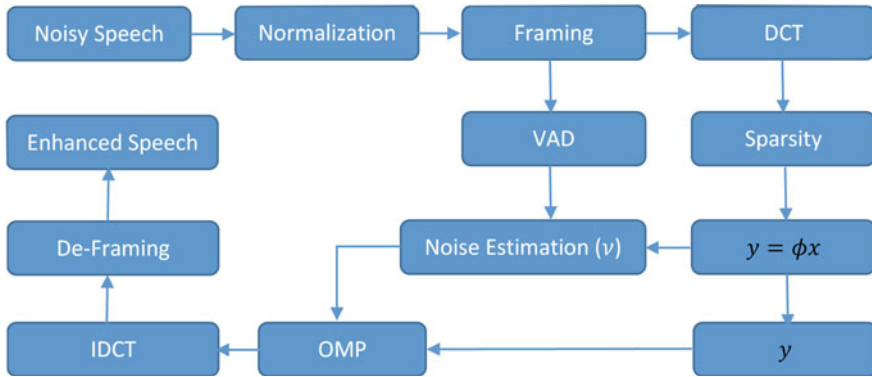


Fig. 3 Block diagram of the proposed algorithm

DCT for each frame  $f$  is computed. Next, VAD is performed, and depending on the VAD output, a logical signal (0 or 1) is generated. The signal frame is considered as speech if the  $VAD(f) = 1$ . Noise analysis can be done from the non-speech frames. The frames corresponding to non-speech are commonly regarded as silence; thus they contain only noise. These frames are used to determine/update the noise  $\nu$  by averaging the variance of all such frames. The estimated noise is used as termination criteria for orthogonal matching pursuit and finally, the improved speech is obtained by CS recovery. IDCT is used to return to the time domain and enhanced speech is obtained after de-framing the processed frames.

### 4 Performance Evaluation Metrics and Result Analysis

Three objective measures are utilized to evaluate the performance of the proposed speech enhancement algorithm based on compressive sensing: Signal to Noise ratio (SNR), the segmental SNR (SNRseg), and the perceptual evaluation of speech quality (PESQ). The mean of SNRs of each segment/frame is defined as the segmental SNR [8]. The PESQ provides an objective measure of the quality of speech that actually measures its goodness [24]. It provides an estimation of the mean opinion score (MOS). The PESQ score ranges from  $-0.5$  (bad) to  $4.5$  (excellent).

The performance comparison is shown in Table 1. White Gaussian noise is added in the SNR range from  $-2$  to  $5$  dB. Results of the proposed algorithm are compared with that of OMP and CoSaMP [25] enhancement methods. From these results, it can be noted that better performance in terms of SNRseg is achieved using the proposed VAD-based CS speech enhancement method. In terms of SNR, CoSaMP provides better results than the proposed method at lower SNRs, but at higher SNRs, the proposed method provides better results. SSNR performance is best for most of the values of the proposed method except for the case when SNR is  $3$  dB where OMP

**Table 1** Performance comparison

Measure	Method	Additive white gaussian noise (dB)							
		-2	-1	0	1	2	3	4	5
SNR (dB)	Noisy signal	-2	-1	0	1	2	3	4	5
	OMP	3.34	3.98	4.74	6.47	7.10	7.87	8.35	9.28
	CoSaMP	6.93	7.07	7.68	7.41	7.70	8.74	8.80	8.81
	Proposed	6.11	6.82	7.61	8.20	8.64	8.79	9.85	10.18
SSNR (dB)	Noisy Signal	0.86	1.13	1.41	1.68	1.95	2.27	2.59	2.95
	OMP	2.69	3.05	3.19	3.75	4.18	4.44	4.75	5.20
	CoSaMP	2.57	2.67	3.02	2.91	3.08	3.36	3.89	4.14
	Proposed	2.93	3.25	3.73	4.12	4.32	4.37	5.13	5.45
PESQ	Noisy signal	1.47	1.53	1.58	1.65	1.72	1.79	1.87	1.94
	OMP	1.88	1.96	2.07	2.11	2.14	2.15	2.23	2.32
	CoSaMP	1.37	1.41	1.47	1.31	1.34	1.57	1.45	1.53
	Proposed	1.90	1.94	2.10	2.03	2.14	2.17	2.12	2.32

shows a slight improvement. PESQ parameter shows better results for most of the noisy conditions but for a few values OMP performs better.

## 5 Conclusion

This paper presents a speech enhancement technique that is based on compressive sensing of speech signals. Speech and non-speech (silence) frames are separated by using a voice activity detector. Noise is measured from the silence frames of the speech and is used to dynamically estimate the OMP termination threshold. The performance of the proposed speech enhancement algorithm is evaluated for three different evaluation parameters and compared with two state-of-the-art techniques. The experimental results show that the proposed algorithm performs better than OMP and CoSaMP algorithms and provides much more stable performance throughout the noise variations. In terms of the PESQ parameter, OMP shows better results for some of the noisy conditions, but for most of the values, the proposed method performs better. In terms of SNR, CoSaMP provides better results than the proposed method at lower SNRs, but at higher SNRs, the proposed method provides better results. SSNR for the proposed method is better for almost all noise values.

## References

1. Candes EJ, Romberg J (2006) Quantitative robust uncertainty principles and optimally sparse

- decompositions. *Found Comput Math* 6:227–254
2. Donoho DL (2006) Compressed sensing. *IEEE Trans Inform Theory* 52:1289–1306
  3. Candes E, Tao T (2006) Near-optimal signal recovery from random projections and universal encoding strategies. *IEEE Trans Inform Theory* 52:5406–5425
  4. Mohtashemi M, Smith H, Walburger D, Sutton F, Diggans J (2010) Sparse sensing DNA microarray-based biosensor: is it feasible? In: *Sensors applications symposium (SAS), 2010*, IEEE. IEEE, pp 127–130
  5. Parvaresh F, Vikalo H, Misra S, Hassibi B (2008) Recovering sparse signals using sparse measurement matrices in compressed DNA microarrays. *IEEE J Sel Top Sign Proces* 2:275–285
  6. Ahmad F (2013) Through-the-wall human motion indication using sparsity-driven change detection. *IEEE Trans Geosci Remote Sens*, 881–890 (IEEE)
  7. Mishali M, Eldar Y (2009) From theory to practice: sub-nyquist sampling of sparse wideband analog signals. *IEEE Sel Top Signal Process*
  8. Bajwa W, Haupt J, Sayeed A, Nowak R (2010) Compressed channel sensing: a new approach to estimating sparse multipath channels. *Proc IEEE* 98(6):1058–1076
  9. Wang Y, Pandharipande A, Polo Y, Leus G (2009) Distributed compressive wide-band spectrum sensing. *IEEE Proc Inf Theory Appl*, 1–4
  10. Zhang P, Hu Z, Qiu RC, Sadler BM (2009) A compressed sensing based ultra-wideband communication system. In: *IEEE international conference on communications, 2009. ICC'09*, pp 1–5
  11. Fu S, Kuai X, Zheng R, Yang G, Hou Z (2010) Compressive sensing approach based mapping and localization for mobile robot in an indoor wireless sensor network. In: *2010 international conference on networking, sensing and control (ICNSC)*. IEEE, pp 122–127
  12. Duarte M, Davenport M, Takhar D, Laska J, Sun T, Kelly K, Baraniuk R (2008) Single-pixel imaging via compressive sampling. *IEEE Signal Process Mag* 25:83–91
  13. Nagesh P et al (2009) Compressive imaging of color images. In: *Proceedings of the 2009 IEEE international conference on acoustics, speech and signal processing*, 00:1261–1264 (IEEE Computer Society)
  14. Chan W, Charan K, Takhar D, Kelly K, Baraniuk R, Mittlefán D (2008) A single-pixel terahertz imaging system based on compressed sensing. *Appl Phys Lett* 93:121105
  15. Hennenfent G, Herrmann FJ (2008) Simply denoise: wavefield reconstruction via jittered undersampling. *Geophysics*, 19–28
  16. Lustig M, Donoho D, Santos J, Pauly J (2008) Compressed sensing MRI. *IEEE Signal Process Mag* 25:72–82
  17. Gamper U, Boesiger P, Kozerke S (2008) Compressed sensing in dynamic MRI. *Mag Reson Med* 59(2):365–373
  18. Sreenivas TV, Kleijn WB (2009) Compressive sensing for sparsely excited speech signals. In: *IEEE international conference on acoustics, speech and signal processing*, Taipei, pp 4125–4128
  19. Giacobello D, Christensen MG, Murthi MN, Jensen SH, Moonen M (2010) Retrieving sparse patterns using a compressed sensing framework: applications to speech coding based on sparse linear prediction. *IEEE Signal Proc Lett* 17:103–106
  20. Ramadan MH (2005) Compressive sampling of speech signal, Sebha. Sebha University, Libya
  21. Sabir A (2011) Compressive sensing for speech signals in mobile systems. Texas University
  22. Hamid OK (2018) Frame blocking and windowing speech signal. *J Inf Commun Intell Syst (JICIS)* 4:87–94
  23. Van Segbroeck M, Tsiartas A, Narayanan SS (2013) A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice. In: *INTERSPEECH*, pp 704–708
  24. Swami PD, Sharma R, Jain A, Swami DK (2015) Speech enhancement by noise driven adaptation of perceptual scales and thresholds of continuous wavelet transform coefficients. *Speech Commun* 70:1–12
  25. Needell D, Tropp JA (2009) CoSaMP: iterative signal recovery from incomplete and inaccurate samples. *Appl Comput Harmon Anal* 26:301–321

# A Comparative Study on Single Image De-Raining Using Convolutional Neural Network



Poornima Shrivastava, Roopam Gupta, and Asmita A. Moghe

**Abstract** Most of the algorithms for computer vision require a clear image for its processing. For finding the proper solution to the visual degradation of the image due to streaks caused by rain, an effective de-raining algorithm needs to be developed. De-raining is the task of removing streaks of rain and their effect from the distorted image that contains rain. The huge success of methods based on learning and that too CNN-based methods in different domains have prompted its use in de-raining. For effectively obtaining the automatic feature information, deep CNN methods are used. The efficiency of various methods of de-raining is applied and tested on datasets such as Rain1200, Rain100H, Rain100L, Rain1400, and many more that are discussed and analyzed here. The quantitative metrics for the comparison are PSNR and SSIM.

**Keywords** Single image de-raining · Rain streaks removal · Deep learning · Convolutional neural network (CNN) and image restoration

## 1 Introduction

Rain is a natural phenomenon that causes blurring and unclarity in images. The main reason behind this blur effect is the light that gets reflected and scattered due to rain streaks after falling on the image. Due to complex and changeable weather conditions at the time of rain, it sometimes led to an unclear background with some sort of haze and uneven rain streaks. This requires the image to be initially preprocessed before being used in the applications of artificial intelligence like tracking of objects, semantic segmentation, detection of objects, and some of the other high-level visions.

---

P. Shrivastava (✉)

Department of Electronics and Communication, UIT, RGPV, Bhopal, India

e-mail: [poonamshrivastava62@gmail.com](mailto:poonamshrivastava62@gmail.com)

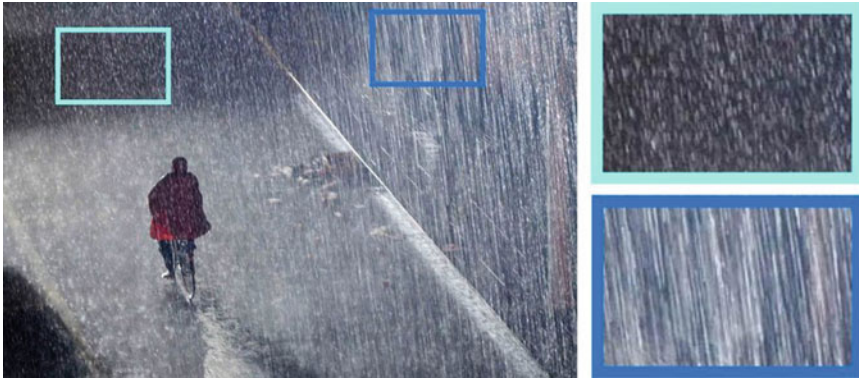
R. Gupta (✉) · A. A. Moghe

Department of Information Technology, UIT, RGPV, Bhopal, India

e-mail: [roopamgupta@rgtu.net](mailto:roopamgupta@rgtu.net)

A. A. Moghe

e-mail: [Aamoghe@rgtu.net](mailto:Aamoghe@rgtu.net)



**Fig. 1** Image shows rain model at same time frame different sized rain streaks. Image courtesy [28]

De-raining is one of the preprocessing steps in computer vision tasks. There are several de-raining methods proposed so far and they are classified into methods based on a single image, multiple images, and video processing. Video-based methods are said to be less complex to de-rain as compared to methods based on single images because, in this case, to remove streaks of rain as shown in Fig. 1 and to identify rain areas between adjacent frames, inter-frame information is used, whereas in single image de-raining, there is no inter-frame information between adjacent frames, and the information available is also less [1].

As shown in Fig. 1, rain models show distinct rain types like raindrops, rain streaks, and rain with haze. All these need a different type of specialized model. At present, models are promising for their particular rain type. Some models are prior-based, some are data-driven, and others are a combination of data-driven with prior-based methods.

So, the removal of rain streaks is a more challenging problem with a single image [1]. Most of the methods on the single image are based on the following Eq. (1) and shown in Fig. 2:

$$R = B + S \quad (1)$$

In which a rainy image is denoted by  $R$ , a clear image is denoted by  $B$ , and rainy streaks are denoted by Wang et al. [2] Thus the restoration work of the original image calls for the removal of rain streaks from the rainy image and restores the original image  $B$  as given in Eq. (2).

$$B = R - S \quad (2)$$

As given in Eq. (2) the models of de-raining networks designed by various works assumed that rain streak exists more in the high-frequency part. So de-raining of images is done by decomposition of rain image into two parts, i.e., a detailed layer

and base layer, further determining the difference between these detailed layers with and without the streaks of rain.

In the detail layer, most of the decomposition of rain is done but there are remains of rain streaks in the base layer and this limits the performance of de-raining. So, in [1], a clean image is restored from the rainy image. Suppose  $H(\theta)$  denotes the de-raining model,  $\theta$  denotes the model's input, then the de-rained output is directly the clear image. This is represented in Eq. (3)

$$\hat{S} = H(R) \tag{3}$$

for the clean image  $B$  where  $B$  is the de-raining model's prediction and for the rain streaks,  $S$  is the prediction.

From the model, the rain streak's predictions are learned from the rainy image and the predicted rain streaks is subtracted to estimate the clean image as shown in Eq. (4).

$$\hat{B} = R - H(R) \tag{4}$$

In the air, the raindrops have various appearances that complicate the case of real weather conditions. In heavy rain conditions when the streaks of rain get accumulated in the atmosphere, it leads to an increase in the diversity of rain streaks, and thus affects brightness, scattering of light, and attenuation. A cleaner image is much more complex than the layer of rain streak. So, the simpler part is to learn the rain streak layer as compared to directly learning the clear image from the rain image [1].

In the matter of single image rain removal, the most ill-posed problem is that only a single type of rainy image is available. To solve this type of problem efficiently, different algorithms make use of prior knowledge for obtaining a clear image by removing streaks of rain. For this, some of the prior-based methods like sparse representation, low-rank prior and Gaussian mixture model, etc., were used. Though the de-raining performance of this method is good, it cannot be used in some of the cases [2].

## 2 Work Done on Prior-Based Methods on Single Image De-Raining

Rain removal methods for the single image can further be categorized into methods that are based upon priors and deep learning. The research work for de-raining initially started from prior-based methods. Some of the most widely known methods are the Gaussian mixture model [3], low-rank model [4], sparse codes [5], non-local filter [6], and image decomposition [7]. Kang et al. [7] proposed a new method for rain



removal in a single type of image, which is based on morphological component analysis which is an image decomposition problem. This was done by properly formulating rain removal in which the image is separated into parts of high-frequency and low-frequency with the help of a bilateral filter. Further sparse coding and dictionary-based learning are used for the decomposition of the high-frequency region into a non-rain and rain type of component. Kim et al. [6] proposed a technique, in which the rotation angle and the aspect ratio are analyzed for the elliptical kernel. Firstly, rain streak regions at each pixel location are detected on it, they performed the non-local means filtering by the selection of the pixels of non-local neighbor and their weights. Luo et al. [5] proposed a de-raining technique for a single image by dictionary learning-based algorithm which comprises over a learned dictionary by very high discriminative codes and the patches of two layers were sparsely approximated. Chen et al. [4] have worked on the removal of rain streaks from rain images by working on a model with repeated and similar patterns of rain streaks that are correlated to obtain a lower rank of the model from the matrix structure to tensored structure. Li et al. [3] proposed a technique that is based on prior information which is built on Gaussian type of mixture models. This method was applicable for both, i.e., for background and the layers of rain. It used a simple type of patch-based priors and the streaks of rain accommodated multiple directions and scales.

### 3 Some Previous Works on De-Raining on Deep Neural Network

Fu et al. [8, 9] proposed a deep network of a learning-based method for the removal of streaks of rain from a single image. Yang et al. [10] proposed a different method that detects and jointly eliminates the rainy streaks by the usage of a deep recurrent dilated network. Li et al. [11] de-rained an image by multiple stages and used recurrent neural networks to exchange information across stages. Zhang et al. used a method in which the degeneration of background is prevented by the generative adversarial network (GAN) when it got extracted from a rainy image. Further to ensure the good visual quality, they utilized perceptual loss. Liu et al. proposed a network that is symmetry enhanced, in which the tilted streaks of the rain are removed from images [2]. Li et al. [12] proposed a de-raining method that introduces a framework of the network that is non-locally enhanced by encoder–decoder and designed to not only correctly capture the structural information and long-distance dependencies, but also from all the convolutional layers, it fully exploits hierarchical features, utilizes and designs no locally enhanced dense blocks. Similarly, a basic structure of encoder–decoder block using some intermediate layers is given here. Either the prior method or learning method can be applied to this basic block.

## 4 Some of the Recent Works in Single Image De-Raining Using Deep CNN Learning Models

Lan et al. [1] proposed a technique for the removal of rainy streaks from a single type of image and to upgrade better reuse of information on rainy streaks by using a double recurrent scheme. As this was the basic model for rain streaks, the Dense Net blocks are used for iteratively cascading the LSTM. The relative clean image predicted for the input in—the basic model by subtracting the streaks of the rain from the rainy image, this model makes detailed information of the image and effectively eradicates rain streaks. To get good rain removal performance, the loss functions used here are a combination of L1 loss, L2 loss, and SSIM loss. Wang et al. [2] proposed a module in which information aggregation of spatial contextual type is used for the learning of the streaks of rain of many dissimilar sizes from the perspective of the feature maps and Kernel. Here the multi-scaled features of rainy streaks are obtained and fused using the pyramid network module onto this network. Modules included in it are of two types, i.e., pyramid network module (PNM) and spatial contextual information aggregation module (SCIAM). Skip connections are also used for feature activation of proceeding layers. This signifies that for a single image de-raining, multi-scale information is important. Wang et al. in [13] proposed a framework that is a learning-based end-to-end deep network for rain removal. This method tried to figure out the issue like loss of details, by taking into consideration different receptive fields with the correlation between different layers. In this network, the basic unit is a residual block of multi-level guided type to solve the problem of information loss. The objective here is to obtain it along with a smaller receptive field layer and to obtain a different receptive field. In this block, for the guidance of the learning process, larger receptive fields are utilized.

Xiang et al. in [14] have developed a neural network of deep type based on supervised feature generative adversarial network (GAN) for optimization purposes. It is basically for de-raining but can also be applied to dehazing and other applications. They imposed supervision from the ground truth on various generative layers of this network. It is basically for single image removal of rain, using the deep network type architecture based on feature-supervised GAN on generator layers. For the training purpose, their basic idea was to impose supervision from the ground truth on the different layers of the generative network. Wang et al. [15] proposed de-raining using two modules, one is feature maps of multi-scale rain-free layer and the other is kernels, and then fused these two modules to learn the streaks structures of rain that are the primary requirement. Rain streak information has been dealt with the use of kernels of multi-scale rain-free layers with parameters that are shared at different scales. Then the information is used to maximize the information flow to link scale feature maps of multi-scale rain-free layers by use of dense connections. Peng et al. in [16] proposed W-net which has a strong learning ability for removing rain streaks using cumulative rain-density sensing. The low-frequency rain invariant signals were recovered from an auxiliary residual image. For cumulative rain-density classification, a label-encoding strategy is proposed that is also cost-sensitive. The rain-density

classifier under the multi-task training mechanism is pleasant to refine the selective information power of streaks of high frequency. Wang et al. [17] proposed a network for single-image rain removal. The layer of the components of rain is learned by the transformation mapping from input images. To enhance the representation of feature maps and to adaptively measure the feature response, a method named squeeze-and-excitation was adopted. For extracting the rain components from a residual block of dense non-local leverages spatial contextual information, the usage of non-local mean operations was required. Wang et al. [18] proposed and put forward an end-to-end model for single image type dehazing as well as a de-raining model called DRHNet (Deep Residual Haze Network). By the subtraction of the learning-based residual map which is negative in nature, they restored an image that was damaged by haze. To collect the contextual information, a module having context-aware feature extraction was put forward. To increase the convergence of DRHNet, an activation function called RPRReLU was also proposed. Yang et al. [19] proposed a method for rain streak removal and for the mismatching that arises between the different rain streak sizes, especially in the case of large streaks leading to poor performance during the training and testing phase. To remove this problem, representation in a hierarchical way of the wavelet transform in the recurrent de-raining process is embedded and executed as follows: (1) Removal of rain on lower frequency side component, (2) under the guidance of recovered lower frequency component on high-frequency components, a recurrent detail recovery is applied using the benefits of the recurrent modeling in multi-scale design like the wavelet transform. In this way, the variations in the rain streaks can be handled and the network is trained on the varying streaks of a single size which can be adapted to those with larger sizes for removal of the real streak of rain. In the process of the basic model of recurrent recovery, a dense type of dilated residual network is used. Problems related to heavy rain can also be handled by constructing a detailed appearance of an accumulation of rain removal.

Jin et al. [20] proposed a new general mathematical rain model with a different outlook on the disentanglement of features and redefined the de-raining problem. In this model, which is an end-to-end, de-raining is proposed based on the Asynchronous Interactive type of Generative Adversarial Network and the input is disentangled into two latent spaces. Matsui et al. [21] proposed a de-raining method that is based on GAN. From the training, dataset mapping is learned between the rainy type and residual type of images to train the generator. For the creation of synthetic rainy images, a combination of two composite models is proposed. The U-Net structure is used as a generator. For training, to generate synthetic rain noise, an automatic rain noise generator is also introduced. Sharma et al. [22] proposed a new framework of rain streak prediction model built on deep CNN learning of different types, for rain removal of a single image, in which, instead of the spatial domain, they used transform mechanism for compressed domain coefficients (DFT coefficients) as inputs directly by decomposing the rainy image in the frequency domain. Further, it is derived from the noise signal of pseudo-periodic characteristics type in which traces are left in the frequency domain and can be used for noise signal prediction. Starting calculation of the rain streak map is predicted by D-Net and the final calculation of the rain map is learned by DNet + N-Net. For attaining the DFT coefficients of

the de-rained image, the subtraction of two parameters that is DFT coefficients from the rainy image, and the final calculation of pixel-wise rain map was done. In the transformed domain, rain streaks information is preserved, and to use such kinds of features for image de-raining problems, deep CNN can be trained. Lin et al. [23] proposed a technique where Fuzzy Broad Learning System (FBLs) and two-phase processing methods were merged. At first, the input rainy image is preprocessed by the dehazing algorithm, and through this, it is separated into two layers, i.e., base and detail layers. In the FBLs, the detail layer of the image of the Y-channel was applied to get the channel of a clean image. Further, the channel images of Cb and Cr were combined. Later, to get an initial rain-free image, the base layer and the rain-free detail layer were fused. Then the final result was obtained by superimposing the details drawn out from the haze-free image with some transparent nature on the initial result. Wang et al. [24] proposed a technique for the removal of rain based on the feedback mechanism of clique recursive. Residual clique block (RCB) was constructed to deduce local details, for the consideration of the features in the interaction between distinct convolution layers. To cover more scale components lodge into a scale clique block (SCB), a convolutional unit of multipath dilated was used. For excellent feature representation, the multi-scale features are alternately updated and different scales of complementary correlation were considered essential. During propagation, it maximized the information flowing among SCB and RCB along with the clique recursive.

Ren et al. [25] put forward a dually connected de-raining net using pixel-wise attention for rain removal in single images. To learn the map of residual rain streaks, an encoder–decoder method was employed with the help of skip concatenation connection and skip sum connection. In the de-raining net, the information flowed between layers was promoted by the dual connections, thus allowing it to localize and discriminate the streaks of rain. Details of the image were preserved by the learnable pixel-wise attention that decodes weighted features for adaptively recalibrating their responses. Fan et al. [26] proposed a technique for single image de-raining which was a screen blend deep learning model. The composition has shortcut connections and ended with sibling branches. For the task of heterogeneous joint optimization, this architecture was designed and formulated in sibling branches to learn jointly the background scene and the distributive characteristics of the rain layer. Tasks like feature level were conducted and the isotropic image's gradient operator was set and employed in the first task for the construction of a model of perceptual loss which benefits the preservation of the edge of the output images. The vision tasks in a self-adaptive manner can be optimized without the knowledge of priors. Further, the de-rained image quality can be improved by a framework of post-processing for haze removal. To assign appropriate hyperparameters, an adaptive method was proposed and the implementation of this is done through a genetic algorithm based on population, by solving optimization problems of multi-objective type. Sharma et al. [27] proposed a model to generate a clean rain-free image through a High-Resolution Image De-Raining using Conditional Generative Adversarial Networks (HRID-GAN). The U net network is used here for the reconstruction of the image. As an alternative to using standard transpose convolution, they proposed to use effective

sub-pixel convolution and did this for the removal of the checkerboard artifacts in the generated rain-free image. The framework of encoder–decoder using a deep residual network was used and training was based on CGAN.

## 5 Discussion and Analysis

The performance metric discussed for image reconstruction here in this paper is PSNR, i.e., Peak Signal to noise ratio and SSIM, i.e., Structural similarity index, given in decibels unit estimated between the two images. The value of PSNR should be high for better reconstruction and provides the average of the image. Similarly, SSIM is used for quality measurement of the image, the higher the value, the better is the quality, and its range is from 0 to 1. Theoretical analysis of the various methods is shown in the comparison table, based on which the selection of a better method can be done and that will be used further in the proposed work. It gives clarity about the complications of its usage. There are different rain types/rain models like raindrops, rain streaks, and rain with haze, and all these need a different type of specialized model. Models are promising for their particular rain type. Some models are prior-based, some are data-driven, and others are a combination of data-driven with prior-based methods. But there should be a particular single model and algorithm to deal with all types of rain. To deal with this problem, the hybrid model is required. Another problem that exists is that, among all metrics, there is no single metric based on which various de-raining algorithms can be compared [23]. The recent algorithms as shown here are required to deal with de-raining under various circumstances. With complex types of datasets implemented on synthetic images, these methods give good results but are not always accurate with real images. For real-world rainy images and all such high-level vision problems, there lies a wide scope of developing new algorithms.

Figure 3 shows different de-raining methods for the selection of better techniques by considering all the aspects from different angles. It shows steps that need to be taken into consideration for the proper choice of de-raining/rain-free model. All these are end-to-end types of de-raining/rain-free networks. For getting better results, deep neural networks are used in recent times, and still, there is much scope for further exploration in this area.

GAN [14] FS-GAN is the architecture that is made up of two subnetworks, a generator  $G$  which is a network of convolution type, and a discriminator  $D$ . The advantage of this method is that different level features can be extracted. DRHNet [18] does the work of restoration by using the learning-based deep residual network. The advantage is that, on rainy images of the real world, learning-based methods' performance is better than the traditional methods and non-learning by a much large margin. Many of the drawbacks like antihalation, halo effects, white light, and more were eliminated effectively by these learning-based methods.

(FBLS) [23] It considered the ground truth image and image of Y-channel of YCbCr type color space for training and the time of training and running is less which is an added advantage.

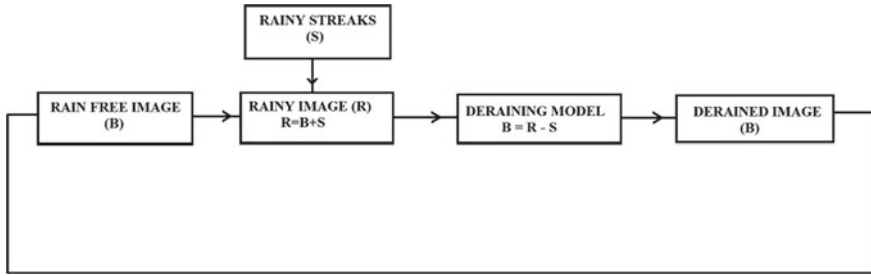


Fig. 2 Shows a basic de-rained model

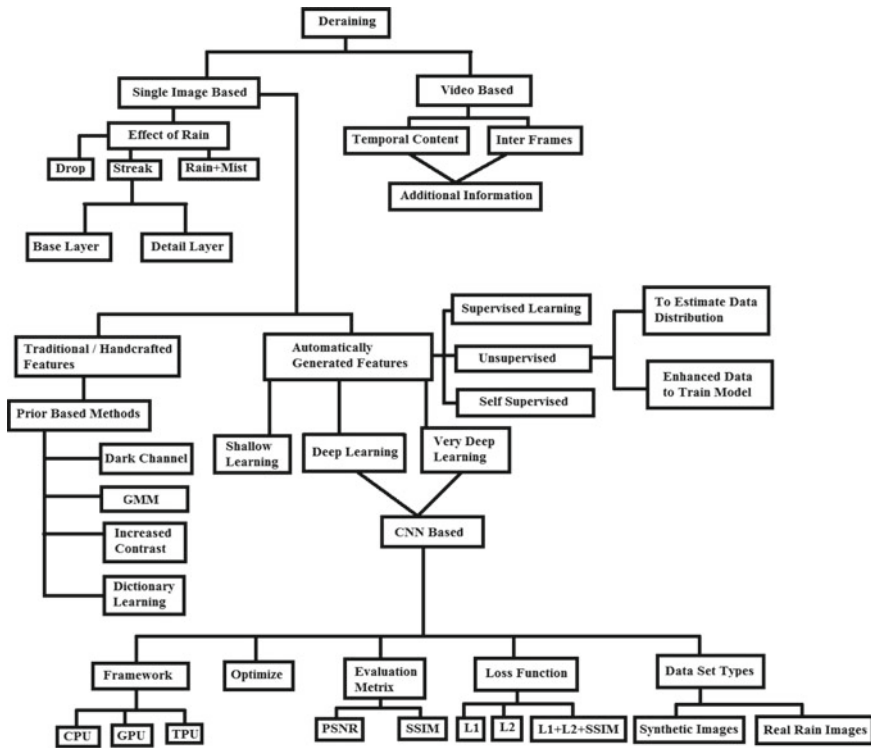


Fig. 3 The figure shows different stages under de-raining

(HRID-GAN) [27]. Here instead of using convolution, sub-pixel convolution was used to avoid the artifacts of checkerboard in the rain-free images that were generated and also used a residual network of deep type and decoder–encoder architecture. Its advantage is that the rain-free image generated at the output was of good quality and with less noise just like its clean image and the discriminator will find it tough to identify the distinction between the generated image and the real image.

BLSTM Standard LSTM [1] have many gates like input, output, forget, and hidden. Its advantage is that it allows the propagation of features beyond many stages and at the same time it connects the rain streak layer and background image layer.

In Table 1, the comparison is shown among the various literature work done in this paper. In Table 2, the comparison is done between the available quantitative results on the synthetic rain dataset of Rain100L and Rain100H. From Table 2, it is given that the better value of SSIM is for Rain 100L and Rain 100H [1], and in the case of PSNR, Rain 100H [13] and Rain [1] gave better results in quantitative analysis.

All these methods are recent works done in this area and have their benefits and limitations, but none of the methods cover all the rain model issues single-handedly.

**Table 1** Comparison of various literature work done

Technique used	Advantages	Limitations
Lan et al. [1] The double recurrent scheme, dense net blocks cascades LSTM	Different intensity rain streaks and directions can be tackled, heavy rain streaks removed on synthetic images	For further study, more exploring experiments are left, performs better in single image de-raining only
Wang et al. [2] Multi-scale information, feature maps, pyramid network module, spatial contextual information aggregation module	(a) Rain streaks of different sizes and scales are captured, (b) Performance on a real-scene dataset was better, to apply other low-level visions like dehazing, deblurring, and denoising	Performance on real-scene images was visually evaluated
Wang et al. [13] Multi-level Guided learning, the correlation between distinct layers with distinct receptive fields	Information loss is less	Evaluated the performance of the real-images data visually
Xiang et al. [14] Generative adversarial network, supervised feature	Can be applied to dehazing, image denoising, restorations, super-resolution	Evaluated the performance of the real images data visually
Wang et al. [15] Kernel and feature of multi-scale nature, parameter shared	Parameters reduced	Different loss functions may influence, affect the behavior of the network was affected by the different acquired manner of multi-scale ma
Peng et al. [16] Cumulative rain-density sensing network, w net	Different frequency information preservation	Evaluated the performance of the real-images data visually
Wang et al. [17] Encoder, Dense Non-Local Residual Block (DNLRB), decoder, spatial contextual details, squeeze-and-excitation	De-raining performance is boosted	Evaluated the performance based on visual comparisons in the case of real-world images

(continued)

**Table 1** (continued)

Technique used	Advantages	Limitations
Wang et al. [18] Residual map (negative), end-to-end structure, context-aware extraction of features	Used in both de-raining and dehazing	Particular parts are not considered that would handle white scenes
Yang et al. [19] Recurrent process, wavelet transform, Residual network	Heavy rain performance is also well	Evaluated the performance of the real-images data visually
Jin et al. [20] Asynchronous Interactive Generative Adversarial Network, two-branch structure, feed-forward information, feedback gradients	Various benefits in Re-identification of person Image/Video Coding, Action Recognition	Efficiency and scalability required to improve
Matsui et al. [21] GAN, residual learning	Overcomes problems like hue change problems, under de-raining, and over de-raining	Evaluated the performance of the real-images data visually
Sharma et al. [22] DNet + N-Net	In deep convolutional neural networks using frequency domain, input achieving comparable performance, as the rain streaks were reduced a visual improvement can be seen in the rain-free image	Evaluated the performance on the real images data visually
Lin et al. [23] Dehazing algorithm, fuzzy broad learning system (FBLS)	Less time for running and training, effective for real rainy images processing	Qualitative comparison on real rainy images not conducted
Wang et al. [24] Residual clique block (RCB), dilated convolutional unit of multipath nature, scale clique block (SCB)	The majority of the rain streaks were removed, achieving promising performance	Acceptable calculation consumption
Ren et al. [25] The recurrent mechanism, using pixel-wise attention a dually connected de-raining net, encoder–decoder, rain-streaks map of residual nature, skip sum connection, and concatenation	The superior performance of our method	Evaluated the performance of the real-images data visually
Fan et al. [26] Shortcut connections, an image enhancement framework, feature-level task, appropriate weighting coefficients, screen blend model, Operator is Sobel–Feldman	It preserves the complex texture of the background layer	Evaluated the performance of the real-images data visually

(continued)



**Table 1** (continued)

Technique used	Advantages	Limitations
Sharma et al. [27] High-resolution image rain removal used conditional, generative adversarial networks (HRID-GAN), the encoder–decoder network, deep residual network, U-Net	Have better resolution, and not contain white rounds artifacts and blur effect	Evaluated the performance of the real images data visually

**Table 2** Quantitative analysis in terms of PSNR and SSIM

Methods	RAIN 100L		RAIN 100H	
	PSNR (dB)	SSIM	PSNR (dB)	SSIM
Lan et al. [1]	37.82	0.9788	31.24	0.9266
Wang et al. [2]	28.24	0.89	38.20	0.98
Wang et al. [13]	27.52	0.86	36.97	0.98
Wang et al. [15]	36.72	0.98	27.08	0.85
Wang et al. [17]	36.64	0.9780	27.12	0.8524
Yang et al. [19]	36.75	0.9754	26.89	0.8406
Jin et al. [20]	37.21	0.982	27.13	0.902
Lin et al. [23]	27.36	0.89	20.13	0.74
Wang et al. [24]	28.26	0.8837	24.86	0.8016
Ren et al. [25]	36.27	0.9725	29.11	0.8691
Fan et al. [26]	–	<b>0.9412</b>	–	<b>0.8521</b>

## 6 Conclusion

This paper discussed the concept of single image de-raining from all perspectives and special attention is given to the deep learning models. Discussed single image rain removal/de-raining models, especially deep CNN-based learning networks and also discussed the de-rain datasets with different intensities of rain like a raindrop, rain streaks, rain with mist, and haze. Most of the techniques discussed in this paper de-rained the image using an end-to-end model. This paper discussed the disadvantages and advantages of these de-raining methods both quantitatively and qualitatively. Also compared the existing results available and the conclusion is drawn from this is that for a different type of rain that is from small drop to complex real rain streaks one single method is not applicable. The metrics for evaluation used here are PSNR and SSIM. The analysis here shows that many of the huge learning-based techniques of single image rain removal/de-raining techniques such as HRID-GAN, RRSIFT (DNet + N-Net), BLSTM, Residual clique block (RCB) are some of the recent techniques used for de-raining and came up with better performance outcome in terms of evaluation metrics in both the synthetic type of images and real rainy images.

## References

1. Lan Y, Xia H, Li H, Song S, Wu L (2020) Double recurrent dense network for single image deraining. *IEEE Access* 8:30615–30627. <https://doi.org/10.1109/ACCESS.2020.2972909>
2. Wang C, Wu Y, Cai Y et al (2020) Single image deraining via deep pyramid network with spatial contextual information aggregation. *Appl Intell* 50:1437–1447. <https://doi.org/10.1007/s10489-019-01567-5>
3. Li Y, Tan RT, Guo X, Lu J, Brown MS (2016) Rain streak removal using layer priors. In: *CVPR*, pp 2736–2744. <https://doi.org/10.1109/CVPR.2016.299>
4. Chen Y, Hsu C (2013) A generalized low-rank appearance model for spatio-temporally correlated rain streaks. In: *ICCV*, pp 1968–1975. <https://doi.org/10.1109/ICCV.2013.247>
5. Luo Y, Xu Y, Ji H (2015) Removing rain from a single image via discriminative sparse coding. In: *ICCV*, pp 3397–3405. <https://doi.org/10.1109/ICCV.2015.388>
6. Kim J, Lee C, Sim J, Kim C (2013) Single-image deraining using an adaptive nonlocal means filter. In: *ICIP*, pp 914–917. <https://doi.org/10.1109/ICIP.2013.6738189>
7. Kang L, Lin C, Fu Y (2012) Automatic single-image-based rain streaks removal via image decomposition 21(4):1742–1755. <https://doi.org/10.1109/TIP.2011.2179057>
8. Fu X, Huang J, Ding X, Liao Y, Paisley J (2017) Clearing the skies: a deep network architecture for single-image rain removal 26(6):2944–2956. <https://doi.org/10.1109/TIP.2017.2691802>
9. Fu X, Huang J, Zeng D, Huang Y, Ding X, Paisley J (2017) Removing rain from single images via a deep detail network. In: *CVPR*, pp 1715–1723. <https://doi.org/10.1109/CVPR.2017.186>
10. Yang W, Tan RT, Feng J, Liu J, Guo Z, Yan S (2017) Deep joint rain detection and removal from a single image. In: *CVPR*, pp 1685–1694. <https://doi.org/10.1109/CVPR.2017.183>
11. Li G, He X, Zhang W, Chang H, Dong L, Lin L (2018) Nonlocally enhanced encoder-decoder network for single image deraining. In: *ACM MM*, pp 1056–1064. <https://doi.org/10.1145/3240508.3240636>
12. Li X, Wu J, Lin Z, Liu H, Zha H (2018) Recurrent squeeze and-excitation context aggregation net for single image deraining. In: *ECCV*, pp 262–277. <https://doi.org/10.1007/978-3-030-01234-216>
13. Wang C et al (2019) Learning a multi-level guided residual network for single image deraining. *Signal Process: Image Commun* 78:206–215
14. Xiang P, Wang L, Wu F, Cheng J, Zhou M (2019) Single-image de-raining with feature-supervised generative adversarial network. *IEEE Signal Process Lett* 26(5):650–654. <https://doi.org/10.1109/LSP.2019.2903874>
15. Wang C, Zhang M, Su Z et al (2020) Densely connected multi-scale de-raining net. *Multimed Tools Appl*. <https://doi.org/10.1007/s11042-020-08855-0>
16. Peng L, Jiang A, Yi Q, Wang M (2020) Cumulative rain density sensing network for single image derain. *IEEE Signal Process Lett* 27:406–410. <https://doi.org/10.1109/LSP.2020.2974691>
17. Wang C, Fan W, Zhu H et al (2020) Single image deraining via nonlocal squeeze-and-excitation enhancing network. *Appl Intell*. <https://doi.org/10.1007/s10489-020-01693-5>
18. Wang C, Li Z, Wu J, Fan H, Xiao G, Zhang H (2020) Deep Residual Haze network for image dehazing and deraining. *IEEE Access* 8:9488–9500. <https://doi.org/10.1109/ACCESS.2020.2964271>
19. Yang W et al, Scale-free single image deraining via visibility-enhanced recurrent wavelet learning. *IEEE Trans Image Process Draft*
20. Jin X, Chen Z, Li W (2020) AI-GAN: asynchronous interactive generative adversarial network for single image rain removal. *Pattern Recogn* 100:107143
21. Matsui T, Ikehara M (2020)GAN-based rain noise removal from single-image considering rain composite models. *IEEE Access* 8:40892–40900. <https://doi.org/10.1109/ACCESS.2020.2976761>
22. Sharma PK, Basavaraju S, Sur A (2020) Deep learning-based image de-raining using discrete Fourier transformation. *Vis Comput*. <https://doi.org/10.1007/s00371-020-01971-w>

23. Lin X, Ma L, Sheng B, Wang Z, Chen W, Utilizing two-phase processing with FBLS for single image deraining. *IEEE Trans Multimed.* <https://doi.org/10.1109/TMM.2020.2987703>
24. Wanga X, Chen J, Jianga K, Hana Z, Ruana W, Wanga Z, Lianga C (2020) 417:142–154
25. Ren W, Tian J, Wang Q, Tang Y (2020) Dually connected deraining net using pixel-wise attention. *IEEE Signal Process Lett* 27:316–320. <https://doi.org/10.1109/LSP.2020.2970345>
26. Fan Y, Chen R, Li Y et al (2020) Deep neural de-raining model based on dynamic fusion of multiple vision tasks. *Soft Comput.* <https://doi.org/10.1007/s00500-020-05291-y>
27. Sharma PK, Basavaraju S, Sur A (2020) High-resolution image de-raining using conditional GAN with sub-pixel upscaling. *Multimed Tools Appl.* <https://doi.org/10.1007/s11042-020-09642-7>
28. Li R, Cheong L-F, Tan R (2017) Single image deraining using scale-aware multi-stage recurrent network

# A Hybrid Translation Model for Pidgin English to English Language Translation



Saviour Oluwatomiyin, Sanjay Misra, John Wejin, Akshat Agrawal,  
and Jonathan Oluranti

**Abstract** The African continent is made up of people with rich diverse cultures and spoken languages. Despite the diversity, one common point of unification, especially among the West African communities is the spoken pidgin-English language. With the development in web technology and the English language dominance of web content, this growing population stands disadvantaged in understanding content on the web. To proffer a solution, researchers in machine translation from Pidgin English to the English language have leveraged only unsupervised and supervised Neural Machine Translation-based models. In this paper, we propose a hybrid-strategic model that improves the accuracy of the baseline Neural Machine Translation Model (NMT) in translating pidgin English to the English language. From the JW300 public dataset, we used 22,047 sentence pairs for training our model, 1000 for tuning, and 2520 for testing. The Bi-Lingual Evaluation Understudy (BLEU) score was employed as a metric of measurement. From our findings, our hybrid model outperforms the baseline NMT model with a BLEU score of 1.05 on two-level translation. This indicates that the accuracy is dependent on the level and type of hybrid used. Studies that look at in-depth pre-translation strategies for developing translation machine model are green areas for pidgin-English translation.

**Keywords** Machine translation · BLEU · Supervised learning

---

S. Oluwatomiyin · J. Wejin · J. Oluranti  
Center of ICT/ICE Research, Covenant University, Ota, Ogun State, Nigeria  
e-mail: [john.wejinpgs@stu.cu.edu.ng](mailto:john.wejinpgs@stu.cu.edu.ng)

J. Oluranti  
e-mail: [jonathan.oluranti@covenantuniversity.edu.ng](mailto:jonathan.oluranti@covenantuniversity.edu.ng)

S. Misra  
Department of Computer Science and Communication, Østfold University College, Halden,  
Norway  
e-mail: [sanjay.misra@hiof.no](mailto:sanjay.misra@hiof.no)

A. Agrawal (✉)  
Amity University Haryana, Gurgaon, India  
e-mail: [akshatag20@gmail.com](mailto:akshatag20@gmail.com)

## 1 Introduction

The Pidgin English language which has its root in the trading contacts of the Portuguese on the coast of the Niger Delta [1], has become widely spoken in West Africa today. In Nigeria, it is considered a second language spoken by over 75 million Nigerians [2]. As 65% of the Nigerian population is expected to have internet access in the year 2025 [3], the growing population stands disadvantaged, as most of the content over the internet is written in the English language. For maximum benefit of the development in web technology by these fast-growing pidgin-English speaking people, the need for translation of English content on the internet to pidgin English becomes important.

This need makes machine language translation a paramount tool in achieving this goal. Machine language translation is a field in computational linguistics by which computer algorithms are used to translate text or voice from one natural language to another.

This research work explores the subject of translation between a low-resource language and a high-resource creole of that language, with translation from Nigerian Pidgin English to the English Language as a case study. It further investigates different approaches to machine translation to determine what methods may be most suitable for translating from a low-resource creole of a high-resource language to a high-resource language.

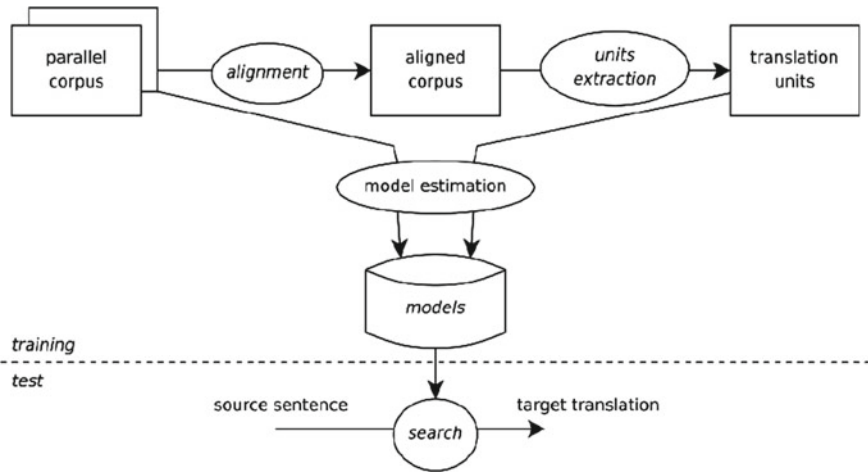
The paper is structured in 5 sections. The next section provides the background of the work and related studies. Section three presents the methodology. Results are summarized in Sect. 4. Finally the conclusion is drawn and future work is discussed in Sect. 5.

## 2 Types of Machine Translation

### 2.1 *Statistical Machine Translation (SMT)*

This model describes a group of approaches that apply statistical methods to the problem of automatic translation. Statistical Machine Translation models human languages as a machine learning task [4]. Sets of parameters that optimize a model such that sequences input into the model produce a sequence with the equivalent in some target language, T. As shown in Fig. 1, optimization is done by applying a learning algorithm to a body of human-translated text commonly called a parallel corpus. In Ref. [5], the researchers modeled translation as a Bayesian probability problem. The model uses n-gram as its language model while the translation model was premised on fertility and distortion.

An accuracy of 48% with decoding French sentences from Hansard to English was obtained. The drawback in this approach is the persistence constrained of words in source sentences to independently produce translation in target sentence [6]. To



**Fig. 1** Pipeline of a statistical machine translation

provide a solution to the weakness in SMT, authors in Refs. [6] and [7] proposed Phrase-Based Statistical Machine Translation (PBSMT). This model uses focus phrases (in this context, mere sequences of words), as a unit of translation and employs a log-linear model [8]. The framework of a statistical machine translation [9] is given in Fig. 1.

## 2.2 Unsupervised Statistical Machine Translation (UNMT)

This machine translation does not require a parallel corpus for training. In Ref. [10], the researchers modeled translation as a decipherment task. It uses an iterative variant of the Expectation–Maximization algorithm and Bayesian decipherment to deal with scale. Evaluating this method on the Time Corpus shows a score of 48.7% BLEU points. Though this model looks promising, there are model severe limitations in vocabulary size due to the computational expense in training a model with a larger vocabulary.

In Ref. [11], the authors implemented an unsupervised approach that automatically created shallow transfer rules (transfer rules are usually hand-coded in RBMT). Mikolov et al. [12] explored a means to harness multilingual data and a small starting dictionary using distributed representations of words. They used the Continuous Bag of Words (CBOW) representation model to learn translations of words and phrases. Although this approach increases accuracy, it relies on large monolingual data.

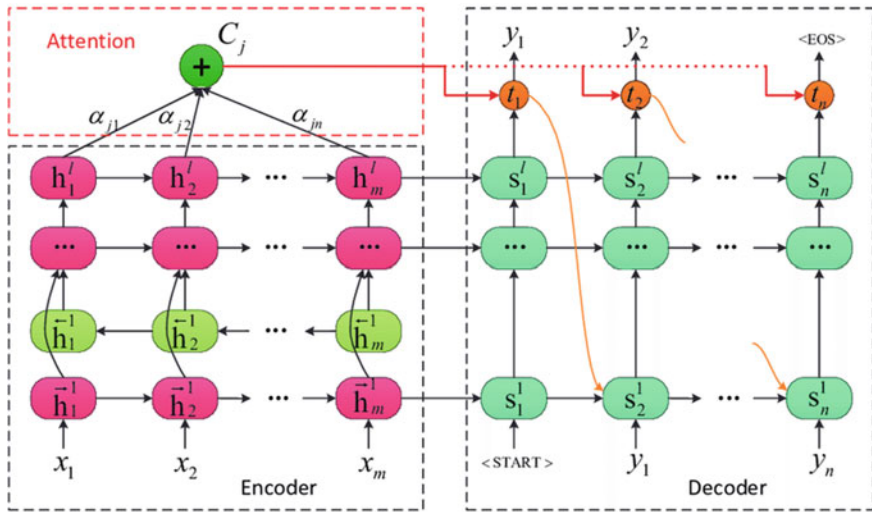


Fig. 2 Structure of a neural machine translation

### 2.3 Neural Machine Translation (NMT)

This model translates language as a direct sequence-to-sequence transformation task, using a deep neural network [13]. NMT uses an encoder–decoder fashion of translation as shown in Fig. 2. Sutskever et al. [13] proposed a model that used two Long-Short Term Memory (LSTM) networks—one to generate a high dimensional fixed-length vector representation of the input sequence and the second as a language model, to decode that vector representation and generate the output sequence.

A BLEU accuracy score of 34.81 on a translation from English–French was observed. Cho et al. [14] introduced the RNN encoder–decoder as a part of the log-linear model of a Statistical Machine Translation System. In their work, the RNN encoder–decoder was used to score phrase pairs in the phrase table of an SMT. The model they proposed achieved a BLEU score of 33.87 on an English- > French translation task (WMT14 data set), compared to a 33.30 BLEU score for a baseline system trained with MOSES. This model was more computationally intensive than SMT and could not handle unknown words effectively.

A Structure of a neural machine translation [15] is given in Fig. 2.

### 2.4 Machine Translation for Nigerian Pidgin English

Ogueji and Ahia [16] implemented the first known machine translation system for the Nigerian pidgin English (NPE)–English language translation task. They explored an unsupervised approach [17] which shows a BLEU score of 7.93 translating from

Nigerian Pidgin English to the English Language and a BLEU score of 5.18 translating from the English Language to Nigerian Pidgin English. Ogueji and Ahia [18] trained a supervised MT system, using the transformer architecture [19]. The model they trained achieved a BLEU of 24.29 on a translation task from the English Language to Nigerian Pidgin English (using byte-pair encodings) and a BLEU score of 24.67 translating from Pidgin English to the English Language.

## 2.5 Evaluation Metrics for Machine Translation Systems

The Bi-Lingual Evaluation Understudy (BLEU) score introduced in [20] was the first automatic evaluation metric for machine translation. It is calculated as a measure of the frequency of the appearance of expected n-grams (typically up to 4-g) in a translation made by an automatic translation system. While the BLEU score provides an easy mechanism to evaluate and train machine translation systems in faster iterations, it performs poorly at evaluating translation at the sentence level.

To address the limitation inherent in the BLEU metric, various metrics have been proposed. One such metric is the NIST metric (named after the American National Institute of Standards and Technology). NIST is based on BLEU but assigns a higher weight to n-grams with a low probability of appearing in an automatically generated translation. METEOR (Metric for Evaluation of Translation with Explicit Ordering) [21] was another metric proposed for the automatic evaluation of translation quality. A METEOR score is generated from an evaluation of translation quality using explicit word alignments produced by a hierarchically built word-alignment model. METEOR considers word stems and synonyms when building said word alignments. This makes it more robust than the BLEU score for evaluating translations that may use words that are not in the expected target. Consequently, METEOR performs better than BLEU on sentence-level translation quality scoring [22, 23].

Regardless of the existence of the METEOR and NIST evaluation metrics, the BLEU score has remained the de-facto standard for the evaluation of translation quality due to the fact that it is easier to calculate and it provides an easy means to benchmark translation systems against each other.

## 3 Methodology

The approach used in this paper is shown in Fig. 2. It begins with the regularization of bible references taken from the dataset. This was achieved using python code that generates a special token ‘scptr’ to replace bible references as shown in Table 1 Samples from the JW300 corpus [24] showing scripture references regularization.

We then tokenized the replaced references and lowercased them for consistency in building a translating system. After lower casing, we cleaned the data by eliminating sentence pairs with 8 times the number of tokens in the corresponding sequence in



**Table 1** Reference replacement using python

<p>A1. <i>True true , we go gain from their example because we know sey their story really happen . — Rom .</i></p> <p>A2. <i>15 : 4 .</i></p> <p>EN</p> <p>B1. <i>Indeed , all of this helps us to consider how we should or should not deal with similar i ssues . — Rom .</i></p> <p>B2. <i>15 : 4 .</i></p>
---

the order side of the corpus, and sentence pairs for which a sequence on any side of the corpus contained more than 60 tokens. The cleaned data is passed to the various models with appropriate sequences taken as shown in Fig. 3.

### 3.1 Dataset

The JW300 corpus [22] whose content information is shown in Table 1 was used for this project. The corpus consists of publications made by the Jehovah’s Witnesses Organization in multiple languages. The train set contained 22,047 sentence pairs, while the validation contained 1000 sentence pairs. All models were evaluated on a test set of 2529 sentences preprocessed by the *Masakhane1* group.

### 3.2 Model

The Phrase-based statistical machine translation (PBSMT), Transformer-based neural machine translation which used a Transformer of 2 encoders/2 decoders and

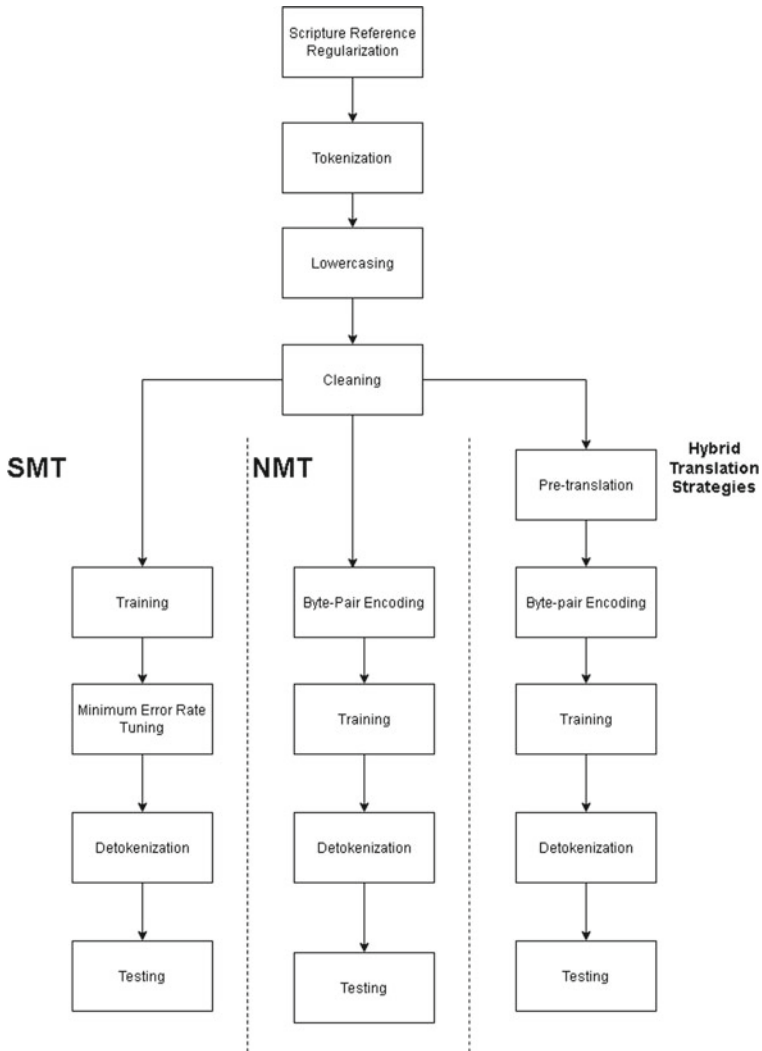


Fig. 3 Process description

trained by the free to use JoeyNMT toolkit were used. We also created a hybrid translation system by subjecting a two-level pre-translation for an NMT as a model for comparison.

**Table 2** BLEU scores attained for the pre-translation strategy

Test	BLEU
Pre-translation Strategy 1 (simple replacement)	21.63
Pre-translation Strategy 2 (context-aware replacement)	23.23

**Table 3** BLEU score comparison

Machine translation system	BLEU score
pidgin-pbsmt	29.43
pidgin-nmt	22.81
pidgin-psnmt	23.23

## 4 Result

### 4.1 Pre-Translation Strategy

Table 2 below shows the BLEU score for our hybrid two-level pre-translation strategy model. It is shown that when using a two-level strategy for not, the BLEU score is increased to 23.23 on the second pre-translation. This shows that the pre-translation strategy allows more information to be encoded, thus increasing accuracy. This also shows that more pre-translation levels will lead to more accuracy.

Table 3 shows the summary of the three models used in this research. It can be seen that pidgin-pbsmt outperforms pidgin-nmt pre-translation model (pidgin-psnmt), with the BLEU score of 29.43.

## 5 Conclusion and Future Work

Researches are going on various languages [25, 26] and in each country, English is mixed up with local languages and causes for creating pigeon language. In this research work, three approaches to machine translation from Nigerian Pidgin English to the English Language have been considered. The performance of each of these systems has been analyzed and compared based on their performance on an automatic evaluation metric (the BLEU score).

The results show that Phrase-based Statistical Machine Translation is the most effective of the methods considered (attaining a higher BLEU score than all other approaches considered), as is known to often be the case in low-resource scenarios. More interesting is the relative performance of the purely transformer-based system and the transformer-based system with pre-translation applied. Studies that look at in-depth pre-translation strategies for developing translation machines are green areas for pidgin-English translation.

## References

1. Bbc starts pidgin digital service for west Africa audiences (2017). <https://www.bbc.com/news/world-africa-40975399>
2. Statista. Mobile internet user penetration in Nigeria from 2015 to 2025". Available: Nigeria mobile internet user penetration 2025 | Statista. Accessed 15 March 2021
3. Lopez A (2007) A survey of statistical machine translation. <https://doi.org/10.21236/ada466330>
4. Brown PF et al (1990) A statistical approach to machine translation. *Comput Linguist* 16(2):79–85
5. Marcu D, Wong W (2002) A phrase-based, joint probability model for statistical machine Translation Daniel Marcu. <https://doi.org/10.3115/1118693.1118711>
6. Koehn P, Och FJ, Marcu D (2003) Statistical phrase-based translation. In: Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics. pp 127–133. <https://www.aclweb.org/anthology/N03-1017>. Accessed 29 Apr 2020
7. Artetxe M, Labaka G, Agirre E (2020) Unsupervised statistical machine translation. ArXiv180901272 Cs. <http://arxiv.org/abs/1809.01272>. Accessed 28 Apr 2020
8. Karan S (2015) Methods for leveraging lexical information in SMT. M.S Thesis, Comp Sci, IIIT, Hyderabad, India. [https://www.researchgate.net/publication/279181014\\_Methods\\_for\\_Leveraging\\_Lexical\\_Information\\_in\\_SMT](https://www.researchgate.net/publication/279181014_Methods_for_Leveraging_Lexical_Information_in_SMT). Accessed 15 March 2021
9. Ravi S, Knight K (2011) Deciphering Foreign Language. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, Oregon, USA, pp 12–21. <https://www.aclweb.org/anthology/P11-1002>
10. Sánchez-Martínez F, Forcada ML (2009) Inferring shallow-transfer machine translation rules from small parallel corpora. *J Artif Intell Res* 34:605–635. <https://doi.org/10.1613/jair.2735>
11. Mikolov T, Le QV, Sutskever I (2020) Exploiting similarities among languages for machine translation. ArXiv13094168 Cs. <http://arxiv.org/abs/1309.4168>. Accessed: 16 May 2020
12. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. ArXiv E-Prints, ArXiv:1409.3215
13. Cho K, van Merriënboer B, Gülçehre C, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. *CoRR* abs/1406.1078. <http://arxiv.org/abs/1406.1078>
14. Zhou L, Zhang J, Zong C (2018) Look-ahead attention for generation in neural machine translation. In: Huang X, Jiang J, Zhao D, Feng Y, Hong Y (eds) *Natural Language Processing and Chinese Computing*. NLPCC 2017. *Lect Notes Comput Sci* 10619. Springer, Cham. [https://doi.org/10.1007/978-3-319-73618-1\\_18](https://doi.org/10.1007/978-3-319-73618-1_18)
15. Ogueji K, Ahia O (2019) PidginUNMT: Unsupervised Neural Machine Translation from West African Pidgin to English
16. Lample G, Ott M, Conneau A, Denoyer L, Ranzato M (2018) Phrase-based & neural unsupervised machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing, Brussels, Belgium. pp 5039–5049. <https://doi.org/10.18653/v1/D18-1549>
17. Ahia O, Ogueji K (2020) Towards supervised and unsupervised neural machine translation baselines for nigerian pidgin. ArXiv200312660 Cs. <http://arxiv.org/abs/2003.12660>. Accessed 12 May 2020
18. Vaswani A et al. (2017) Attention Is All You Need. ArXiv170603762 Cs. <http://arxiv.org/abs/1706.03762>. Accessed 24 Apr 2020
19. Papineni K, Roukos S, Ward T, Zhu W-J (2002) Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA. pp 311–318. <https://doi.org/10.3115/1073083.1073135>
20. Lavie A, Sagae K, Jayaraman S (2004) The significance of recall in automatic metrics for MT evaluation, vol. 3265, pp. 134–143. [https://doi.org/10.1007/978-3-540-30194-3\\_16](https://doi.org/10.1007/978-3-540-30194-3_16)

21. Lavie A, Agarwal A (2020) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the Second Workshop on Statistical Machine Translation, Prague, Czech Republic. pp 228–231. <https://www.aclweb.org/anthology/W07-0734>. Accessed 24 July 2020
22. Agić Z, Vulić I (2019) JW300: A wide-coverage parallel corpus for low-resource languages. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy. pp 3204–3210. <https://doi.org/10.18653/v1/P19-1310>
23. Adubi SA, Misra S (2016) Syllable-based text compression: a language case study. Arab J Sci Eng (Springer) 41(8):3089–3097
24. Ojumah S, Misra S, Adewumi A (2017). A database for handwritten yoruba characters. In: International Conference on Recent Developments in Science, Engineering and Technology. Springer, Singapore, pp 107–115
25. Sharma I, Anand S, Goyal R, Misra S (2017). Representing contextual relations with sanskrit word embeddings. In: International conference on computational science and its applications. Springer, Cham, pp 262–273
26. Akman A, Bayindir H, Ozleme S, Akin Z, Misra S (2011) A lossless text compression technique using syllable-based morphology. Int J Inf Technol 8(1):1–9

# An Incremental Load Balancing Algorithm in Federated Cloud Environment



Nzanzu Vingi Patrick, Sanjay Misra, Emmanuel Adetiba,  
and Akshat Agrawal

**Abstract** With the development of cloud computing, load balancing issues have substantially become prominent, which is of concern as well to the industry as to academia. Load balancing contributes to a high degree of customer satisfaction and the use of resources in the cloud by verifying an accurate, secure, and equitable allocation of all computing resources. In this paper, we review the research progress on load balancing issues from the perspective of several incremental and dynamic algorithms in federated cloud computing. First, we present some obstacles to load balancing algorithms in cloud computing and suggest a corresponding incremental algorithm for load balancing in a federated cloud. Last, we offer current challenges and highlight possible future directions for research in this field.

**Keywords** Federated cloud · Cloud computing · Ant colony optimization · Load balancing

## 1 Introduction

In recent years, monumental advancements in information technology have allowed people to create, process, and vast exchange volumes of information [1]. Cloud

---

N. V. Patrick · E. Adetiba

Covenant Applied Informatics and Communication African Center of Excellence, Covenant University, Ogun State, Ota, Nigeria

e-mail: [emmanuel.adetiba@covenantuniversity.edu.ng](mailto:emmanuel.adetiba@covenantuniversity.edu.ng)

S. Misra

Department of Computer Science and Communication, Østfold University College, Halden, Norway

e-mail: [sanjay.misra@hiof.no](mailto:sanjay.misra@hiof.no)

A. Agrawal (✉)

Amity University Haryana, Gurgaon, India

e-mail: [akshatag20@gmail.com](mailto:akshatag20@gmail.com)

computing as an Internet-based solution offers flexibility, efficiency, location independence, accessibility, delivery agility, wide network access, elasticity, volunteerism, complexity, scalability, dynamics, rapid service conformation, lower investment expenditure, distribution and pooling of resources for a variety of services and applications [2, 3]. Cloud computing also allows service providers to have network access and share fully customizable computing resources [4]. It also offers certain functionalities for the universal, distributed, and omnipresent handling of information [5]. A cloud environment may supply users with a variety of services and computing capabilities, such as databases, networks, equipment, storage, and software applications, all of this within an on-demand paradigm [6]. Cloud computing offers significant cost benefits to companies, corporations, and independent people while also providing tenants with high-level interactive capabilities [7].

In contrast, various types of IT-based systems have indicated that the cloud computing trend has shifted from a single Cloud Services Provider (CSP) to a federation of CSPs, which should comprise several dispersed public and private cloud platforms [8]. Rochwerger et al. [9] report that federated clouds allow both public and private clouds to exchange resources. Within the federation in order to maximize their pools of resources at peak times. Only the federation of clouds can provide practically limitless processing capacity, parallel storage, and scale economies said Buyya et al. [10]. The restricted physical resources within a single provider's resource pool are the reason for cloud federation [11, 12].

Any cloud computing framework is composed of three main components: client, datacenters, and distributed servers.

Task scheduling is one of the critical issues that affect a federated cloud environment. The benefit of cloud computing is its powerful and efficient use of all the computational power that should be available precisely when processes are correctly scheduled [13, 14]. Load balancing, Known as the method of segmenting workloads and processing resources, is an essential player in the cloud computing environment when planning loads [14, 15].

An algorithm to balance the load in a federated cloud infrastructure is implied to provide an equilibrium on the usage of the resources for all the tenants of the federation in this paper. Here the artificial Ant Colony optimization (ACO) algorithm helps to reduce the workload of each node within the infrastructure, enhance the performance of the entire environment, and finally improve the scheduling of processes for all the nodes within the federated cloud.

The rest of this paper is organized as follows. Section 2 illustrates the related works of load balancing in the cloud computing environment. Section 3 presents the methodology and the proposed technique to balance the workload in the federated cloud. The experiments and analysis of the given technique are illustrated in Sect. 4. Finally, the conclusion and future work is presented in Sect. 5.

## 2 Related Works

A federated cloud is defined as a distributed virtual computer environment that uses the Internet to share its resources among its tenants (users and CSPs) across a vast region. Many cloudlets will be able to request and apply the resources simultaneously. Clients can access the services via the Internet at any time, anywhere, and CSPs can also access and suggest services.

CloudSim, a simulator simulating cloudlets on a computer network, was introduced by Buyya et al. [16]. The load balancers delegate functions of the cloudlets to computer servers while describing all the workload and all those groups that are present in scheduling. It follows the theory of arranging the tasks from one node to the next. This task planning process is based on how much energy a node requires to cope with the workload. It results in underbalanced nodes when the load is transmitted to the Virtual Machines (VM) [17].

Load scheduling is the easiest way to distribute and run cloudlets on VM to lower the cost of computing. This refers to execution time, transfer time, waiting time, response time, as well as operating expenses. Chaudhary and Kumar [18] thought that scheduling could be static or dynamic when allocating the load. Dynamic techniques use bio-inspired algorithms to schedule workloads. Based on swarm algorithms are used for the assignment of workload in cloud computing to address the successive dependent on particle velocity and location, reported Devaraj et al. [14]. This model of scheduling is reinforced with a round-robin algorithm. This method of scheduling tasks operates from meta-heuristic models. After multiple iterations and fitness values, the roulette selection function is determined on the basis of random timing. The static algorithm for load balancing in the cloud may not fall under the group of meta-heuristic techniques. This strategy entails low running costs compared with primitive models like Round Robin and First Come First Serve (FCFS) scheduling techniques [19].

Rahmeh et al. [20] use biased random sampling to present a compelling and distributed Grid Load Balancing Network. The developed network system is a kind of self-organized and only depends on local data for the delivery of loads and the exposure of resources. They show that adding a regional sensitivity factor to the random walk selection will lessen the impact of contact latency in the network environment Grid. The Biased Random Sampling Algorithm is a distributed load balancing approach that uses random machine domain profiling to accomplish self-organization and therefore load distribution over all network nodes [1, 20].

Rashedi and Zarezadeh [21] present the Binary Gravitational Search Algorithm (BGSa). This technique is specialized in the optimization of the scheduling operation generated from different platforms. Khatibinia and Khosravi [22] proposed a hybrid Gravitational Search Algorithm using an orthogonal crossover and pattern search to schedule the workloads in the infrastructure of the cloud.

An improved variant of Ant Colony Optimization (ACO) is provided by Nishant et al. [24]. ACO is utilized for balancing the workload in a cloud computing environment to help the network work well, even during its rush hours of operation. The



proposed algorithm is a solution based on Artificial Intelligence (AI) having a head node intended for the production of ants. Those Ants go through the cloud network's width and length in the way any under-loaded or over-loaded node is located. They then refresh a pheromone table to hold the details about resource use.

Chen et al. [25] endorse the notion that the computational capacity of virtual computers declines when a whole host of consumers access products in cloud computing. To solve this problem, they built an optimized load balancing system named EUQoS framework for virtual machine scheduling. The UEQoS framework is realized using two open cloud systems, Eucalyptus and Hadoop. In the architecture described here, the load balancer module has two main components: a load balancer and an agent-based controller [26]. The study reports that three important mechanisms are provided by the load balancer module: equilibrium triggering, EUQoS scheduling, and VM power. A weighted round-robin load-balancing algorithm allows the slicing of the workload. In this algorithm approach, tasks are selected in a fixed sequence for execution [27].

Ma et al. [28] present a VM load balancing distributed model in a cloud data center with the usage of a method named Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS). TOPSIS is one of the most popular ways of Multi-Criteria Decision-Making (MCDM). Every entity in the infrastructure runs a duplicate VM control module that tracks the local resource usage according to the architecture provided. If the monitoring detects an indication that resources are overloaded or underloaded, it will make two decisions:

- Which VM to move into a new physical machine should be chosen?
- What physical machine to pass selected VM, too, should be selected?

To minimize the amount of probabilistic migrations of VMs, the algorithm is going to sort all VMs on the problematical machine in declining order of actual usage and then select the topmost use for migration to a new correct physical system [26].

Ghafari et al. [29] are proposing a new framework to balance the load, relating to power-aware, named Bee-MMT (artificial bee colony algorithm- Minimal migration time), to lessen power usage in cloud computing. To this end, an artificial bee colony (ABC)-based algorithm was developed to detect overused nodes and then move one or even more VMs from them to minimize their use. This then detects underserved hosts and migrates all VMs assigned to them if necessary, and finally switches the unused host to sleep mode.

Ni et al. [30] present an algorithm for VM mapping. This takes into account several tools and seeks to alleviate the crowding of loads, based on a probabilistic framework for adjusting unbalanced loads. These instigators concentrate on concurrent users, and on the scheme. At the same time, multiple users can need the same resource from the same server, which will quickly overload the target host and lead to degradation of performance. The algorithm applies Proportional Selection to evaluate each host's selection probability where the host is best placed and would most likely admit VMs [31].

Tordsson et al. [32] present a virtual machine optimization scheme in a multi-cloud environment. The proposed algorithm for optimizing VM placement is intended for a multi-objective schedule that involves workload balance, performance, and expenses. As in a multidomain, different CSPs support various infrastructures and have specific types of VMs; the searchers put a lot of effort into managing multi-cloud heterogeneous resources.

To distribute the load within the intra-cloud, Yi and Wenlong [33] present a distributed load balancing algorithm based on comparison and balance (DLBA-CAB) via an effective adaptive migration of VMs. The purpose of this technique is to ensure that each host achieves a balance of processor use and I/O use. The authors model a cost function taking into consideration the weighted use of the Processor and I/O, and each host decides individually the values of that function.

## ***2.1 Problem Formulation***

In the field of networking and communication in data centers and cloud computing system, scheduling  $n$  tasks into  $m$  nodes is a complete problem. It even becomes complex when considering the federated cloud. We have to allocate the entire process to a host in such a way that cloud users can execute their tasks in minimum time and maximum average usage of resources. Although there are many works that tackle the load balance in cloud computing, fewer are those that concentrate on federated cloud load balancing. The current research proposes an improved technique of splitting workloads in order to obtain better results. It applies the Ant Colony based-algorithm to determine the suitable node, within the federation, to correctly balance the workload.

## **3 Methodology**

In this work, we design a load balancing algorithm and demonstrate its application in a federated cloud. We will test the proposed algorithm in a simulation tool that will allow us to evaluate the competitiveness of the algorithm. Furthermore, we show how it should be integrated and combined with the supervising components of the federated cloud. The big concern we face is to show that this algorithm is competitive, especially in the context of a federation of clouds.

### 3.1 *CloudSim*

Simulation methods are suitable alternative methodologies for testing algorithms, software, and policies before the cloud and high-performance computing technologies production mode [34]. Simulation tools open the possibility of evaluating different experiments in an environment where we can perform tests. CloudSim is a simulator that enables testbeds to be built in Java programming language using concepts such as task loading and VM computing capabilities. CloudSim helps us to perform algorithms of allocation and migration policies of different VMs. The simulator can simulate and model large-scale noisy computing data centers and federated clouds, as well as virtualized server hosts and different rules for delivering virtual machine host services, containers, energy-conscious computing resources, and dynamic simulation features [35].

### 3.2 *Suggested Load Balancing Algorithm*

Load balancing [38, 39] is accomplished in two steps within a cloud environment: the first is to spread the task between the node, and the second is to control the virtual machine and perform the necessary operations to balance the loads. To adjust the loads, we use a modified well-known Ant Colony Optimization (ACO) algorithm. ACO is a method for tackling combinatorial optimization issues based on biological fundamentals [36]. Marco Dorigo first mentioned it in his Ph.D. [37], dissertation in the early '90 s with the intention of determining the best path across a graph. In the natural world, a colony of ants would seek the shortest route between its food source and its nest, and this is how the artificial ACO algorithm works. Using artificial ants that mimic wild ants' foraging behavior, the ACO algorithm tries to figure out what the problem is. The growing ant leaves the nest in search of a food source, then returns to the nest to finish an iteration. A material called pheromone is left behind by growing ants along the path they travel. The distance to the route and the quality of the available food supply affect the amount of pheromones in each route. There's a significant factor in route selection: the number of pheromones on the trail. If there are too many pheromones on a path, more ants will be attracted to it. Using pheromone concentration and other heuristic meanings, each ant probabilistically selects a path. The simplified ACO-algorithm pseudocode is as follows (Table 1):

Adopting ACO's problem-solving paradigm, we propose an incremental load balancing algorithm with the objective of determining the workload of each node in the federation and balancing the workload among all the federation nodes. We consider two sets of mobile agents for creating call connections in the circuit-switched network. Those two groups create, exploit, and monitor their own routing tables to establish links between federation gateways. Each collection of mobile agents in the ACO algorithm fits a colony of ants, and the routing table for each group corresponds to a table of pheromones for each colony. And though both classes of mobile entities

have their own mapping intentions, they also take into account the other party's routing preferences.

Figure 1 shows that all of the awnings have characteristics such as the ant walking path being a guided acyclic graph and that all of the ants begin at point  $N_s$  and choose the next node based on the itinerary transition likelihood. The probability of the direction depends on the distance from the track and the amount of pheromone in the trac. When a particular ant walks to the endpoint  $N_e$ , the ant has completed its journey (all nodes traversed), and the ants leave a pheromone on the way to walk. The ACO algorithm is improved, has high precision, and is used for the optimal solution of load balancing. Figure 2 shows the flowchart of the suggested algorithm.

## 4 Simulation Results

Experiments are driven to test the performance of the proposed load balancing algorithm. This describes the simulation setup as well as performance and efficiency metrics.

### 4.1 Simulation Setup

By importing CloudSim into the NetBeans Integrated Development Environment (IDE), we design a model for the simulation and check the proposed load balancing algorithm. The federated cloud system model and related load balancing model are set up, and the ACO algorithm is introduced to address the load balancing problem. Table 2 lists the parameters of the device model and the simulation.

**Table 1** System algorithm

<b>Algo: Ant Colony Optimization</b>
The device parameters are initialized
<b>While</b> <i>Condition for termination not met</i> <b>do</b>
<i>Build solutions</i>
<i>Apply Scanning Path</i>
<i>Update pheromones</i>
<b>End</b>

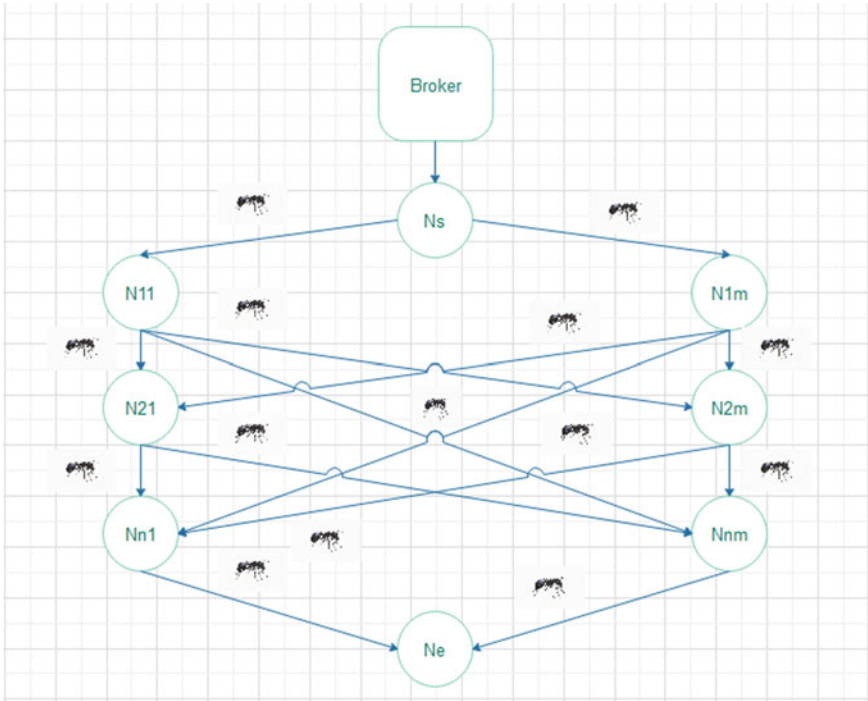


Fig. 1 The ACO algorithm structure towards workload scheduling

### 4.2 Performance Metrics

We adopt three performance metrics. First, the time of completion, it refers to the time of processing of all tasks. Secondly, energy use relates to energy processing for all tasks. The third factor is reliability, which relates to the task’s success rate within the constraints of the maximum tolerance time and available residual energy. We compare the performance metrics with the performance metrics of a single cloud architecture with ten nodes using the improved ACO algorithm.

#### 4.2.1 Completion Time Evaluation

The completion time is used to check the load balancing algorithm performance. Figure 3 shows the result.

As the number of tasks grows, so does the time it takes to do them. However, the difference in completion time is evident between the federated cloud environment and the single cloud environment.

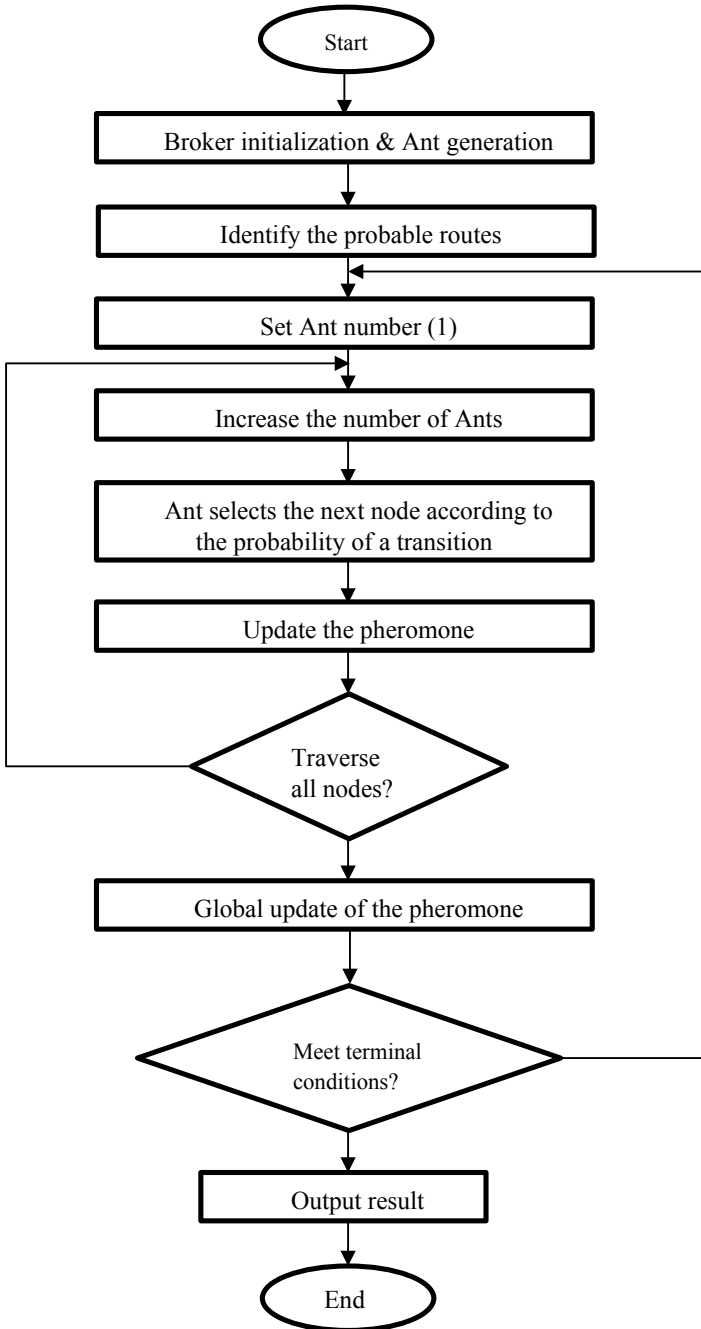
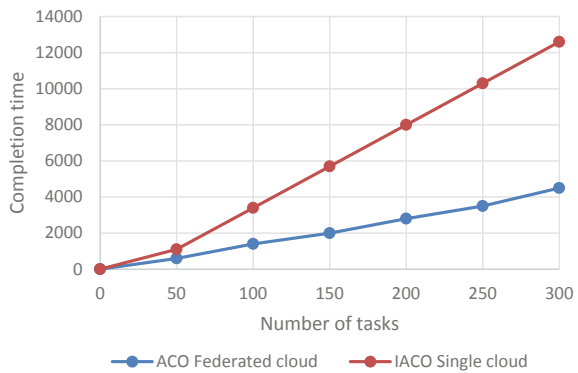


Fig. 2 Flowchart of the algorithm

**Table 2** Simulation parameter values

Parameters	Values
Number of clouds	10
Number of nodes	100
The computing capacity of each node	1–2 G cycles/s
Computing intensity	300 cycles/bit
Link bandwidth	100 MHz
Data size	10–50 MB
Real-time satisfaction weight	0.7
Maximum iteration	100
Accelerate factor	2
Pheromone weight coefficient	1
Pheromone Volatilization coefficient	0.5

**Fig. 3** Completion time comparison graph



### 4.2.2 Evaluation of Energy Consumption

The simulation of energy use results of the two algorithms in both environments is shown in Fig. 4.

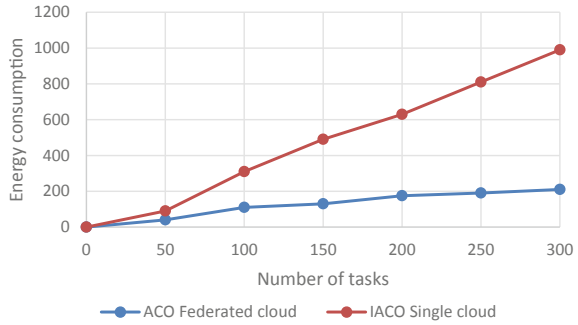
It can be concluded from Fig. 4 that the lowest energy use is within the federated cloud environment.

### 4.2.3 Evaluation of Reliability

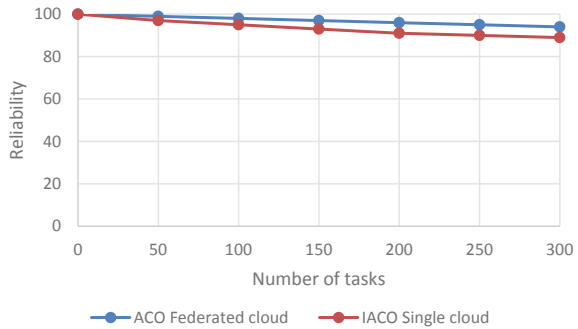
Figure 5 shows the results of the simulation of the load balancing algorithm’s reliability in both different environments.

The proposed algorithm performs best compared to the improved ACO algorithm in single cloud architecture, given completion time, power consumption, and reliability.

**Fig. 4** Energy consumption comparison graph



**Fig. 5** Reliability comparison graph



## 5 Conclusion and Future Work

We developed an incremental load balancing algorithm to make effective use of federated cloud resources. In a cloud environment, there are actually several algorithms, like hybrid algorithms, heuristic-based algorithms, metaheuristic-based algorithms, conventional approach-based algorithms, etc. Each algorithm works on various parameters like completion time, response time, use of resources, reliability, energy consumption, output, etc. We proposed a heuristic approach to workload balancing in a federated cloud environment by looking for the optimum solution. The load-balancing algorithm being evaluated is validated by simulation and compared to an improved algorithm in the single cloud environment. We demonstrated that using three metrics of performance.

Future research work must focus on the average ratio of resource utilization in the federated cloud by taking into consideration the makespan times. Moreover, researchers must focus on providing a hybrid-based approach to enhance the ability to balance workload in a federated cloud architecture.

**Acknowledgements** The authors appreciate the Covenant University through its Centre for Research, Innovation and Discovery for Financial assistance.



## References

1. Kumar S, Singh D (2015) Various dynamic load balancing algorithms in cloud environment: a survey. *Int J Comput Appl* 129(6):14–19. <https://doi.org/10.5120/ijca2015906927>
2. Chiregi M, Jafari Navimipour N (2018) Cloud computing and trust evaluation: a systematic literature review of the state-of-the-art mechanisms. *J Electr Syst Inf Technol* 5(3):608–622. <https://doi.org/10.1016/j.jesit.2017.09.001>
3. Rajabion L, Shaltook AA, Taghikhah M, Ghasemi A, Badfar A (2019) Healthcare big data processing mechanisms: the role of cloud computing. *Int J Inf Manag (Elsevier Ltd)* 49:271–289. <https://doi.org/10.1016/j.ijinfomgt.2019.05.017>
4. Khodaei H, Hajiali M, Darvishan A, Sepehr M, Ghadimi N (Jun. 2018) Fuzzy-based heat and power hub models for cost-emission operation of an industrial consumer using compromise programming. *Appl Therm Eng* 137:395–405. <https://doi.org/10.1016/j.applthermaleng.2018.04.008>
5. El-Seoud SA, AboGamie EA, Salama M (Apr. 2017) Integrated Education Management System via Cloud Computing. *Int J Interact Mob Technol* 11(2):24. <https://doi.org/10.3991/ijim.v11i2.6560>
6. Vakili A, Navimipour NJ (2017) Comprehensive and systematic review of the service composition mechanisms in the cloud environments. *J Netw Comput Appl* 81:24–36. <https://doi.org/10.1016/j.jnca.2017.01.005>
7. Abedinia O, Bekravi M, Ghadimi N (2017) Intelligent controller based wide-area control in power system. *Int J Uncertain, Fuzziness Knowl-Based Syst* 25(01):1–30. <https://doi.org/10.1142/S0218488517500015>
8. Liaqat M et al (Jan. 2017) Federated cloud resource management: review and discussion. *J Netw Comput Appl* 77:87–105. <https://doi.org/10.1016/j.jnca.2016.10.008>
9. Rochwerger B et al (2011) An Architecture for Federated Cloud Computing. In: *Computing C (ed) Hoboken. John Wiley & Sons Inc., NJ, USA*, pp 391–411
10. Buyya R, Yeo CS, Venugopal S, Broberg J, Brandic I (2009) Cloud computing and emerging IT platforms: vision, hype, and reality for delivering computing as the 5th utility. *Futur Gener Comput Syst* 25(6):599–616. <https://doi.org/10.1016/j.future.2008.12.001>
11. Ahmad A, Alzahrani AS, Ahmed N, Ahsan T (2020) A delegation model for SDN-driven federated cloud. *Alexandria Eng J*. <https://doi.org/10.1016/j.aej.2020.06.018>
12. M. Liaqat et al. (2017) Federated cloud resource management: Review and discussion. *J Netw Comput Appl (Academic Press)* 77:87–105. <https://doi.org/10.1016/j.jnca.2016.10.008>
13. M. Armbrust et al. (2010) A view of cloud computing. *Commun ACM* 53(4):50–58. <https://doi.org/10.1145/1721654.1721672>
14. Devaraj AFS, Elhoseny M, Dhanasekaran S, Lydia EL, Shankar K (2020) Hybridization of firefly and improved multi-objective particle swarm optimization algorithm for energy efficient load balancing in cloud computing environments. *J Parallel Distrib Comput* 142:36–45. <https://doi.org/10.1016/j.jpdc.2020.03.022>
15. Hsieh HC, Chiang ML (2019) The incremental load balance cloud algorithm by using dynamic data deployment. *J Grid Comput* 17(3):553–575. <https://doi.org/10.1007/s10723-019-09474-2>
16. Buyya R, Pandey S, Vecchiola C (2009) Cloudbus toolkit for market-oriented cloud computing. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) (LNCS)* 5931:24–44. [https://doi.org/10.1007/978-3-642-10665-1\\_4](https://doi.org/10.1007/978-3-642-10665-1_4)
17. Yu J, Buyya R, Ramamohanarao K (2008) Workflow scheduling algorithms for grid computing. *Stud Comput Intell* 146:173–214. [https://doi.org/10.1007/978-3-540-69277-5\\_7](https://doi.org/10.1007/978-3-540-69277-5_7)
18. Chaudhary D, Kumar B (2014) An analysis of the load scheduling algorithms in the cloud computing environment: a survey. In *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, pp 1–6. <https://doi.org/10.1109/ICIINFS.2014.7036659>
19. Pacini E, Mateos C, García Garino C (2014) Distributed job scheduling based on Swarm Intelligence: a survey. *Comput Electr Eng* 40(1):252–269. <https://doi.org/10.1016/j.compeleceng.2013.11.023>

20. Rahmeh OA, Johnson P, Taleb-Bendiab A (2020) A dynamic biased random sampling scheme for scalable and reliable grid networks. *INFOCOMP J Comput Sci* 7(4):1–10. <http://professores.dcc.ufla.br/ojs/index.php/INFOCOMP/article/view/233%5Cnhttp://professores.dcc.ufla.br/ojs/index.php/INFOCOMP/article/download/233/218%5Cnhttp://professores.dcc.ufla.br/ojs/index.php/INFOCOMP/article/view/233/218>. Accessed 06 Jul 2020
21. Rashedi E, Zarezadeh A (2014) Noise filtering in ultrasound images using Gravitational Search Algorithm. In: 2014 Iranian Conference on Intelligent Systems (ICIS), pp 1–4. <https://doi.org/10.1109/IranianCIS.2014.6802559>
22. Khatibinia M, Khosravi S (2014) A hybrid approach based on an improved gravitational search algorithm and orthogonal crossover for optimal shape design of concrete gravity dams. *Appl Soft Comput* 16:223–233. <https://doi.org/10.1016/j.asoc.2013.12.008>
23. Mishra R (Apr. 2012) Ant colony optimization: a solution of load balancing in cloud. *Int J Web Seman Technol* 3(2):33–50. <https://doi.org/10.5121/ijwest.2012.3203>
24. Nishant K et al. (2012) Load balancing of nodes in cloud using ant colony optimization. In: 2012 UKSim 14th International Conference on Computer Modelling and Simulation, pp 3–8. <https://doi.org/10.1109/UKSim.2012.11>
25. Chen JL, Larosa YT, Yang PJ (2012) Optimal QoS load balancing mechanism for virtual machines scheduling in eucalyptus cloud computing platform. In: 2012 2nd Baltic Congress on Future Internet Communications, BCFIC 2012, pp 214–221. <https://doi.org/10.1109/BCFIC.2012.6217949>
26. Mesbahi M, Masoud Rahmani A (2016) Load balancing in cloud computing: a state of the art survey. *Int J Mod Educ Comput Sci* 8(3):64–78. <https://doi.org/10.5815/ijmecs.2016.03.08>
27. Tan J, Fu X (2017) Addressing hardware reliability challenges in general-purpose GPUs. In: *Advances in GPU Research and Practice*. Elsevier Inc., pp 649–705
28. Ma F, Liu F, Liu Z (2012) Distributed load balancing allocation of virtual machine in cloud data center. In: *ICSESS 2012—Proceedings of 2012 IEEE 3rd International Conference on Software Engineering and Service Science*, 2012, pp 20–23. <https://doi.org/10.1109/ICSESS.2012.6269396>
29. Ghafari SM, Fazeli M, Patooghy A, Rikhtechi L (2013) Bee-MMT: a load balancing method for power consumption management in cloud computing. In: 2013 6th International Conference on Contemporary Computing, IC3 2013, pp 76–80. <https://doi.org/10.1109/IC3.2013.6612165>
30. Ni J, Huang Y, Luan Z, Zhang J, Qian D (2011) Virtual machine mapping policy based on load balancing in private cloud environment. In: 2011 International Conference on Cloud and Service Computing, pp 292–295. <https://doi.org/10.1109/CSC.2011.6138536>
31. Xu M, Tian W, Buyya R (Jun. 2017) A survey on load balancing algorithms for virtual machines placement in cloud computing. *Concurr Comput* 29(12):e4123. <https://doi.org/10.1002/cpe.4123>
32. Tordsson J, Montero RS, Moreno-Vozmediano R, Llorente IM (Feb. 2012) Cloud brokering mechanisms for optimized placement of virtual machines across multiple providers. *Futur Gener Comput Syst* 28(2):358–367. <https://doi.org/10.1016/j.future.2011.07.003>
33. Yi Z, Wenlong H (2009) Adaptive distributed load balancing algorithm based on live migration of virtual machines in cloud. In: *NCM 2009—5th International Joint Conference on INC, IMS, and IDC*, pp 170–175. <https://doi.org/10.1109/NCM.2009.350>
34. Goyal T, Singh A, Agrawal A (Jan. 2012) Cloudsim: simulator for cloud computing infrastructure and modeling. *Proc. Eng.* 38:3566–3572. <https://doi.org/10.1016/j.proeng.2012.06.412>
35. Barbierato E, Gribaudo M, Iacono M, Jakóbič A (2019) Exploiting CloudSim in a multiformalism modeling approach for cloud based systems. *Simul Model Pract Theory* 93:133–147. <https://doi.org/10.1016/j.simpat.2018.09.018>
36. Mandloi M, Bhatia V (2017) Symbol detection in multiple antenna wireless systems via ant colony optimization. *Handb Neural Comput*, pp 225–237
37. Colomi A, Dorigo M, Maniezzo V (1992) An investigation of some properties of an ‘Ant algorithm. Elsevier Publishing.

38. Kazeem Moses A, Joseph Bamidele A, Roseline Oluwaseun O, Misra S, Abidemi Emmanuel A (2021) Applicability of MMRR load balancing algorithm in cloud computing. *Int J Comput Math: Comput Syst Theory* 6(1):7–20
39. Shri, Devi ML, Balusamy EG, Kadry B, Misra S, Odusami M (2018) A fuzzy based hybrid firefly optimization technique for load balancing in cloud datacenters. In: *International Conference on Innovations in Bio-Inspired Computing and Applications*. Springer, Cham, pp 463–473

# An Intelligent Hydroponic Farm Monitoring System Using IoT



Jalani H. Naphtali, Sanjay Misra, John Wejin, Akshat Agrawal,  
and Jonathan Oluranti

**Abstract** Traditional farming is a process in which man works the soil for planting seeds or rearing animals, either to provide food for himself and his family or for business. The traditional farming system faces myriads' challenges ranging from difficulty in getting good soil, high cost of plowing and weed removal, difficulty in pest control, and plant monitoring. To solve these myriad's challenges, technological innovation has ushered in the various eco-friendly system of farming. Prominent amongst this new method of farming is hydroponics. Hydroponics uses nutrients in solution instead of soil to grow a plant. Practicing such a system of farming among peasant farmers from developing countries demands huge capital investment. In this paper, we propose a cost-effective and free resource Internet of Things (IoT) based system that smartly monitors a hydroponic farm. The designed solution uses an ATmega328P microcontroller connected to a sensor network with connectivity to the internet. These sensors read the humidity, temperature, pH value, dissolved solids, and water level of the farm for efficient monitoring. In addition, the system provides an alert to a remote user through email, and a buzzer sound for on-site supervisors when there is a change in measured parameters that may be detrimental to plant growth and yield.

**Keywords** IoT · Sensor network · Hydroponic

---

J. H. Naphtali · J. Wejin · J. Oluranti  
Center of ICT/ICE Research, Covenant University, Ogun State, Ota, Nigeria  
e-mail: [john.wejinpgs@stu.cu.edu.ng](mailto:john.wejinpgs@stu.cu.edu.ng)

J. Oluranti  
e-mail: [jonathan.oluranti@covenantuniversity.edu.ng](mailto:jonathan.oluranti@covenantuniversity.edu.ng)

S. Misra  
Department of Computer Science and Communication, Østfold University College, Halden,  
Norway  
e-mail: [sanjay.misra@hiof.no](mailto:sanjay.misra@hiof.no)

A. Agrawal (✉)  
Amity University Haryana, Gurgaon, India  
e-mail: [akshatag20@gmail.com](mailto:akshatag20@gmail.com)

## 1 Introduction

Traditional farming is a process in which man works the soil for planting seeds or rearing of animals to provide food for himself and his family, or commercially for business. The traditional farming system faces myriad challenges ranging from difficulty in getting good soil, high cost of plowing and weed removal, difficulty in pest control and plant monitoring, and possible difficulty in land availability due to the high estimation of 9.6 billion human population in the year 2050 [1]. To solve these myriads' challenges, technological innovation has brought the new eco-friendly and cost-effective system of farming. Prominent among such farming innovations is Hydroponic farming.

Hydroponic is coined from the Greek words “Hydro” meaning water, and “ponic” meaning working. This system of farming uses nutrients in solution instead of using soil to grow plants [2]. Hydroponic farming is rapidly been adopted by so many individuals and companies because it provides; (1) yields within a shorter period compared to the conventional system of farming, (2) no soil needed, (3) occupies less space, (4) not location dependent, (5) not limited by season, and (6) complete minerals for growth [3]. To handle the needed monitoring of the hydroponic farming system for maximum yield, an intelligent Internet of Things (IoT) system becomes paramount.

Internet of things (IoT) makes it possible to connect different devices via the internet and smartly controlling them for the effective remote device-to-device and device-to-human communication without human intervention [4]. In this paper, we put forward a low-cost intelligent IoT-based system for monitoring a hydroponic farm. The proposed system monitors the condition of the plant solution and automatically takes the necessary steps needed to maintain normalcy without any human intervention.

The paper is organized into five sections. The next Sect. 2 provides the fundamentals and background of the work. Section 3 proposes the proposed model. Methodology, results and discussion, and conclusion and future works are summarized in Sects. 4–6.

## 2 Categories of Hydroponics

Hydroponic farming systems are classified based on the techniques employed. The common types as depicted in Fig. 1 are as follows; (1) Wick, (2) Nutrient film technique, (3) Deep Water Culture (DWC), (4) Aeroponics, (5) Ebb and Flow, and (6) Drip.

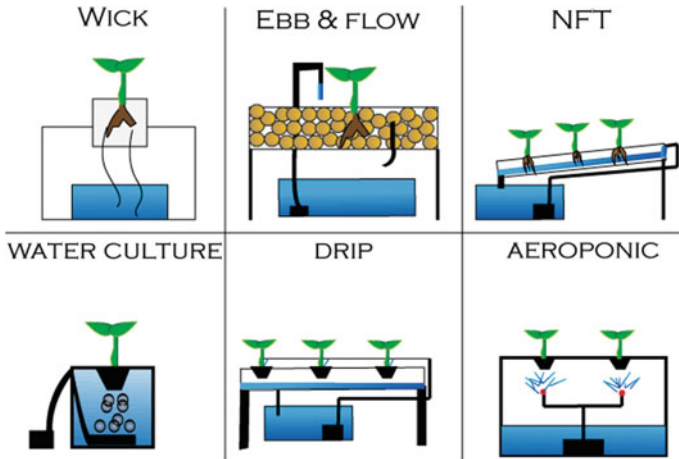


Fig. 1 Types of hydroponics

### 2.1 Wick Hydroponic

This type of technique uses a wick to supply nutrient solutions to the plants. Due to the small nature of the wick, wick hydroponic is best suited for plants with lesser need of water.

### 2.2 Nutrient Film Technique (NFT)

NFT supplies plant nutrients to a tilted growth tray. The solution is drained back to the reservoir using the other end of the tray. Here, the system needs no timer and growth medium and requires no air pump. Plants grown using this method are susceptible to wilting [5].

### 2.3 Deep Water Culture (DWC) Hydroponics

This is also called “water culture” hydroponics. The roots of plants are wholly submerged in the solution and provided with oxygen through air stones and air pumps. This technique is suited for growing lettuce and other plants that need much water [6].

## **2.4 *Aeroponics***

In this technique the roots are suspended above the solution in the reservoir. Intermittent spray of solution to the roots is done by a pump and as such consumes less nutrients.

## **2.5 *EBB and Flow (Flood and Drain) Hydroponics***

This system uses a timer-based pump to splash up nutrients solution and oxygen to plants. Plant roots are prone to be infected with diseases easily in this system [7].

## **2.6 *Drip Hydroponic:***

As the name applies, this method uses drip to supply nutrients from the reservoir to the plants. As nutrients are supplied to the plants, nutrients are re-injected back to the reservoir (recovery drain hydroponics), or drained through the process of evaporation (non-recovery hydroponics) [8].

# **3 Intelligent Hydroponic Monitoring System**

Intelligent hydroponic systems are smart systems that require no manpower for an effective hydroponic that provides a large yield or throughput. Various researchers have put forward various approaches to monitoring. In [9] a data acquisition system for hydroponics farming was developed. The system monitors vital parameters such as air temperature, relative humidity, water temperature, water level, pH level, and light intensity. Owing to the large amount of data generated by sensors, the use of SD cards for storage in this system becomes a weakness. Researchers in [10], implemented a system that monitors the hydroponic environment. In this system the management of the farm solely depends on the farmer as no actuators were used. To avoid the need of recalibration of pH sensors, in [11] developed an auto-calibration-based system for measuring pH. Though this system is effective, the reliability reduces due to the large use of moving parts. In [12] a system that uses smartphones to monitor temperature and water level was proposed.

The authors in [13], developed smart hydroponic farming that is being monitored using telegram messenger. The system uses various sensors to automatically monitor the vital parameters. Research is going on for improving the productivity in various types of farming [14–17], due to the limitation of the work(space), we are not discussing and including them in details.

## 4 Methodology

- Getting Materials.
- Soldering of component on veror board.
- Constructing the outer frame.
- Designing the app interface and database.
- Testing the system.

### 4.1 Hardware and Software Components

#### Hardware Requirement:

The hardware circuit diagram drawn with frizzing is shown in Fig. 2 and some of the key components are explained below.

*Power Supply Unit:* This unit is made up of a 12 V, 4.5A step-down transformer that steps down 220–240 V Alternating Current (AC) to 12 V. We used other discrete devices such as bridge rectifier, Ac varistor, transistors, switch, and diodes to provide the needed power requirements for the microcontroller, sensors, and display units.

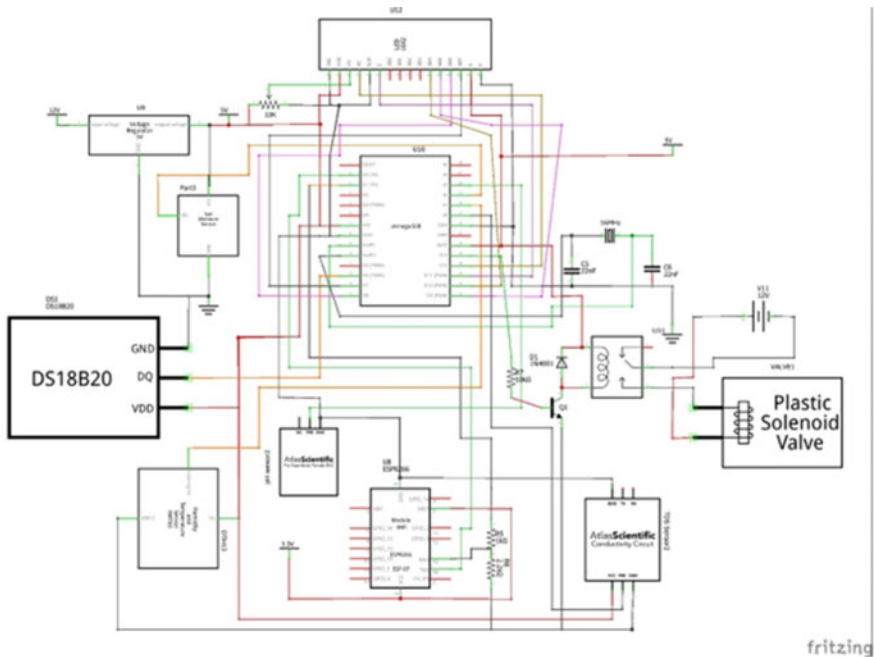


Fig. 2 Circuit diagram showing connection with MCU



*Atmega328P-PU microcontroller:* The Atmega328p-pu microcontroller is a 28-pin AVR chip used to program, sense, monitor, and control all the sensors used in this research.

*Liquid Crystal Display (LCD):* We used a low-cost  $20 \times 4$  LCD display in this project. It was used in a 4-bit configuration interfacing with the MCU. The LCD displays all the commands made and decisions are taken by the 'brain', the microcontroller unit (MCU) is displayed.

*pH sensor:* We used the E250 pH sensor module to monitor the pH level of the nutrient solution. It uses 5.00 V, has a pH and temperature measuring range of 0–14 pH, and 0–60 °C respectively, and a response time of less than 1 min.

*TDS Sensor:* In place of the Electrical Conductivity (EC) sensor, we used a low-cost but efficient TDS sensor. This sensor was used to measure the cleanliness of the water solution and quality test. It supports a 3.3–5 v voltage range, the current range of 3–6 mA, TDS measurement range of 0–1000 ppm. This makes it compatible with most microcontroller boards.

*Relative Humidity and Temperature Sensor (DHT11):* This is ultra-low-cost moisture and temperature sensor. It is made up of two subunits, a thermistor that measures temperature and a capacitive humidity unit for measuring humidity. It supports a 3–5 v range, 2.5 mA current, 20–80% humidity readings, and 0–50 °C temperature range.

*Water Level Sensor:* This sensor uses an analog to digital converter (ADC) to read the level of voltage at the conductive electrode probes and maps the reading on the microcontroller. It supports 3–5 v voltage levels and a mapping range of 0–1023.

## Software Requirement

*ThingSpeak:* In this project, we used Thingspeak as a platform for accessing, analyzing, and visualizing data recorded by the sensors. This is a free license (IoT) application and API used in storing and retrieving data from things via HTTP and MQTT protocol over the Internet. It enables logging creation for sensors that enables tracking, aggregation, analyses, and visualization of data. The sensors were programmed and codes uploaded to the thingspeak platform as an AIA file. In case of any parameter abnormality or threshold, we programmed the system to send an email alert and trigger a buzzer. The algorithm for our proposed system is shown table while the block diagram is shown in Fig. 3.

The proposed system uses an Atmega328P microcontroller to process data coming in from the sensors for onward delivery to the data bank on ThingSpeak. Our system uses a two-way system communication strategy for monitoring. The first phase is communication between the sensors and the environment. The second communication pathway is between the microcontroller unit and the ThinkSpeak IoT-based data telemetry. The user gets an update about the hydroponic farm condition from the Thinspeak repository via an app and gets an email and a buzzer sound at the farm sit in case of any abnormality (Table 1).

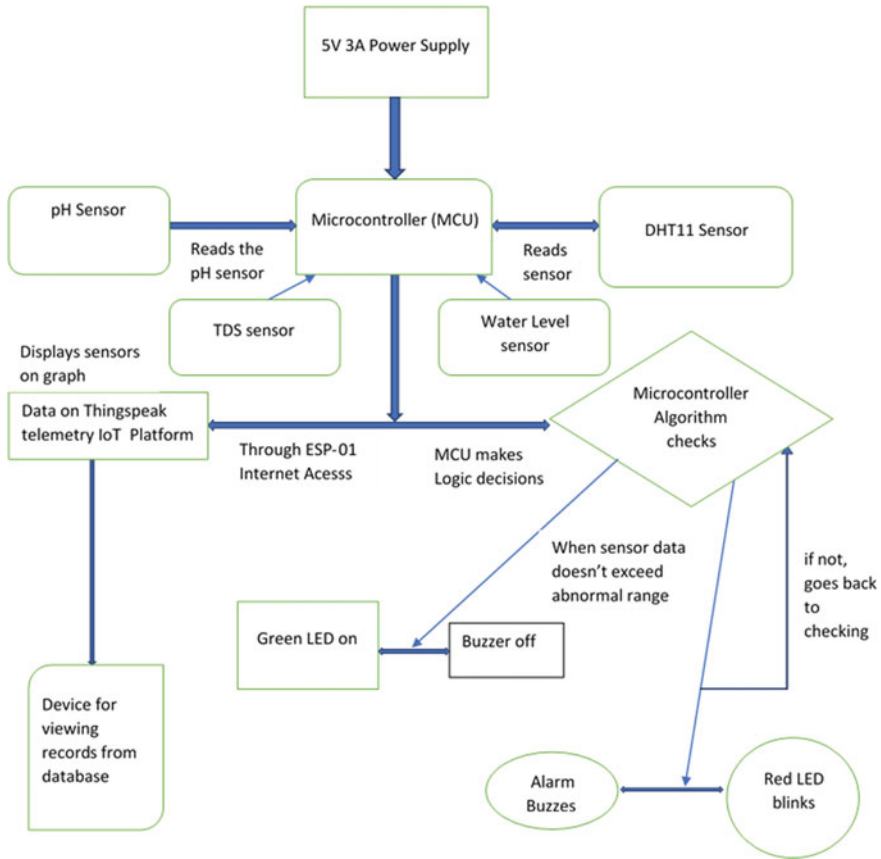


Fig. 3 Block diagram of the proposed system

Table 1 System algorithm

5. begin
6. connect microcontroller to internet
7. read data from sensor
8. upload data to thingspeak database
9.     if sensor value >= threshold
10.         send email and initialize buzzer
11.     else
12.         goto step 3
13.     end

## 5 Results

This research monitors four key parameters that are vital for the huge yield of plants grown using hydroponic systems. These parameters are water level, pH value, temperature, and humidity. The cased system as shown in Fig. 4 below was designed. Data



**Fig. 4** The proposed system in a case

were collected, analyzed, and visualized using the thingspeak open-source platform. Figure 5 shows the MCU writing sensor readings to the thingspeak database. Figure 6 shows the air temperature reading from the DTH11 sensor. It indicates a reading range of 25–28 °C, with 26 °C as optimum for plant growth. Figure 7 indicates an optimum pH level throughout the monitoring period, while Fig. 8 shows the readings from the TDS sensors. The highest reading of 42.5 ppm by around 18:27, and a drop to about 37.5 ppm at the end of the day was observed. This drop is due to the addition of more water into the system.

## 6 Discussion

In this paper, a cost-effective IoT monitoring system for the hydroponic farm is proposed. The low cost of this system is a result of the careful choice of equipment used. The DTH11 sensor is a very low-cost sensor, but yet efficient in measuring temperature. The ATMEGA328P is a low-cost and versatile microcontroller. It uses RISC and has a high-temperature resistant feature, making it ideal for extreme temperature measurement. These lead to a reduction in the total cost of the monitoring system. Table 2 below gives the comparison in terms of cost between our proposed

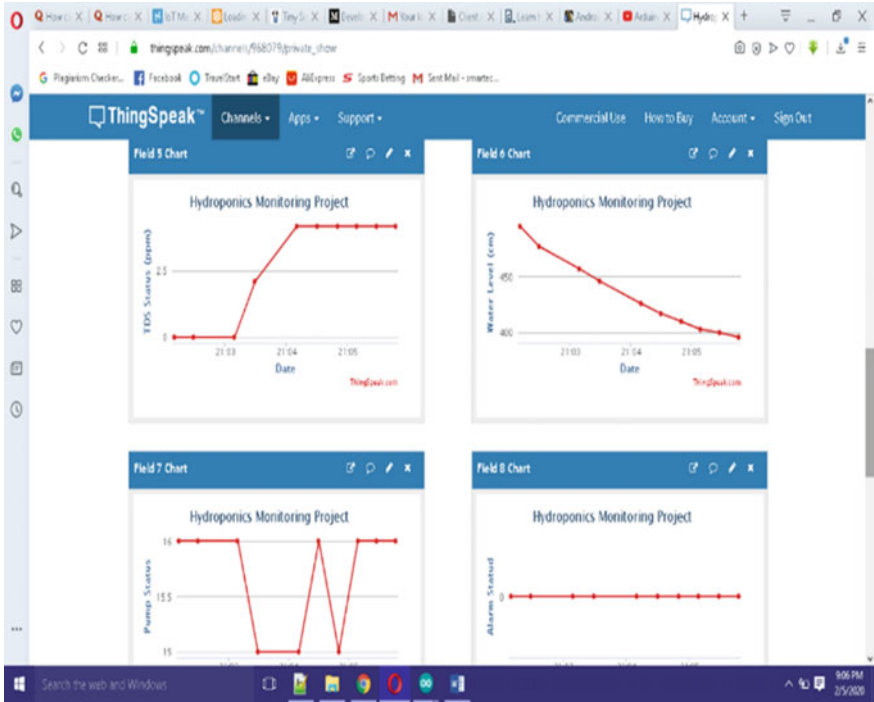


Fig. 5 MCU uploading readings to Thingspeak

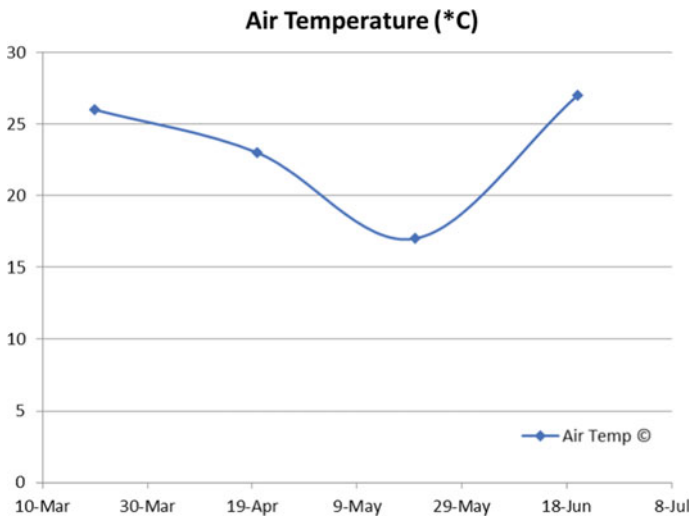


Fig. 6 Air temperature reading from DTH11 sensor

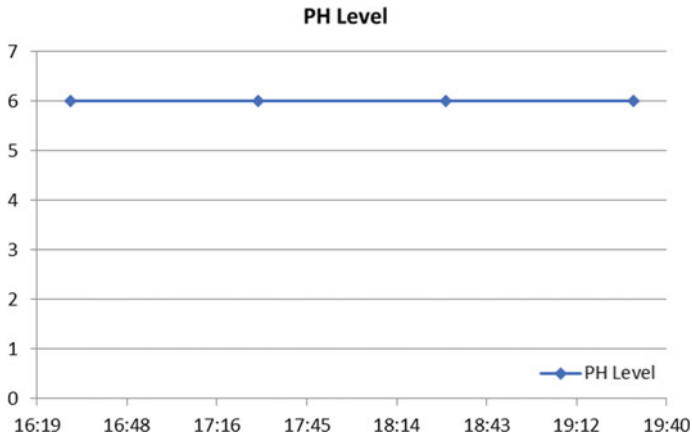


Fig. 7 pH level reading

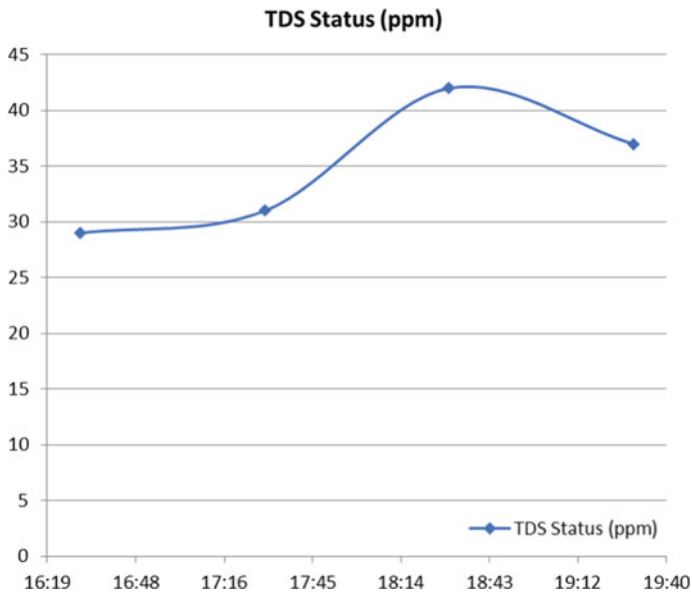


Fig. 8 Reading showing dissolved solid in solution with time

system and those in the literature. It is evident that Raspberry Pi and Arduino mega 2560 provide good computation power, they are costlier compared to the microcontroller used in our proposed system, and need external ADC devices to get analog signals for measurement. This shows that the proposed system is cost-effective.

**Table. 2** Controller price comparison

Paper	Features	Intelligent device	Price
[9]	High computation power	DFRobot Arduino Mega 2560	\$16.99
[10]	Good computation but no ADC	Raspberry pi	\$49.8
Proposed system	Good computational power and Integrated ADCs	ATmega328P	\$3.8

## 7 Conclusion and Future Works

In this paper, we investigate the studies on the methods used in monitoring hydroponic farming. A low-cost but intelligent monitoring system for hydroponic is designed. An implementation of the prototype shows the performance of this system. Various sensors, a controller, and a monitoring algorithm are described. A low-cost strategy to reduce the cost of the instruments is undertaken. The experimental testbed shows that the high performance of a hydroponic farm in respect to air temperature, water level, relative humidity, pH level, and dissolved solids can be achieved with low-cost instruments.

Integrating various types of crops in a hydroponic farm, and implementing a low-cost monitoring system that leverages IoT and machine learning to enhance, generate reports, and graph for the various crops in the farm will be a good future research endeavor.

## References

1. Bakhtar N, Chhabria V, Chougale I, Vidhrani H, Hande R (2018) IoT based hydroponic farm. In: 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp 205–209. <https://doi.org/10.1109/ICSSIT.2018.8748447>
2. Modu F, Adam A, Aliyu F, Mabu A, Musa M (2020) A survey of smart hydroponic systems. *Adv Sci Technol Eng Syst (ASTES)* 5(1):233–248. <https://doi.org/10.25046/aj050130>
3. Baras T (2018) *DIY Hydroponic gardens: how to design and build an inexpensive system for growing plants in water*. Cool Springs Press. <https://books.google.com.sa/books?id=rwlMDwAAQBAJ>
4. Al-fugaha A, Guizani M, Mohammadi M, Aledhari M, ayyash M (2015) Internet of Things: A survey on enabling technology, protocols and applications. *IEEE Commun Surv Tutor*, 17(4):2347–2376. <https://doi.org/10.1109/comst.2015.24440>
5. Ibarra MJ, Alcarraz E, Tapia O, Atencio YP, Mamani-Coaquira Y, Huilcen Baca HA (2020) NFT-I technique using IoT to improve hydroponic cultivation of lettuce. In: 2020 39th International Conference of the Chilean Computer Science Society (SCCC). Coquimbo, Chile, pp 1–7. <https://doi.org/10.1109/SCCC51225.2020.9281277>
6. Butcher JD, Laubscher CP, Coetzee JC (2017) A study of oxygenation techniques and the chlorophyll responses of pelargonium tomentosum grown in deep water culture hydroponics. *HortScience* 52(7):952–957
7. Jones J (2016) *Hydroponics: a practical guide for the soilless grower*. CRC Press. <https://books.google.com.ng/books?id=ybKBQAAQBAJ>

8. JOSHI N (2018) GREEN SPACES: CREATE YOUR OWN. Notion Press. <https://books.google.com.sa/books?id=CA9tDwAAQBAJ>
9. Tagle S, Pena R, Oblea F, Benozza H, Ledesma N, Gonzaga J, Lim LAG (2018) Development of an automated data acquisition system for hydroponic farming. In: 2018 IEEE 10th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM), pp 1–5.
10. Belhekar P, Thakare A, Budhe P, Shinde U, Waghmode V (2018) Automated system for farming with hydroponic style. In: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), pp 1–4.
11. Cambra C, Sendra S, Lloret J, Lacuesta R (2018) Smart system for bicarbonate control in irrigation for hydroponic precision farming. *Sensors* 18(5):1333
12. Sihombing P, Karina NA, Tarigan JT, Syarif MI (2018) Automated hydroponics nutrition plants systems using arduino uno microcontroller based on android. *J Phys: Conf Ser* 978:012014. <https://doi.org/10.1088%2F1742-6596%2F978%2F1%2F012014>
13. Sisyanto REN, Kurniawan NB et al. (2017) Hydroponic smart farming using cyber physical social system with telegram messenger. In: 2017 International Conference on Information Technology Systems and Innovation (ICITSI). IEEE, pp 239–245
14. Abayomi-Alli AA, Misra S, Akala MO, Ikotun AM, Ojokoh BA (2021) An Ontology-based information extraction system for organic farming. *Int J Semant Web Inf Syst (IJSWIS)* 17(2):79–99
15. Arogundade O, Qudus R, Abayomi-Alli A, Misra S, Agbaegbu J, Akinwale A, Ahuja R (2021) A mobile-based farm machinery hiring system. In: *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security*. Springer, Singapore, pp 213–226
16. Golubenkov A, Alexandrov D, Misra S, Abayomi-Alli O, Leon M, Ahuja R (2021) Decision support system on the need for veterinary control of passing livestock and farm produce. In: *Evolving Technologies for Computing, Communication and Smart World*. Springer, Singapore, pp 517–526
17. Abayomi-Alli O, Odusami M, Ojinaka D, Shobayo O, Misra S, Damasevicius R, Maskeliunas R (2018) Smart-solar irrigation system (SMIS) for sustainable agriculture. In: *Int Conf Appl Inform*. Springer, Cham, pp 198–212

# IoT and Machine Learning Based Anomaly Detection in WSN for a Smart Greenhouse



Molo Mbasa Joaquim, Abednego Wamuhindo Kamble, Sanjay Misra, Joke Badejo, and Akshat Agrawal

**Abstract** Agriculture is the most crucial sector which raises the economy of every country; several techniques have been developed to control and monitor the environment in which a particular crop is growing. Farmers need efficient support in terms of monitoring the temperature, the humidity, the water supply etc. However, the measurements provided by a wireless sensor network within a smart greenhouse are an essential aspect to take into consideration when it comes to evaluating the performance of sensor nodes used for controlling and monitoring the climatic condition (temperature, humidity, water supply, etc.). Therefore, this paper proposes a machine learning-based anomaly detection approach with the help of the DBSCAN algorithm of clustering to determine whether an unusual event has been found in the data. This approach allows farmers to ensure the reliability of the network. In this paper, we presented the description of the DBSCAN algorithm; we used an existing dataset that incorporates information about rose cultivation. With the used dataset, we introduced some noise, and we used MATLAB and Python to analyse and predict whether the introduced data is noise or not with DBSCAN. The performance of the algorithm after performing the prediction is 100% for two chosen features of the dataset and 75.4% for five features of the dataset in terms of precision.

**Keywords** Anomaly detection · Wireless sensor network · Smart greenhouse

---

M. M. Joaquim · J. Badejo  
Center of ICT/ICE, Covenant University, Ota, Nigeria  
e-mail: [joke.badejo@covenantuniversity.edu.ng](mailto:joke.badejo@covenantuniversity.edu.ng)

A. W. Kamble  
Politecnico Di Milano, Milan, Italy

S. Misra (✉)  
Department of Computer Science and Communication, Østfold University College, Halden, Norway  
e-mail: [sanjay.misra@hiof.no](mailto:sanjay.misra@hiof.no)

A. Agrawal (✉)  
Amity University, Gurgaon, India  
e-mail: [akshatag20@gmail.com](mailto:akshatag20@gmail.com)



# 1 Introduction

The demand in terms of food is continually growing day by day. Though, the sector of agriculture has to be enlarged to prevent the starvation and the need of food. To face several challenges such as soil salinity due to the excess of irrigation, the immoderate utilisation of fertilisers, pesticides and insecticides which makes the soil dependent on them, and different requirement of moisture, humidity, light wavelength, and temperature. Also, the lack of awareness of relevant information about the agricultural environment, plants cannot reach their maturity or can die earlier [1].

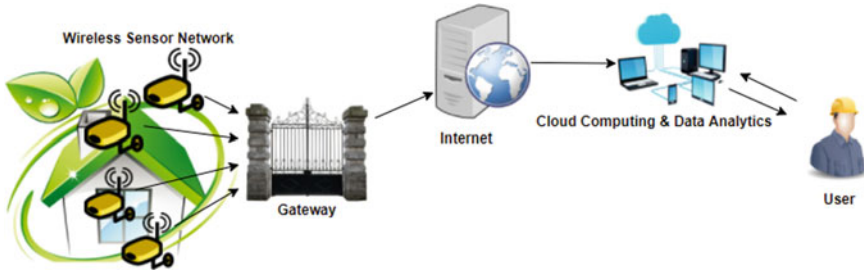
Nowadays, the greenhouse is the most popular industry of agriculture that allows farmers to separate crops from the farm so that it may have its environment [2]. The advantage of a greenhouse is that farmers can adjust the environmental condition according to the requirements of the crops. Thus, greenhouses are places of controlling physical inspection of plants; therefore, men and machines supervise this process to ensure that the plants are receiving all needs for their growth [3].

Internet of Things (IoT) has been developed in so many domains areas such as agriculture, health care, smart houses etc. Using IoT in agriculture eases farmers for monitoring plants in a large-scale in a way that it releases farmers to manual action for any control [4]. The IoT is intelligently connected systems and devices which offer the capability of sharing information across a platform by enabling interaction between smart machines through a sensor network within an environment developing a common goal to enable innovative application [5, 6]. Considering the use of IoT in a smart greenhouse, the WSN faces many challenges, among them we can cite: prolonged network lifetime, quality of service, fault tolerance, fidelity, network information processing and so on [7].

WSN refers to a large number of low-power, low-cost, low memory capacity and weak computer capability wireless nodes which helps to control and monitor any environment [8, 9]. WSNs are vulnerable; the primary concern about WSNs is that nodes may dwell in a dangerous or hostile environment where they are not well protected. Detecting sensors issues is very important so that it allows to increase the performance of the system. One of several ways of detecting abnormal functioning of a system is anomaly detection. Anomaly-based detection is the most popular technique that helps to detect anomalies based on the unusual behaviour of the system. It can identify abnormal patterns that are spectrally distinct from their surrounding neighbourhood, knowing that the system behaviour is strange [9, 10].

Following Fig. 1 is the example of the implementation of a smart greenhouse based on IoT.

The paper is structured as follows. The second section provides the state of art in the field. Methodology and experimentations are provided in the third section. Section 4 provides the result and discussion, and finally conclusion is drawn on the work in Sect. 5.



**Fig. 1** Example of smart greenhouse based on IoT

## 2 State of the Art

The use of IoT allows for solving so many issues in several domains. One previous work developed an environmental monitoring and disease detection in a smart greenhouse as in the reference [11]. In this work, a greenhouse has been transformed into a smart greenhouse by applying sensors and actuators within the greenhouse. The monitoring of the temperature, the humidity and water irrigation was implemented using sensors connected to a microcontroller (Arduino) linked with a software-based portal to receive the data from the sensory devices for analysis.

Similarly, the work discussed the image of the leaf of the plant is being captured by applying image processing and computer vision algorithms for early plant disease detection. In the same context as in the reference [11], as well as the reference [12] is also about performing the monitoring of the environment condition but they use the computer vision for detecting the tomato growth state by using a raspberry pi microcontroller both for the image analysis and IoT functions. The fundamental objective of the reference [12] is for detecting a tomato fruit's maturity state as ripe and sending a notification to the user either via an Android application or via email. Another monitoring in the reference [13] is for controlling water irrigation by collecting data from sensors by using Arduino Uno for data aggregation with the help of Zigbee for multi-hop communication and raspberry pi for the transmission of data with the support for Time Division Multiple Access (TDMA).

In Ref. [4] a monitoring system is implemented, and the tomato detection and growth classification have been performed by training the data with MATLAB R2017a. The reference [14] discusses the implementation of a smart greenhouse made of three operation modes: Time Based, Sensor-Based, and User-Based. A certain amount of module has been implemented to perform the operations previously cited, which are data collection, data processing, system configuration, IoT cloud, End-User Web Application.

In Ref. [15], the proposed model in the greenhouse uses the sensor network ZigBee having sensors like temperature, humidity and soil controls which allow to collect and analyse data sent to a cloud platform.

In Ref. [1, 16] and [17], the monitoring of the greenhouse uses a WSN and actuators controlled by a microcontroller for the collection and analysis of data. In

Ref. [18], a group sensor is uniformly distributed in the greenhouse depending on its size. One coordinator node manages all sensors and control actuators. Under normal conditions, sensors collect data in real-time, and when detecting parameters more or less from the limit, the coordinator turns on or off the actuator.

The Ref. [19] discusses the monitoring of the environment parameters (temperature, relative humidity and carbon-di-oxide) of a smart greenhouse. For the monitoring, sensors are connected to the embedded wireless system cc3200. Authors Ref. [20] presents a WSN which consists of MicaZ nodes used to perform the greenhouse measurement (temperature, light, pressure and humidity). Research are going on for improving the productivity in various types of farming [25–28], due to limitation of the work(space), we are not discussing an include them in details.

Considering the monitoring of greenhouses as discussed in state of the art, the management of the environment within the greenhouse takes account only of the values coming from the sensors assuming that they are performing well. Moreover, taking into account that a sensor may provide a wrong measurement at a certain point for time is a critical aspect to consider. Therefore, our research is based on the question of knowing how we can automate the detection of when a particular sensor is not performing very well by providing wrong values during/at a specified period.

### 3 Methodology

Considering an anomaly in a WSN, the detection of abnormality can be divided into three types [21]:

- The anomaly concerns an individual measurement, i.e. the observation is that some events are unusual with respect to the rest of the data at a particular sensor node
- The whole data at one specific sensor node is anomalous with respect to the surrounding nodes, i.e. this node is abnormal
- The behaviour of a certain number of nodes is anomalous within the network.

In this work, the anomaly detection is performed within the entire dataset. We do not consider the provenance of anomalies specifically. The idea is to provide some information regarding the performance of the network.

Also, in the context of our work, let us consider the DBSCAN (Density-based Spatial Clustering of Application with Noise) algorithm which assumes that areas of high density (also called core) form a cluster. In contrast, data points in the less dense area are considered as anomalies [22].

The DBSCAN algorithm uses two parameters of its implementation [23] (Fig. 2):

- **eps**: for two points to be neighbours, eps is the minimum distance to consider. If this distance is significant, we might form only one big cluster, yet, if it is too small, we may end up with any cluster.

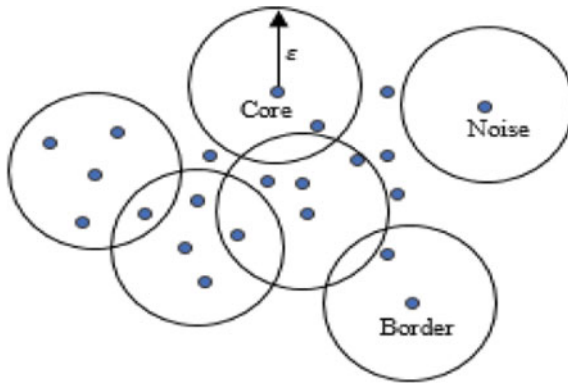


Fig. 2 DBSCAN clustering with min\_point  $\theta = 5$

- **min\_point:** is the minimum number to form a cluster. If this number is small, we might form so many small clusters; however, a significant value can stop the algorithm for creating any cluster, ending up with a dataset form only by anomalies.

To be able to achieve the anomaly for a real smart greenhouse, we took the dataset in Ref. [24] for the training of our model in which we introduced some anomalies. Following are the result of our analysis:

- Determination of suitable values for eps and min\_point for the DBSCAN algorithm:

Calculate and plot the k-Nearest Neighbour Distance. For that, we use python. First, we compute the k-nearest neighbour distances in a matrix of points. Then the plot is used to help find a suitable value for the eps neighbourhood. The knee is found at the eps distance of approximately 0.8 for  $K = 4$ . Then the parameters chosen for DBSCAN are min\_point = 4, eps = 0.8 (Fig. 3).

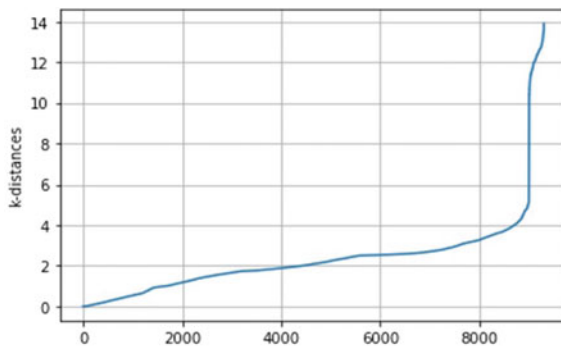


Fig. 3 Determination of eps and min\_point

- Dataset:

The used dataset incorporates information about roses cultivation in greenhouses. It is aimed at identifying adequate actions to improve the roses state [24] The target variables are (1) Soil without water, (2) Environment correct, (3) Too much hot and (4) very cold. Data acquisition was made with an autonomous robot incorporating sensors. The data coming from the sensors give us five features which are soil humidity in analogue–digital conversion, light in lux, the temperature in °C and humidity relative, and CO<sub>2</sub> in analogue–digital conversion.

In our work, we removed the label, so we did not use the target variable. Also, we introduced five anomalies by moving one or more parameters out of the range of the application.

We used the `gplotmatrix` MATLAB’s function to display an array of all the bivariate scatters plots between our five variables, along with a univariate histogram for each variable (Fig. 4).

We chose all the given five features to perform the clustering and see the result. One might consider temperature and the humidity as the most important variables for the application in order to make the clustering with two features, or can choose even three or four features. In our work, we chose to do with the two variables.

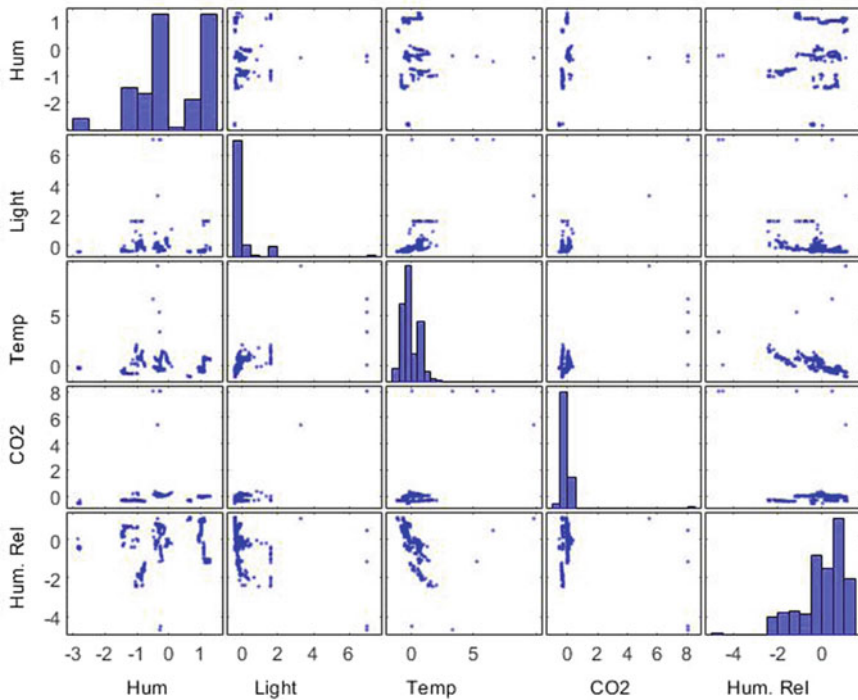


Fig. 4 Scatterplot of the features of the dataset

We are going then to see how our algorithm performs in both cases in the result and discussion section.

## 4 Result and Discussion

As mentioned above, the following are the result of our experimentation:

### 4.1 DBSCAN Using Two Features: Temperature and Humidity

The following plot shows the result of the DBSCAN algorithm on our dataset while considering two features: the temperature and the relative humidity. It can be seen that the data are grouped in our cluster with the cyan colour and the red colour which is used for the noise. The detailed results about the performance are described in the performance evaluation section (Fig. 5).

### 4.2 DBSCAN Using Five Features

While using the five features described above, we resort to the following plot. The anomalies are all in red colour whereas all the other colours are used for the clusters

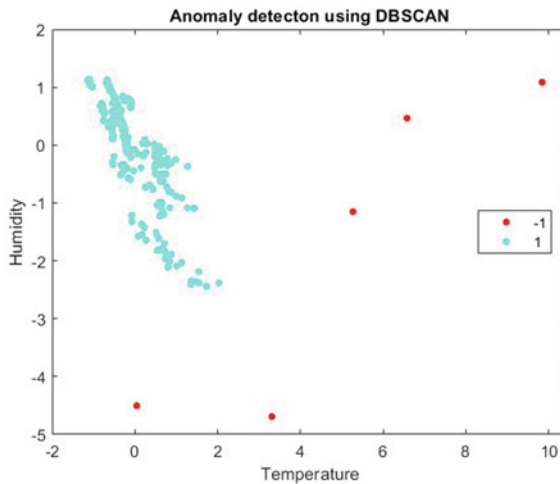
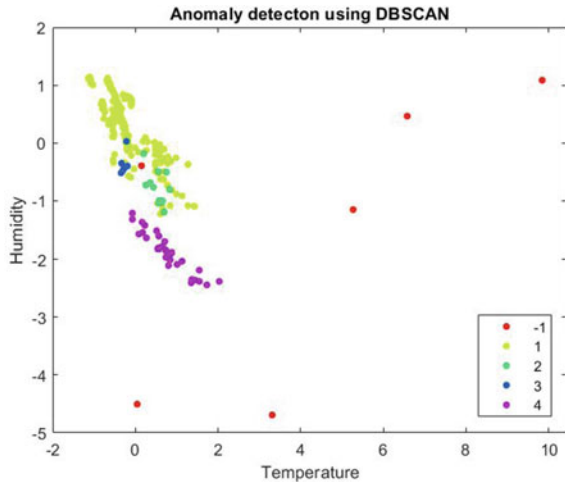


Fig. 5 Anomaly detection with two features

**Fig. 6** Anomaly detection with five features



found by our algorithm. Our aim is to detect anomalies while considering the nominal data we will not dig into the details of the clusters found. As said in the previous line, the detailed result of the performance is described in the performance evaluation section (Fig. 6).

### 4.3 Performance and Evaluation of the Model

To evaluate the performance of the used algorithm, we will be using the confusion matrix. We are going to group our cluster into two general clusters in both cases, the one with two features and the other with five features. In the case of the five features, we are not interested in each specific cluster among the 4 clusters because there are all considered as the nominal data. Our focus is on the anomaly, which in our case is the noise detected by the algorithm.

The confusion matrices are presented as follows:

- Confusion matrix for the DBSCAN algorithm with two features (Table 1):

In this case, the algorithm has been able to find all five anomalies introduced in the data, i.e. there is no false positive or true negative found.

In this case, the algorithm has been able to find five anomalies introduced in the data with two false-positive data (Table 2).

**Table 1** Confusion matrix for two features

	Actual cluster: -1	Actual cluster: 1
Predicted cluster: -1	5	0
Predicted cluster: 1	0	300

**Table 2** Confusion matrix for five features

	Actual cluster: -1	Actual cluster: 1
Predicted cluster: -1	5	0
Predicted cluster: 1	0	298

**Table 3** Performance of the algorithm

	Clustering with two features	Clustering with five features
Accuracy (%)	100	99,3
Precision (%)	100	71.4
Recall (%)	100	100

$$Accuracy = \frac{true\ positive + true\ negative}{Total\ number\ of\ data\ samples} \tag{1}$$

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \tag{2}$$

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \tag{3}$$

Table 3 shows the performance of the algorithms in both cases, with two features and five features.

## 5 Conclusion and Future Work

Building a greenhouse is an advantage that allows producing healthier crops, increases yield and minimises time successfully. This work discussed an anomaly detection approach which can be utilised in greenhouses to predict an abnormal event that may occur when monitoring the crops remotely. For better control, this work proposed a mean to detect unusual behaviour in the WSN in order to evaluate the performance of the network. In this work, the anomaly detection had been implemented with the help of the DSCAN algorithm for the clustering.

Analysing and predicting abnormal events in a set of data is the first step to take into consideration when using a WSN. Moreover, real-time monitoring is also another aspect that has to be considered in the field of IoT. Therefore, developing a cloud computing mobile application is the next direction to focus on. With the result obtained, this work will be extended by developing a real-time mobile app to help to alert the users when an unusual event occurs.



## References

1. Kodali RK, Jain V, Karagwal S (2016) IoT based smart greenhouse. In: 2016 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp. 1–6. <https://doi.org/10.1109/R10-HTC.2016.7906846>
2. Khaldun A, Arif I, Abbas F (2015) Design and implementation a smart greenhouse. *Int J Comput Sci Mob Comput* 48(8):335–347
3. Dedeepya P, Srinija USA, Gowtham Krishna M, Sindhusa G, Gnanesh T (2018) Smart greenhouse farming based on IOT. In: 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp 1890–1893. <https://doi.org/10.1109/ICECA.2018.8474713>
4. Kitpo N, Kugai Y, Inoue M, Yokemura T, Satomura S (2019) Internet of things for greenhouse monitoring system using deep learning and bot notification services. In: 2019 IEEE International Conference on Consumer Electronics (ICCE), pp 1–4. <https://doi.org/10.1109/ICCE.2019.8661999>
5. Polepaka S, Swami Das M, Ram Kumar RP (2020) Internet of things and its applications: an overview. *Lect Notes Electr Eng (IEEE)* 643:67–75
6. Risteska Stojkoska BL, Trivodaliev KV (2017) A review of Internet of Things for smart home: challenges and solutions. *J Clean Prod* 140:1454–1464. <https://doi.org/10.1016/j.jclepro.2016.10.006>
7. Patel NR, Kumar S (2018) ‘Wireless sensor networks’ challenges and future prospects. In: 2018 International Conference on System Modeling & Advancement in Research Trends (SMART), pp 60–65. <https://doi.org/10.1109/SYSMART.2018.8746937>
8. Amrizal MA, Guillen L, Sukanuma T (2019) Toward an optimal anomaly detection pattern in wireless sensor networks. In: 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), vol 1, pp 912–913. <https://doi.org/10.1109/COMPSAC.2019.00137>
9. Zhang K (2009) A danger model based anomaly detection method for wireless sensor networks. In 2009 second international symposium on knowledge acquisition and modeling, vol 1, pp 11–14. <https://doi.org/10.1109/KAM.2009.7>
10. Wang Y et al. (2017) Iterative anomaly detection. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), vol 2017-July, 1st edn, pp 586–589. <https://doi.org/10.1109/IGARSS.2017.8127021>
11. Khan FA, Ibrahim AA, Zeki AM (2020) Environmental monitoring and disease detection of plants in smart greenhouse using internet of things. *J Phys Commun* 4(5):055008. <https://doi.org/10.1088/2399-6528/ab90c1>
12. Bano F, Simran Baseer S, Professor A (2020) Detection of tomato growth state and surveillance system using computer vision and internet of things. [www.jetir.org](http://www.jetir.org). Accessed 13 Jul 2020
13. Abhishek L, Rishi Barath B (2019) Automation in agriculture using IoT and machine learning. *Int J Innov Technol Explor Eng* 8(8):1520–1524
14. Shah NP (2017) Greenhouse automation and monitoring system design and implementation. *Int J Adv Res Comput Sci* 8(9):468–471. <https://doi.org/10.26483/ijarcs.v8i9.4981>
15. Vatari S, Bakshi A, Thakur T (2016) Green house by using IOT and cloud computing. In: 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), pp 246–250. <https://doi.org/10.1109/RTEICT.2016.7807821>
16. Siddiqui MF, ur Rehman Khan A, Kanwal N, Mehdi H, Noor A, Khan MA (2017) Automation and monitoring of greenhouse. In: 2017 International Conference on Information and Communication Technologies (ICICT), vol. 2017-Decem, pp 197–201. <https://doi.org/10.1109/ICICT.2017.8320190>
17. Pavithra G (2018) Intelligent monitoring device for agricultural greenhouse using IOT. *J Agric Sci Food Res* 9(2):2–5
18. Kareem OS, Qaqos NN (2019) Real-time implementation of greenhouse monitoring system based on wireless sensor network. *Int J Recent Technol Eng* 8(2) Special Issue 2:215–219. <https://doi.org/10.35940/ijrte.B1039.0782S219>

19. Muthupavithran PRS, Akash S (2016) Greenhouse monitoring using internet greenhouse monitoring using Internet of Things. *Int J Innov Res Comput Sci Eng* 2(June):12–19
20. Akkaş MA, Sokullu R (2017) An IoT-based greenhouse monitoring system with Micaz motes. <https://doi.org/10.1016/j.procs.2017.08.300>
21. Suthaharan S, Alzahrani M, Rajasegarar S, Leckie C, Palaniswami M (2010) Labelled data collection for anomaly detection in wireless sensor networks. In: 2010 Sixth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp 269–274. <https://doi.org/10.1109/ISSNIP.2010.5706782>
22. Wang Y, Gu Y, Shun J (2020) Theoretically-efficient and practical parallel DBSCAN. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, pp 2555–2571. <https://doi.org/10.1145/3318464.3380582>
23. Tran TN, Drab K, Daszykowski M (2013) Revised DBSCAN algorithm to cluster data with dense adjacent clusters. *Chemom Intell Lab Syst* 120:92–96. <https://doi.org/10.1016/j.chemolab.2012.11.006>
24. EAF-H and Rosero-Montalvo PD (2019) Roses Greenhouse Cultivation Database Repository (ROESGREENHDB). <https://doi.org/10.21227/899y-zh32>
25. Abayomi-Alli AA, Misra S, Akala MO, Ikotun AM, Ojokoh BA (2021) An ontology-based information extraction system for organic farming. *Int J Semant Web Inf Syst (IJSWIS)* 17(2):79–99
26. Arogundade O, Qudus R, Abayomi-Alli A, Misra S, Agbaegbu J, Akinwale A, Ahuja R (2021) A mobile-based farm machinery hiring system. In: Proceedings of Second International Conference on Computing, Communications, and Cyber-Security. Springer, Singapore, pp 213–226
27. Golubenkov A, Alexandrov D, Misra S, Abayomi-Alli O, Leon M, Ahuja R (2021) Decision support system on the need for veterinary control of passing livestock and farm produce. In: *Evolving Technologies for Computing, Communication and Smart World*. Springer, Singapore, pp. 517–526
28. Abayomi-Alli O, Odusami M, Ojinaka D, Shobayo O, Misra S, Damasevicius R, Maskeliunas R (2018) Smart-solar irrigation system (SMIS) for sustainable agriculture. In: *International Conference on Applied Informatics*. Springer, Cham, pp 198–212

# Content Based Deep Factorization Framework for Scientific Article Recommender System



Akhil M. Nair, Oshin Anto, Anchana Shaji, and Jossy George

**Abstract** With the advancement in technology and the tremendous number of citations available in the digital libraries, it has become difficult for the research scholars to find a relevant set of reference papers. The accelerating rate of scientific publications results in the problem of information overload because of which the scholars spend their 70% of the time finding relevant papers. A citation recommendation system resolves the issue of spending a good amount of time and other resources for collecting a set of papers by providing the user with personalised recommendations of the articles. Existing state of art models do not take high-low order feature interactions into consideration, due to which the recommendations are not up to the desired level of performance. In this paper, we propose a content-based model which combines Deep Neural Network (DNN) and Factorization Machines (FM) where no pre-trainings are required for providing the citation recommendations.

**Keywords** Citation recommendation · Factorization machines · Deep neural network

## 1 Introduction

The rapid development in information technology has caused an exponential expansion in scientific publications, which in turn has challenged research scholars in finding relevant sets of papers satisfying their citation recommendations. Therefore, personalised paper recommendations have become a significant and timely research topic. The most common method of collecting a set of papers is by following the citations from reference articles. Recommendation Systems (RS) helps users to filter the most relevant set of items based on their interest such as a movie, music and papers for research work as well. A real-time example is Amazon.com which recommends items and books and movie recommendations by Netflix based on the user-preferences [1].

---

A. M. Nair (✉) · O. Anto · A. Shaji · J. George  
Department of Computer Science, Christ University, Bangalore, India  
e-mail: [akhil.nair@christuniversity.in](mailto:akhil.nair@christuniversity.in)

Research paper recommender systems help users in finding a set of papers based on their interests by mitigating information overload. They also improve the quality of a decision by suggesting relevant and adequate items according to the user's preference. Literature in the area of RS can be broadly classified as Content Based Filtering (CBF) and Collaborative Filtering (CF) based recommendation systems. CBF makes recommendations based on user's preferences where CF mimics user-user recommendations. The third category of algorithm, Hybrid algorithm combines content based and collaborative filtering. They produce separate lists of recommendations and finally merge them together for the final list of recommendations [2].

Deep learning has gained much popularity in recommender systems especially when it comes to citation recommendations. It is proven that deep learning-based models are able to provide a better level of recommendations compared to state-of-the-art models. They are able to capture the semantic representation which ultimately leads to significant improvement in providing final recommendations to the user [3]. Factorisation Machines (FM) have become one of the most powerful and efficient models for recommender systems due to their ability to accurately estimate parameters under very sparse data. FM maps real-valued features into low dimensional latent factor space and does predictions. In this work, we propose a model in which a deep neural network would be combined with FM for recommending citations to users.

## 2 Related Work

Although the research scholars contribute many models for personalized citation recommendations, the quality of these recommendations is improved by deep learning models. A survey of 35 citation recommendation models considers the information used, data representations, methodologies used, the recommendation types, the problem addressed, and personalization of the users [3]. Literature show that the most popularly used dataset is the ACL anthology, which gives access to the papers, authors, citation relations, venues, and contents. Mainly, the data is utilized to generate BERT and Graph Convolutional Networks (GCN) for content-based recommendations [4, 5]. It takes the semantic representations of papers and utilizes the data information, exploiting research's preferences and providing recommendations.

To improve the data reusability, relevant literature articles corresponding to the datasets in Gene Expression Omnibus (GEO) are utilized for the literature recommendation system. Moreover, the semantic features are considered from the abstract and title of MEDLINE articles [6]. The VSM baseline model is evaluated using different distributional methods such as TF-IDF, a ranking function as BM (Best Matching) [25], Latent Semantic Analysis, Latent Dirichlet Allocation, word2vec, and doc2vec. Thus, the recommendation of similar top papers is obtained using cosine similarity within the dataset's vector and paper's vector representation. Moreover, SCM re-ranking and normalization methods were used in word2vec embeddings

to improve the recommendations. The proposed model ensures 10 of 0.833 strict precision and a partial precision at 10 of 0.90 by using BM (Best matching)25.

The AAN and PeerRead datasets were used to form a FullTextPeerRead dataset to propose a deep learning-based model that includes a document encoder and a context encoder that utilizes the Graph Convolutional Networks (GCN) and Bidirectional Encoder Representations from Transformers (BERT) [7]. The proposed model ensures a mean average precision and recall@k of more than 28%. A hybrid model of the Latent Factor Model (LFM) and Bidirectional Gated Recurrent Unit (BIGRU) were proposed using research favorites records and the literature context that leads to obtaining the research paper recommendation based on the researcher's interest. The semantic information is obtained by using the BERT model that outputs the word embedding vectors. Later, the user latent factor vector uses to develop a recommendation list of the papers. The proposed recommendation model outperforms compared to the traditional baseline recommendation models.

The recommendation system for Click Through Rate (CTR) prediction is optimized by feature interactions based on the user's behaviors. A deep FM model is proposed to learn low and high-order interactions based on factorization machines for recommendations and deep learning for learning the feature. The DeepFM model shows more efficiency than the CTR prediction both in the benchmark (Criteo) and commercial (Company\*Dataset) data with an AUC of 87% and LogLoss of 0.026 [8]. A fast and more accurate generic-based Discrete Factorization Machine (DFM) recommendation model is generated by Liu et al. [9]. The model learns in a binary space for each feature and is evaluated in two real-time data, i.e., Yelp and Amazon. The results show that DFM outperforms other than the binarized recommendation models and obtains a recommendation accuracy approximately equivalent to FM.

A novel FM framework combined with Generalized Metric Learning (GML-FM) by utilizing Amazon and MovieLens data in focusing the feature correlations in FMs. One of the proposed framework methods is GML-FMmd, which adopts the linear correlations between Mahalanobis's features and GML-FMdn by DNN-based distance function for non-linear feature correlations [10]. Data sparsity and cold start in recommendation systems are addressed by collecting a large-scale second-hand trading dataset. A sequential recommendation system of extended Factorization Machines that decreases the computation's complexity to linear time is proposed by Wen and Zhang [11]. Further, a novel of the model elaborates non-linear and higher-order feature interactions under the neural network framework. The developed model is preferable in predicting the sequential behaviors of the users for the recommendation system.

A Hierarchical Attentional Factorization Machines (HaFMRank) framework using attention mechanism and factorization machines for the recommendation of newly posted questions in Community Question Answering (CQA) is proposed [12]. A pre-training process for content-based feature embeddings that considers the weighted feature interactions is designed along with the proposed framework. Integration of matrix factorization technique with tag-aware document embeddings is considered in the personalized recommendation of research papers. The embeddings are retrieved from a social tag prediction model to collect the significant words

and sentences on the particular tags with collaborative filtering to increase the general representation and visualization of the key concepts in research papers.

The deep learning-based hybrid recommendation method, i.e., ACNN-FM, combines two parallel CNN-attention mechanism-based models by focusing on the semantic representation of the user's and items from their comments is extracted and proposed [13]. Moreover, the ratings are also considered for the user-item association, predicting the user's preferences for new items. Event-based online Social Networks (EBSNs) allows users to take part in events and join event groups. It is more convenient to the users based on event-group recommendation systems [14]. Hence, a Content-based Co-Factorization Machines (CC-FM) recommendation is proposed by utilizing the user features, content information, and event features to model the user's preferences.

### 3 Methodology

Matrix Factorization (MF) learns low-dimensional representations or embeddings. The interaction matrix is factorized into two matrices of embeddings. Moreover, the latent features are learned and the similarity of the embeddings is computed and compared. This technique is used in the prediction of regression, ranking and classification. It estimates the parameters more accurately under sparse data and trains with linear complexity. Thus, influences the recommendation system for the MF technique.

The study considers  $m$  paper by  $n$  paper matrix which results on the behaviour of citations, i.e., cited or non-cited corresponding to paper\* paper ( $m*n$ ). Matrix factorization generates latent features by paper\*paper interactions. It factorizes the real matrix  $R_{m*n}$  into the inner product of two matrices.

$$R_{m*n} \approx P_{m*k} \bullet QT_{n*k} \quad (1)$$

(InteractionMatrix) (PaperMatrix) (PaperMatrix)

where  $k$  is the hyperparameter of the factorization model that represents the length used to embed both Paper\_1 and Paper\_2 into real vectors. Each paper\_1 and paper\_2 is represented by a real vector of length  $k$  as embedding and the resultant dot product of paper embedding and paper embedding gives the interaction i.e., the citation, result of the corresponding Paper-Paper pair. Hence, the Matrix Factorization technique learns the embeddings of all corresponding sets of papers such as Paper\_1 and Paper\_2. Thus, the inner product of the embeddings represents the citation interaction.

In this study, the Matrix Factorization technique has been used to recommend the cited papers to the research scholars. A model known as CR-MF (Citation Recommendation with Matrix Factorization) has been proposed to identify the cited and non-cited papers by using the Matrix Factorization technique. Moreover, the predicted

adjacency matrix by the proposed model has been compared and evaluated with the adjacency matrix of the considered data.

ACL Anthology Network (AAN) is an open source dataset containing more than 20,000 papers and details. It also has citation and collaborative networks which are used by various recommendation models. The dataset consists of many sub-folders including `author_affiliations`, `paper_text` and `release`. The AAN 2014 release data includes details of 23,766 papers and 1,24,857 paper citations. In this model we make use of the paper citation network dataset, which consists of two columns indicating the citation network of each paper.

The workflow of the Citation Recommendation-Matrix Factorization (CR-MF) model is elaborated in Fig. 1. The features from the AAN dataset are renamed as “Paper 1” and “Paper 2”. An attribute, “Result” has been added to the considered data to specify the cited papers respectively. With the obtained dataset, a paper-by-paper adjacency matrix (paper\*paper) of cited papers as “1” and non-cited papers as “0” has been generated as the result. The resultant adjacency matrix is used in the Matrix Factorization technique to obtain the adjacency matrix which predicts the cited and non-cited papers.

The primary step of the model is to obtain the relevant data from the paper citation network. The dataset contains details on the paper citation indicating which paper is being cited to which all papers from the collection. In the second stage, a new feature named “Result” is added to the dataset with values as “1”. The presence of “1” indicates that the corresponding papers are cited together. With this data, an adjacency matrix is generated with “1” s and “0” s.

The corresponding adjacency matrix is passed onto the Matrix Factorization algorithm with alpha and beta values indicating the learning rate and regularization parameter. Matrix factorization decomposes the paper–paper interaction matrix into the product of two lower dimensionality matrices. The dot product of the interaction predicts the cited and non-cited paper. Hence, the predicted and considered adjacency matrix has been compared and evaluated by Root Mean Square Error (RMSE) and Mean Absolute Error (MAE).

Table 1 illustrates the sequential steps in the proposed model. The adjacency matrix (paper\*paper) is obtained using Matrix factorization technique and later is compared with the actual adjacency matrix from the AAN dataset. The proposed model’s performance is evaluated using the RMSE and MAE metrics.

## 4 Experimental Study

The implementation of the scientific paper recommender system using the matrix factorization technique is done using ACL Anthology Network (AAN) dataset. The data is released in the year 2014 and has details about the scientific publications and paper citations. For the experimental study, we used 3000 rows from the paper citation network data which indicated the citation network of each paper in the dataset.

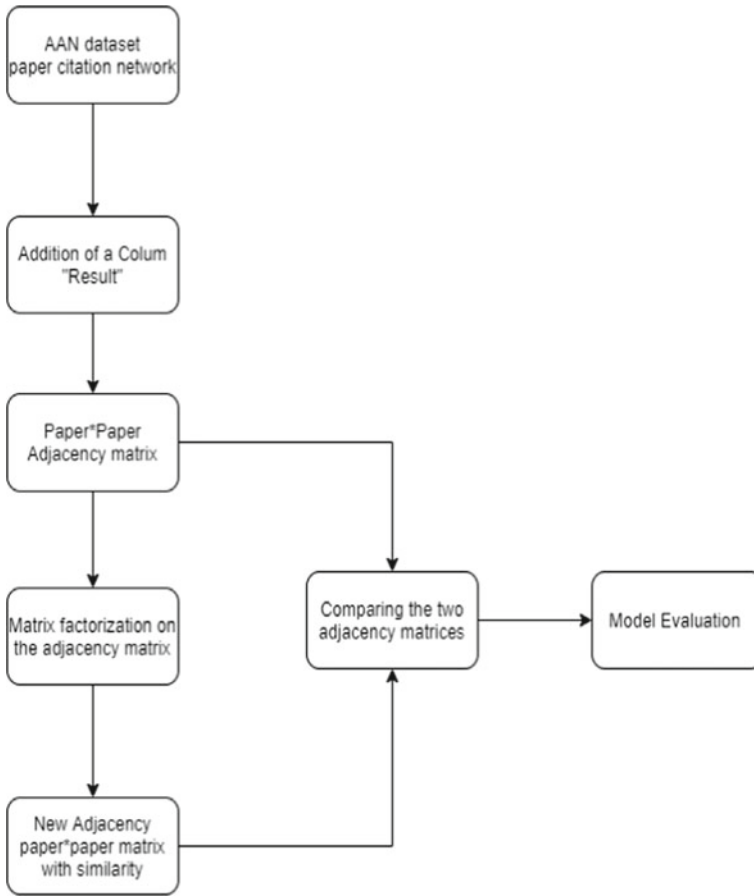


Fig. 1 CR-MF model

Table 1 Steps of the proposed CR-MF model

Step 1	Paper citation network is extracted from AAN dataset
Step 2	Addition of a new feature "Result" with value 1 indicating each paper is cited to the corresponding paper
Step 3	Creation of paper*paper adjacency matrix with "Result" values and "0" for null values
Step 4	Matrix Factorization on the created adjacency matrix with calculation of gradient with alpha and beta parameter
Step 5	Generation of new adjacency matrix for papers
Step 6	Compare the result of both adjacency matrices
Step 7	Evaluation of the model using RMSE and MAE



**Table 2** AAN dataset

	Paper 1	Paper 2	Result
0	C08-3004	A00-1002	1
1	D09-1141	A00-1002	1
2	D12-1027	A00-1002	1
3	E06-1047	A00-1002	1
4	H05-1110	A00-1002	1

Table 2 shows the sample AAN dataset which is used for the implementation. The dataset being used has two columns namely ‘paper\_1’ and ‘paper\_2’. The IDs in ‘paper\_1’ are the source nodes and ‘paper\_2’ has the ids with target nodes. A feature named “Result” is added to the data frame to indicate that the corresponding papers are cited together. The value of Result has the value 1 is it is cited together and 0 in case of not being cited. An adjacency matrix based on the three columns is created and the non-cited combinations of papers are replaced with “0” as the “Result”. Matrix factorization is applied to the adjacency matrix which then replaces the values with the calculated dot product indicating the levels to which two papers can be considered to be cited together. For example, the two papers “C08-3004” and “A00-1002” are cited and hence the “Result” is 1, whereas the non-cited papers “A00-1004” and “A00-1002” would have “Result” as 0.

## 5 Results and Discussions

The proposed scientific article recommendation system based on the matrix factorization technique provides satisfying results in terms of predictions. Matrix Factorization based recommender systems have always proved to be more efficient than the existing state-of-the-art models for Click through Rate (CTR) prediction and other related recommendations. Matrix factorization is applied to the obtained adjacency matrix with alpha and beta values. The final result is a new adjacency matrix with the values that are generated using the MF technique as shown in Table 3.

The matrix in Table 3 provides the probability value of the paper-by-paper citation ranking provided the model considers the number of features as  $k = 5$ . The values above 0.70 can be considered for the prediction of the corresponding paper\_Id. For example, the most considerable citation for the paper ‘C00-1007’ is ‘A00-1002’ and ‘A00-1004’. The value 0.70 is the threshold considered for the prediction from the probability matrix obtained from the matrix factorization.

**Table 3** Predicted adjacency paper\*paper matrix

	AOO-1002	AOO-1004	AOO-1005	AOO-1006	AOO-1007	AOO-1008	AOO-1009
AOO-1004	0.891952	0.856158	0.676546	0.733713	0.710824	0.673904	0.799573
AOO-1030	0.926993	0.889793	0.703124	0.762538	0.738749	0.700379	0.830984
AOO-2024	0.884883	0.849373	0.671184	0.727899	0.705191	0.668564	0.793236
AOO-2027	0.903707	0.867442	0.685462	0.743383	0.720192	0.682786	0.810110
COO-1007	0.890983	0.855228	0.675810	0.732916	0.710052	0.673172	0.798704

## 6 Conclusion

Due to the ever-increasing number of research papers that are published by scholars, it is very important to have recommender systems that can provide the best recommendation of papers to the ones who are in need. Much advancement has been made in these areas which have looked into various types including content based and collaborative filtering methods. Deep learning techniques are also gaining popularity due to its better efficiency in providing the recommendations. In this work, we tried to develop a scientific article recommendation system based on matrix factorization method. The MF is implemented on the paper\*paper adjacency matrix which predicted the score of citation similarity. This is purely a trial method in order to understand how the dataset would react when the MF method is used for the prediction of cited papers. In future, better models can be derived from this idea by incorporating deep neural networks along with the Matrix Factorization technique. The addition of features in the model would also improve the model's performance.

## References

1. Lee J, Lee K, Kim JG (2013) Personalized academic research paper recommendation system. <http://arxiv.org/abs/1304.5457>
2. LKB et al. (2019) Advances in information retrieval (ECIR '19) Part (II). Proc. ACM/IEEE Jt. Conf. Digit. Libr. 11438(1):267–274. <https://doi.org/10.1007/978-3-030-15719-7>
3. Feng X et al. (2019) The deep learning-based recommender system 'pubmender' for choosing a biomedical publication venue: development and validation study. J Med Internet Res 21(5). <https://doi.org/10.2196/12957>
4. George JP (2020) Similarity analysis for citation recommendation system using binary encoded data. pp 12–13
5. Nair AM, Wagh RS (2018) Similarity analysis of court judgements using association rule mining on case citation data—a case study. Int J Eng Res Technol 11(3):373–381
6. Patra BG et al (2020) A content-based literature recommendation system for datasets to improve data reusability—A case study on Gene Expression Omnibus (GEO) datasets. J Biomed Inform. <https://doi.org/10.1016/j.jbi.2020.103399>
7. Jeong C, Jang S, Shin H, Park E, Choi S (2019) A context-aware citation recommendation model with BERT and graph convolutional networks. <http://arxiv.org/abs/1903.06464>
8. Guo H, Tang R, Ye Y, Li Z, He X (2017) DeepFM: a factorization-machine based neural network for CTR prediction. <https://doi.org/10.24963/ijcai.2017/239>
9. Liu H, He X, Feng F, Nie L, Liu R, Zhang H (2018) Discrete factorization machines for fast feature-based recommendation. <https://doi.org/10.24963/ijcai.2018/479>
10. Guo Y, Cheng Z, Jing J, Lin Y, Nie L, Wang M (2020) Enhancing factorization machines with generalized metric learning. arXiv. <https://doi.org/10.1109/tkde.2020.3034613>
11. Wen N, Zhang F (2020) Extended factorization machines for sequential recommendation. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2977231>
12. Tang W, Lu T, Li D, Gu H, Gu N (2020) Hierarchical attentional factorization machines for expert recommendation in community question answering. IEEE Access. <https://doi.org/10.1109/ACCESS.2020.2974893>

13. Pang G et al (2019) ACNN-FM: a novel recommender with attention-based convolutional neural network and factorization machines. Knowledge-Based Syst. <https://doi.org/10.1016/j.knosys.2019.05.029>
14. Zhao Y, He Y, Li H (2018) Content-based co-factorization machines: modeling user decisions in event-based social networks. [https://doi.org/10.1007/978-3-030-00767-6\\_72](https://doi.org/10.1007/978-3-030-00767-6_72)

# On Discrimination Power of Image Feature Vector



Sushila Palwe

**Abstract** Image Analysis using systems like Content based Image Retrieval is the information retrieval system which helps in retrieving similar images, found very useful in many areas. Feature vectors are the compact and precise representation of Images which effectively needs to be used in Image Retrieval System. The performance of Image Retrieval is majorly dependent on the quality of feature vectors. Many intelligent algorithms are proposed in literature to retrieve the features of images and represent it in an optimal way. Feature Vectors with reduced dimensions are always preferred for image retrieval systems. Quality of Feature Vectors is always checked using the output of image retrieval. There has always been a trade-off between the size of the feature vector and retrieval performance. Common measures used for checking retrieval system performance are precision, recall. This paper discusses the important measure: “discrimination power” of Image feature vector to check feature vector quality in context of image matching and retrieval. The discrimination power of Image Feature Vector, and its use-case with our feature extraction method is discussed here.

**Keywords** Discrimination power · Image retrieval and analysis · Dimensionality reduction · Feature extraction · Low-level features · Image analysis

## 1 Introduction

In today’s information era, Image retrieval systems are proved very useful in many of the domain varying from social, technical, commercial, agriculture, medical, educational, and many more. Image retrieval systems retrieves the images similar to sample case images as provided. In many of the system, one can tally the already known scenario presented in terms of images with the current scenario and infer the action

---

S. Palwe (✉)  
School of CET, MITWPU, Pune, India  
e-mail: [sushila.palwe@mitwpu.edu.in](mailto:sushila.palwe@mitwpu.edu.in)

or effects which may happen with such scenario. In such systems images are represented and modeled as image feature vectors, also called as image signatures. Images in image-database stored as Image-Feature vectors.

It is so important that these feature vector must possess the following properties.

Retain the low-level and high-level features of image Compact in Size i.e. Dimensionally Optimal. Various Feature vector extraction algorithms are proposed by many of the researches and tested for image retrieval, object detection, shape detection, etc. With these proposed techniques of feature vectors, the first and second property of the feature vector are experimented using the performance measures like precision and recall of the system in which these feature vectors are used for particular task.

Discrimination power is an essential property that needs to be addressed and experimented for the image feature vector. In this paper, we propose a novel approach to check the discrimination power of the feature vector of image. Proceeding section of paper discuss the overview, methodology, and results. Image Retrieval Performance is analyzed using various performance measures depending on the methodology used in finding similar images. In literature the discriminative feature creation attempted were made by many researchers.

Feature discrimination depends on the aspects that target features in image has a stronger difference than others. With the features with the highest distinguishing features, it is used to achieve a great dimensionality reduction and a compact representation of image features The compact feature vector thereby provides a better way to perform better in Image Retrieval Applications [1–5].

In Literature many approaches are discussed to retrieve the similar images based on feature vector. Major Algorithms like Classification, Clustering, Template matching, and Feature matching are used for image retrieval and matching.

In literature, Under Supervised learning, SVM, KNN, Bayesian Classification, Decision Tree algorithms are used to retrieve domain specific similar images. neural network for simultaneous clustering and feature discrimination has been proven useful.

Under unsupervised learning, K means algorithm is found useful in retrieving similar images from image clusters.

Though many feature extraction algorithms and image retrieval methods are discussed in literature, very few attempts were made to check the discrimination power of feature vector. This is the major research gap identified with the domain of Image Retrieval [6–9].

## 2 Implementation

We proposed a multimodal approach for Feature vector creation using Color-encoded bins. In this approach we attempt to create a dimensionally reduced feature vector using energy compaction property of various image transforms [10–12].

### 2.1 Multimodal Feature Extraction Algorithm

1. As in Fig. 1 in Multimodal feature extraction, the image features are extracted using Transform Methods and then Histogram Method. Algorithm for Multimodal feature extraction is stated below.
2. Algorithm 1: Multimodal Feature Extraction
3. Input  $\rightarrow$  Database Images  $D = \{D1, D2, D2, \dots, Dn\}$ , Query Image  $\rightarrow I$
4.  $FV(D)_{Intermediate} \rightarrow \{Energy\ Compaction(DT)\}$
5.  $FV(D) \rightarrow \text{Binning}(FV(D)_{Intermediate}, CG)$  For each  $D = \{D1, D2, D2, \dots, Dn\}$ , where Center Graylevel as given in Eq. (1)

$$Centre\ Graylevel = \frac{\sum \text{pixel Count per Graylevel} \times \text{Graylevel}}{\text{Total no of Pixels}} \quad (1)$$

6. Function Binning ( $FV(D), CG$ )  
 $\{$

For each plane Label the pixels as “0” if at left of Centre Graylevel else “1” if they are right of CG.

Combine these Codes of each plane, form the bins with address

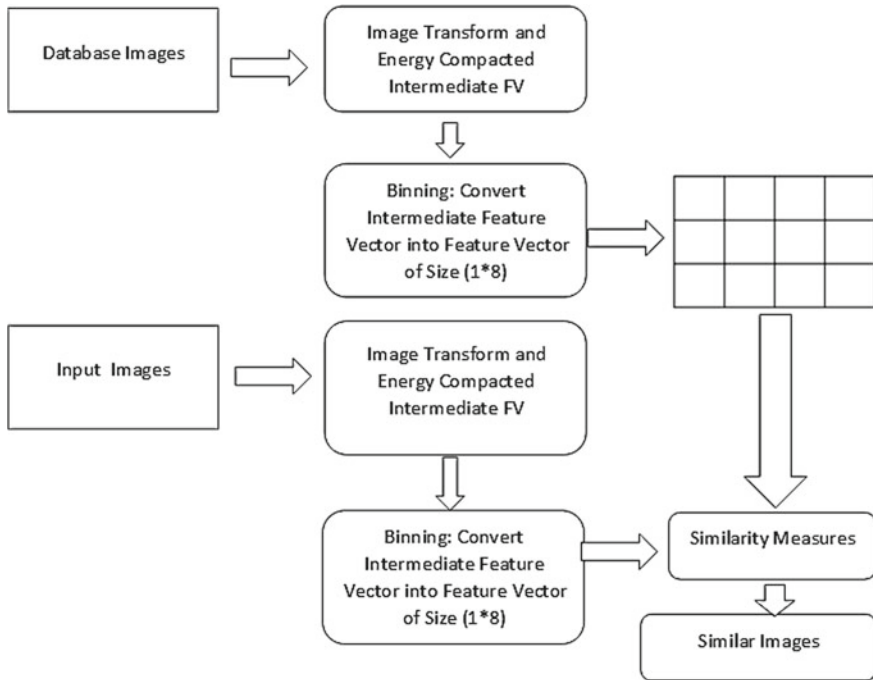


Fig. 1 CBIR using multimodal feature extraction

000,001,010,100,101,011,110,111.  
 Calculate no of pixels falling in each Bin.  
 $N*N*3$  pixels are addressed using only 8 addresses.  
 }

7.  $FV(I)_{intermediate} \rightarrow$  Energy Compaction (IT)  
 $FV(I) \rightarrow$  Binning( $FV(I)_{intermediate}$ , CG)
8. Sort  $Z = \text{Dist}(FV(I), FV(D_{i..n}))$ , For each  $D = \{D_1, D_2, D_2, \dots, D_n\}$
9. Display Images with top K Dist.

With this multimodal algorithm the size of Feature vector is reduced from  $N*N*3$  to  $8*1$ . Thus, the computation for similarity measures is also reduced.

### 3 Image Retrieval

To find out the similar images of the query image, the input query image needs to be converted into a feature vector using Multimodal Feature Extraction. The steps of CBIR Query execution are [13–18]:

1. Apply the algorithm for all images to create feature vector database.
2. Apply the algorithm for query image
3. Apply Similarity Measure: Euclidian Distance.
4. Set the threshold value of similarity distance.
5. Display the “Similar” images with maximum similar to Query Image.

### 4 Discrimination Power of Feature Vector

Most of the Image feature extraction methods get verified based on their capability to find out the similarity amongst the images. While focusing on index of similar image retrieval, the other affecting component, i.e. capability of feature vector to check discrimination amongst image feature vectors is ignored [1].

To improve the image retrieval performance, the image feature vector must hold the good discrimination capability known as discrimination power of feature vector. Discrimination Power is a measure to check the ability of image feature vector, to distinguish between dissimilar images using their feature vectors.

With our approach of Feature Extraction and Image Retrieval, discrimination power is discussed as below.



### 4.1 Discrimination Power

Discrimination of Image feature Vector is defined as the probability that the system identifies and assign the different image-class to two unrelated images.

To check the discrimination power of FV using CBIR output, the following Eq. (2) is used.

$$Discrimination\ Power = 1 - \frac{\sum_{i=0}^n \cdot \sum_{j=0}^m CSI(i, j)}{N} \tag{2}$$

where

- CSI Confusion String of Images in a particular output.
- I Input Image.
- J Output Image Class.
- N No of Observations.

CSI (Eq. (3)) represents the Confusion String of Images and it is calculated by observing the output

$$CSI = \frac{Count(irrelavant\ Images\ belonging\ to\ same\ class)}{Images\ in\ Output} \tag{3}$$

## 5 Understanding the Discrimination Power

Discrimination Power of Image Feature Vector is calculated by observing the Image Retrieval Output for the database of Images. Database images used in experimentation are from the WANG database [19]. It contains 1000 images of 10 image classes, each with 100 images, as shown in Table 1.


With input as flower Image the Image Retrieval output (100 Images), with Feature Vector Extraction Algorithm is shown in Fig. 2

Confusion String of Images are as follows with the observation mentioned in Fig. 2 are mentioned in Table 2

With such N observations of each n class, Eq. (2) is calculated for identifying the discrimination power.

Higher the value of Discrimination Power suggests the good quality of Image Feature Vector.

Table 1 Database images

Wild People		Elephant	
Beach		Flower	
Monuments		Horse	
Bus		Mountain	
Dinosaur		Food Items	

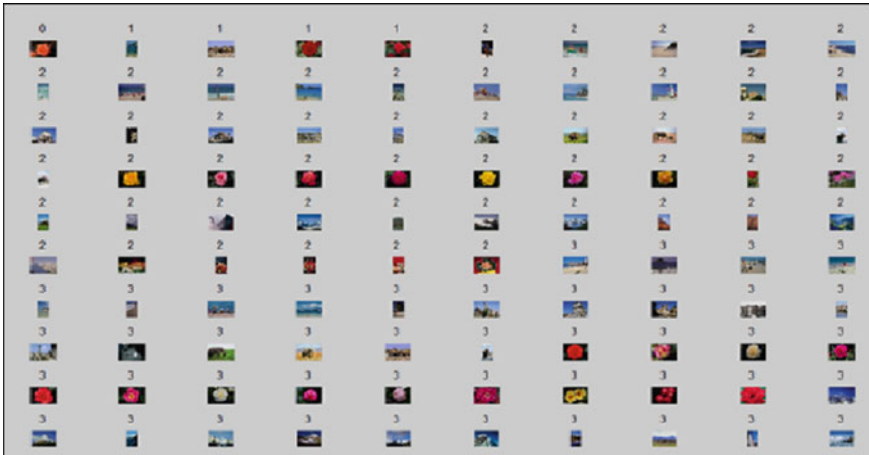


Fig. 2 Images retrieved similar to input image of flower class

**Table 2** CSI with observation 1 (flower image)

Image classes	CSI
Wild people	0.06
Elephant	0.08
Beach	0.09
Flower	(Input Image) 0.35
Monument	0.07
Horse	0.04
Bus	0.05
Mountain	0.04
Dinosaur	0.13
Food Item	0.09

## 6 Evaluation of Discrimination Power

Image Analysis and Image Retrieval Performance is totally based on the quality of Image Feature Vector. With our Feature Vector Creation approach, the feature vector size is dimensionally reduced to 8\*1. But reduced dimensionality size of Feature Vector may degrade the Image Analysis and Retrieval performance. Generally, the Retrieval Performance is measured using Precision and Recall. These two measures are useful in analysing the output across expected input. But it becomes necessary to observe the quality of Image feature vector with respect to its capacity to differentiate between two different class of images. With discrimination analysis, it is possible to observe the quality of feature vector with respect to it differentiation capability amongst the irrelevant class of Images. With Eq. (2), discrimination power of the Feature Vector which is derived using Algorithm 1, is compared with respect to various image classes as input.

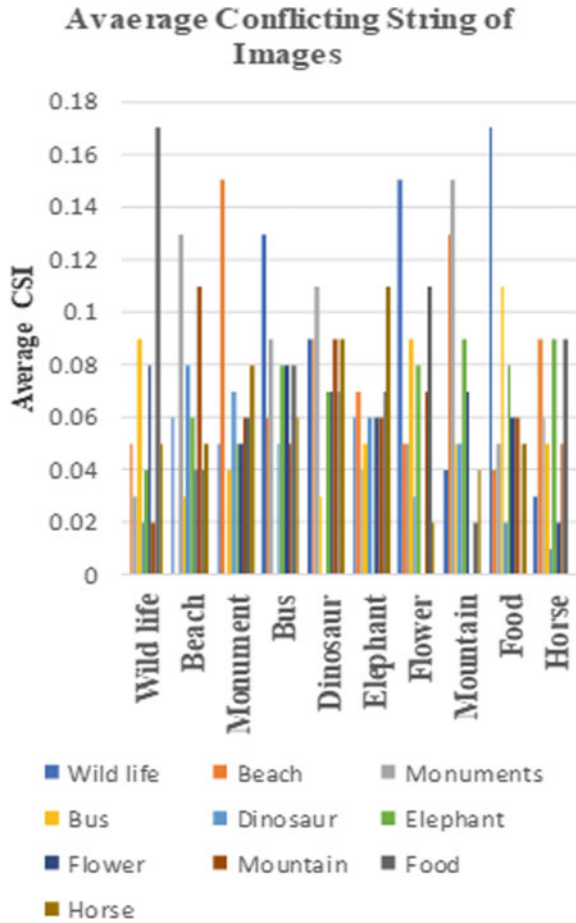
The Performance Analysis is carried out for different sets of images as input and 100 images in image retrieval output. Discrimination Power in each observation is recorded, and Average is calculated for N observations.

With Fig. 3, CSI of N input Images, with respect to all unrelated classes are calculated and taken into the consideration for Discrimination power calculation.

As per the observation in Fig. 3, the Image classes like Wild Life, Food, which are with diverse background, are resulting with maximum CSI.

With our experimental set of 1000 images of 10 classes and with 10 observations, with 100 images in the output Image, the discrimination power of the Multimodal Feature vector is 0.38.

**Fig. 3** Average CSI with respect to input image class and output image class



## 7 Conclusion

Dimensionally reduce Image Feature Vector with multimodal approach is created and tested for the discriminative power, with our proposed way of discrimination analysis for Feature Vector. While deriving the discrimination power, the Conflicting String of Images are taken into the consideration. The conflicting string of images are those images which belongs to unrelated image class of that of input class. With N observation of various input images, we could get the discrimination power of 0.38 for the proposed feature vector with multimodal approach. It has been observed that Images with diverse background outputs the most no of CSIs of various classes as compared to the images with a plain background. While designing the feature vector also while finding the similarity distance as feedback and taking corrective action.

## References

1. Meyer-Baese A, Schmid V (2014) Analysis of dynamic susceptibility contrast MRI time-series based on unsupervised clustering methods, pattern recognition and signal analysis in medical imaging, 2nd edn
2. Ashraf R (2018) Content based image retrieval by using color descriptor and discrete wavelet transform. *J Med Syst* (Springer)
3. Sravani N, Veera Swamy K (2018) CBIR using slant transform using DC & AC coefficients. *Int J Eng Technol* 7(3.6):276–280
4. Sapiecha K, Lukawski G, Krechowicz A (2014) Enhancing throughput of scalable distributed two – layer data structures. Parallel and Distributed Computing (ISPDC). In: 2014 IEEE 13th International Symposium, pp 103–110
5. Kekre HB, Mishra D (2011) Sectorization of Walsh and Walsh Wavelet in CBIR. *Int J Comput Sci Eng (IJCSSE)* 3(6)
6. Shih JL, Chen LH (2002) Colour image retrieval based on primitives of colour moments. *IEE Proc* 149(6):370–374
7. Aghav-Palwe S, Mishra D (2020) Statistical tree-based feature vector for content-based image retrieval. *Int J Comput Sci Eng* 2020 <https://doi.org/10.1504/IJCSE.2020.106868>
8. Goshtasby AA (2012) Similarity and dissimilarity measures. In: image registration. *Adv Comput Vis Pattern Recognit* (Springer) 6(2):224–239
9. Kimutai G, Cheruiyot P, Otieno D (2018) A Content Based Image Retrieval Model for E-Commerce. *Int J Eng Comput Sci* 7(11):24392–24396
10. Kekre HB, Mishra D (2010) Performance comparison of density distribution and sector mean in Walsh transform sectors as feature vectors for image retrieval. *Int J Image Process (IJIP)* 4(3)
11. Sciascio ED, Celentano A (1996) Similarity evaluation in image retrieval using simple features. *J Comput Inf Technol* 4(3)
12. Datta R, Joshi D, Li J, Wang JZ (2008) Image retrieval: idea, influences and trends of the new age. *ACM Comput Surv* 40(2)
13. Kato T (1992) Database architecture for content based image retrieval in Image Storage and Retrieval Systems. *Proc SPIE* 2185:112–123
14. Kekre HB, Mishra D (2010) Content based image retrieval using weighted hamming distance image hash value. In: International conference on contours of computing technology
15. Afifi AJ, Ashour WM (2012) Image retrieval based on content using color feature. *ISRN Comput Graph*
16. Porat M, Zeevi YY (1988) The generalized Gabor scheme of image representation in biological and machine vision. *IEEE Trans Pattern Anal Mach Intell* 4(10)
17. Palwe S, Budkule C (2015) Comparison of feature extraction techniques in cotton leaf disease classification using CBIR. *Int J Comput Technol* 2(7)
18. Rao MB, Rao BP, Govardhan A (2011) CTDCIRS: content based image retrieval system based on dominant color and texture features. *Int J Comput Appl* 18(6):40–46
19. Wang JZ (2010) Wang database.

# Design a Mechanism for Opinion Mining



Samir N. Ajani and Parul Bhanarkar

**Abstract** Opinion Mining is the task of Information Extraction. One of the important parts of opinion mining is to collect the opinions of the focus groups. The collection of this information is very important as it helps to improve the customer-aligned services and improve the product quality through user and customer opinions. In order to collect the opinions there are various sources are available such as social media blogs, online product reviews, and trending discussions. As these information are available very easily there are new challenging dimensions that come across. Data authenticity is the key element in generating opinion mining. The purpose of this project is to learn the concepts of opinion mining and semantic analysis. The application takes input as opinions of students about various engineering colleges based on parameters such as Branch, faculty, fieldwork/practical, canteen, academics, placements, sports/extracurricular activities, performing subject classification of views using the concept of semantic similarity, and creating a semantic structure of views and then performing sentimental analysis on those received opinions. Finally, presenting the information in an effective manner by categorizing the opinions as positive, negative, and neutral comments along with a semantic net. The project will help students with proper decision-making in order to take admissions in engineering colleges for their bright future.

**Keywords** Text categorization · Semantic structure · Sub-node · Opinion mining · Semantic analysis · Machine learning · Opinion mining · Tool · Semantic · Corpus · Tagging

---

S. N. Ajani (✉)

Shri Ramdeobaba College of Engineering and Management, Nagpur, Maharashtra, India  
e-mail: [samir.ajani@gmail.com](mailto:samir.ajani@gmail.com)

P. Bhanarkar

Jhulelal Institute of Technology, Nagpur, Maharashtra, India

## 1 Introduction

Semantic understanding is a study of meaning which includes the most complex tasks such as: finding synonyms, disambiguated words, building answering systems questions, translate from one NL to another, fill the knowledge base. Semantic Analytics is the use of an ontology to analyze content in web resources and is a composite of “semantic analysis” and “computing” components. “Semantic analysis” refers to formal meaning analysis, and “computing” refers to an approach that in principle supports effective implementation. It requires detection and classification of semantic relations mentioning in a set of artifacts, usually from the text. This is a statistical technique for using Word that allows the comparison of semantic similarities between textual information pieces.

## 2 Literature Survey

Yoon et al. [1] proposed the idea of designing algorithms based on opinion mining and graph-based semi-supervised learning for decision making. This algorithm is proposed to help in stock investment decision-making tasks. The important research part here includes tasks of filtrations of fake information, assessment of credit risk and detection of the risk signs, and prediction of credit events, This can be done with the help of sentiment analysis, word2vec involving graph-based semi-supervised learning, and logistic regression. The proposed idea can help future investors with large historical data and detect risk events beforehand. The algorithm uses large volumes of opinion data to detect the risk signals which will directly affect the decision of buying or selling stocks. The valuable information for the research was taken from the online communities and websites related to stocks investment. Here the classification of the sentiment in words is done using the Graph-based semi-supervised learning. Logistic regression is applied for the prediction of risk events like bankruptcy.

Jamshidi Nejad et al. [2] observed that people experiences of service or product review can be used for learning how to make an appropriate purchase. Algorithms based on opinion mining and NLP could be used for this type of automated analysis. In this work, the authors have worked on Persian language problem aspects of single and multi-words. They have constructed a directed-weighted graph based on information collected using the FP-growth frequent pattern Algorithm applied over the Persian sentence. The work can be extended by polarity detection assigning numeric polarities to sentiments. This work can be done using deep learning and AE neural Networks.

Sánchez-Núñez et al. [3] have concentrated on methods that help finding relevant ads at scale. In this work, the authors have shown how smart and contextual advertising can be achieved through opinion mining and sentiment analysis. A biblio-metric analysis of the methods including computer vision, natural language

processing and also neuroscience here provides an idea about the usage of mining for studying customer reviews & other similar systems. The study also focuses on the information retrieval using deep learning, neural network-based user experience analysis, and other methods also. They also presented the concept of augmented reality used in marketing and also real-time video annotation using scalable mobile image recognition. They also stated that ontology's can be used in the future to properly structure the knowledge used for marketing decisions.

Abas et al. [4] have proposed a new method for opinion mining namely fine-grained aspect opinion mining. It applied a deep neural network where Bidirectional Encoder Representations from Transformers (BERT) is trained and then used for extracting local and global context features. Multithread Self attention model is also proposed here for semantic text representation. The FGAOM model gives more performance as compared to other deep learning models. The Multiple attention mechanism applied here provides better results and produces improved aspect specific sequence representations for text. The future work here can be done in cross domain and multi linguistic analysis models.

Yu et al. [5] have proposed a sentiment unification model for information evaluation called as Fine-grained Topic Sentiment Unification (FG-TSU) model. This hybrid model is designed as an improvement over the Latent Dirichlet Allocation model. The method first divides the text into local and global categories, then using the sliding window lowers co-occurring information to sentence level. Over this, fine grained extraction is applied to indicate aspects and opinions differently. The LDA model is then used to generate the sentiment polarity over aspects. The model here is tested for topic discovery and sentiment & opinion analysis. The LDA model can be extended in future for emotional tendency analysis when combined with sentiment dictionary. The proposed FG-TSU model can also integrate elements of time and user to make it better.

Wu et al. [6] have focused on the voting data and used the model as a discrete random process. The discuss-then-vote model is proposed which is designed using the DeGroot opinion dynamics. The maximum—a-posterior estimator is introduced opinions and influence matrix. Then the convex optimization problem is designed using the derived tractable approximation. The proposed idea here focuses on network dynamic parameters and vote prediction procedure considering the derived matrix. The Bayesian framework is used here to formulate the inference problem which helped in vote prediction.

Zuo et al. [7] have proposed a method for finding opinions about products or events. Here the authors stressed the online public opinion analysis. The proposed method called Aspect based opinion mining concentrates to work on opinionated text and cross-media opinions on the web about products and events. The proposed CAMEL model here performs complementary opinion mining across both common and specific aspects of opinions. The model provides high-quality aspects and opinion mining and can adapt to all types of opinionated text. The evaluation of the model is done with real-world multi-collection review data.



Clavel and Callejas [8], have discussed the opinion mining and human-agent interaction for sentiment analysis. The human-agent interaction deals with different detection and dialog management methods which can be applicable in opinion mining. Here the authors have mentioned the socio-affective human-agent strategies to be applied to sentiment-related phenomena and the sentiment detection methods. They derived that the development of interaction models is possible that consider sentiment analysis over socio-emotional ECA's.

Hai et al. [9] concentrated on the opinion feature extraction over the word distributional characteristics across the different corpora. Here, the authors have proposed a new novel method for opinion feature extraction from the online reviews. For the model, one domain specific corpus and one domain independent corpus is being used. The disparity is captured by domain relevance (DR) that actually characterizes term to text collection relevance. The set of syntactic dependence rules is used to find the list of candidate opinion features, then for each candidate its intrinsic domain relevance IDR & extrinsic-domain relevance EDR scores are taken for both corpora and evaluated. The features that are found to be less generic and are more domain specific are termed as opinion features. To make the approach work better, good quality domain independent corpus is much needed.

Ren and Wu [10] have worked over the twitter data which is the resource providing spontaneous emotional information from the user tweets. This data can be then used for learning and analysis in various domains. The learned knowledge can be used to predict the opinions on interest topics. The authors have proposed a matrix factorization (ScTcMF) model to demonstrate that social context and topical context can be used to improve the user topic opinion prediction. The opinion homophile theory is used here to design a hypothesis about the social context from the extracted data. The task is evaluated using the regularization constraint of social context by evaluating the similarities of the user opinions. The evaluations show that the social and topical context can help improve the performance for user topic opinion prediction. The future work suggests to predict the multiple opinions and emotion states (Table 1).

### 3 Proposed Approach and System Architecture

The proposed algorithm addresses the limitations of these existing methods by forming the word vector dynamically based entirely on the words in the compared sentences. The dimension of our vector is not fixed but varies with the sentence pair and, so, it is far more computationally efficient than existing methods. Our algorithm also considers word order, which is a further aspect of primary syntactic information [1]. Here we identifying the specific class and subclass that a comment belongs to. For example, if a comment is related to price of food then the statement should be categorized under canteen class and rate subclass. For this we needed to first identify the classes and subclasses a prior. For processing the comments there were a number of ways that could have been used. A similarity based approach has been used to develop our project (Fig. 1).

**Table 1** Comparison of existing work

Sr. No	Author	Technique	Description	Pro's	Con's
1	B. Yoon, Y. Jeong and S. Kim	Algorithm based on opinion mining and graph based semi-supervised learning for decision making	The processes of filtrations of fake information, assessment of credit risk and detection of the risk signs, prediction of credit events is done using sentiment analysis, word2vec involving graph-based semi-supervised learning and logistic regression	Effective in decision making, can easily detect hidden events, detects risk events in advance	Handling fake opinions, aggregation of data is time consuming
2	S. Jamshidi Nejad, F. Ahmadi- Abkenari and P. Bayat	Design of Frequent Pattern Mining and Graph Traversal Algorithm for Aspect Elicitation in Customer Reviews for decision making	Methodology for explicit problematic aspect extraction from Persian reviews of single-word and multi-word aspects through weighted directed graph	Robust technique	Deep learning can be used for more better results
3	P. Sánchez-Núñez, M. J. Cobo, C. D. L. Heras-Pedrosa, J. I. Peláez and E. Herrera-Viedma	A Bibliometric Analysis over Opinion Mining, Sentiment Analysis and Emotion Understanding used in Advertising	Smart and contextual advertising through opinion mining and sentiment analysis	Mining is suited for working with marketing and advertisement	Deep Neural Nets also are more useful

(continued)

**Table 1** (continued)

Sr. No	Author	Technique	Description	Pro's	Con's
4	A. R. Abas, I. El-Henawy, H. Mohamed and A. Abdellatif L	A novel deep learning model for fine-grained aspect-based opinion mining, named as FGAOM	Novel cognitive analysis model where Bidirectional Encoder Representations from Transformers (BERT) is trained & then used for extracting local and global context features. Design of Multi-head Self-Attention (MSHA) to fuse internal semantic text representation	Gives more accuracy as compared to other deep learning models, improved aspect-specific sequence representations	The performance of the proposed model in Neutral class prediction is low compared to Positive and Negative classes
5	L. Yu, L. Wang, D. Liu and Y. Liu	Fine-grained Topic Sentiment Unification (FG-TSU) model is proposed based on the improvement of LDA (Latent Dirichlet Allocation) model	Here, text is divided into local and global words. The sliding window is then introduced to lower co-occurrence information from document to sentence level, to implement fine-grained extraction of local words. The indicator variables are then used to distinguish aspects and opinions. Finally, we incorporate the sentiment layer into LDA model to obtain the sentiment polarity of the review and specific aspects	More feasible to complete aspect extraction	LDA makes clustering granularity coarser and unable to identify the evaluated entities properly, definition of opinions could involve the opinions holder and time

(continued)

**Table 1** (continued)

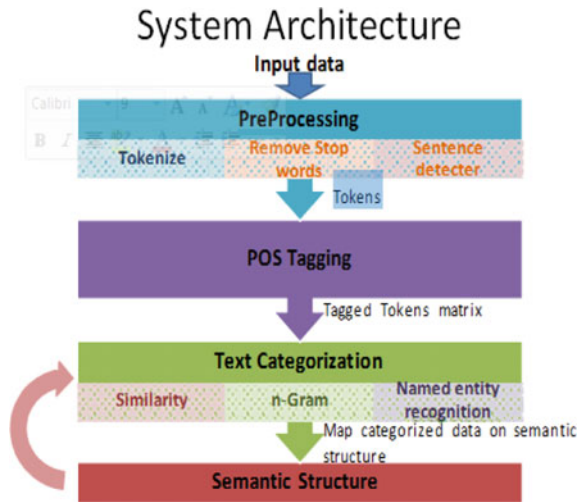
Sr. No	Author	Technique	Description	Pro's	Con's
6	S. X. Wu, H. Wai and A. Scaglione	Estimation of Social Opinion Dynamics Models from Voting Records	Voting data is used to design model based on discrete random process. The discuss-then-vote model is proposed which uses DeGroot opinion dynamics	The influence matrix inferred from the real data is consistent	Upper and lower bounds on the likelihoods of vote prediction are to be decided beforehand
7	Y. Zuo, J. Wu, H. Zhang, D. Wang and K. Xu	Mining aspect-level opinions using Aspect based opinion mining	CAMEL, a novel topic model for complementary aspect-based opinion mining across asymmetric collections	Capable of integrating complementary information from different collections in both aspect and opinion levels	Model can be improved for fragmented information
8	C. Clavel and Z. Callejas	Human-agent interaction for opinion mining and sentiment analysis	The opinion mining and human-agent interaction for sentiment analysis is applied. The human-agent interaction deals with different detection and dialog management methods which can be applicable in opinion mining	More significative sentiment analysis can be done using human-agent interactions	Accuracy of the system is not evaluated
10	Z. Hai, K. Chang, J. Kim and C. C. Yang	Identifying Features in Opinion Mining via Intrinsic and Extrinsic Domain Relevance	A novel method to identify opinion features from online reviews by exploiting the difference in opinion feature statistics across two corpora, one domain-specific corpus (i.e., the given review corpus) and one domain-independent corpus (i.e., the contrasting corpus)	Provides better opinion mining for different corpus	Good quality domain-independent corpus is quite important here

(continued)

**Table 1** (continued)

Sr. No	Author	Technique	Description	Pro's	Con's
11	F. Ren and Y. Wu	Prediction of User-Topic Opinions from the Twitter with Social and Topical Context	Matrix factorization(ScTcMF) model to demonstrate that social context and topical context can be used to improve the user topic opinion prediction. The opinion homophile theory is used here to design a hypothesis about the social context from the extracted data	Help improve the performance for user topic opinion prediction	Also need to consider additional fine-grained features in the future, no unified standard for labelling hash tag producing redundant topics!

Fig. 1 System architecture



A. **Pre-Processing:**

In the pre-processing phase we have performed three main tasks: Sentence detection, Tokenize, Stop words removal.

B. **POS tagging:**

In this part, the parts of speech of the obtained keywords are identified.

Canteen—noun , food—noun, tasty—adjective

Rates—noun, reasonable—adjective.

For POS tagging, we have used penNLP.

i. *OpenNLP:*

The Apache OpenNP library is a toolbox dependent on AI for regular language text preparing. It underpins the most widely recognized NLP errands, like tokenization, division of sentences, grammatical forms labels, named extraction elements, parsing, and coreference goal. These assignments are normally expected to fabricate further developed content preparing administrations. OpenNLP additionally incorporates most extreme entropy AI and perceptron. **Text Categorization:**

For text categorization, we have performed 3 steps:

- Similarity
- N-grams
- Named entity recognition

For similarity we have used LESK and Wordnet.

ii. *LESK:*

Disambiguation is the way toward finding the most proper word sense utilized in specific sentences. Lek calculations use word reference definitions (sparkle) to drop the word polyvis with regard to the sentence. The

fundamental target of the thought is to compute the quantity of words appropriated between two gloses. The more words, progressively identified with the faculties. To deliver the word, shine of every indra contrasted with sparkle of one another word in the expression. A word is relegated to the arrangement that the shine has similar number of similar words as gleam of different words. A model that is frequently utilized depicts this calculation is for the setting of "Pine Cone". The accompanying word reference definition is utilized.

iii. *N-Gram*

In the field of phonetics and the likelihood of figuring, the N-gram is the adjoining grouping of things n from the request for text or welcome given. Things can be phonemes, syllables, letters, words or bases as per the application. N-gram is typically gathered from text or discourse corpus. At the point when the thing is words, n-grams can likewise be called shingles.

C. **Semantic Structure**

After categorizing the data, we have represented the data in the form of a semantic structure. For this, we have the root node which is the main class, the sub-nodes which are the values that belong to the main class and then the values for the sub-nodes which will classify whether the comment is positive or negative.

## 4 System Working

The system consists of following modules:

- A. Tokenizer
- B. POS Tagging
- C. Semantic structure
- D. Categorization
- E. Sentiment analysis

A detailed description of different modules is as follows:

A. ***Tokenizer:***

A tokenizer divides the text into a sequence of tokens, which roughly correspond to "words". Input from the user is taken in the form of paragraphs. Sentence detection is done using the OpenNLP tool. For the separation of sentences, "." character is used. Following is the code snippet for the same.

B. ***POS Tagging:***

Part of the speech tagger denotes a token with the kind of word they are fitting dependent on the symbolic itself and the symbolic setting. Tokens may have a few postal labels relying upon the token and setting. OpenNLP Pos Tagger utilizes the likelihood model to foresee the right post tag of the set tag. To restrict potential labels to tokens, label word references can be utilized that upgrade the runtime checking and execution of Tagger.

String[]posTags = posTagger.tag(tokens);

C. **Semantic structure:**

Knowledge representation is a framework through which we can represent semantic network which can represent semantic relation between concepts related to a particular context. To represent the co-relation between different facilities (classes) in the college, a semantic structure was designed. The design of the structure was done by observing the feedbacks of students. This pre-defined semantic structure has 38 different nodes and sub-nodes. Part of this semantic structure is shown below (Fig. 2):

As it can be seen, College is the root node, which has sub-nodes viz. Faculty, Canteen, Sports, etc. These sub-nodes represent different facilities provided by the college. But these facilities may further have different aspects, depending upon which the facility can be rated. Therefore, each sub-node in turn may have further sub-nodes. E.g. Placement may have sub-nodes as number of companies that visit, packages offered by these companies, industry exposure given to the students, preparation and counselling given for placement, etc. Each sub-node may have different no. children. This structure is helpful in the process of categorization of the sentence. The structure is designed such that each sentence can be made to fit into the structure. Finally each slot will have values such as good, excellent, poor, very bad, etc

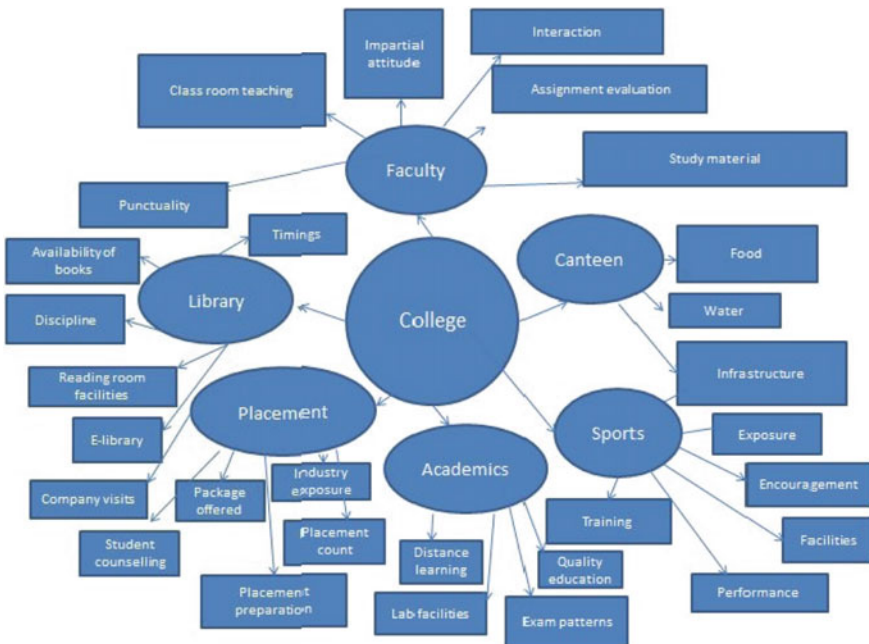


Fig. 2 Semantic structure



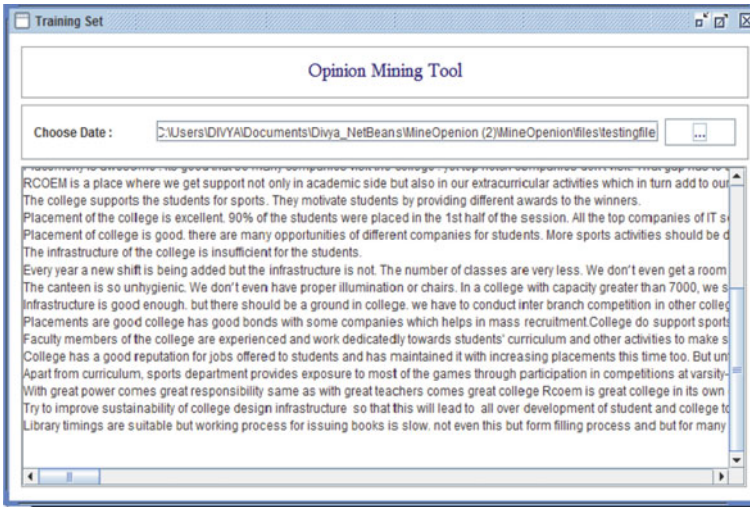


Fig. 3 Home screen

D. **Text Categorization:** Sentences coming from the comments of the students belong to different facilities. There is a need to categorize each sentence into a proper category i.e the facility about which the given comment is to achieve this, similarity between words is calculated. As stated above, the pre-defined semantic structure has total 38 different nodes and sub-nodes. Here, we call them as classes. For every token of each sentence (after the removal of stop words), the similarity of that word (token) with the available classes is calculated.

## 5 Results and Discussion

Following screenshot shows the home screen of the application (Fig. 3):

Result after processing of comments is shown below (Fig. 4):

The final result is shown in the form of a tree structure. The following screenshot shows the result (Fig. 5).

## 6 Conclusion

Here we conclude that this system provides better facilities to its users by allowing them to post their comments without selecting the listed parameter thus saving their time. If there are only a few features available for comments then the students cannot comment on other features they wish to. But here, the students can comment on any number of parameters of the college. In future the project can be extended to work

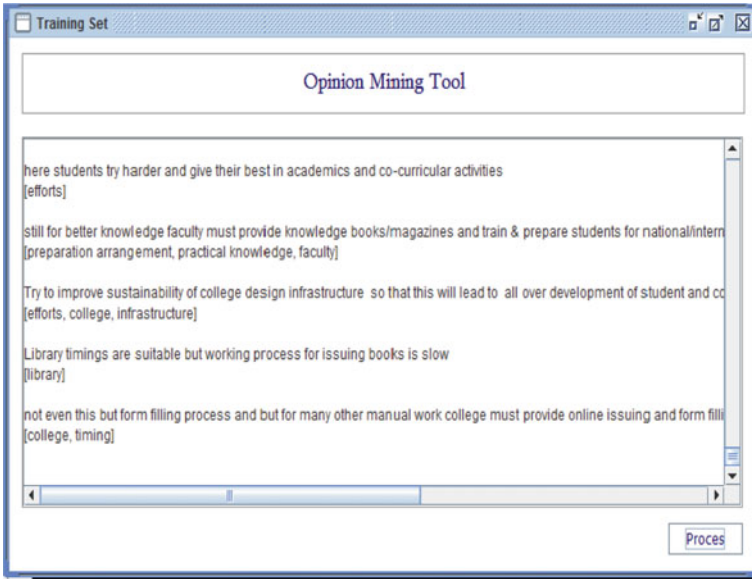


Fig. 4 Result after training

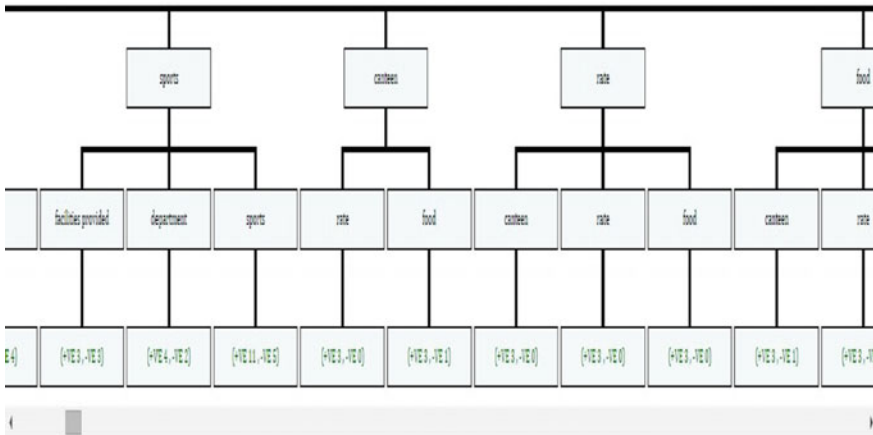


Fig. 5 Resultant tree structure

as an opinion mining tool, which can extract opinions related to any field/area. At present, the application works only for the college domain, but it can be extended so that it can work for any domain. Feature extraction can be added.

## 7 Limitations

Performs classification by ignoring the suggestions given by the users. The suggestions are not properly classified so it becomes difficult to handle the suggestions. Ignores comments posted by users in language other than English. The application works only on those comments that are in English language. The final result shows opinions about a selected parameter, for a selected college. For a given parameter, a direct comparison between different colleges cannot be done. This is a big limitation because students cannot directly compare the reviews of various colleges. At a time, they can view the reviews of only one college. Creates a semantic structure of opinions that are classified under one parameter only. Use of jargons in the opinion is not useful since those words cannot be identified as valid English words also spelling mistakes in the opinion will not provide proper classification of opinion. If there are any spelling mistakes also, It has to be corrected manually because the application cannot work on words having spelling mistakes.

## References

1. Yoon B, Jeong Y, Kim S (2020) Detecting a Risk signal in stock investment through opinion mining and graph-based semi-supervised learning. *IEEE Access* 8:161943–161957. <https://doi.org/10.1109/ACCESS.2020.3021182>
2. Jamshidi Nejad S, Ahmadi-Abkenari F, Bayat P (2020) A combination of frequent pattern mining and graph traversal approaches for aspect elicitation in customer reviews. *IEEE Access* 8:151908–151925. <https://doi.org/10.1109/ACCESS.2020.3017486>
3. Sánchez-Núñez P, Cobo MJ, Heras-Pedrosa CDL, Peláez JI, Herrera-Viedma E (2020) Opinion mining, sentiment analysis and emotion understanding in advertising: a bibliometric analysis. *IEEE Access* 8:134563–134576. <https://doi.org/10.1109/ACCESS.2020.3009482>
4. Abas AR, El-Henawy I, Mohamed H, Abdellatif A (2020) Deep learning model for fine-grained aspect-based opinion mining. *IEEE Access* 8:128845–128855. <https://doi.org/10.1109/ACCESS.2020.3008824>
5. Yu L, Wang L, Liu D, Liu Y (2019) Research on intelligence computing models of fine-grained opinion mining in online reviews. *IEEE Access* 7:116900–116910. <https://doi.org/10.1109/ACCESS.2019.2931912>
6. Wu SX, Wai H, Scaglione A (2018) Estimating social opinion dynamics models from voting records. *IEEE Trans Signal Process* 66(16):4193–4206. <https://doi.org/10.1109/TSP.2018.2827321>
7. Zuo Y, Wu J, Zhang H, Wang D, Xu K (2018) Complementary aspect-based opinion mining. *IEEE Trans Knowl Data Eng* 30(2):249–262. <https://doi.org/10.1109/TKDE.2017.2764084>
8. Clavel and Callejas (2016) Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Trans Affect Comput* 7(1):74–93. <https://doi.org/10.1109/TAFFC.2015.2444846>
9. Hai Z, Chang K, Kim J, Yang CC (2014) Identifying features in opinion mining via intrinsic and extrinsic domain relevance. *IEEE Trans Knowl Data Eng* 26(3):623–634. <https://doi.org/10.1109/TKDE.2013.26>
10. Ren and Wu (2013) Predicting user-topic opinions in twitter with social and topical context. *IEEE Trans Affect Comput* 4(4):412–424. <https://doi.org/10.1109/T-AFFC.2013.22>

11. Jung K, Heo W, Chen W (2012) IRIE: scalable and robust influence maximization in social networks. In: 2012 IEEE 12th International Conference on Data Mining, Brussels, Belgium. pp 918–923. <https://doi.org/10.1109/ICDM.2012.79>
12. Yu X, Liu Y, Huang X, An A (2012) Mining online reviews for predicting sales performance: a case study in the movie domain. *IEEE Trans Knowl Data Eng* 24(4):720–734. <https://doi.org/10.1109/TKDE.2010.269>
13. Conover MD, Goncalves B, Ratkiewicz J, Flammini A, Menczer F (2011) Predicting the political alignment of twitter users. In: 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA. pp 192–199. <https://doi.org/10.1109/PASSAT/SocialCom.2011.34>
14. Chen and Zimbra (2010) AI and opinion mining. *IEEE Intell Syst* 25(3):74–80. <https://doi.org/10.1109/MIS.2010.75>
15. Zhu J, Wang H, Zhu M, Tsou BK, Ma M (2011) Aspect-based opinion polling from customer reviews. *IEEE Trans Affect Comput* 2(1):37–49. <https://doi.org/10.1109/T-AFFC.2011.2>
16. Farra N, Challita E, Assi RA, Hajj H (2010) Sentence-level and document-level sentiment mining for arabic texts. In: 2010 IEEE International Conference on Data Mining Workshops, Sydney, NSW, Australia. pp 1114–1119. <https://doi.org/10.1109/ICDMW.2010.95>
17. Neri F, Aliprandi C, Capecci F, Cuadros M, By T (2012) Sentiment analysis on social media. In: 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, Turkey. pp 919–926. <https://doi.org/10.1109/ASONAM.2012.164>
18. Khan K, Baharudin BB, Khan A, Fazal-e-Malik (2009) Mining opinion from text documents: a survey. In: 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, Istanbul, Turkey. pp. 217–222. <https://doi.org/10.1109/DEST.2009.5276756>
19. Binali H, Potdar V, Wu C (2009) A state of the art opinion mining and its application domains. In: 2009 IEEE International Conference on Industrial Technology, Churchill, VIC, Australia. pp 1–6. <https://doi.org/10.1109/ICIT.2009.4939640>
20. Binali HH, Wu C, Potdar V (2009) A new significant area: Emotion detection in E-learning using opinion mining techniques. In: 2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, Istanbul, Turkey. pp 259–264. <https://doi.org/10.1109/DEST.2009.5276726>

# Event Detection in Live Twitter Streams Using Tf-Idf and Clustering Algorithms



Tavishi Jain, Bhavya Singh, and Rupesh Kumar Dewang

**Abstract** Twitter is a fast emerging form of social media. People use Twitter as a platform to report real-life events. The present paper focuses on extracting those events by analyzing the text stream on Twitter. This paper presents methods that take the tweets in real time as input and generate clusters of tweets denoting different communities as output. The tweets are collected using spark streaming and then pre-processed, and a key graph of keywords is constructed using the tf-idf method. Further community detection is applied on this key graph to generate clusters of tweets as a result.

**Keywords** Event detection · Live tweets · Tf-Idf · K-mean clustering algorithms · Key graph

## 1 Introduction

Twitter has been the most promising news delivery and social media platform for the internet user. As time is passing, the number of users on Twitter is also increasing exponentially. Since the requirements are too high, so we need some technology to satisfy those requirements. A large amount of data in the form of tweets, images, videos and animation is being generated very frequently. There are social media platforms like Facebook, Instagram and YouTube which have changed the way of satisfying needs by using technology. All these social media giants have changed the traditional way of connecting people over the internet. Since Twitter is one of the most popular platforms, that's why we are using Twitter for Event Detection purposes. Twitter is a platform where tweets related to any local or global events are written by users in a few seconds. Natural events like earthquakes, disease outbreaks, etc. are significantly detected by performing event detection on Twitter data [2]. Twitter easily figure out what is happening right now. Tweets are real time data for analysis covering different subjects from various sources. Event detection about emerging

---

T. Jain · B. Singh · R. K. Dewang (✉)

Department of Computer Science & Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India

e-mail: [rupeshdewang@mnnit.ac.in](mailto:rupeshdewang@mnnit.ac.in)

events is highly important if it is performed on real time data [9]. Different people are interested in different types of news and events. Some people want to know about local events [18]. Business oriented organizations are interested in sponsoring their product and services to their favourite customers [4, 5, 13]. Event detection can fulfill the requirements of these people. Event detection can also help in the detection of natural disasters and warning people faster than any other media [7]. Event detection is also helpful in crime detection for example bomb blast or terrorist attack. Event detection in Twitter is effected due to a large amount of meaningless data [15]. Data which is being generated on Twitter is not of high quality, So various high scalable and efficient techniques are being for the pre-processing of the data to make it usable for event detection. Users prefer to write short tweets and they generally use distorted words [12]. Different events consist of different numbers of participants, different time periods and relationships [14]. Real time collection is a big challenge in real time event detection because every time a large amount of data is created on Twitter. Real time event detection has different applications and with technology support. In the present paper, we have proposed the detection of events in live Twitter streams. For construction, the key graph tf-idf scheme is used. The key graph is further used for detecting the communities using the Betweenness Centrality score and then finally clusters are formed using k-mean with cosine similarity. The rest of this paper is structured as follows: Sect. 2 introduces the most relevant related works. Section 3 provides the used dataset description. Section 4 provides a detailed description of the proposed approach. Experimental Setup and Results are discussed in Sect. 5. Section 6 finally concludes the paper with future work.

## 2 Related Work

New Event Detection models do a single step incremental clustering algorithm. When a new document is collected similarity is found between the document along with computation on known events and selection is done on the basis of maximum similarity. A threshold is set, if similarity is more than threshold then the document belongs to a known event else considered as a new event. A Key Graph based approach is used by Hassan Sayyadi, Matthew Hurst and Alexey Markov [6]. They built a key graph based on co-occurrence and used betweenness centrality algorithm for clustering keywords. In this algorithm, they count all keywords in a single community, though a subgroup of the keyword may be better. Modified version of TF/IDF was developed by Allen et al. [1] also penalized the threshold in which the threshold is amended by the time distance between the event and the document. Calculation of IDF using online clustering algorithm is required as to know future document features. The same method is used by Sayyadi et al. in [14]. A supplementary data set is used by Allen et al. [1] in order to find IDF while Yang et al. [16] proposed an incremental IDF factor. Time difference was the basis to find the similarity between documents and events along with consideration of time window and a decay factor that Yang et al. used. To take out past event Yang et. al [16] put forwarded an agglomerative

clustering, GAC (augmented Group Average Clustering). In order to handle cluster quality, they used looping bucketing and re-clustering. A probabilistic model was put forwarded by (Li et al. 2005) for RED and in order to maximize log-likelihood of the distributions they used the Expectation Maximization (EM) algorithm. A significant number of events is required by such algorithm which is not feasible in practice. Li et al. found an approximation of event counts from the article count- time distribution [11]. Event detection models usually use similar algorithms, many different document representation, distance or similarity metrics, and clustering algorithms mentioned in the theory [3, 8, 10, 17].

### 3 Dataset Used

In this paper, we have collected the tweets from Twitter based on a search keyword (hashtag) of “Facebook” and then, this data is given to the spark client where the data is processed in batches. The batch interval used is 1 min in which 80–90 tweets are collected in one batch of spark. So, our one iteration of algorithm runs on approximately 80–90 tweets.

### 4 Proposed Approach

We have collected the data set from Twitter in form of batches in real time using Spark streaming API. Preprocessing is performed on each batch of data which includes Tokenization, removal of Stop words and Stemming. Our approach comprises the following phases: (1) Construction of Key Graph; (2) Identifying Communities in Key Graph; (3) Document Clustering using K-Mean with Cosine similarity.

#### 4.1 Construction of Key Graph

Key graph is made using pre-processed data. We first calculate term frequency (TF), document frequency (DF) and inverse document frequency (IDF). Remove keywords with low DF because these keywords are not useful. Key graph is constructed by taking the remaining keyword as a node of the graph. Co-occurrence of keyword represents the relationship between keywords. Make an edge between keywords which co-occur in the same document. Some edges are present as noise in the key graph. We remove edges which are not satisfying following two conditions-

- (1) If the co-occurrence of two keywords is below some specified threshold value then the edge between these two keywords is removed.
- (2) For the edge of the node ‘X’ and ‘Y’, if there is a potential probability of viewing.

‘X’ in the documents if ‘Y’ is present in the document and if ‘X’ exists then the condition of ‘Y’ in the documents are calculated and if both are smaller than the threshold, then the edge will be removed.

---

**Algorithm 1 Construction of Key graph from tweets**

---

Input: Keyword (hashtag) of tweets

---

Output: Key graph (kgi) of keywords  $k_i$

---

1	BEGIN
2	FOR Extracted all tweets $t_i$ (specified time limit) in real time for corresponding hashtag
3	Pull out words $w_i$ , noun $n_i$ , noun phrases $n_{pi}$ , adverb $a_i$ adjective $a_{di}$ and named entities $nei$ from $t_i$ as keywords $k_i$
4	END FOR
5	FOR all keywords $k_i$ extracted from tweets $t_i$ do
6	Calculate Document frequency $df$ of all tweets $t_i$
7	END FOR
8	Remove keywords $k_i$ with low Document frequency $df$
9	Construct the keyword $k_i$ co-occurrence graph $cgi$ as follows:
	– Generate single vertex $v_k$ for each keyword $k_i$
	– Generate single link $l_{ij}$ for every set of co-occurring keywords $k_i$ and $k_j$
10	END

---

## 4.2 Identifying Communities in Key Graph

Edges of the graph represent the relationship of keywords and the whole key graph presents a social network of keywords. Communities of keywords are represented by highly dense graphs. Communities have a large number of edges because there exist more relationships between keywords in the community. Between different communities, there are few links. Betweenness Centrality score is used to find the edges between communities. Betweenness centrality score for any node is the number of shortest paths between all pairs of graphs which pass through that node. Edges between communities come across many shortest paths so these edges have a high betweenness centrality score. These edges are connecting the different communities so after removing edges with high betweenness centrality score we will get different clusters of keywords and each cluster represent different communities. Each community is the hypothesis of an event. This is an iterative process. In each iteration, we remove edges with low BC scores. For this, we first calculate the BC score by finding the shortest path between all node pairs of graphs. We use BFS to find the shortest path. If two edges have the same BC then the edge with low conditional probability is removed. If the edges which are removed belong to two different communities and have high conditional probability then edge is duplicated in both communities and



then removed. We again calculate BC score for the next iteration and perform the same operation until all edges with a high score are not removed. Finally, we get a cluster of keywords representing events.

---

**Algorithm 2 Identifying Topic Features**

---

Input: Set of key graphs (kgi)

---

Output: Topic ( $T_i = t_1; t_2; t_3::: t_n$ ) feature sets ( $F_i = f_1; f_2; f_3::: f_m$ )

---

1	Separate Key Graph (kgi) in communities coi of strongly related ( kiekoi) keywords ki as topics $T_i$
2	FOR all topic tieT do
3	Place all keywords ki in the related feature vector ( $FV_i = f v_1; f v_2; f v_3::: f v_n$ ) of related keyword partition fp
4	END FOR

---

### 4.3 Document Clustering Using K-Mean

Each group of keywords make a synthetic document called as Key Document. All documents can be grouped in the original collection, similar to this synthetic document, thus a group of periodical documents is obtained. To discover document clusters, k-mean with cosine similarity is used. Sometimes keywords are common in an important document, or the main document has few keywords. It makes a general class of documents. These important documents make a set of documents related to higher levels. For example, a key document containing only “Modi”, “Rahul”, “Votes”, and “Election” indicate the Indian general elections which is not a separate event. We can find such important documents with help of the similarity of documents related to an important document. Documents are data points, and the variance of documents related to such key documents will be very large. So, we calculate the variance of documents for every main document and then filter those key documents that have a large variance. This makes it easier to find those key documents which truly represent events. However, documents allocated to key clusters are based on the cosine similarity of documents to the key documents, some of the documents are filtered later to reduce noise. Apart from these, the similarity between the documents and the centroid of each cluster are also calculated and filter those documents which have low cosine similarity to the centroid.

**Algorithm 3 Document Clustering (K-mean with cosine similarity) using Topics**

Input: Topic (Ti = t1;t2;t3:::tn) feature sets (Fi = f1; f2; f3::: fm)

Output: Topics (Tl = t1;t2;t3:::tm) under the cluster documents (cd1;cd2;cd3:::cdn)

1	FOR all topic ti
2	Initialize the k value and select as the initial centroid value ci&cj
3	Calculate the similarity (s) of topic ti for document di: s(ti = di) = cos(di;FVi) $\sum(tieT)\cos(di;FVi)$
4	Recalculate the centroid of each cluster until centroid value not changed
5	END FOR
6	FOR complete topic ti and tl do
7	IF the topic T overlap of ti and tl
8	Combined ti and tl into a new topic NT where FVi = f vti + f vtl
9	END IF
10	END FOR

**4.4 Example of Construction Keygrpah, Identifying Topic Features and Document Clustering (K-means with Cosine Similarities) Using Topics**

Consider the graph in the above Fig. 1, the nodes represent the keywords and the edges represent the co-occurrence of the two keywords between which it is present. Consider all the edge weights to be equal to 1. The following table gives the values of betweenness centrality scores calculated for every edge using the Breadth First Search algorithm.

**Cosine Similarity**

Let us say we have multiple documents and we need to determine the similarity between those documents. Let us name the two documents as document1 and document2 respectively. A document can be represented by a bag of terms or a long vector, with each attribute recording the frequency of a particular term (such as word, keyword or phrase) in the document. We will be having two term frequency vectors (d1 and d2). Table 2 shows that d1 and d2 denote term frequency in doc1 and doc2 respectively.

$$\cos\theta = \frac{d1.d2}{|d1||d2|}$$

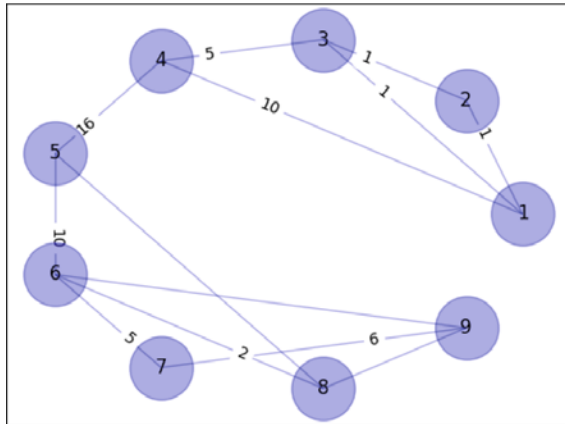


Fig. 1 Initial keyword graph

**Finding similarity between document d1 and d2:**

$d1 = 5,0,3,0,2,0$  and  $d2 = 3,0,2,0,1,1$

1. First, calculate the dot product of these two documents

$$d1.d2 = 5 * 3 + 0 * 0 + 3 * 2 + 0 * 0 + 2 * 1 + 0 * 1 = 23$$

2. Then calculate  $kd1$  and  $kd2$

$$||d1|| = \sqrt{5 * 5 + 0 * 0 + 3 * 3 + 0 * 0 + 2 * 2 + 0 * 0} = 6.164$$

$$||d2|| = \sqrt{3 * 3 + 0 * 0 + 2 * 2 + 0 * 0 + 1 * 1 + 1 * 1} = 3.873$$

3. Calculate cosine similarity:

$$\cos(d1, d2) = \frac{23}{6.164 * 3.873} = 0.96$$

**5 Experimental Setup and Results**

We have used the above presented algorithms on batches of tweets formed by passing the live tweets coming from the spark streaming API by using a search keyword. Batch interval used for the experiment is 1 min in which 80–90 tweets are collected in one batch of spark. So, our one iteration of the algorithm runs on approximately 80–90 tweets. In the first step of the algorithm, a key graph is formed which contains nodes and edges between any two nodes if they co-occur. Here we obtained approximately

40 keywords(nodes). After that, clusters of keywords are formed from the key graph and then tweets are categorized into these clusters of keywords using cosine similarity (Tables 1 and 2).

**Table 1** Betweenness centrality score

Node 1	Node 2	Betweenness centrality
1	2	1
1	3	1
1	4	10
1	5	0
1	6	0
1	7	0
1	8	0
2	3	1
2	4	0
2	5	0
2	6	0
2	7	0
2	8	0
3	4	0
3	5	0
3	6	0
3	7	0
3	8	0
4	5	16
4	6	0
4	7	0
4	8	0
5	6	10
5	7	0
5	8	5
6	7	6
6	8	2
7	8	0

**Table 2** Calculation example

Doc	Team	Coach	Hockey	Baseball	Soccer	Penalty
d 1	5	0	3	0	2	0
d 2	3	0	2	0	1	1

### 5.1 Batch wise Key graphs And Categorization of Tweets

In this section, we have shown the results obtained by running the above approach in batches. The algorithm is applied on each batch and the results are hence depicted batch-wise.

Batch 1 Results In Fig. 2, the initial key graph and final graph of batch 1 results. The figure shows the graph formed initially after pre-processing of tweets and also filtering of keywords based on document frequency. Here, the keywords represent the node of the graph and the edges between them show their co-occurrence. The final key graph is the formed after clustering of keywords phase of the algorithm. It contains edges only between those nodes which belong to a particular cluster. Table 3 shows the tweets belonging to a particular category. For example, number 2 represents the tweet belonging to this category.

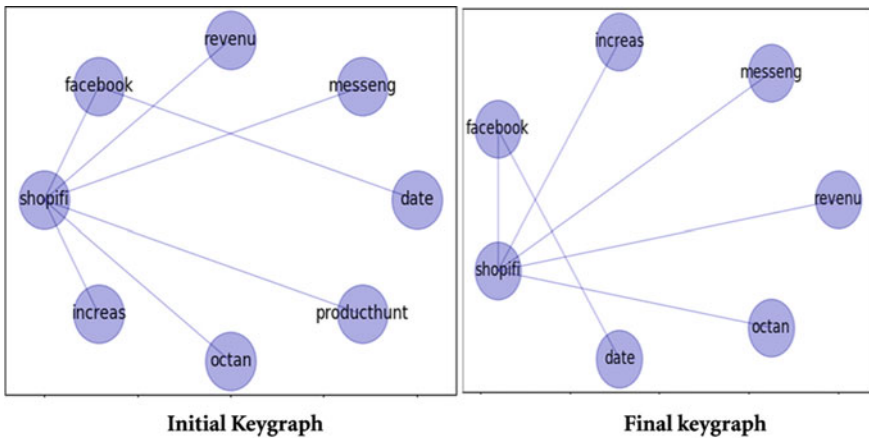
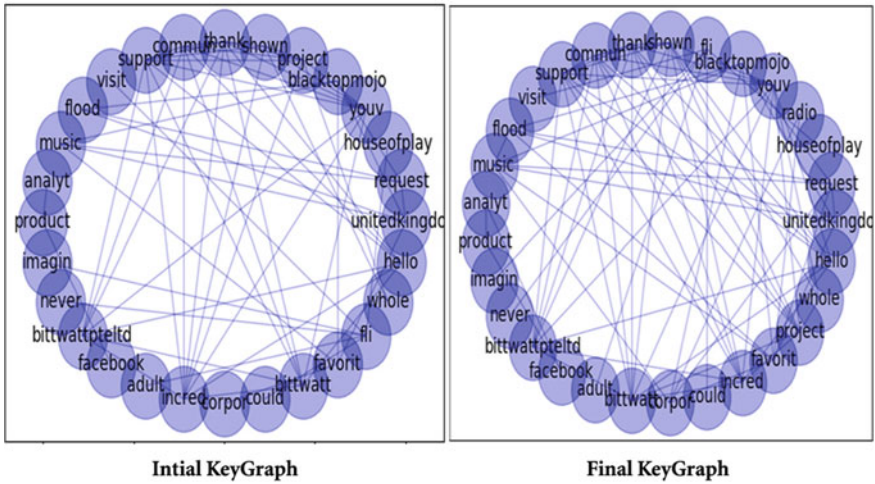


Fig. 2 Initial and Final Key Graph of Batch 1 Results

Table 3 Categorization of tweets btach-1

Category No	Tweets Belonging to this category
2	[‘WhatsApp co-founder flees Facebook just as fb gets underway <a href="http://t.co/yF6wsKEnOi">http://t.co/yF6wsKEnOi</a> ’]
6	[‘TR @Mdproductionsmd: HELLO MAY- What a month it is going to be with some exciting things happening. Check out what we have on at MD HQ th’]
16	[‘Facebook partners with RED to develop a high-end, professional VR camera <a href="https://t.co/MSVSdADKVg">https://t.co/MSVSdADKVg</a> ’]
35	[‘Dont worry Son day of judgement is for a reason.’]
37	[‘#facebook #facebookmarketing #zuck #page #b’]
41	[‘RI @blackmirror: 1000 simulations completed. <a href="https://t.co/mpxcDitBVQ">https://t.co/mpxcDitBVQ</a> ’]

Batch 2 Results Similarly, for batch 2, Fig. 3 represents the initial key graph and final key graph. Table 4 shows the categorization of tweets.



**Fig. 3** Initial and Final Key Graph of Batch 2 Results

**Table 4** Categorization of tweets btach-2

Category No	Tweets Belonging to this category
12	[‘And that is why heaven and hell are for ETERNITY! <a href="https://t.co/k2RQZA74KT">https://t.co/k2RQZA74KT</a> ’]
17	[‘Worth a quick refresher, for all of us parents, <a href="https://t.co/ptGLaWp5kZ">https://t.co/ptGLaWp5kZ</a> ’]
22	[‘Facebook. Into twitter. http’]
23	[‘Facebook is trying to close the book on Cambridge Analytica: <a href="https://t.co/SYcxIfCRcY">https://t.co/SYcxIfCRcY</a> ’]
24	[‘Seven year old girl owns riding in the rain’]
25	[‘Reminder- in 1 h Online Marketing Facebook Series, click on the link an d register! <a href="https://t.co/XgoeDMuNJO">https://t.co/XgoeDMuNJO</a> ’]
33	[1- red coloured vinyl I <a href="https://t.co/rRiGAUatt6">https://t.co/rRiGAUatt6</a> ’]
40	[‘RT @blackmirror: 1000 simulations completed. <a href="https://t.co/mpxcDitBVQ">https://t.co/mpxcDitBVQ</a> ’]
71	[Spotify, Soundcloud, and GoPro can now let their users to share to facebook and Instagram stories <a href="https://t.co/mqkISPEHmv">https://t.co/mqkISPEHmv</a> ]
72	[‘I want this!! <a href="https://t.co/fa1Xz8YF8P">https://t.co/fa1Xz8YF8P</a> ’]
82	[Not new product on Product Hunt: Facebook Analytics Facebook’s new analytics app on iOS and Android <a href="https://t.co/cRFz82Akw8">https://t.co/cRFz82Akw8</a> ’]

## 6 Conclusion and Future Work

In this paper, we have used a key graph based approach for clustering keywords to find out events from live tweets. Presently, the batch interval used is small due to the amount of memory spark needed. Therefore, further work can be done to run the above algorithm on a distributed system so that algorithm can be run on a large dataset and can produce better results. Apart from that, the complexity of the algorithm becomes too high when used for live processing therefore, another approach needs to be formulated to reduce the complexity.

## References

1. Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: *Sigir*, vol 98, pp 37–45. Citeseer
2. Boyd DM, Ellison NB (2007) Social network sites: definition, history, and scholarship. *J Comput-Mediat Commun* 13(1):210–230
3. Brants T, Chen F, Farahat A (2003) A system for new event detection. In: *Proceedings of the 26th annual international ACM SIGIR conference on research and development in informaion retrieval*. ACM, pp 330–337
4. Dewang RK, Singh AK (2018) State-of-art approaches for review spammer detection: a survey. *J Intell Inf Syst* 50(2):231–264
5. Farzindar A, Inkpen D (2012) *Proceedings of the workshop on semantic analysis in social media*. In: *Proceedings of the workshop on semantic analysis in social media*
6. Hasan M, Orgun MA, Schwitter R (2018). Real-time event detection from the twitter data stream using the twitternews+ framework. *Inf Process Manag*
7. Java A, Song X, Finin T, Tseng B (2007) Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on web mining and social network analysis*. ACM, pp 56–65
8. Kumaran G, Allan J (2004) Text classification and named entities for new event detection. In: *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval*. ACM, pp 297–304
9. Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: *Proceedings of the 19th international conference on world wide web*. ACM, pp 591–600
10. Lam W, Meng H, Wong K, Yen J (2001) Using contextual analysis for news event detection. *Int J Intell Syst* 16(4):525–546
11. Li Z, Wang B, Li M, Ma W-Y (2005) A probabilistic model for retrospective news event detection. In *Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval*. ACM, pp 106–113
12. Nallapati R, Feng A, Peng F, Allan J (2004) Event threading within news topics. In: *Proceedings of the thirteenth ACM international conference on information and knowledge management*. ACM, pp 446–453
13. Sabeeh A, Dewang RK (2018) Comparison, classification and survey of aspect based sentiment analysis. In: *International conference on advanced informatics for computing research*. Springer, pp 612–629
14. Sayyadi H, Hurst M, Markov A (2009) Event detection and tracking in social streams. In: *Third international AAAI conference on weblogs and social media*
15. Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from twitter posts. In: *International conference on social computing, behavioral-cultural modeling, and prediction*. Springer, pp 231–238

16. Yang Y, Pierce T, Carbonell JG (1998) A study on retrospective and on-line event detection
17. Yang Y, Zhang J, Carbonell J, Jin C (2002) Topic-conditioned novelty detection. In: Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 688–693
18. Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Who, where, when and what: discover spatio-temporal topics for twitter users. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 605–613



# Event Detection and Summarisation of Live Tweets Using SCAN Algorithms



Tavishi Jain, Bhavya Singh, and Rupesh Kumar Dewang

**Abstract** Twitter is one of the profitable sources of event-related data, to be specific, breaking news and local and global event reports. Due to its capacity to transmit data progressively, Twitter is exploited to detect events and generate a summary. Such detected events and their summaries can help the users and organisations to acquire actionable knowledge and respond to the situations accordingly. The present paper focuses on extracting events by analysing the text stream in Twitter in real-time. Our approach includes 3 stages: Forming key graph, forming clusters of keywords using SCAN (Structural Clustering Algorithm for Networks) and then summarising events. In our proposed approach, all the above processing is done on tweets collected in batches using spark streaming to analyse their results in real-time, enabling users and organisations to respond immediately.

**Keywords** Event detection · Live Tweets · Tf-Idf · K-mean clustering algorithms · Key graph

## 1 Introduction

An event is an unusual activity that happens in a particular area and at a particular duration of time. Kwak et al. studies show that Twitter is one of the first mediums to broadcast important events such as earthquakes and tsunamis, often within a very short span of their occurrence [6]. Thereby tweets are an important and interesting source of "real-time" data. Here, the phrase "real-time" indicates the most relevant data at the current point in time. The freshness of data collected is very much a consideration. This event-related information can be highly valuable if analysed in real-time [4]. Studying those data can provide us with useful information. What is happening right now? It is a fascinating question that is asked by many people

---

T. Jain · B. Singh · R. K. Dewang (✉)

Department of Computer Science & Engineering, Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India

e-mail: [rupeshdewang@mnit.ac.in](mailto:rupeshdewang@mnit.ac.in)

every day. Local events are of greatest interest to the people [15]. Corporations are interested in sponsoring their product to favourable customers [2]. Event detection can answer this question. Besides that, Twitter might detect natural disasters and warn people even faster than other media [5]. Some predictions can also be completed from Twitter data, such as the crime prediction [12]. Hence, the Live streaming analysis of data from Twitter is motivated by the fact that this stream processing allows the processing of data in real-time, thereby allowing users to analyse that information as soon as it is generated and take decisions accordingly. The basic purpose of the paper is to perform Event Detection and summarisation of live tweets for identifying important events relevant at the current point. To process the live tweets, we have used spark streaming API to process tweets collected in batches. And to perform Event Detection, we have used the concept of a key graph given in the paper "Event Detection And Tracking in Social Streams by Hassan Sayyadi, Mathew Hurst and Alexey Maykov". Sayyadi et al. have discussed that the tweets are represented as a graph in which keywords as nodes and their co-occurrence in any tweet can be represented as an edge between them. Hence, to detect events, it is sufficient to identify clusters of keywords in that key graph because keywords that belong to any particular event must belong to any one cluster of keywords. The modularity-based approach has been used to identify clusters of keywords in the graph [3, 7–9]. In the present paper, we have used an approach called SCAN: A Structural Clustering Algorithm for Networks, whose concept have been given by Xu et al. [13].

SCAN has been used for the formation of clusters in any network. So, we have used this concept on the key graph to form clusters of keywords that can represent an event. After determining tweets belonging to a particular event, it is important to analyse them to determine which tweets of a particular event are significant and not. We have generated summaries to find out the most representative tweet of a particular event to fulfil this objective. The concept used for the purpose has been taken from the paper "Real-time Timeline Summarisation for High-Impact Events in Twitter by Yiwei Zhou, Nattiya Kanhabua and Alexandra I. Cristea". We have used the approach given in this base paper to generate summaries of detected events [16]. The rest of this paper is structured as follows: Sect. 2 introduces the most relevant related works. Section 3 provides a detailed description of the proposed approach. Experimental Setup and Results are discussed in Sect. 4. Section 5 concludes the paper with future work.

## 2 Related Work

Event detection has gained much research nowadays [17] into graph partitioning widely used to detect events. Therefore, several methods have been developed for the same. The objective of these methods is to partition the graph such that the edges within a cluster are dense and the edges between the clusters are rare. Modularity-based algorithms and normalised cuts are some examples. The min–max cut strategy segments a graph  $G$  where  $G \rightarrow (v,x)$  into two clusters  $X$  and  $Y$  [16] max clustering

limits the number of associations among X and Y and amplifies the number of associations inside each cluster. A downside of the min–max cut technique is that, on the off chance that one removes a solitary vertex from the diagram, one likely accomplishes the ideal. In this manner, practically speaking, the enhancement must go with some imperative; for example, X and Y ought to be of equivalent or comparable size. Such imperatives are not constantly suitable; for instance, in informal communities, a few networks are a lot bigger than the others [1, 2, 10].

Later, better clustering algorithms under the umbrella of modularity-based algorithms were proposed. One of the modularity-based techniques is the betweenness centrality  $m$  of  $d$  [11, 13]. In this method, the base clustering decision is that the clusters' keywords are more densely connected than those belonging to different clusters. In other words, intra-cluster connections (edges) are denser, and inter-cluster connections are rarer. This method can be node-based or edge-based. Betweenness centrality for an edge is defined as the shortest paths for all pairs of nodes in the network that pass through that edge. All edges with high betweenness centrality are removed in an iterative process, thus cutting all inter-community connections and splitting the graph into several components, each corresponding to one community [12]. The modularity-based approach based on the betweenness centrality score is not very efficient because their complexity is very high as in every iteration of the algorithm, the path between all pairs of nodes is to be calculated. Therefore, the basic aim of using the SCAN algorithm for the key graph is to reduce the complexity of forming clusters of keywords for detecting events [14].

### 3 Proposed Approach

This section gives the details of the used dataset, the proposed approach of the key graph-based clustering algorithm and the summarisation algorithm.

#### 3.1 Dataset Used

In this paper, we have collected the tweets from Twitter based on a search "Facebook" keyword, and then, this data is given to the spark client, where the data is processed in batches. The batch interval used is 5 min, in which 400 tweets are collected in one batch. So our one iteration of the algorithm runs on approximately 400 tweets.

### 3.2 Proposed Approach

Our approach comprises the following phases. (1) Construction of KeyGraph (2) Construction of Clusters of keywords using SCAN Algorithm (3) Summarization of Events.

**Construction of Key Graph.** In this phase, a key graph is formed by using keywords as nodes and edges between those nodes are formed when those terms co-occur in a tweet. For this, we first extracted the keywords and then for each keyword, we calculated the term frequency (TF), document frequency (DF) and the inverse document frequency (IDF). After that, we filter the keywords with low document frequency and then create edges for each remaining keyword. To further reduce the noise in the data, the following two criteria are followed: 1) We remove an edge if the keywords associated with its nodes co-occur below some minimum threshold. 2) And, for a particular edge between ' A ' and ' B ', if the conditional probability of the occurrence (the probability of seeing ' A ' in a document if ' B ' exists in the document), and the conditional probability of the occurrence (the probability of seeing ' B ' in a document if ' A ' exists in the document) are calculated and if both of them are smaller than the defined threshold, the edge removed. Hence, the output of this phase is the graph of keywords. The pseudo-code for the above is shown below:

**Algorithm 1 Construction of Keygraph from tweets**

Input: Keyword (hashtag) of tweets	
Output: Keygraph (k <sub>gi</sub> ) of keywords k <sub>i</sub>	
1	BEGIN
2	FOR Extracted all tweets t <sub>i</sub> (specified time limit) in real-time for the corresponding hashtag
3	Pull out words w <sub>i</sub> , noun n <sub>i</sub> , noun phrases np <sub>i</sub> , adverb ai adjective adi and named entities nei from t <sub>i</sub> as keywords k <sub>i</sub>
4	END FOR
5	FOR all keywords k <sub>i</sub> extracted from tweets t <sub>i</sub> do
6	Calculate Document frequency df of all tweets t <sub>i</sub>
7	END FOR
8	Remove keywords k <sub>i</sub> with low Document frequency df
9	Construct the keyword k <sub>i</sub> co-occurrence graph c <sub>gi</sub> as follows:
	– Generate a single vertex v <sub>k</sub> for each keyword k <sub>i</sub>
	– Generate single link l <sub>i,j</sub> for every set of co-occurring keywords k <sub>i</sub> and k <sub>j</sub>
10	END

Construction of Clusters of Keywords Using SCAN Algorithm In this phase, we apply the SCAN algorithm on the key graph obtained from the above phase. SCAN

algorithms form clusters of keywords based on the direct connectivity of keywords. The terminologies and the algorithm for SCAN are described below. Hence, at the end of this phase, we have clusters of keywords representing the keywords of an event. SCAN: A Structural Clustering Algorithm for Networks Here, we used the SCAN algorithm, which applies the search for clusters, hubs and outliers [13]. Suppose that graph  $G$  is such that  $G = fN, Eg$ , where  $N$  represents the set of vertices or nodes and  $E$  represents the set of edges or links. So, the basic Terminologies used in the SCAN algorithm are:

- **Node Structure:** Let  $n \in N$ , is the structure of  $n$ . It is explained by its locality, denoted by  $\Gamma(n)$

$$\Gamma(n) = w \in N \mid (n, w) \in E \cup n \tag{1}$$

- **Structural Comparability:** Structural comparability into two nodes is given by:

$$\sigma(n, w) = \frac{|\Gamma(n) \cap \Gamma(w)|}{\sqrt{|\Gamma(n)| |\Gamma(w)|}} \tag{2}$$

The vertices that belonged to one cluster have high structural comparability among them.

- **$\epsilon$ -locality definition:**  $\epsilon$ -locality for a node  $n$  is given by:

$$M_\epsilon(n) = w \in \Gamma(n) \mid \sigma(n, w) \geq \epsilon \tag{3}$$

A node is defined as a core vertex or node when it shares structural comparability with a sufficient locality. Hence, main nodes or vertices have at minimum  $\mu$  locality with structural comparability that exceeds the limit  $\epsilon$ . From main nodes or vertices, we form the clusters.

- **Core C:** A node or vertex  $n \in N$  is defined as a core w.r.t  $\epsilon$  and  $\mu$ , if its  $\epsilon$ -locality contains at minimum  $m$  nodes or vertices, proper:

$$C_{\epsilon, \mu}(n) \leftrightarrow |M_\epsilon(n)| \geq \mu \tag{4}$$

We form clusters through main nodes or vertices as we continue. A node or vertex is in  $\epsilon$ -locality of a main, and it must be further in a similar cluster.

- **Direct Structure Reachability (DSR):** The direct structure reachability can be distinguished as follows:

$$DSR_{\epsilon, \mu}(v, w) \leftrightarrow C_{\epsilon, \mu}(v) \wedge w \in M_{\epsilon, \mu}(v) \tag{5}$$

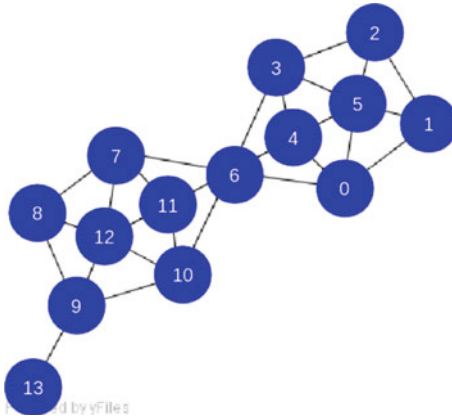
**Algorithm 2: Construction of Clusters of Keywords Using SCAN Algorithm**

Input:	Keygraph (N, E) where N is the representative set of nodes or vertices and E is the represent a set of links or edges between nodes of the key graph
Output:	Clusters of keywords as a set of keywords that together represents keywords of an event
1	BEGIN
2	all nodes or vertices in n are labelled as unclassified
3	FOR each unclassified node or vertex $n \in N$ , do
4	IF n is a core c, then
5	Create new cluster_ID cid to expand new cluster Cn
6	insert all nodes or vertices N in epsilon-locality of n in Queue Q
7	WHILE queue Q is now empty, do
8	pop out a vertex y from queue Q
9	find out the vertices directly reachable from y, say the set as R
10	FOR each node or vertex x in R, do
11	IF x is unclassified or non-member
12	assign current cluster_ID to x
13	END IF
14	IF x is unclassified
15	insert x into queue Q
16	END IF
17	END FOR
18	remove y from Q
19	END
20	ELSE
21	IF n is not a core, it is labelled as a non-member
22	END IF
23	END FOR

Example: Let us apply the SCAN algorithm to the following graph.

Let us take the values of  $\epsilon$  and  $m$  as follows:  $\epsilon = 0.7$ ,  $\mu = 2$  According to the algorithm, initially all the vertices are unclassified, hence, are blue coloured. Now, we find out for each vertex whether it is a core or not. For identifying the core, we calculate the following metrics for each vertex as mentioned above in the algorithm, Vertex Structure, Structural Similarity,  $\epsilon$ -Locality. Following colour notation is used in this example: (1) Blue colour represents unclassified vertex. (2) Orange colour shows the calculation for that vertex. (3) Brown colour represents that vertex has been processed. (4) Green colour represents the vertices of a particular cluster. Let us take vertex number 13: First, we calculate  $\Gamma$  for all the vertices as follows:

$$\begin{aligned} \Gamma(13) &= 9,13 & \Gamma(9) &= 9,13,8,12,10 & \Gamma(8) &= 8,9,12,7 & \Gamma(7) &= 7,8,12,11,6 \\ \Gamma(10) &= 10,9,12,11,6 & \Gamma(12) &= 12,8,7,11,10,9 & \Gamma(11) &= 11,7,6,10,12. \\ \Gamma(6) &= 6,7,11,10,3,4,0 \end{aligned}$$



**Fig. 1** Example of SCAN algorithm

Similarly, we find out  $\Gamma$  for all the vertices. Now, we calculate the structural similarity of vertex number 13 with every other vertices:

$$\begin{aligned} \sigma(13, 13) &= 1 \quad \sigma(9, 13) = 0.632 \quad \sigma(8, 13) = 0.356 \quad \sigma(12, 13) = \frac{1}{\sqrt{12}} \\ &= 0.288 \quad \sigma(10, 13) = 0.316 \quad \sigma(13, 7) = 0 \quad \sigma(13, 6) = 0 \quad \sigma(13, 11) = 0 \end{aligned}$$

Similarly, we find out its structural similarity with the rest of the vertices. Now, we find out  $\varepsilon$ -neighbourhood for vertex number 13: As the structural similarity of vertex number 13 with any other vertex is not greater than  $\varepsilon$ , so:  $N_\varepsilon(13) = \{13\}$  Hence:  $|N_\varepsilon(13)| = 1$  as:  $|N_\varepsilon(13)| \ll \mu$ , hence, 13 is not core. Calculations for the rest of the nodes or vertices are shown in Figs. 1 and 2.

Summarisation of events each cluster formed above contains keywords for that cluster. Hence to find out whether the tweets belonged to a particular cluster of keywords, we use cosine similarity. After clustering original tweets, we apply a summarization algorithm in which each cluster of documents is given as input and produces the summary of each cluster as output. Hence, the results are displayed as a summary of each cluster of tweets. The summary represents the representative tweet belonging to that cluster.

Summarisation Algorithm because every one of the tweets in the same cluster is close copy tweets, we select the most agent tweet in each cluster as the outline of the event.

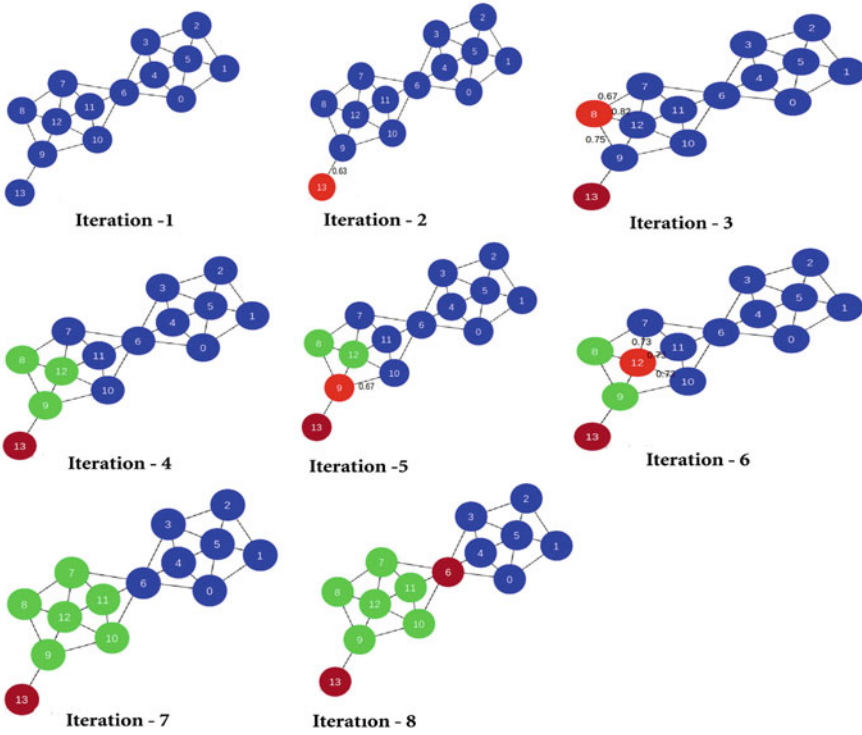


Fig. 2 All Iteration of construction of keygraph

---

**Algorithm 3: Summarisation Algorithm**

---

Input: Clusters of tweets  $c$

---

Output: Summary of events

---

1. BEGIN

---

2. Initialise  $MaxSimilarity = 0$

---

3. FOR each new tweet  $t_i$  in cluster  $c$  do

---

4. Initialise a cluster  $c_i$  with  $t_i$ 's key terms  $K_i$

---

5. IF  $GetSimilarity(c_i, c) > MaxSimilarity$  do

---

6. //  $GetSimilarity(c_1, c_2) = J_{jw}(K_1, K_2) = \frac{k_1 \cap k_2}{k_1 \cup k_2}$

---

Where:  $c_1$  and  $c_2$  denote two clusters,  $K_1, K_2$  denote the key terms mentioned in  $c_1; c_2$

---

7.  $MaxSimilarity = GetSimilarity(c_i, c)$

---

8.  $Tsc \leftarrow Tt_i$

---

9. END IF

---

10. END FOR

---



## 4 Experimental Setup and Results

We have experienced the above algorithm on batches of tweets formed by passing live tweets from the spark streaming API using a search keyword. The Batch interval used for the experiment is 5 min, in which 400 tweets are collected in one batch of sparks. So our one iteration of the algorithm runs on approximately 400 tweets. In one iteration of the algorithm or one batch of spark processing, first of all, the following 3 preprocessing steps are performed:

- Keywords are extracted from tweets.
- Stopwords are removed to remove irrelevant words.
- Then, the remaining words are stemmed to convert to their root word.

Then, a key graph has been formed, which contains approximately 250 keywords as nodes. After that, when the SCAN algorithm was applied, we got approximately 20 clusters containing approximately 15 tweets. According to the SCAN algorithm, the rest of the tweets remain unclustered, termed hubs and outliers. Then, a summarization algorithm has been applied, which finds out the most representative tweet of a cluster and summarises the events.

### 4.1 Batch Wise Keygraphs and Summary

As we have used Spark streaming, so, we got the results in batches. Therefore, the results are shown in batches. The initial graph shown is the graph formed after phase 1 of the key graph's formation, and the final graph shown is the graph formed after phase 2 of the formation of clusters of keywords. And the summary represents the summary of each cluster of tweets.

**Batch 1** Graph in Fig. 3 is formed in batch 1, which contains the keywords shown as nodes and edges representing their co-occurrence. In this graph, if any two keywords are connected via an edge, they must have co-occurred in at least 1 tweet.

In the graph shown in Fig. 3, keywords are the same as in the initial graph, but some edges have been removed. So, the keywords that are connected belong to one particular cluster. Table 1 shows the actual results generated after running the above algorithm. The enumeration of the summary of events is shown. After that, it shows the keywords that belonged to that event, e.g. video and post are keywords that belonged to the event numbered. After that, it shows all the tweets that belonged to that event. And then, it shows the most representative tweet of that event as a summary of that event.

**Batch 2** Similar to batch 1 results, the graph in Fig. 4 is formed in batch 2, which contains the shown keywords as nodes and edges representing their co-occurrence.

The graph shown in Fig. 4 shows that keywords are the same as the initial graph, but some edges have been removed. So, the keywords that are connected belong to one particular cluster. Similarly, Table 2 shows the actual results generated in batch 2

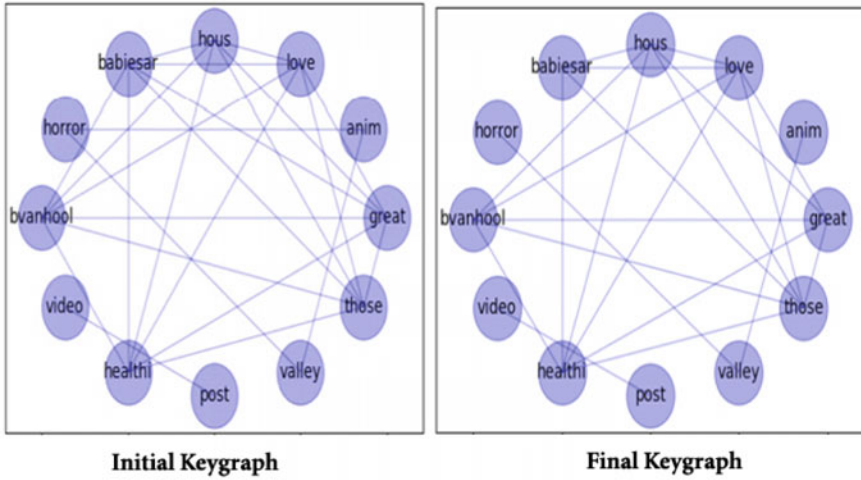


Fig. 3 Initial and final keygraph of batch 1 results

Table 1 Batch 1 summary

Sr No	Keywords of a particular event	Summary of the particular event
1	video, post	I posted a new video to Facebook <a href="https://t.co/XDZzhpN8e">https://t.co/XDZzhpN8e</a>
2	great, healthy, love, hous, bvanhool, those babiesar	@Cover Shadow is 100% true. As if those didn't love Trump's agenda or didn't Brexit idea. What Facebook did <a href="https://t.co/M6dy3hSaak">https://t.co/M6dy3hSaak</a>

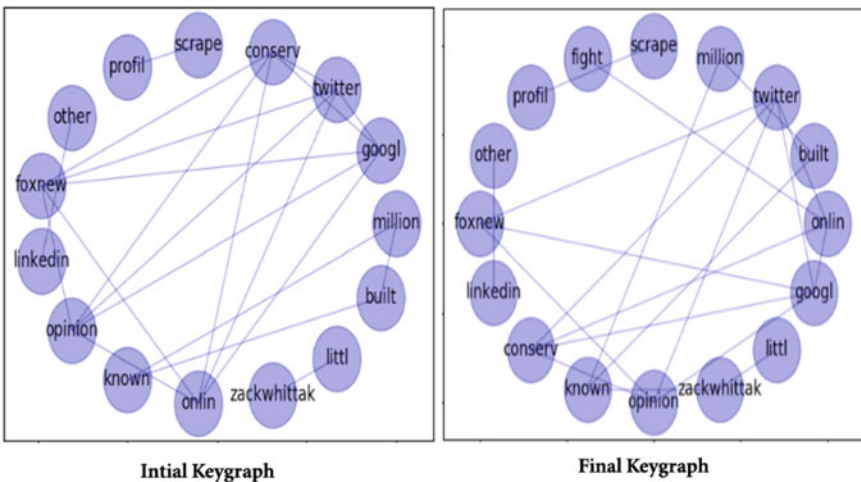


Fig. 4 Initial and final keygraph of batch 2 results

**Table 2** Batch 2 summary

Sr No	Keywords of a particular event	Summary of the particular event
1	foxnew, conserv, googl, opinion, onlin, twitter	Stay tuned to ISNeduction Twitter or Facebook page for all the detail
2	million, known, built	Amazing, 3-D homes are being built by The hundreds, saving thousands! Take a look! <a href="https://t.co/ADC1vp">https://t.co/ADC1vp</a>
3	linkedin, other	Monday:Open 11:00–16:00 Soup: lentil and bacon[veggies fear not, I have other home-cooked home-cooked home-cooked delights for you in] <a href="https://t.co/FEDOIQGp0z">https://t.co/FEDOIQGp0z</a>

after running the above algorithm. It shows that all the tweets belonged to that event. And then, it shows the most representative tweet of that event as a summary of that event.

## 5 Conclusion

In this paper, we have worked upon a more optimal graph clustering algorithm for clustering events than traditional event detection algorithms, i.e. event detection using the co-occurrence of keywords. The results obtained through this algorithm help to find trending topics during the live streaming of tweets from Twitter. Hence, these results can help companies with a news website and feature site content immediately relevant to their readers. Therefore, they can gain insight into events or topics that are very popular among their customers and, hence, generate additional profits. In future work, anyone can use the location and time of events so that case of events that needs urgent action can be handled accordingly. Also, work can be done to display total outcomes at once from the streaming analysis instead of displaying it in batches. Further, we can also make the Graphical User Interface to make it an interactive application.

## References

1. Dewang RK, Singh AK (2018) State-of-art approaches for review spammer detection: a survey. *J Intell Inf Syst* 50(2):231–264
2. Farzindar A, Inkpen D (2012) Proceedings of the workshop on semantic analysis in social media. In: *Proceedings of the workshop on semantic analysis in social media*
3. Hasan M, Orgun MA, Schwitter R (2018) A survey on real-time event detection from the Twitter data stream. *J Inf Sci* 44(4):443–463
4. Hasan M, Orgun MA, Schwitter R (2019) Real-time event detection from the Twitter data stream using the twitternews+ framework. *Inf Process Manage* 56(3):1146–1165

5. Java A, Song X, Finin T, and Tseng B. Why we Twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis. ACM, pp 56–65
6. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or news media? In: Proceedings of the 19th international conference on World wide web. ACM, pp 591–600
7. Lam W, Meng H, Wong K, Yen J (2001) Using contextual analysis for news event detection. *Int J Intell Syst* 16(4):525–546
8. Li Z, Wang B, Li M, Ma W-Y (200) A probabilistic model for retrospective news event detection. In: Proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 106–113
9. Nallapati R, Feng A, Peng F, Allan J (2004) Event threading within news topics. In: Proceedings of the thirteenth ACM international conference on information and knowledge management. ACM, pp 446–453
10. Sabeeh A, Dewang RK (2018) Comparison, classification and survey of aspect-based sentiment analysis. In: International conference on advanced informatics for computing research. Springer, pp 612–629
11. Mewada A, Prafful G, Shamaila K, Udayapal Reddy M (2010) Network intrusion detection using multiclass support vector machine. *Special Issue of IJCCT* 1(2–4):172–175
12. Wang X, Gerber MS, Brown DE (2012) Automatic crime prediction using events extracted from Twitter posts. In: International conference on social computing, behavioural-cultural modelling, and prediction. Springer, pp 231–238
13. Xu X, Yuruk N, Feng Z, Schweiger TA (2007) Scan: a structural clustering algorithm for networks. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 824–833
14. Yang Y, Zhang J, Carbonell J, Jin C (2002) Topic-conditioned novelty detection. In: Proceedings of the Eighth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 688–693
15. Yuan Q, Cong G, Ma Z, Sun A, Thalmann NM (2013) Who, where, when and what: discover Spatio-temporal topics for Twitter users. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 605–613
16. Zhou Y, Kanhabua N, Cristea AI (2013) Real-time timeline summarization for high-impact events on Twitter. In: Proceedings of the Twenty-second European conference on artificial intelligence, pp 1158–1166. IOS Press, 2016. of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, pp 605–613
17. Sayyadi H, Hurst M, Maykov A (2009) Event detection and tracking in social streams. In: Third international AAAI conference on weblogs and social media

# Spam Review Detection Using Okapi Relevance Method for Negative Reviews



Saloni Juneja, Shubham Goyallal, Sonali Agarwal, Saransh Agrawal, Rohit Kumar, Rupesh Dewang, and Arvind Mewada

**Abstract** Human nature is that one takes others' opinions before deciding to use the product and service. Hotel reviews provide a great deal of help to customers in online hotel booking. Negative reviews have a more devastating effect than positive ones. This paper presents research on a negative reviews dataset; half are fake, and the other half are non-fake. Early research represents that semantic relations are lost if the Bag-of-Words model is replaced by the Lexical Chain-Based Semantic Similarity (LCBSS) algorithm. Further, the okapi relevance method to generate feature vector works better than the TF-IDF method. Thus, the combination of LCBSS and okapi BM25 works best in detecting spam reviews.

**Keywords** Lexical chain · TF-IDF · Okapi BM25 · Spam review · Semantic similarity

## 1 Introduction

Nowadays, e-commerce platforms popular sources to purchase goods and services for humans. Almost every e-commerce website has set up a customer review section to write their involvement with the products or services they purchased. Online booking of hotels increases because sites compare hotels based on price, customer reviews, and ratings. Reviews provide a valuable source of information on any hotel service. They help the potential customers choose based on their personal experience of the hotel and the services offered during their stay. Reviews can have only two types of sentiment, either positive could imply buying a hotel service, or negative could improve a hotel's reputation and position. However, the negative reviews could affect badly [1]. They can put a customer in doubt and deter them from choosing that hotel while making an online booking. Thus, they can adversely affect a hotel's

---

S. Juneja · S. Goyallal · S. Agarwal · S. Agrawal · R. Kumar · R. Dewang · A. Mewada (✉)  
Motilal Nehru National Institute of Technology Allahabad, Prayagraj, India  
e-mail: [arvindmewada@mnnit.ac.in](mailto:arvindmewada@mnnit.ac.in)

R. Dewang  
e-mail: [rupeshdewang@mnnit.ac.in](mailto:rupeshdewang@mnnit.ac.in)

reputation (or even entire hotel chains). Unfortunately, this gives strong motivation for sentiment mining, which mentions unlawful diversions that mislead readers by giving undeserving positive reviews of target substances to support the substances and giving negative reviews to some other substances to break their reputation market. Fake review spam is becoming prevalent nowadays. Spam is defined as an unsolicited message sent over the internet to a large number of recipients. Many researchers admit that as reviews on the web are increasingly used in the process by customers, systems and employments are answerable.

Spamming develops into bad, and detecting spam reviews develop into increasingly critical [2, 3]. Some researchers have worked on fake negative reviews because the negative impact is more than the positive [1, 4–6]. Jindal and Liu proposed the first Fake review detection method in 2008 by classifying the spam review into three parts [7]. Then on-wards, many researchers have done much work on spam review and spammer detection [8, 9]. Sandulescu used the Bag-of-Word model in 2015 for spam review detection; then, many researchers used it [2, 3, 10]. This paper shows how the bag-of-words model and lexical chain-based semantic similarity (LCBSS) algorithm proposed by [3] can be used on negative reviews to detect spam.

Further, it uses TF-IDF and Okapi BM25 models for feature vector generation. Thus, our proposed LCBSS and okapi combination can prove to work much better at detecting spam reviews. The Remainder of this paper is presented as follows: In Sect. 2, discussion about the related works; in Sect. 3, discussed proposed approaches; in Sect. 4, is the experiment setup and results in discussion and last Sect. 5, provided the conclusion.

## 2 Related Work

Spamming has been a critical issue all these years. A collection of research works have been done associated with spam detection. It started with email spam detection. Now spam detection techniques are being applied in numerous other fields like web spam, SMS spam, chat spam, blog spam, Internet gathering spam, file sharing spam and review spam. In the year 2008, Jindal and Liu defined three forms of review spam:

- Type 1: These are the deceitful opinions where the purpose of a sole review is to benefit or dumb a product by posting fake reviews.
- Type 2: The reviewer writes only about the brand instead of the specific product intending to break its prestige.
- Type 3: These are non-reviews that contain advertisements, questionnaires or some random text.

In 2010, Jindal et al. [7] published a paper analyzing uncommon review markings, producing reviewers' doubtful behaviours. They used a domain-independent technique to study the Amazon.com review dataset and found many unexpected behaviours indicating spam activities.

Lau et al. [11] used the semantic concept to detect review spam for the first time in 2011. They proposed two concepts: the detailed general system architecture and the other was the high-order concept association mining module for detecting words changed with their semantic meaning.

In 2013, Ott et al. [12] studied negative sentiment reviews because negative deceptive opinion spam was largely not focused too much. They found out that the n-gram text categorization techniques work much better than the human evaluation of spam. They analyzed for possible collaboration of sentiment and deceptive text. There's a major drawback of such sentiment analysis based models. If a spammer replaces some words in a review with their semantic meaning or synonyms, then the model won't identify whether the new review is spam or not.

In 2015, Sandulescu and Ester [2] tackled the issue of exposing fake reviews written by the same reviewer using different names while posting each review. They proposed two methods to detect similar reviews. The first method used semantic similarity between words to review level. The second method used topic modelling with bag-of-words and bag-of-opinion-phrases.

The main limitation of bag-of-words is that it does not take into consideration the semantic meaning of words. Thus, the proposed model uses lexical chain-based semantic similarity algorithms to consider the semantic meaning and relation of words while constructing lexical chains.

### 3 Approach

This section has given the used dataset's details, the proposed approach of lexical chain-based semantic similarity (LCBSS) algorithm, and feature vector generation methods, namely TF-IDF and Okapi BM25.

#### 3.1 Dataset Used

This paper used a publicly available dataset of 800 negative reviews built by Ott in 2011. It contains a total of 20 hotels in reviews. These reviews are pre-labelled as spam and non-spam reviews. There are a total of 400 fake and 400 non-fake reviews in the dataset.

#### 3.2 Bag-of-Words

The bag-of-words model [13] is a very popular scheme used for representing documents. But it has a major limitation, i.e., it ignores many of the semantic features contained in the document. In this model, first, tokenize the review data and make a set

of words and the multiplication of term frequency and inverse document frequency is fed as an input to the classifier. The accuracy achieved is less using the BOW feature, So lexical chains capture semantic relationships between applied words and improve accuracy and f-measure.

The Bag-of-words approach is placed on the selection of raw words. These raw words do not provide as much information as bounded with semantic meaning, making it difficult to identify the topic and decreasing accuracy. The utmost familiar type of feature adjusted from the Bag-of-words model is term frequency. However, term frequency is not the best representation of the text because some common words appear more frequently.

The BOW scheme was initially designed for the information retrieval domain to index the documents and not necessarily the topic distribution. It represents features as an unordered set of words contained in the text along with their frequency count. The scheme ignores information such as position, relation and co-occurrences among words. It also results in the generation of feature space which is very large in time and space costs.

### 3.3 Lexical Chain Model

For the computation of feature vectors, TF-IDF and Okapi BM25 on lexical chains are used. Lexical chain-based semantic similarity algorithms (LCBSS) for generating lexical chains. First, perform data pre-processing.

#### 3.3.1 Pre-processing of the Dataset

For this, first of all, combine all reviews to make a single document. Then perform tokenization, stop words removal and word sense disambiguation(WSD). WSD is a technique to identify the sense of ambiguous words. For this, Roget’s thesaurus used word sense already stored in 1044 head files.

<b>Algorithm-1 Pre-processing of reviews dataset</b>		
<i>Input</i>	<i>List of reviews where each review is a string of words</i>	
	$R = \{R_1, R_2, \dots, R_n\}$	
<i>Output</i>	<i>The list of important candidate words for each review</i>	
	1	<i>Begin</i>
	2	<i>Combine all reviews to make a mixed document</i>
	3	<i>For</i> <i>mix document do</i>
	4	<i>perform tokenization</i>
	5	<i>Perform removal of stop words</i>
	6	<i>select noun, adjective, adverbs as candidate words</i>

(continued)



(continued)

<b>Algorithm-1 Pre-processing of reviews dataset</b>			
	7		<i>Perform word sense disambiguation (WSD)</i>
	8		<i>Store the candidate words for lexical chain Generation</i>
	9	<i>End for</i>	
	10	<i>End</i>	

The output of algorithm-1 is the set of candidate words that are used to construct lexical chains. Candidate words are nouns, adjectives, verbs and adverbs; consider all of these because only nouns can identify a document’s topic.

<b>Algorithm 2 Generating lexical chains</b>				
<i>Input</i>	<i>1. List of reviews R where <math>R = \{R_1, R_2, \dots, R_n\}</math></i>			
	<i>2. List of candidate words obtained from Algorithm-1</i>			
<i>Output</i>	<i>Set of lexical chains</i>			
	1	<i>BEGIN</i>		
	2	<i>Maintain a Global set that is initialized to null initially</i>		
	3	<i>For</i>	<i>each document do</i>	
	4		<i>For</i>	<i>each candidate word from algorithm-1</i>
	5		<i>if</i>	<i>candidate word has a repetition of the same word or in the same paragraph</i>
	6		<i>then</i>	
	7			<i>store the candidate word in the respective chain</i>
	8		<i>end if</i>	
	9		<i>else</i>	
	10			<i>if no chain is identified, then</i>
	11			<i>Create a new potential chain for this word</i>
	12		<i>end if</i>	
	13			<i>Add identified/created chain to the global set</i>
	14		<i>End For</i>	
	15	<i>End For</i>		
	16	<i>End</i>		

### 3.3.2 Lexical Chains Generation

Lexical chains are a group of words that exhibit cohesive relations [14]. Lexical cohesion is mostly built by the general nouns (superordinates, subordinates) or the same word's repetition in the same reference. These relations are used to decide if a candidate word has a relation with a lexical chain or not. In 1976, Haliday and Hasan classified these relations as:

- Reiteration with the existence of reference (e.g., bus and bus)
- Reiteration without the existence of reference (e.g., bus and automobile)
- Reiteration utilizing superordinate (e.g., car and vehicle)
- Systematic semantic relation
- Non-systematic semantic relation.

The first three relations are a type of reiteration, a scheme of lexical cohesion that affects the recurrence of the lexical item at one side of the scale; the adoption of a natural word refers back to a lexical item at another end scale. It can be a synonym, superordinate or the same word, and with these relations, insertion of a word in a lexical chain is certified. The last two relations include semantic relations between words that co-happen regularly (e.g., goal and football).

Our algorithm is based on WordNet lexical database, which is used in the ELKB. WordNet is used to identify the relationships among words, and it uses only identity and synonymy relations to compute the chains representing summarized topics. It works by maintaining a set of global lexical chains where each chain represents a topic. For each candidate word, it looks up the synsets for the word from WordNet. The global list of lexical chains is traversed to identify the relation. If the candidate has a relation, the word is added to that respective chain; otherwise, a new chain is created if not related to any existing chain. Our input of 800 review documents received an output of 219 lexical chains on applying this algorithm.

## 3.4 Feature Extraction

Once the computing lexical chains task is complete, compute the feature vector using the TF-IDF and Okapi BM25 algorithms.

### 3.4.1 TF-IDF Algorithm

Term-frequency and inverse document frequency, in each review, is considered as an individual document. The classical TF-IDF [15] is commonly used to reduce the effect of words that appear more repeatedly in documents. This paper's concept on lexical chains calculates the TF-IDF [3] for each document corresponding to each lexical chain. Since the dataset has 800 reviews and 219 lexical chains, the feature vector's size is  $800 \times 219$ .

**Algorithm 3 Feature vector Generation using TF-IDF for lexical chains**

<i>Input</i>		1. List of reviews $R$ where $R = \{R_1, R_2, \dots, R_n\}$	
		2. List of lexical chains obtained from Algorithm 2	
<i>Output</i>		Feature vector	
1		Begin	
2		Initialize feature vector	
3	For	each document do	
4		For	each lexical chain for algorithm-2 do
5			compute the term frequency (tf)
6			suppose for lexical chain $Lc$ , $f(Lc)$ is the number of times words of lexical chain appear in the document then
7			$tf = \log[1 + f(Lc)]$
8			Calculate the review frequency(rf), which is the total of documents/reviews in which LC appears
9			Calculate the inverse document frequency (idf)
10			$idf = 1/\log[1 + rf]$
11			$Tf-idf = \text{term frequency}(tf) * \text{inverse document frequency}(idf)$
12			Store it in the feature vector
13		End For	
14	End For		
15	End		

**3.4.2 Okapi BM25 Algorithm**

Okapi BM25 (BM stance for Best Matching) [16] is an estimate function in information retrieval worn by search engines to rank documents acquiesce to their applicability for a query. BM25 is a bag-of-words retrieval function that ranks a set of documents based on a document’s query terms. Set of a query  $Q$ , accommodate keywords  $q_1, q_2, \dots, q_n$ , the score of document  $D$  for this query is:

$$score(D, Q) = \sum_{i=1}^n IDF(q_i) * \frac{f(q_i, D) * (K_1 + 1)}{f(q_i, D) + k_1 * \left(1 - b + b \cdot \frac{|D|}{avgdl}\right)} \tag{1}$$

where,

$f(q_i, D)$	the term frequency of term $q_i$ in $D$
$ D $	the length of the document in words
$avgdl$	the average document length in a text collection
$IDF(q_i)$	the inverse document

The average document length in text collection  $k1$  and  $b$  are parameters, usually chosen without advanced optimization as  $k1$  is between 1.2 and 2.0 and  $b$  equals 0.75. The inverse document frequency for the document, which is calculated as:

$$IDF(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2}$$

where,  $N$ : total number of documents in the collection  $n(q_i)$  containing term  $q_i$ . The feature vector obtained as output is  $800 \times 219$ .

**Algorithm 4 Feature vector Generation using Okapi BM25 for lexical chains**

<b>Input</b>	1. List of reviews $r$			
	$R = \{R_1, R_2, \dots, R_n\}$			
	2. Set of lexical chains obtained from Algo.- 2			
<b>Output</b>	Feature vector			
	1	Begin		
	2	Initialize feature vector		
	3	For	each lexical chain (LC) computed from algorithm-2 do	
	4		For	each document $r$ in $R$ do
	5			Initialize sum = 0
	6		For	each term $q$ in LC do
	7			Compute term frequency (tf) of word $q$ in document $r$
	8			Compute the inverse frequency $n(q)$ of word $q$ in document $r$
	9			Compute the Okapi Score for that term by the above formula
	10			Add that score in sum
	11		End For	
	12		Store this sum in the feature vector	
	13		End For	
	14	End For		
	15	End		

### 3.5 Classification

The feature vector obtained from applying algorithms discussed in Sect. 3.3 is given input to the tool's classifiers. The classifiers are trained and then tested to get accuracy and F-measure.

## 4 Experimental Setup and Results

This experiment with our algorithms on the hotel reviews was divided into spam and ham(non-spam) classes. Roget's ELKB, an open-source tool, generates the set of global lexical chains by capturing the semantic relationship between the words of reviews for better results. The feature vector is then generated using the Okapi BM25 and TF-IDF algorithms. The feature vector size is 800\*219, used as input to the Weka Tool's classifiers. The same evaluation (feature vector generation and classification) is also done for the Bag-of-Words model with TF-IDF and Okapi BM25 for comparison. The given methods divide the dataset into two parts training dataset, 70% and 30% testing dataset, and computed the dataset's accuracy and f-measure. Figure 1 shows the comparison of accuracies obtained with various models using various classifiers. It can be seen that the accuracy which is acquired by the lexical chains for any classification algorithm is greater than the simple BOW model. It is also seen that the Okapi BM25 method of vector generation leads to more accurate classification, be it any classifier. Accuracy and F-measure are calculated as follows: Suppose  $H_p$  is the number of ham (not spam) reviews that are correctly classified,  $H_n$  is the number of ham reviews that are not correctly classified,  $S_p$  is the number of spam reviews that are correctly classified, and  $S_n$  is the number of spam reviews which are not correctly classified then:

$$Accuracy = \frac{H_p + S_p}{H_p + H_n + S_p + S_n} \quad (3)$$

F-measure is a combination of Precision and Recall and is calculated as:

$$F - measure = \frac{2 * Precision * Recall}{Recall + Precision} \quad (4)$$

where,  $Precision = \frac{S_p}{S_p + S_n}$  and  $Recall = \frac{S_p}{S_p + H_n}$

The Okapi BM25 is worked on the relevance feedback method, and the query term appeared in the document. This property of the Okapi BM 25 method is different from other used methods. The main reason behind high accuracy. Table 1 shows that the Okapi BM 25 has the highest accuracy with BOW and Lexical Chain models (Table 2 and Fig. 2).

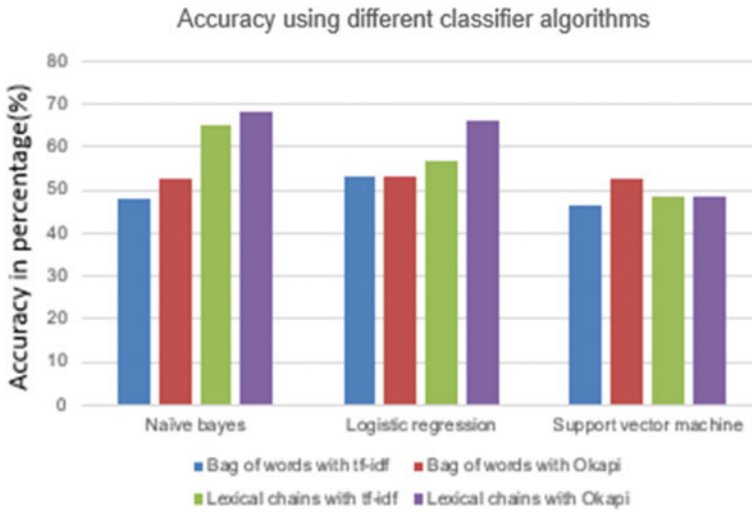


Fig.1 Accuracy Graph for different classifier algorithms with different Models

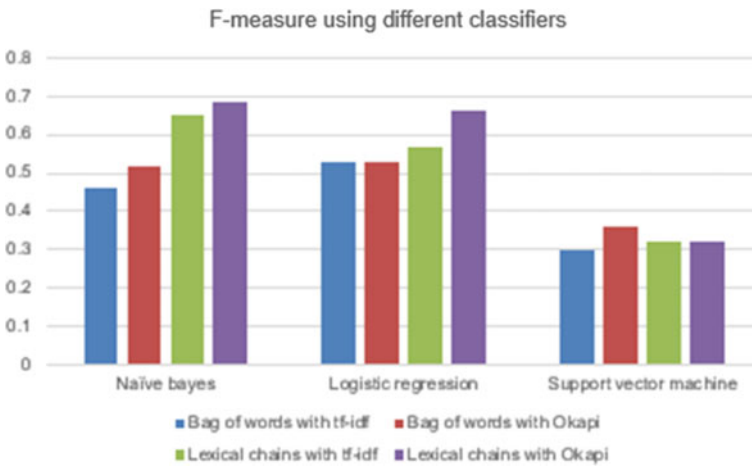


Fig. 2 F-measure graph for different models with different classifiers

Table 1 Accuracy obtained with various models using different classifiers

Algorithm used	Naïve (%)	Log. Reg. (%)	SVM (%)
BOW + TF-IDF	47.9	53.33	46.67
BOW + Okapi BM25	52.92	53.33	52.92
Lex Chain + TF-IDF	65.92	57.08	48.75
Lex Chain + okapi BM25	68.17	6.25	48.75

**Table 2** F-measure obtained with various models using different Classifiers

Algorithm used	Naïve	Log. Reg	SVM
BOW + TF-IDF	0.46	0.52	0.29
BOW + Okapi BM25	0.52	0.53	0.52
Lex Chain + TF-IDF	0.65	0.57	0.32
Lex Chain + okapi BM25	0.69	66.25	48.8

## 5 Conclusion

Thus, the conclusion is that the Lexical Chain-Based Semantic Similarity (LCBSS) algorithm is a much better approach than the Bag-of-Words model. We also observe that the Okapi BM25 gives better results than the TF-IDF model. In comparison, we find that combining the LCBSS algorithm and Okapi BM25 gives the most accurate results with up to 65% accuracy. We can work to achieve better accuracy. Review Spam detection is a critical issue, so more accuracy must be achieved.

## References

- Banerjee S, Chua AY (2014) A study of manipulative and authentic negative reviews. In: Proceedings of the 8th international conference on ubiquitous information management and communication, ACM, 2014, p 76
- Sandulescu V, Ester M (2015) Detecting singleton review spammers using semantic similarity. In: Proceedings of the 24th international conference on World Wide Web. ACM, 2015, pp 971–976
- Dewang RK, Singh AK (2016) Spam review detection through lexical chain based semantic similarity algorithm (LCBSS) for negative reviews. *Int J Eng Technol (IJET)* 8(6):2946–2955
- Banerjee S, Chua AY (2014) Applauses in hotel reviews: genuine or deceptive?. In: 2014 science and information conference. IEEE, 2014, pp 938–942
- Hu N, Bose I, Koh NS, Liu L (2012) Manipulation of online reviews: an analysis of ratings, readability, and sentiments. *Decis Support Syst* 52(3):674–684
- Zheng R, Li J, Chen H, Huang Z (2006) A framework for authorship identification of online messages: writing-style features and classification techniques. *J Am Soc Inform Sci Technol* 57(3):378–393
- Jindal N, Liu B (2008) Opinion spam and analysis. In: Proceedings of the 2008 international conference on web search and data mining, ACM, 2008, pp 219–230
- Jindal N, Liu B, Lim E-P (2010) Finding unusual review patterns using unexpected rules. In: Proceedings of the 19th ACM international conference on information and knowledge management. ACM, 2010, pp 1549–1552
- Dewang RK, Singh AK (2018) State-of-art approaches for review spammer detection: a survey. *J Intell Inf Syst* 50(2):231–264
- Etaiwi W, Awajan A (2017) The effects of features selection methods on spam review detection performance. In: 2017 international conference on new trends in computing sciences (ICTCS). IEEE, 2017, pp 116–120
- Lau RY, Liao S, Kwok RCW, Xu K, Xia Y, Li Y (2011) Text mining and probabilistic language modelling for online review spam detecting. *ACM Trans Manag Inf Syst* 2(4):1–30

12. Ott M, Cardie C, Hancock JT (2013) Negative deceptive opinion spam. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, 2013, pp 497–501
13. Zhang Y, Jin R, Zhou Z-H (2010) Understanding bag-of-words model: a statistical framework. *Int J Mach Learn Cybern* 1(1–4):43–52
14. Jayarajan D, Deodhare D, Ravindran B (2007) Document clustering using lexical chains
15. Aizawa A (2003) An information-theoretic perspective of TF–IDF measures. *Inf Process Manag* 39(1):45–65
16. Robertson S, Zaragoza H, Taylor M (2004) Simple BM25 extension to multiple weighted fields. In: Proceedings of the thirteenth ACM international conference on Information and knowledge management



# Predicting Time-Series Data Using Linear and Deep Learning Models—An Experimental Study



Ahmad Alsharef, Sonia, Monika Arora, and Karan Aggarwal

**Abstract** Analyzing time-series data have gained significant attention in modern research works. Its significance lies in different applications such as weather forecasting, economic forecasting, sales forecasting, etc. This project aims to explore the efficiency of various approaches in predicting time-series cryptocurrency prices data. This work is an experimental study on the potential of different approaches. These approaches included Auto-Regressive (AR), Moving Average (MA), Auto-Regressive Integrated Moving Average (ARIMA), Long Short-Term Memory (LSTM), Independently RNN (IndRNN), and Fine-tuned IndRNN. The dataset was a quantitative secondary historical time-series data of Ethereum cryptocurrency prices collected from a reliable bulletin. The suggested approaches were implemented and tested on the Ethereum Cryptocurrency historical prices where Ethereum is the second-largest cryptocurrency by market share at our time. IndRNN and its fine-tuned version were proved to yield higher prediction potential than the other presently utilized methods with MSE of 239 and 213, respectively. Also, deep learning models, outperformed the linear models, in terms of accuracy of prediction. The tested approaches can be utilized as a standard in predicting cryptocurrency data with acceptable accuracy. These models can be used by investors to help them make good decisions in buying or selling cryptocurrency stocks.

**Keywords** Time-series · Cryptocurrency · ARIMA · Deep learning · Linear prediction · Machine learning models · Ethereum

---

A. Alsharef · Sonia (✉)

Yogananda School of AI, Computer and Data Science, Shoolini University, Solan, India

A. Alsharef

e-mail: [ahmadalsharef@shooliniuniversity.com](mailto:ahmadalsharef@shooliniuniversity.com)

M. Arora

Apeejay School of Management, Dwarka, Delhi, India

K. Aggarwal

Electronics and Communication Engineering Department M.M. (Deemed to be University), Haryana, India

## 1 Introduction

Time-series data is a chronological sequence of numerical data points in consecutive order that represent a collection of observations for a specific phenomenon that were recorded at different times and usually occurred at uniform intervals. A time-series can be considered as a historical record that can be constructed to perform future forecasting [1]. Forecasting models usually fail to recognize patterns in time-series data and complex relations between its values. However, with the recent progress in machine learning, data patterns became able to be detected accurately. Previously, linear time-series methods such as the ARIMA model were applied for cryptocurrency price and trend prediction [2, 3]. However, it was incapable to discover non-linear patterns in complicated prediction problems. On the other hand, Machine Learning and Deep Learning models achieved greater performance [4, 5]. Many researchers have utilized these two types of models and evaluated the performance in different tasks [6, 7]. Most recent works proved the efficiency of deep neural networks in forecasting the prices of Bitcoin and other cryptocurrencies [8, 9].

Cryptocurrencies' rate behavior is still in large part unexplored. Predicting how the financial exchange will perform is a very difficult challenge where there are infinite elements that affect the exchange rate. It should be noted that forecasting cryptocurrency is different from forecasting ordinary currencies where the prices are relatively stable. These issues motivate us to further explore this novel research area. Moreover, it is worth mentioning that in this paper, results from the study applied on different models assist the financiers and investors in sound decision making during the cryptocurrency stock business. The main contributions of this work include the experimental study exhibited on cryptocurrency data using a very popular and effective linear and deep learning prediction model. They are substantially useful in determining that:

- Whether it is possible for deep learning algorithms to forecast time-series cryptocurrency prices accurately or not and up to what extent these algorithms are suitable for the cryptocurrency stock market. As cryptocurrency is a new field and requires every aspect to be examined carefully.
- What is the impact of linear and non-linear prediction models on the accuracy of predicting cryptocurrency prices? Since accurate predictions can support cryptocurrency traders towards the right investment choices and lead to potential profits.

Moreover, this research contributes by filling the gap in between the earlier studies conducted in this area of research by using more advanced deep learning approaches and novel models. The present paper can act as a reference guide for the growing body of research since it is exploring the potential of six different time-series data prediction models in-depth. Further, in this study, these models were implemented and tested on Ethereum prices which are the second-largest cryptocurrency by market capitalization.

By using the models, investors can make reasonable decisions on selling or buying cryptocurrency stocks, especially since cryptocurrencies are being seen as the future of payments and a way of investing after the continuous drop in the prices of other stocks with its ease of use and transfer that cryptocurrency provided.

The layout of this research paper is streamlined as follows. Section 2 presents the literature review done by eminent authors in this field. The methodology for conducting this work detailing various models and algorithms is discussed in Sect. 3. Section 3 followed by the next section that is, Sect. 4 that elaborates on the experimental work and the results drawn out from the study. Section 5 discusses the conclusion and future prospects in this area of research.

## 2 Related Works

Many prominent researchers have done a lot of work in the field of time-series data analysis. However, research on the application of Cryptocurrency data for time-series analysis is scarce to date. With technological advancement, nowadays, Cryptocurrency is gaining too much popularity. But being a new concept not many studies and results can be seen in this area. Assessing this research gap, this paper conducted a study on cryptocurrency data of several time-series algorithms. To start with, related work in this field is being determined. Thus, this section presents the in-depth study of related work conducted in this area.

Baharammiraze et al. (2010) [10] conducted a survey on financial prediction using AI, expert systems, and hybrid models to predict the high-variant non-linear behavior of the financial market.

Li et al. [11] conducted a review on using AI and ANNs for stock market prediction and found that artificial neural networks are efficient prediction models in financial economics due to the properties of learning, generalization, and non-linear behavior.

Long et al. [12] investigated the use of deep learning to predict stock market prices and trends. Results found that bidirectional LSTM could predict the stock market with a great performance. Additionally, Rekha et al. [13] explored the potential of RNN and Convolutional Neural Network (CNN) in predicting the stock market. Some researchers have further explored the area and have taken the risk aspect with the learning models [14] as stock exchanges are a very crucial field so considering the risk features is very critical.

Dutta et al. [15], inspected different machine learning models to predict Bitcoin prices like LSTM and GRU which were shown to perform better than the other models. They compared these different models' efficiency using the root mean squared error (RMSE) [15]. Their results were as shown in Table 1.

From the previous results, they concluded that GRU with a dropout yielded the best accuracy and GRU yielded better accuracy than LSTM in predicting Bitcoins prices.

Other works included: Phaladisailoed et al. [16] who worked on a Bitcoin-USD prices dataset which included the Bitcoin exchange rate in USD. They used two

**Table 1** RMSE for different models of Aniruddha Dutta et al. [15]

Model	Test RMSE
LSTM	0.024
GRU	0.19
GRU-Dropout	0.017

regression models, Theil-Sen and Huber. Besides, two deep learning models are LSTM and GRU. The results showed that deep learning-based models gave better accuracy compared to Theil-Sen regression and Huber regression.

Ji et al. [1] compared various deep learning models such as Deep Neural Network (DNN), LSTM, CNN, residual networks (ResNets), and their combination for Bitcoin price prediction these researchers also analyzed the Bitcoin prices by observing the blockchain transactions. Their results showed that LSTM based models slightly outperformed the other models whereas DNN based models performed the best.

McNally et al. [4] aimed to examine with what accuracy the direction of the Bitcoin exchange rate in US dollars can be predicted through implementing a Bayesian optimized RNN and an LSTM network. The LSTM achieved the best classification accuracy. The non-linear deep learning techniques outperformed the ARIMA model which performed poorly.

Uras et al. [17] forecasted the closing rate of Bitcoin, Litecoin, and Ethereum cryptocurrencies, using historical data of previous days using a simple linear model for univariate series prediction and a multiple-linear regression depending on both price and volume data. Also, they implemented Multilayer Perceptron (MLP) and LSTM neural networks. Their models outperformed previous studies in terms of time complexity and provided better overall results.

Politis et. al. [18] used LSTM, GRU, Temporal Convolutional Networks (TCN), and hybrid models to forecast Ethereum prices based on historical data bulletin and a set of 13 additional features, that were considered important based on domain knowledge, on a daily and weekly basis. All models performed well in the regression. In general, Hybrid models outperformed individual models in general, whereas an assembly of LSTM, Hybrid LSTM-GRU, and Hybrid LSTM-TCN models performed the best. The quality of weekly forecasts was poorer than daily forecasts, showing difficulty in predicting Ethereum prices in the long-term.

The previous works didn't study the potential of advanced neural networks that are used with time-series data like IndrRNN [19] and Fine-tuned IndrRNN [20] which were proved, in this work, to be better potential in predicting Ethereum prices. This work implemented and tested the efficiency of six different models of Ethereum cryptocurrency which is the second-largest cryptocurrency by market share.

### 3 Methodology

#### 3.1 Linear Prediction Algorithms

Auto-Regressive (AR): A statistical model that predicts the future based on historical past values. Like, an auto-regressive model might seek to predict a cryptocurrency’s future prices based on the model’s past performance. AR(1) indicates that the current value is based on the immediately preceding value, while AR(2) indicates that the current value is based on the preceding two values. This work used AR(10) which means each forecasted day’s price depends on the previous 10 days.

Moving Average (MA): A method that creates a series of means of subsets of a complete dataset. It is commonly used in time-series data forecasting [21].

Auto-Regressive Integrated Moving Average (ARIMA): It is a linear speculation model that considers that previously time-series data values may be employed unaccompanied in predicting future values for the same data [21]. ARIMA’s full formula is given as.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + .. + \alpha_p Y_{t-p} + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \dots + \phi_q \epsilon_{t-q}$$

*Predicted value = Constant + Weighted sum of the recent p data values + Weighted sum of the recent q predictions’ error values (difference between the actual values and predicted values in the past predictions).*

ARIMA configurations are represented in the form ARIMA (p,d,q) [22]. Where d is the number of non-seasonal differences needed for stationarity, here a shift in time doesn’t cause a change in the shape of the distribution. In other words, d is the times of differences that were performed between the current values and the previous ones where current values will be replaced by these differences.

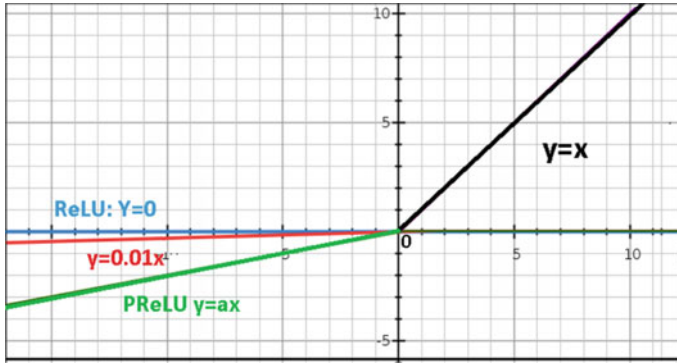
Using the Grid search optimization algorithm [23] which automatically discovers the optimal order for an ARIMA model, the p, d, and q coordinates of best fit in our case were as the following (p = 7, d = 0, q = 7).

#### 3.2 Deep Learning Prediction Models

Recurrent Neural Network (RNN): It is a neural network that receives inputs from two sources, the current state, and the past state. In a sense, it has a memory.

Long Short-Term Memory (LSTM): Designed to solve long-term dependence by keeping long-term data back. Contains memory cells. Each has three gates to adjust the data flow. In the process of guessing the next word in the sentence, LSTM exceeded RNN due to its memory block [24].

Independently Recurrent Neural Network (IndRNN): It was proposed by Shuai Li et al. [19]. It is different from RNN in that each neuron receives data only from its



**Fig. 1.** ReLU versus PReLU

input and its hidden state at the previous step as content data (instead of fully linking to all other neurons in the previous layer as in RNN).

Fine-tuned Independently Recurrent Neural Network (Fine-tuned IndRNN): IndRNN yielded a higher accuracy when replacing its default ReLU activation function with a more advanced activation function called Parametric ReLU (PReLU) [20]. The difference between ReLU and PReLU activation functions is represented in the following line graph of Fig. 1.

ReLU outputs a slight slope for the negative input values whereas ReLU outputs zero for any negative input. The problem of outputting zeros for all negative input values becomes a disadvantage in case of a neuron gets stuck on the negative side causing zeros as the output and once it keeps getting negative values, it becomes improbable to recover and hence the neuron becomes inactive.

## 4 Experimental Work

### 4.1 Experimental Workflow

The workflow of the work conducted through this research is represented through a flow chart as shown in Fig. 2.

The experiment started by collecting Ethereum cryptocurrency, which is the second-largest cryptocurrency by market capitalization after Bitcoin, and historical prices in USD between the years 2015 and 2020 from the Yahoo Finance datasets [25]. The dataset included 1193 entries including general features like Date, Open, High, Low, and Close prices. The dataset head has been elaborated in Fig. 3.

We extracted the Close price attribute from this dataset, and it was the parameter we depended on to predict the future close prices. In other words, the project calculates the future close price values depending on the previous close price values since

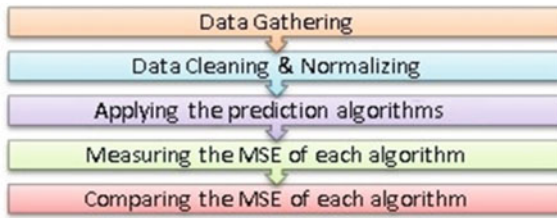


Fig. 2 Experimental Workflow Diagram

	Date	Open	High	Low	Close	Adj Close	Volume
0	2015-08-07	2.831620	2.521120	3.536610	2.772120	2.772120	16432900
1	2015-08-10	0.713989	0.636546	0.729854	0.708448	0.708448	40528300
2	2015-08-11	0.708087	0.663235	1.131410	1.067860	1.067860	146310000
3	2015-08-12	1.058750	0.883608	1.289940	1.217440	1.217440	215062000
4	2015-08-13	1.222240	1.171990	1.965070	1.827670	1.827670	406868000

Fig. 3 ETH-USD Dataset Head

investors usually take their decisions on selling or buying depending on their view about the closing price.

The data was cleaned by converting all the data value types to integer numbers and by filling in missing values by copying the previous value. Furthermore, a Min–Max scaler was used to adjust data in the range, 0 to 1.

The following Fig. 4 describes the close prices in the used dataset.

The following Fig. 5 shows the probability plot and skewness of the dataset. We realize that the data is skewed where skewness is a general case that applies to most cryptocurrency datasets.

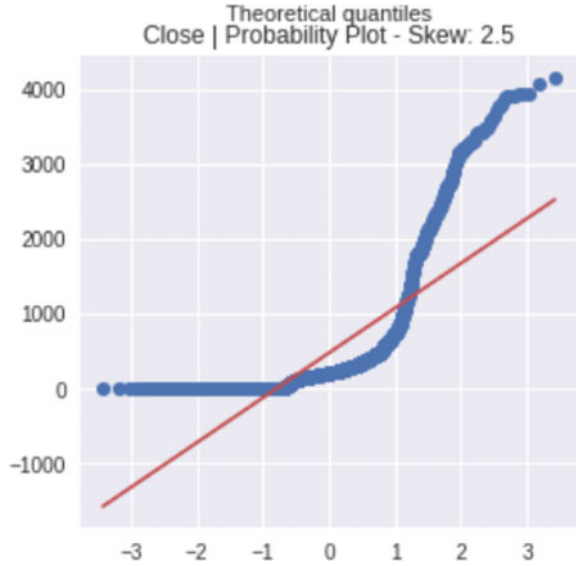
Six algorithms have been tested on the previous data:

**Close prices data description**

MOST FREQUENT VALUES			SMALLEST VALUES			LARGEST VALUES		
11.65279	2	<0.1%	0.434828	1	<0.1%	4168.701	1	<0.1%
11.95160	2	<0.1%	0.447329	1	<0.1%	4079.057	1	<0.1%
2.772119	1	<0.1%	0.489013	1	<0.1%	3952.293	1	<0.1%
179.7871	1	<0.1%	0.517733	1	<0.1%	3952.133	1	<0.1%
175.9928	1	<0.1%	0.522967	1	<0.1%	3940.614	1	<0.1%
174.2171	1	<0.1%	0.536494	1	<0.1%	3928.844	1	<0.1%
169.9561	1	<0.1%	0.539656	1	<0.1%	3928.379	1	<0.1%
178.2626	1	<0.1%	0.547177	1	<0.1%	3902.647	1	<0.1%
181.3555	1	<0.1%	0.561678	1	<0.1%	3887.828	1	<0.1%
181.1494	1	<0.1%	0.563389	1	<0.1%	3834.828	1	<0.1%
178.7254	1	<0.1%	0.567701	1	<0.1%	3790.989	1	<0.1%
178.3473	1	<0.1%	0.582885	1	<0.1%	3785.848	1	<0.1%
181.0160	1	<0.1%	0.607654	1	<0.1%	3715.148	1	<0.1%
181.1097	1	<0.1%	0.609387	1	<0.1%	3638.122	1	<0.1%
188.1055	1	<0.1%	0.616038	1	<0.1%	3615.282	1	<0.1%

Fig. 4 Description of close prices in the dataset

**Fig. 5** Description of close prices in the dataset



- AR depending on the prices of the previous 10 days
- MA depending on the prices of the previous 10 days.
- ARIMA model with the configurations of  $(p = 7, d = 0, q = 7)$ . Where  $p = 7$  is the number of previous days to consider when predicting the price of the next day,  $q = 7$  is the number of previous days to consider their prediction error values (the difference between the anticipated value and the real value) in forecasting the next day's stock price,  $d = 0$  is the number of times the differences between the current data values and the previous values are performed where data current values will be replaced by these differences. These values were identified as the best parameters after conducting a grid search ARIMA optimization function on the dataset [26–28].
- LSTM, IndRNN, and Fine-tuned IndRNN by predicting the prices of the next 10 days depending on the previous 40 days. The parameters configurations of LSTM, IndRNN, and Fine-tuned IndRNN were chosen as follows: Time-Steps: 40, Optimizer Algorithm: Adam, Learning-Rate: 0.001, Batch-Size: 40, Epochs: 200, Number of Hidden Layers: 5.

The explored models are summarized in the following Table 2.

The accuracy of each prediction model was measured using the MSE method.

Finding the best fit and comparison of efficiency was given depending on the MSE of each prediction model.



**Table 2** Configurations of used models

Model	Configuration
AR	P = 10
MA	10-day moving average
ARIMA	(p = 7, d = 0, q = 7)
LSTM IndRNN Fine-tuned IndRNN	Time-Steps: 40 Optimizer Algorithm: Adam Learning-Rate: 0.001 Batch-Size: 40 Epochs: 200 Number of Hidden Layers: 5

### 4.2 Result Analysis

After applying the experiments on the cleaned data using linear algorithms and deep learning models, the mean squared error obtained for each approach is given in the following Table 3.

LSTM outperformed the linear models. IndRNN and its fine-tuned version worked the best. Auto-regressive (AR) was found non-eligible in predicting time-series data achieving exponential MSE.

The moving average (MA) achieved a low accuracy with high MSE where it is known to be not suitable for high volatile data with frequent changes in trends like historical data of unstable Ethereum cryptocurrency with fluctuating prices. Moving average shows a consistent change in the price over time. However, since every stock/asset has different price histories and levels of volatility, this uniform linear formula can't be approved across all markets.

ARIMA, although being a linear algorithm, achieved an acceptable accuracy and was able to predict the prices with low error rates by examining the differences between values in the series instead of through actual values and using an unsupervised learning approach to learn patterns from data. However, it failed to prove efficiency when compared to non-linear techniques like artificial neural networks. While the model is adept at modeling seasonality and trends, some outliers or extremely volatile values are difficult to forecast for the reason that they lie outside of the general trend as captured by ARIMA.

**Table 3** MSE of the tested models when predicting Ethereum prices

Model	MSE
AR	78,843
MA	14,438
ARIMA	314
LSTM	298
IndRNN	239
Fine-tuned IndRNN	213

LSTM neural networks which were built to deal with sequences like time-series data achieved very good results with an MSE of 298 outperforming the ARIMA linear model because LSTM is able to store past information. This was important in our case because the previous price of a stock is crucial in predicting its future price.

IndRNN worked better than LSTM with an MSE of only 239 and this is an acceptable accuracy in predicting volatile data like Ethereum prices where, for instance, the stock price doubled several times in less than two years, it moved from 170 USD in January 2020 to more than 2000 USD in July 2021 including many different fluctuations up and down within.

IndRNN in its fine-tuned version [20] performed the best with a low MSE of 213 since it uses the PReLU activation function instead of the traditional ReLU which is used by IndRNN.

It has been determined that the linear approaches aren't useful in predicting strongly volatile time-series data like cryptocurrency prices that contain extreme values. The results showed that although the prediction of price and trending is extremely challenging and depends on unpredictable events, it is possible to make an intelligent system that can predict the cryptocurrency with acceptable accuracy.

## 5 Summary

Deep learning and machine learning models became useful tools for forecasting statistical time-series financial data like cryptocurrency prices. There is a considerable number of studies on time-series-data prediction using deep learning. However, these studies didn't test the efficiency of one of the most recent important deep learning algorithms like IndRNN. This work explored the potential of financial data prediction depending on different linear statistical and deep learning models. From the results obtained, it's clear that deep learning architectures outperform linear models. IndRNN and its fine-tuned version have been used to explore their potential and found to yield higher accuracy than the other deep learning presently utilized methods with MSE of 239 and 213, respectively. This work hasn't explored the advantage of utilizing a hybrid network that combines neural network models to produce a new prediction model. The work will be further extended by collecting behaviors and trends of the currency market from financial newspapers using Natural Languages Processing techniques due to the numerous unpredictable factors that might impact the price such as political events. We can obtain these events from News and analyze them to determine the trend of the currency price.

## References

1. Ji S, Kim J, Im H (2019) A comparative study of bitcoin price prediction using deep learning. *Mathematics* 7(10):898
2. Bakar NA, Rosbi S (2017) Autoregressive integrated moving average (ARIMA) model for forecasting cryptocurrency exchange rate in high volatility environment: a new insight of bitcoin transaction. *Int J Adv Eng Res Sci* 4(11):237311
3. Siami-Namini S, Tavakoli N, Namin AS (2018) A comparison of ARIMA and LSTM in forecasting time series. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA), pp 1394–1401
4. McNally S, Roche J, Caton S (2018) Predicting the price of bitcoin using machine learning. In: 2018 26th Euromicro international conference on parallel, distributed and network-based processing (PDP), 2018, pp 339–343
5. Taneva G (2019) An analysis and a forecast of the cryptomarket based on the ARIMA model. *Econ Thought J* 4:66–84
6. Tahiri P, Sonia S, Jain P, Gupta G, Salehi W, Tajjour S (2021) An estimation of machine learning approaches for intrusion detection system. *Int Conf Adv Comput Innov Technol Eng (ICACITE)* 2021:343–348. <https://doi.org/10.1109/ICACITE51222.2021.9404643>
7. Alsharif SA, Jain P, Arora M, Zahra SR, Gupta G (2021) Cache memory: an analysis on performance issues. In: 2021 8th international conference on computing for sustainable global development (INDIACom), 2021, pp 184–188. <https://doi.org/10.1109/INDIACom51348.2021.00033>
8. Ahmadi F, Sonia, Gupta G, Zahra SR, Baglat P, Thakur P (2021) Multi-factor biometric authentication approach for fog computing to ensure security perspective. In: 2021 8th international conference on computing for sustainable global development (INDIACom), 2021, pp 172–176. <https://doi.org/10.1109/INDIACom51348.2021.00031>
9. Salehi AW, Gupta G (2021) A prospective and comparative study of machine and deep learning techniques for smart healthcare applications. In: 2021 mobile health: advances in research and applications, pp.163–189. Scopus | ID: covidwho-1316123
10. Bahrammirzaee A (2010) A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system, and hybrid intelligent systems. *Neural Comput Appl* 19(8):1165–1195
11. Li Y, Ma W (2010) Applications of artificial neural networks in financial economics: a survey. In: 2010 international symposium on computational intelligence and design, 2010, vol 1, pp 211–214
12. Long J, Chen Z, He W, Wu T, Ren J (2020) An integrated framework of deep learning and knowledge graph for prediction of stock price trend: an application in Chinese stock exchange market. *Appl Soft Comput* 91:106205
13. Rekha G, Sravanthi BD, Ramasubbareddy S, Govinda K (2019) Prediction of stock market using neural network strategies. *J Comput Theor Nanosci* 16(5–6):2333–2336
14. Awoke T et al. (2021) Bitcoin price prediction and analysis using deep learning models. In: *Communication software and networks*. Springer, Singapore, pp 631–640
15. Dutta A, Kumar S, Basu M (2020) A gated recurrent unit approach to bitcoin price prediction. *J Risk FinancManag* 13(2):23
16. Phaladisailoed T, Numnonda T (2018) Machine learning models comparison for bitcoin price prediction. In: 2018 10th international conference on information technology and electrical engineering (ICITEE), 2018, pp 506–511
17. Uras N, Marchesi L, Marchesi M, Tonelli R (2020) Forecasting Bitcoin closing price series using linear regression and neural networks models. *Peer J Comput Sci* 6:e279
18. Politis A, Doka K, Koziris N (2021) Ether price prediction using advanced deep learning models. In: 2021 IEEE international conference on blockchain and cryptocurrency (ICBC). IEEE, 2021

19. Li S, Li W, Cook C, Zhu C, Gao Y (2018) Independently recurrent neural network (IndRNN): building a longer and deeper RNN. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp 5457–5466
20. Alsharef A, Bhuyan P, Ray A Predicting stock market prices using fine-tuned IndRNN.”
21. Hamilton JD (1990) Time series analysis (Vol. 2). Princeton, Princet.
22. Pankratz A (2009) Forecasting with univariate Box-Jenkins models: concepts and cases, vol 224. Wiley
23. Brownie J (2017) How to grid search ARIMA model hyperparameters with Python. Mach Learn Mastery Saatavissa. Hakupäivä 2:2019. <https://machinelearningmastery.com/grid-search-arima-hyperparameters-with-python/>
24. Hiransha M, Gopalakrishnan EA, Menon VK, Soman KP (2018) NSE stock market prediction using deep-learning models. Procedia Comput Sci 132:1351–1362
25. Finance Y (2020) Eth/USD real-time cryptocurrency price in USD. <https://finance.yahoo.com/quote/ETH-USD/>
26. Lahmiri S, Bekiros S (2021) Deep learning forecasting in cryptocurrency high-frequency trading. Cogn Comput 13:485–487
27. Alsharef A, Aggarwal K, Sonia S, Koundal D, Alyami H, Ameyed D (2022) An automated toxicity classification on social media using LSTM and word embedding. Comput Intell Neurosci 8. Article ID 8467349
28. Gupta G, Salehi AW, Sharma B, Kumar N, Vaidya SP COVID-19: automated detection and monitoring of patients worldwide using machine learning. In: Azar AT, Hassanien AE (eds) Modeling, control and drug development for COVID-19 outbreak prevention. Studies in systems, decision and control, vol 366. Springer.

# A Review on Community Detection Methods and Algorithms in Social Networks: Open Trends and Challenges



Ranjana Sikarwar, Shashank Sheshar Singh, and Harish Kumar Shakya

**Abstract** Community detection in social networks is the most widely studied topic of research direction in complex networks. Among other challenging issues of social networks like link prediction, influence maximization and information propagation through diffusion models, community detection has shown substantial growth in popularity gained in the field of technical research. Investigating social structures through clustering and identification of communities has a plethora of applications in the scientific panorama. The identified community structures help in analyzing the common interest, behavior, and psychology of people connected through social ties irrespective of cultural barriers. This paper conducts a detailed review of several evolutionary and swarm intelligence-based algorithms used more recently and widely for finding the formulated community structures in social networks.

**Keywords** Social area networks · Community detection · Genetic algorithms · Metaheuristics · Complex Networks

## 1 Introduction

A complex network is a web containing a collection of nodes connected through edges, for instance, the world wide web, technological networks, biological networks, brain networks, collaboration networks, online social networks, etc. Community detection (CD) problem deals with finding groups of nodes that have strong intracommunity connections and weak intercommunity connectivity. Investigating important nodes in such networks through community detection can provide better insights to analyze the quality of interconnections between different nodes. Community detection problem is considered to be NP-Hard due to the high complexity of the

---

R. Sikarwar · S. S. Singh · H. K. Shakya (✉)

Department of Computer Science and Engineering, Amity University Gwalior, Gwalior, Madhya Pradesh 474005, India

e-mail: [hkshakya@gwa.amity.edu](mailto:hkshakya@gwa.amity.edu)

S. S. Singh

e-mail: [shashank.sheshar@thapar.edu](mailto:shashank.sheshar@thapar.edu)

network structure [1]. It has numerous applications in social networks, healthcare, modeling of epidemic spreading on networks, business, fraud detection, communication networks, biological networks, etc. [1]. Previous studies done on community detection problems found in the literature have worked upon graph partitioning methods, hierarchical clustering approaches, genetic algorithms, and many other evolutionary algorithms and swarm intelligence-based techniques. As shown in Fig. 1 community detection problems can be studied for disjoint communities (no nodes common in 2 or more different communities) and overlapping community detection (nodes common in two or more communities). Many approaches used for detecting community structure in the literature have focused on static communities satisfying the Modularity fitness function for assessing the quality of partitions. Later on, algorithms for detecting dynamic communities concerning the temporal smoothness along with community partitions have also been examined [2, 3]. CD is considered as an optimization problem, so approaches used for community detection have used either single-objective function for optimization or multiobjective functions. Some of the metaheuristic approaches which employs random searching algorithm have been implemented with great efficiency resulting in global optimal solutions which are discussed in this paper as those using genetic algorithms and PSO. These metaheuristic approaches with heuristic operators are used by many researchers for detecting communities [4, 5].

### ***1.1 Classification of Various Types of Algorithms Used in Literature for Community Detection Problem***

Community Detection is an important direction of research in multidisciplinary areas. So many algorithms are classified and proposed in the literature by the scientists and researchers according to the dimension of the work chosen as enlisted in Fig. 1. The algorithms implemented for community detection can be broadly categorized into Graph partitioning, Clustering, and Genetic algorithms for disjoint communities and clique-based algorithms for overlapping community detection.

### ***1.2 Classification of Various Methods of Community Detection Based on Social Network***

As social networks are so vast and widespread according to the applications, the detection of community structure in different types of social networks needs to exploit different algorithms for the analysis. The community detection problem is intended to identify the highly interrelated nodes or vertices in a network within a group which is strongly communicating with each other.

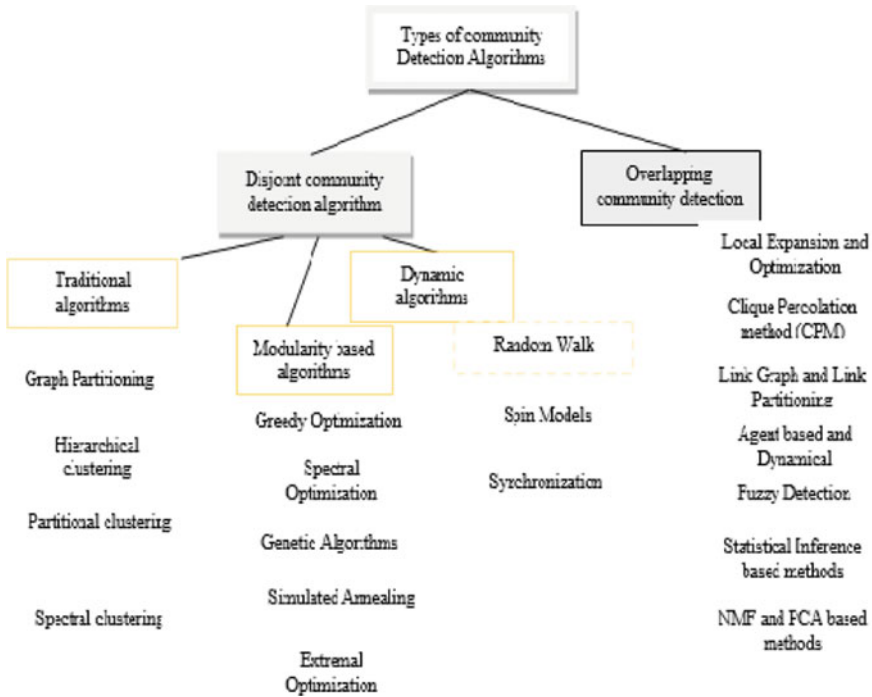


Fig. 1 Classification of Community Detection algorithms [7]

Initially, most of the networks were static but later on with the widespread use of social networking, the networks became more dynamic in nature. The different types of social networks to be investigated in research work are static, signed, positive, dynamic, directed, and heterogeneous [6]. So, the aforementioned task is to:

- Detect communities in static networks
- Community finding in signed networks
- Community detection in dynamic networks
- Detection of community in positive networks
- Detecting communities in heterogeneous networks
- Community detection in a directed network.

The sequence flow of the paper is as follows “**Literature Survey**” discusses the background or the related work, “**Contribution**” describes the novel work done by previous researchers in the latest approaches used, and the “**Conclusion**” the final viewpoint concluding the review done in the paper.

## 2 Literature Survey

Traditional methods used for community detection inspired by clustering methods are those using hierarchical and partitioning methods [8, 9]. But these methods require high computational time and are inefficient to generate optimal solutions in a reasonable time. Also not found efficient for implementation in large-scale networks. The aforementioned issues are addressed very well by evolutionary algorithms using heuristic search. The primary methods used for discovering communities are those by optimizing single-objective function (SO) and using modularity(Q) as an evaluation function which computes intracommunity edges [10]. But using only one objective function may direct the evolving population to form a particular type of community structure or result in a resolution limit problem [11]. Thus, this issue can be addressed by using a multiobjective function (MO) for optimization. MO methods find an optimal solution by establishing a trade-off between different objectives [12]. The concept of Pareto optimality is employed by many evolutionary methods utilizing MO functions. Here in the case of community detection problem MO can find communities with dense intracommunity links and sparse intercommunity connections by optimizing two objective functions simultaneously.

A short glimpse of the previous work done on community detection reveals its societal impact using social networks. The valuable knowledge which can be drawn from studying community structures has led many researchers to investigate the literature behind it. Many comprehensive surveys focus on detecting community structures in multilayer networks [13].

The work presented by (Che et al., 2021) focused on community detection in two modes (bipartite graphs). Their work proposal includes an algorithm which is known as IABC-BN (artificial bee colony algorithm) for detecting communities in bipartite graphs. The experimental results have proved the ABC method to be an excellent algorithm for the discovery of clusters in two-mode graphs. The main contribution of this new algorithm seen is cluster partition for bipartite graphs [14].

Yin et al., (2020) approached the real problem occurring in dynamic networks. The proposed method used DYN-MODPSO for dynamic community detection is an improved evolutionary clustering framework. The multiobjective method is devised for large-scale dynamic networks using PSO. The basic idea is to detect the evolving community structures based on temporal intervals [15].

Reference [16, 17] proposed the use of genetic algorithms with multiobjective criteria to detect communities in complex networks using the algorithm MOGA-Net [18]. His work contributed to the first proposal of using multiobjective GA to discover communities. This algorithm used two objective functions which were optimized to identify partitions in the network structure. The first one uses a community score to evaluate meaningful partitions in the network called communities. A high value of community score corresponds to dense clustering. Another objective function called community fitness is used to analyze the fitness of the nodes confined to a certain group. Further, they extended their work for the application



in dynamic networks using DYNMOGA optimizing modularity and Normalized Mutual Information(NMI) as fitness functions [19].

A particle swarm intelligence-based algorithm called MOPSO-Net was proposed by the authors [20]. Kernel k-means (KKM) and Ratio Cut (RC) are the objective functions to be minimized here. In each iteration, the swarm moved in the direction to achieve the global best solution using the NMI criterion. A Locus-based encoding scheme is used for representation and effective exploration of the solution space. In [21], the authors proposed a many objective(MaOPs) approach for community detection to address the challenges faced by multiobjective methods(using only 2 or 3 quality metrics) in community detection in multi-structural networks. Each quality measure has its specific property for detection thus ignoring other important features to be detected. For example, using only modularity as a quality metric, small communities are left unseen. This issue is addressed by using at least four or larger number of objective functions for identifying community structures.

## ***2.1 Datasets Description***

The datasets often used by many researchers for conducting experimental studies in research work for community detection can be categorized as real datasets (Zachary's karate club, Political blogs, Less Miserables, American college football, Books about US politics, Internet, Coauthorships in Network Science) as well as artificial datasets. They are also known as the benchmark datasets (Lancichinetti et al., Girvan and Newman). These network datasets are in GML format which can be interpreted by many network analysis packages like NetworkX, Cytoscape, etc. (link to download <http://www-personal.umich.edu/~mejn/netdata/>).

## ***2.2 Network Analysis Packages and Tools Used for Identification of Communities***

Some of the popular social network analysis frameworks and tools used for analyzing social network data and graphs are Igraph, Cytoscape, SocNetV, Stanford Network analysis platform (SNAP), Network workbench, NetMiner, NetworkX, Gephi, Graphviz, Neo4j, etc. These social network analysis tools accept network data as GraphML, CSV, GML, and Graphviz file formats and can analyze any type of network data and files. Also, they analyze social networks and outputs important network statistics such as link strength, node density, node strength, visual representation of data, etc. The output file of analyzed network data or graph can be saved or exported in the form of GraphML, GML, BMP, PNG, etc.

### 2.3 Community Detection (CD) as an Optimization Task

In most of the research papers, CD is formulated as an optimization task solved using either a single-objective function or multiobjective function. For instance, reference [12] used two objective functions Negative Ratio Association(NRA) and Ratio cut(RC)(sum of the density of intercommunity links) to be minimized. NRA corresponds to negative RA (sum of internal edge densities of the communities identified). Some of the papers have used modularity as single-objective function [22–25] and many of them used more than one objective function like modularity and NMI(when ground-truth communities are known in advance) [19, 26, 27]. Reference [28] used two objective functions Kernel k-means and Ratio cut with PSO algorithms. Kernel k-means finds solutions with maximum intracommunity edges density and Ratio Cut tries to approach solutions with minimized intercommunity links. The authors [29] have used different variants of objective functions like (Kernel k-means, Ratio Cut, Modularity) as the first variant and (community score, community fitness and modularity) as the second variant with a non-dominated sorting genetic algorithm(NSGA-III). Reference [21] used many objective quality functions such as modularity, NMI, Community Score, Normalized Cut, Conductance, Purity and Rand Index for evaluating the structural properties and quality of the detected communities.

### 2.4 Representation of the Solution

The success of any algorithm depends on the encoding scheme used for representing a solution in the computational search space. Some of the most widely used solution representation schemes used for addressing the community detection problems are discussed below [30].

**Label-based encoding**—Label-based encoding scheme represents the population in the computational space as an integer vector of size(position)  $n$ . Here  $n$  stands for the number(genotype) of nodes. Each location in this vector  $1 \leq l \leq n$ . Suppose if  $k$  is the number of communities in  $\{1, 2, \dots, k\}$ , the  $i$ th position(gene) corresponds to the  $i$ th node. Provided that a genotype has  $k$  number of communities then each gene has a value in the set  $\{1, \dots, k\}$  which is actually the label identifying the community to which the node  $i$  belongs to, thus known as label-based representation. The network in Fig. 2 is partitioned into 3 individual communities as  $\{1, 2, 3\}$ ,  $\{4, 5, 6\}$ ,  $\{7, 8, 9\}$ . Figure 3 below shows the label-representation scheme for Fig. 2.

**Locus-based representation**—This type of solution representation scheme employs an individual  $g$  consisting of  $n$  number of genes  $g_1, g_2, g_3, \dots, g_n$  and each gene  $g_i$  can be mapped to take any adjacent connected node of any node  $i$  as shown in Fig. 4. Thus, in this graph-based representation, a value  $j$  which is assigned to the  $i$ th gene can further be used as a link between node  $i$  and  $j$  in the resultant division of the nodes as communities or partitions of the network. It can be

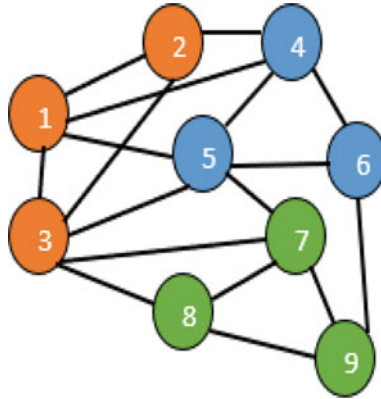


Fig. 2 A network of 9 nodes, 17 edges and 3 communities

<b>position</b>	1	2	3	4	5	6	7	8	9
<b>Label</b>	1	1	1	2	2	2	3	3	3

Fig. 3 Label-based representation of network of Fig. 2

<b>Position</b>	1	2	3	4	5	6	7	8	9
<b>neighbor</b>	2	3	1	5	6	4	8	9	7

Fig. 4 An example of Locus-based representation

concluded that nodes  $i$  and  $j$  belong to the same community. When this representation is decoded all the connected components of the network are identified. The nodes present in one connected component are assigned to one community. The decoding step here helps in finding connected components of the graph. The nodes which form these connected components are assigned to the desired community. This type of representation exhibits redundancy. Label-based representation scheme reduces the complexity of the search space from  $n^n$  (in case of) to  $\prod_{i=1}^n k_i$ ,  $k_i =$  degree of node  $i$ .

**Medoid- based representation** – It’s a prototype-based representation. Here, an  $n$ -dimensional array is used with input elements as the number of communities. For example, from Fig. 2 the partitioned communities are  $\{1,2,3\}$ ,  $\{4,5,6\}$ ,  $\{7,8,9\}$ . Here 1 is the element of the array indicating the prototype of community likewise. This is the medoid-based representation for Fig. 2. These community prototypes coincide with elements of the array. This type of representation scheme shows efficiency for space complexity. However, it has many drawbacks like it is redundant in

nature because medoid can be any element of a particular community and also prior knowledge of  $k$  is required.

Label-based and Locus-based solution representation schemes are the most widely used ones in the literature. The above-described representation schemes refrain a node from becoming a member of more than one community. To overcome this drawback a new representation scheme for overlapping communities was introduced by [31].

## 2.5 Crossover Operator

Although one-point or two-point crossover fits well with label-based representation still it has two main drawbacks. The first drawback is that a community may contain disconnected subgroups of the node means nodes having no connections are placed in the same community. To allay this problem, the idea of one-way crossover was proposed by [32]. But it produces only one child from two parents. Another drawback observed is that the children doesn't receive the genetic characteristics of the parent nodes fully. This issue was encountered by [33]. However, according to the author's observation and view point this crossover enhances the global search fitness of their method but they didn't throw any light on the increase in computational time. While medoid-based representation works with one-point crossover and standard uniform crossover is exploited by locus-based representation. Standard uniform crossover is used by the locus-based representation scheme in which the off springs fully inherit the genetic properties of their parents [16].

## 3 Contribution

### Evolutionary Algorithms (EA)

The category of EA algorithms particularly the genetic algorithms (GA) work on the concept of random population generation. These individuals in the population refer to chromosomes in the case of GA. The structure of chromosomes is organized according to the type of problem GA addresses. An objective function quantifies the quality of chromosomes in the population. This objective function evaluates the fitness value of the chromosomes and a percentage of high-fitness valued chromosomes are selected for the next iteration. Crossover and mutation operators on the chromosomes generate an improved population of individuals until the termination condition is achieved. An optimal solution is produced at the last step of the algorithm. These are the widely embraced techniques to solve NP-complete problems related to optimization due to their robustness in contrast to other traditional methods. GAs that can use different representation schemes are good for solving dynamic problems [34].

**Particle swarm optimization (PSO)**

PSO is used as a population-based stochastic searching algorithm for the community detection problem. It solves the optimization problems simulating the bird flocking behavior which are randomly searching for food in an area. The exact location of the food particle is not known to them. So, they apply the strategy of following those birds which are in close proximity to the food particle. To address any problem, a population or swarm of particles(solutions) is randomly generated initially. These particles search for the optimal solutions in the state space of possible solutions by updating generations. Each particle is associated with a position vector ( $X_i$ ) and a velocity vector ( $V_i$ ). At each iteration, every particle is attracted towards its personal best position ( $Pbest_{id}$ ) and best position of all particles ( $Gbest_{id}$ ) while moving randomly at the same time. [35, 36].

$$v_i^{t+1} = v_i^t + c_1z_1(Gbest_{id} - x_i^t) + c_2z_2(Pbest_{id} - x_i^t) \tag{1}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1} \tag{2}$$

where,  $c_1, c_2$  stands for acceleration parameters known as cognitive and social components  $r_1, r_2$  are random numbers between  $\{0,1\}$ .

**Bat Algorithm (BA)**

Bat Algorithm is also a metaheuristic algorithm that mimics the prey hunting behavior of bats using an echolocation strategy to sense distance and velocity with static variations and loudness frequency. Bat algorithm address the CD as an optimization task where each ‘bat’ represents an individual in the population. It adapts the features of both particle swarm optimization (PSO) and simulated annealing. These combined features make Bat algorithm an outstanding one to achieve global search capability and strong convergence capability. BA simulates the emission rates, loudness and frequency variations of bats when they go for prey hunting. Bats transform their wavelength according to pulse frequency variations to locate the target. The updation rules for position and velocity for BA are similar to those of PSO algorithms. Continuous process of frequency and loudness adjustment maintains a balance between the intensification and diversification operations of the algorithm. BA overcome the drawback of PSO by generating a random solution using random flight behavior to avoid sinking into local optimum [37, 38, 27, 37]. The main equation for updating the bat location based on frequency and velocity is shown below:

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

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

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

where,  $f_{\min}$  is minimum frequency,  $f_{\max}$  is a maximum frequency,  $\beta$  is a random number which takes a value between 0 and 1.  $x_i^t$  is the current location of the  $i$ th bat,  $x_i^{t-1}$  is the previous location of the  $i$ th bat,  $v_i^t$  is the new velocity,  $v_i^{t-1}$  is the previous velocity of the  $i$ th bat.

### Differential Evolution (DE)

DE is a new population-based stochastic search evolutionary algorithm. As compared to the traditional GA algorithm, DE algorithm exhibits some merits: fast convergence, identifies optimized solutions regardless of initial parameters, requiring only a few control parameters. DE initiates the search procedure with a population of NP individuals randomly sampled where each individual signifies the target vector is selected from the population used to generate the mutant vector using the mutation operator. DE's performance depends on the setting of control parameters like the size of the population, crossover, scale factor and the mutation scheme. These parameters should be set properly for the efficient solution of the problem. The mutation scheme of the DE algorithm exploits the genetic information of several individuals to utilize the distributed population characteristics and improve the search ability [39, 40]. Some mutation strategies often used with DE are: DE/rand/1 (known as classical mutation scheme in DE), DE/best/1, DE/best/2, and DE/rand-to-best/1. DE/rand/1 is the most popular mutation strategy used with DE in community detection is as follows:

$$v_{i,m} = x_{r1,m} + F * (x_{r2,m} - x_{r3,m}) \quad (6)$$

where  $i = \{1, 2, \dots, NP\}$ ,  $r1, r2$  and  $r3$  are randomly selected integer values from  $1, 2, \dots, NP$ , satisfying  $r1 \neq r2 \neq r3 \neq I$ , scaling factor  $F$  is a real number between  $\{0,1\}$ .

### Memetic Algorithms (MAs)

Memetic Algorithms (MA) are considered as the hybridization of previous GA based evolutionary algorithms. It is also a population-based approach with separate individual learning or optimization intersperse with the recombination of high-quality solutions. They consider evolution as a baseline principle of working. It relies on the local search improvement procedures for problem search thus reducing the premature convergence. The word "memetic" is stirred by the Dawkin's notion of the word 'meme', an element of social development resulting in local refinement [41]. The meme used in MAs represents a distinct learning procedure which can exhibit local refinements. MA uses the combination of GA and local search procedure to solve the optimization problem. MA outperforms existing genetic algorithms for specific applications of community detection [42–44].

### Ant Colony Optimization

It is a metaheuristic optimization algorithm, basically a simulation of the ants foraging behavior independently communicating with each other through pheromone. It is also considered as a distributed multi-agent system where the search for food begins from different locations at the same time [45]. The population of ants construct solutions iteratively by finding the shortest path using pheromone and leaving the heuristic information behind them by crossing the paths. ACO algorithms are used in finding the community structure in the network. The positive feedback mechanism is used to find optimal solutions. The quality of the solution achieved by each artificial ant is assessed by its modularity. The probability of selecting a route by the ants from vertex  $x$  to  $y$  is given by the following formula below:

$$p_{xy} = \frac{\mu_{xy}^\alpha h_{xy}^\beta}{\sum_{x,y=1}^n \mu_{xy}^\alpha h_{xy}^\beta} \tag{7}$$

where,

- $\mu_{xy}$  is the pheromone concentration of the path between  $x, y$
- $h_{xy}$  is a heuristic function with a likelihood to select an edge from point  $x$  to  $y$ .
- $\alpha, \beta$  determines relative influence of trail information and visibility.

### Firefly Algorithm

It's a population-grounded algorithm where each firefly represents a feasible solution. This algorithm imitates the flashing patterns and activities of the fireflies [46]. The main principle for the sparkle of fireflies is to attract other fireflies. This algorithm was proposed with a few assumptions like a firefly is attracted towards another firefly according to the brightness intensity. With the increase in the distance the brightness of the firefly decreases. The movement of fireflies towards the brightest firefly is to achieve a global optimal solution. FA algorithm depends on the parameters like random movement and attractiveness as performance measures. Community detection problems can be solved using the FA algorithm as an optimization algorithm by maximizing the modularity function. The main update formula [47] for any pair of two fireflies  $x_i$  and  $x_j$  is

$$x_i^{t+1} = x_i^t + \beta_0^{-\gamma r_{ij}^2} (x_j^t - x_i^t) + \alpha_t e_i^t \tag{8}$$

where,  $x_i^t$  represents the  $i^{\text{th}}$  solution (firefly) at iteration  $t$ .

$\beta_0$  is brightness at source.

A solution  $x_i$  will be attracted towards a brighter firefly  $x_j$ , means  $x_i$  moves towards  $x_j$ ,  $\alpha$  randomization parameter,  $e_i^t$  vector of random variables.

## 4 Conclusion

The aim of this comprehensive review is to encompass various evolutionary and swarm intelligent-based algorithms for community detection that have encouraged a flurry of research. The widespread use of the aforementioned algorithms has shown an outstanding performance in detecting communities in static, dynamic, complex or multi-structural networks. Classification of different types of methods and algorithms used in addressing the community detection problem on the basis of social networks is also discussed here. The discussion of evolutionary and nature-inspired (NIA) algorithms based on single-objective or multiobjective has also been covered along with the most commonly adopted evaluation metrics like Modularity and NMI. A detailed description of the most widely used EA and NIA algorithms is statistically broken down and summarized in the tabular form according to the common key components used. These statistics provides a direction to the readers and researchers to select the characteristics of the algorithms like population initialization methods, perturbation operators and types of objective functions. It is observed that most of the research papers have shown a research gap for community detection in overlapping communities, multilayer networks and large-scale networks, implementing the algorithms independent of the increasing network size and substantial improvement in speed and accuracy.

## References

1. Javed MA, Younis MS, Latif S, Qadir J, Baig A (2018) Community detection in networks: a multidisciplinary review. *J Netw Comput Appl* 108:87–111. <https://doi.org/10.1016/j.jnca.2018.02.011>
2. Zeng X, Member S, Wang W, Chen C, Yen GG (2019) A consensus community-based particle swarm optimization for dynamic community detection. *IEEE Trans Cybern* 1–12. <https://doi.org/10.1109/TCYB.2019.2938895>
3. Messaoudi I, Kamel N (2019) A multi-objective bat algorithm for community detection on dynamic social networks. *Appl Intell* 49(6):2119–2136. <https://doi.org/10.1007/s10489-018-1386-9>
4. Hussain K, Mohd Salleh MN, Cheng S, Shi Y (2019) Metaheuristic research: a comprehensive survey. *Artif Intell Rev* 52(4), 2191–2233. <https://doi.org/10.1007/s10462-017-9605-z>
5. Dokeroglu T, Sevinc E, Kucukyilmaz T, Cosar A (2019) A survey on new generation metaheuristic algorithms. *Comput Ind Eng* 137. <https://doi.org/10.1016/j.cie.2019.106040>
6. Pourkazemi M, Keyvanpour M (2013) A survey on community detection methods based on the nature of social networks. In: *International conference on computer and knowledge engineering ICCKE 2013*, no. Ickce, pp 114–120. <https://doi.org/10.1109/ICCKE.2013.6682855>
7. Fortunato S (2010) Community detection in graphs. *Phys Rep* 486(3–5):75–174. <https://doi.org/10.1016/j.physrep.2009.11.002>
8. Lu X, Kuzmin K, Chen M, Szymanski BK (2018) Adaptive modularity maximization via edge weighting scheme. *Inf Sci (Ny)* 424:55–68. <https://doi.org/10.1016/j.ins.2017.09.063>
9. Wu W, Kwong S, Zhou Y, Jia Y, Gao W (2018) Nonnegative matrix factorization with mixed hypergraph regularization for community detection. *Inf Sci (Ny)* 435:263–281. <https://doi.org/10.1016/j.ins.2018.01.008>



10. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E - Stat Nonlinear Soft Matter Phys* 69(22), 1–15. <https://doi.org/10.1103/PhysRevE.69.026113>
11. Fortunato S, Barthélemy M (2007) Resolution limit in community detection. *Proc Natl Acad Sci USA* 104(1):36–41. <https://doi.org/10.1073/pnas.0605965104>
12. Shang J, Li Y, Sun Y, Li F, Zhang Y, Liu J (2021) SS symmetry MOPIO : a multi-objective pigeon-inspired optimization, pp 1–16
13. Huang X, Chen D, Ren T, Wang D (2020) A survey of community detection methods in multilayer networks. Springer, US
14. Che S, Yang W, Wang W (2021) An improved artificial bee colony algorithm for community detection in bipartite networks. *IEEE Access* 9:10025–10040. <https://doi.org/10.1109/ACCESS.2021.3050752>
15. Yin Y et al (2020) Multi-objective evolutionary clustering for large-scale dynamic community detection. <https://doi.org/10.1016/j.ins.2020.11.025>
16. Pizzuti C (2008) GA-Net: a genetic algorithm for community detection in social networks. *Lecture notes in computer science (including Lecture notes in artificial intelligence and lecture notes in bioinformatics)*, vol 5199 LNCS, pp 1081–1090. [https://doi.org/10.1007/978-3-540-87700-4\\_107](https://doi.org/10.1007/978-3-540-87700-4_107)
17. Pizzuti C (2012) A multiobjective genetic algorithm to find communities in complex networks. *IEEE Trans Evol Comput* 16(3):418–430. <https://doi.org/10.1109/TEVC.2011.2161090>
18. Pizzuti C (2009) A multi-objective genetic algorithm for community detection in networks. In: *Proceedings of international conference on tools with artificial intelligence ICTAI*, no. October 2014, pp 379–386. <https://doi.org/10.1109/ICTAI.2009.58>
19. Folino F, Pizzuti C (2014) An evolutionary multiobjective approach for community discovery in dynamic networks. *IEEE Trans Knowl Data Eng* 26(8):1838–1852. <https://doi.org/10.1109/TKDE.2013.131>
20. Rahimi S, Abdollahpouri A, Moradi P (2017) “SC,” and evolutionary computation BASE DATA. <https://doi.org/10.1016/j.swevo.2017.10.009>
21. Tahmasebi S, Moradi P, Ghodsi S, Abdollahpouri A (2019) An ideal point based many-objective optimization for community detection of complex networks. *Inf Sci (Ny)* 502:125–145. <https://doi.org/10.1016/j.ins.2019.06.010>
22. Guerrero M, Montoya FG, Baños R, Alcalde A, Gil C (2017) Adaptive community detection in complex networks using genetic algorithms. *Neurocomputing* 266:101–113. <https://doi.org/10.1016/j.neucom.2017.05.029>
23. Moradi M, Parsa S (2019) An evolutionary method for community detection using a novel local search strategy. *Phys A*. <https://doi.org/10.1016/j.physa.2019.01.133>
24. Guo X, Su J, Zhou H, Liu C, Cao J, Li L (2019) Community detection based on genetic algorithm using local structural similarity. *IEEE Access* 7:134583–134600. <https://doi.org/10.1109/ACCESS.2019.2939864>
25. Li C, Wang R, Li J, and Fei L, *Face detection based on YOLOv3*, vol. 1031 AISC. 2020
26. Ghaffaripour Z, Abdollahpouri A, Moradi P (2016) A multi-objective genetic algorithm for community detection in weighted networks. In: *2016 8th international conference on information and knowledge technology IKT 2016*, pp 193–199. <https://doi.org/10.1109/IKT.2016.7777766>
27. Doush IA, Alrashdan WB, Al-Betar MA, Awadallah MA (2020) Community detection in complex networks using multi-objective bat algorithm. *Int J Math Model Numer Optim* 10(2):123–140. <https://doi.org/10.1504/IJMMNO.2020.106529>
28. Gong M, Cai Q, Chen X, Ma L (2014) Complex network clustering by multiobjective discrete particle swarm optimization based on decomposition. *IEEE Trans Evol Comput* 18(1):82–97. <https://doi.org/10.1109/TEVC.2013.2260862>
29. Shaik T, Ravi V, Deb K (2021) Evolutionary multi - objective optimization algorithm for community detection in complex social networks. *SN Comput Sci*. <https://doi.org/10.1007/s42979-020-00382-x>

30. Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLF (2009) A survey of evolutionary algorithms for clustering. *IEEE Trans Syst Man Cybern Part C Appl Rev* 39(2):133–155. <https://doi.org/10.1109/TSMCC.2008.2007252>
31. Liu J, Zhong W, Abbass HA, Green DG (2010) Separated and overlapping community detection in complex networks using multiobjective Evolutionary Algorithms. In: 2010 IEEE world congress on computational intelligence WCCI 2010 - 2010 IEEE congress on evolutionary computation CEC 2010. <https://doi.org/10.1109/CEC.2010.5586522>
32. Tasgin M, Bingol H (2006) Community detection in complex networks using genetic algorithm, pp 1–6. <http://arxiv.org/abs/cond-mat/0604419>
33. He D, Wang Z, Yang B, Zhou C (2009) Genetic algorithm with ensemble learning for detecting community structure in complex networks. In: ICCIT 2009 - 4th international conference on computer sciences and Convergence Information Technology, pp 702–707. <https://doi.org/10.1109/ICCIT.2009.189>
34. Abduljabbar DA, Hashim SZM, Sallehuddin R (2020) Nature-inspired optimization algorithms for community detection in complex networks: a review and future trends. *Telecommun Syst* 74(2):225–252. <https://doi.org/10.1007/s11235-019-00636-x>
35. Cai Q, Gong M, Ma L, Ruan S, Yuan F, Jiao L (2015) Greedy discrete particle swarm optimization for large-scale social network clustering. *Inf Sci (Ny)* 316:503–516. <https://doi.org/10.1016/j.ins.2014.09.041>
36. Gao C, Chen Z, Li X, Tian Z, Li S (2018) Multiobjective discrete particle swarm optimization for community detection in dynamic networks. <https://doi.org/10.1209/0295-5075/122/28001>
37. Hassan EA, Hafez AI, Hassanien AE, Fahmy AA (2015) A discrete bat algorithm for the community detection problem. *Lecture notes in computer science (including Lecture notes in artificial intelligence and lecture notes in bioinformatics)*, vol 9121, pp 188–199. [https://doi.org/10.1007/978-3-319-19644-2\\_16](https://doi.org/10.1007/978-3-319-19644-2_16)
38. Chunyu W, Yun P (2015) Discrete bat algorithm and application in community detection. *Open Cybern Syst J* 9(1):967–972. <https://doi.org/10.2174/1874110X01509010967>
39. Jia G et al. (2012) Community detection in social and biological networks using differential evolution. *Lecture notes in computer science (including Lecture notes in artificial intelligence and lecture notes in bioinformatics)*, vol 7219. LNCS, pp 71–85. [https://doi.org/10.1007/978-3-642-34413-8\\_6](https://doi.org/10.1007/978-3-642-34413-8_6)
40. Sun H, Ma S, Wang Z (2018) A community detection algorithm using differential evolution. In: 2017 3rd IEEE international conference on computer and communications ICC 2017, pp 1515–1519. <https://doi.org/10.1109/CompComm.2017.8322793>
41. Mu CH, Xie J, Liu Y, Chen F, Liu Y, Jiao LC (2015) Memetic algorithm with simulated annealing strategy and tightness greedy optimization for community detection in networks. *Appl Soft Comput J* 34:485–501. <https://doi.org/10.1016/j.asoc.2015.05.034>
42. Žalik KR, Žalik B (2018) Memetic algorithm using node entropy and partition entropy for community detection in networks. *Inf Sci (Ny)* 445–446:38–49. <https://doi.org/10.1016/j.ins.2018.02.063>
43. Wang S, Gong M, Liu W, Wu Y (2020) Preventing epidemic spreading in networks by community detection and memetic algorithm. *Appl Soft Comput J* 89:106118. <https://doi.org/10.1016/j.asoc.2020.106118>
44. Haque MN, Mathieson L, Moscato P (2018) A memetic algorithm for community detection by maximising the connected cohesion. In: 2017 IEEE symposium series on computational intelligence SSCI 2017 - Proceedings, pp 1–8. <https://doi.org/10.1109/SSCI.2017.8285404>
45. Chen B, Chen L, Chen Y (2012) Detecting community structure in networks based on ant colony optimization. In: International conference on information and knowledge engineering, pp 247–253
46. Jaradat AS, Hamad SB (2018) Community structure detection using firefly algorithm. *Int J Appl Metaheuristic Comput* 9(4):52–70. <https://doi.org/10.4018/IJAMC.2018100103>
47. Del Ser J, Lobo JL, Villar-Rodriguez E, Bilbao MN, Perfecto C (2016) Community detection in graphs based on surprise maximization using firefly heuristics. In: 2016 IEEE congress on evolutionary computation CEC, pp 2233–2239. <https://doi.org/10.1109/CEC.2016.7744064>

# Three-Stage Heterogeneous Data Clustering Using Unsupervised Multiple Kernel and Extreme Learning Machine



Ankit R. Mune and Sohel A. Bhura

**Abstract** In spite of momentous endeavors for regulated various portion ELM, hardly any distributions have tended to the solo case, which is more basic yet trying for handling the practical issue. Different Authors proposed, in particular, two-stage unaided numerous piece outrageous learning machine (TUMK-ELM) a quick unsupervised heterogeneous information learning algorithm to settled this issue on the other hand removes data from various sources and learns the heterogeneous information portrayal with closed structure arrangements, which empowers its very quick speed. With the low computational complexity, TUMK-ELM provides at each stage, and the emphasis of its two phases can be merged inside limited advances. This issue of two phase can be settled by three-Stage unsupervised multiple kernel clustering-based extreme learning machine (TMKC-ELM) for quick unsupervised heterogeneous data learning. This will catch the heterogeneous data from various sources through different portions and incorporate the heterogeneous data into an ideal piece through an iterative three-stage approach directed by an overall unsupervised even-handed. Dataset will be preprocessed to eliminate filthy qualities in it. Information will be built from the various kernels, and pseudo-marks will be doled out using a clustering algorithm as indicated by a learned K-space through the optimal kernel. This research work will be an endeavor to propose an efficient three-stage unsupervised multiple kernels extreme learning approach.

**Keywords** : Heterogeneous data · Unsupervised clustering · Multiple kernel functions

---

A. R. Mune (✉) · S. A. Bhura (✉)  
Babasaheb Naik College of Engineering, Pusad, India  
e-mail: [mune.ankit@gmail.com](mailto:mune.ankit@gmail.com)

S. A. Bhura  
e-mail: [sabhura@rediffmail.com](mailto:sabhura@rediffmail.com)

## 1 Introduction

Recent advances in computing technology, sensor, and communication innovations are producing data at no other time seen multiply and at continually expanding rates. According to the perspective of information examination, the data is made out of a variety of information types and it contains vulnerabilities and deficiency of various degrees, which add an additional segment to the first heterogeneity. Numerous information mining and AI techniques don't deal with heterogeneity well. Extreme learning machines (ELM) are fascinating computational calculations in light of their straightforwardness, their great presentation, and their speed. They can be stretched out for preparing data made out of heterogeneous information types (HT-ELM), fit for tending to order, and relapse issues with complex information [1].

Heterogeneity is one of the significant highlights of large information and heterogeneous information brings about issues in information reconciliation and Big Data investigation. Numerous informational indexes are heterogeneous in type, structure, association, granularity, semantics, openness, and so on the high variety of information sources frequently prompts information storehouses, an assortment of non-coordinated information the executive's frameworks with heterogeneous compositions, APIs, and inquiry dialects. Information types from heterogeneous sources are frequently needed to be brought together during pre-preparing [2].

The heterogeneity of the benchmark dataset is summed up, where every information type is given close by its related inspecting rate and capacity strategy [3].

In recent years, numerous scientists improve model limits by joining learning with an unsupervised model having profound figuring out how to propose unsupervised deep learning models. Be that as it may, a large portion of them neglect to gain from various sources. In view of the difficulties looked by them, one promising approach to uncover data from various sources is to utilize learning from multiple kernels (MKL). In spite of the upsides of MKL, it requires regulated name data to get familiar with the ideal bit mix coefficients. All the more as of late, MKL with unsupervised learning has been concentrated to handle the heterogeneous information learning without administered marks. Heterogeneous information learning requires an unsupervised learning model. In most genuine cases, administered name data for heterogeneous information learning isn't accessible or with high time/human utilization. In these cases, the vast majority of current heterogeneous information learning strategies don't function admirably since the rules for heterogeneous data coordination is absent. The most effective method to characterize an unaided target work that can profit general examination undertakings is basic yet challenging. The information coordinated from various sources might be loud or deficient so this is another test. All this together motivates us to apply statistical machine learning analysis on a heterogeneous set of information [4].

In a previous project, TUMK-ELM accumulates heterogeneous data from various sources through many kernels in a three-phase iterative technique, then merges heterogeneous data into an optimal kernel, with ultimately unsupervised purpose.

It integrates data from several sources successfully using a confined method that more completely describes the data set. TUMK-ELM [5] has the advantages of the kernel k-means and ELM, which optimize the effectiveness of unsupervised heterogeneous data training regarding efficiency and effectiveness. Extreme learning machines (ELMs), due to their structural simplicity, efficiency and rapidity, are particularly fascinating computing technologies. Expanding their breadth by enabling them to analyze heterogeneous facts may enhance their desirability as a modeling method for complicated Massive Data situations. Processes the heterogeneous data using a neuron model that can handle heterogeneous data with artificial neural networks in general [1].

This paper highlights key ideas as follows:

- To study existing Extreme Learning Machine for Heterogeneous information.
- To study two-stage multiple kernel unsupervised extreme learning machine (TUMK-ELM).
- To study and Analyze Multiple heterogeneous datasets.
- To incorporate an iterative three phases methodology to integrate various heterogeneous data.
- Define a Three-Stage unsupervised Clustering with multiple kernels using extreme learning machine (TMKC-ELM).

## ***1.1 Related Work***

Vasileios Christou et al. [6] presented a learning model with a hybrid approach, which consolidates the extreme learning machine (ELM) with a genetic algorithm (GA). The use of this hybrid model empowers the production of heterogeneous single-layer neural networks (SLNNs) with preferred speculation capacity over customary ELM as far as a lower mean square error (MSE) for relapse issues or higher exactness for grouping issues. The engineering of this technique isn't restricted to conventional straight neurons, where each info takes an interest similar to the neuron's enactment, however, is reached out to help higher request neurons which influence the organization's speculation capacity. At first, the proposed heterogeneous cross breed outrageous learning machine (He-HyELM) calculation makes various exclusively made neurons with various construction, which are utilized for the production of homogeneous SLNNs. These organizations are prepared with ELM and an application explicit GA advances them into heterogeneous organizations as per a wellness standard using the uniform hybrid administrator for the recombination cycle. After the finishing of the development cycle, the organization with the best wellness is chosen as the most ideal. Test results show that the proposed learning calculation can improve results than conventional ELM, homogeneous cross breed outrageous learning machine (Ho-HyELM), and ideally pruned outrageous learning machine (OP-ELM) for homogeneous and heterogeneous SLNNs.

Zhou and Mama [7] introduced an extreme learning machine (ELM)-based heterogeneous space transformation (HST) calculation that is proposed for the arrangement

of distant detecting pictures. In the versatile ELM organization, one secret layer is utilized for the source information to give the arbitrary highlights, while two secret layers are set for target information to deliver the irregular highlights just as a change network. DA is accomplished by obliging both the source information and the changed objective information to have similar yield loads. In addition, complex regularization is received to safeguard the neighborhood calculation of unlabeled objective information. The proposed ELM-based HDA (EHDA) technique is applied to the cross-space characterization of distant detecting pictures, and the trial results utilizing multisensor far-off detecting pictures exhibit the viability of the proposed approach.

Xiang et al. [5] proposed a quick solo heterogeneous information learning calculation, in particular two-stage unaided various bit outrageous learning machine (TUMK-ELM). TUMK-ELM then again separates data from numerous sources and learns the heterogeneous information portrayal with shut structure arrangements, which empowers its incredibly quick speed. As supported by hypothetical proof, TUMK-ELM has low computational intricacy at each stage, and the cycle of its two phases can be joined inside limited advances. As tentatively exhibited on 13 genuine informational collections, TUMK-ELM acquires a huge proficiency improvement contrasted and three cutting edge unaided heterogeneous information learning techniques (up to 140 000 times) while it's anything but a practically identical execution as far as viability.

Kong et al. [8] proposed various piece learning (MKL) that was utilized to join the heterogeneous highlights of Landsat-8 and Sentinel-1A information. Impenetrable surface was assessed at a sub-pixel level dependent on help vector relapse (SVR) model. The impenetrable surface rate (ISP) of preparing information was gotten from the high goal picture utilizing the article arranged order approach. The outcomes demonstrate that the consolidated utilization of optical and SAR by utilizing MKL can essentially improve the assessment of impenetrable surfaces since it's difficult to lessen the underestimation and overestimation of ISP in metropolitan regions, yet in addition, well isolates uncovered soils from the impenetrable surface. Contrasted and utilizing optical picture alone, the root mean square blunder (RMSE) is diminished by 5.5% and the coefficient of assurance ( $R^2$ ) is expanded by 8.8%.

Liu et al. [9] propose a novel fluffy-based heterogeneous solo space transformation approach. This methodology maps the component spaces of the source and target areas onto a similar inactive space built by fluffy highlights. In the new component space, the mark spaces of two areas are kept up to diminish the likelihood of negative exchange happening. The proposed approach conveys better execution over current benchmarks, and the heterogeneous unaided area transformation (HeUDA) strategy gives promising methods for giving a learning framework the cooperative capacity to pass judgment on obscure things utilizing related information.

Yang and Zhou [10] propose a novel methodology dependent on the CluS-stream structure for bunching information stream with heterogeneous highlights. The centroid of consistent qualities and the histogram of the discrete properties are utilized to address the miniature groups, and k-model bunching calculation is utilized to make

the miniature bunches and full-scale groups. Trial results on both manufactured and genuine informational indexes show its proficiency.

Salama et al. [11] proposed a framework for solving heterogeneous data analysis problems and reducing the volume of data and focused mainly on the preparation and discovery of the optimum ML model for analyzing heterogeneous text data. In this study, several ML and DL algorithms have been compared throughout our case study to more than 14,500 tweets. We used supervised ML algorithms and deep learning algorithms to improve the precision of the positive, negative, and neutral classification of tweets.

## 2 Analysis of Tumk-ELM

Author proposed two-stage unsupervised multiple kernel extreme learning machine (TUMK-ELM, for short) for the fast unsupervised heterogeneous data learning. TUMK-ELM captures the heterogeneous information from different sources via multiple kernels and integrates the heterogeneous information into an optimal kernel through an iterative two-stage approach guided by a general unsupervised objective. The framework of TUMK-ELM is shown in Fig. 1 [5].

### TUMK-ELM

- Firstly target heterogeneous information into part spaces by numerous pieces.

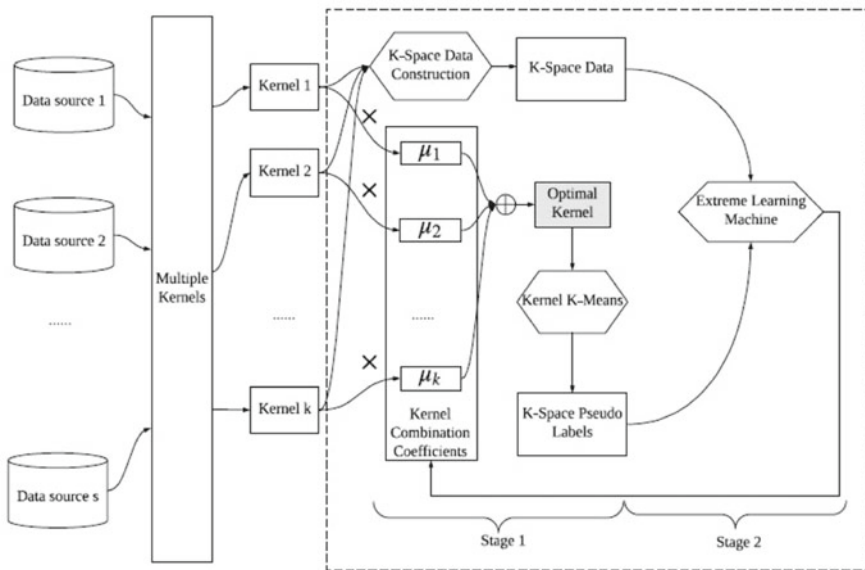


Fig. 1 TUMK-ELM framework



- It then at that point embraces an iterative 2 phases way to deal with coordinate heterogeneous data.
- In the principal stage, TUMK-ELM produces a K-Space, where the information is developed from different piece spaces and the pseudo-names are allotted as indicated by the learned ideal part.
- In the subsequent stage, TUMK-ELM learns ideal part blend coefficients dependent on the produced K-Space. After assembly, the ideal part contains the incorporated data from heterogeneous information that suits the resulting examination undertakings [5].

### Drawbacks of Two-stage

- The Learning speed is exceptionally less.
- As per the perception, the boisterous information is exceptionally huge so need to preprocess the information.
- As per the perception K Means isn't useful for the classification of heterogeneous information.

## 2.1 Study of TUMK-ELM

The TUMK-ELM viability screen concerns bunch productivity and heterogeneous datasets quality. Two measurements evaluate the bunch effectiveness: NMI and virtue. The above measurements utilize the information marks as a grouping premise of the UCI AI store. The joined meaning of a few components is illustrated. In addition, the NMI shows the connection between both the yields of the bunches and the reality of the ground, while immaculateness assesses the inspecting percent for a particular group. A subjectively imagining of information conveyances estimates the heterogeneity level of portrayals.

In this examination, the benchmark informational indexes incorporate Iris, Haberman, Wine, Glass, Seeds, Biodeg, Libras-development, and Image-section. The UCI ML Repository [12] permits recovering all informational collections. Figure 2 shows the dimensionality addressing the quantities of tests, classes, and highlights of every one of these datasets.

Fig. 2 Datasets summary

Data Set	Number of Instance	Number of Attributes	Number of Class
Iris	150	4	3
Haberman	360	4	2
Wine	178	13	3
Glass	214	9	6
Seeds	210	7	3
Biodeg	1055	41	2
Libras-movement	360	90	15
Image-segment	210	18	7



## Evaluation of TUMK-ELM is based on Two Factors

1. Normalized Mutual Information (NMI).
2. Purity.

## 3 Proposed Tmkc-Elm

(Three-Stage unsupervised multiple kernel Clustering-based extreme learning machine).

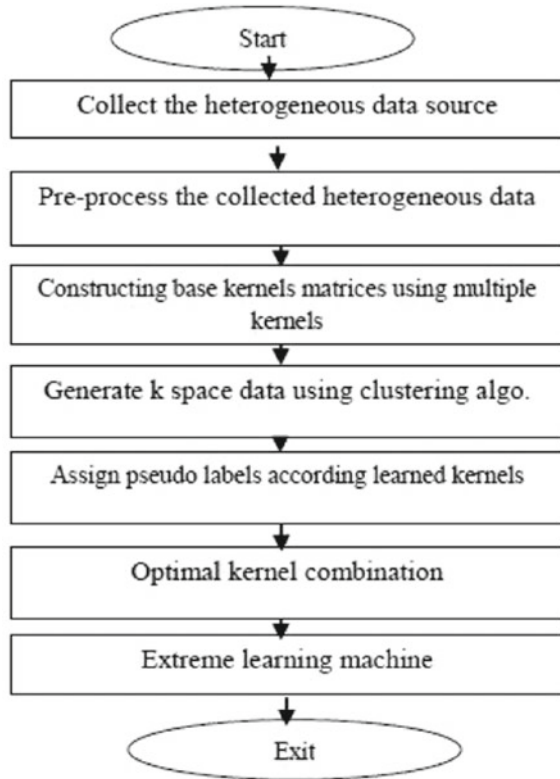
The Point of this examination is to empower the learning without directed marks; it's anything but a three phase of various part-based solo learning objective to get familiar with the ideal portion mix coefficients. Albeit unaided MKL accomplishes amazing execution in solo heterogeneous information learning, the majority of the current solo MKL techniques are with a lethargic learning speed. The lethargic learning speed is mostly brought about by the iterative mathematical arrangement, which is received by these strategies for improving the piece mix coefficients. It doesn't fulfill the prerequisites of dealing with a lot of information and constant learning.

### 3.1 Proposed Stages of TMKC-ELM

Following are the Stages for Three-Stage unsupervised multiple kernel Clustering-based extreme learning machine (TMKC-ELM).

- Stage 1. Gather the heterogeneous dataset from various information sources.
- Stage 2. Embraces an iterative three phases way to deal with coordinate heterogeneous data.
- Stage 3. In the first stage, Pre-measure the gathered dataset to eliminate exceptions.
- Stage 4. Three-Stage solo various part Clustering-based outrageous learning machine (TMKC-ELM) first will extend heterogeneous information into portion spaces by different pieces. Developing base part grids  $K$  by utilizing the information bits to project the heterogeneous information.
- Stage 5. At the Second stage, TMKC-ELM will produce a  $K$ -Space, wherein the information will develop from different portion spaces and the pseudo-names will be allotted as per the learned ideal piece.
- Stage 6. At the third stage, TMKC-ELM will learn ideal piece mix coefficients dependent on the produced  $K$ -Space.
- Stage 7. After union, the ideal piece will contain the coordinated data from heterogeneous information that suits the resulting examination errands (Fig. 3).

**Fig. 3** Proposed TUMK-ELM Flow Diagram



### 3.2 Proposed Methodology

In Proposed TMKC-ELM Methodology will be working in the following phases.

(a) **Data Collection**

The standard benchmark datasets are available in the UCI machine learning repository, which we have used to perform our experiment.

(b) **Data Pre-processing**

The present genuine and original information bases are profoundly susceptible to noise, missing, and conflicting information as a result of their commonly gigantic size and their probable beginning from different, heterogeneous sources. Fragmented information can happen for various reasons. Qualities of interest may not generally be accessible. Information pre-preparing is a demonstrated technique for addressing such issues.

(c) **K-Space Data Construction**

TMKC-ELM will remove heterogeneous data from numerous sources by p piece capacities. These part capacities can be configured as indicated by earlier

information and information qualities. After the part projection, TMKC-ELM gets a bunch of  $k$  base bit grids, which is utilized for the ideal piece age and K-Space information development. Meaning the informational collection in a K-Space as  $Z$ , the change from  $K$  to  $Z$  of a given informational index  $X$  is formalize. TMKC-ELM will allot K-Space pseudo-name by means of grouping calculation. The ideal part will create by a direct blend of the  $k$  base piece frameworks as indicated by a bunch of mix coefficients.

(d) **Clustering**

Bunching can be viewed as the main unaided learning issue; thus, as every issue of this sort. The objective of bunching is to decide the interior gathering in a bunch of unlabeled information. The client should supply this model so that the aftereffect of the bunching will suit their requirements.

(e) **Multiple Kernel Learning**

For  $n_k$  K-Space information and pseudo-names, TMKC-ELM will streamline the given target capacity to figure the ideal piece mix coefficients. For information from plenty of different sources, TMKC-ELM likes to compute the ideal arrangement in a quicker manner.

## 4 Conclusion and Future Work

The primary commitment of this examination is to marking the unlabeled information and lessen the time complexity. This will be further useful for Social Network Analysis and frameworks including Heterogeneous information processing. In this paper, we examined around two-stage extreme unsupervised learning machine with multiple kernel TUMK-ELM with its functioning situation and how it will be helpful to heterogeneous information processing and clustering. The Evaluation of TUMK-ELM depends on NMI and the Purity of Data. The NMI and Purity are exceptionally less so to address this issue, we plan the Three-Stage multiple kernel Clustering using unsupervised learning-based extreme learning machine TMKC-ELM. In future works, the TMKC-ELM will be applied on available benchmark eight datasets for guaranteeing the best NMI and Purity than the TUMK-ELM, the utilization of an extreme learning machine will be upgraded, which accomplishes exceptional improvement, remembering both efficiency and performance for heterogeneous data unsupervised information learning. ELM will gain proficiency with the best coefficients to every multiple kernel combination. Such coefficients will be learned by information in addition to pseudo-marks in K-Space based on an ELM during stage 2 of the TUMKFCM.

## References

1. Valdés JJ (2018) Extreme learning machines with heterogeneous data types. *Neurocomputing* 277:38–52. <https://doi.org/10.1016/j.neucom.2017.02.103>
2. Wang L, Jones R (2017) Big data analytics for disparate data, *American. J Intell Syst* 7(2):39–46. <https://doi.org/10.5923/j.ajis.20170702.01>
3. Stief A, Tan R, Cao Y, Ottewill JR, Thornhill NF, Baranowski J (2019) A heterogeneous benchmark dataset for data analytics: multiphase flow facility case study. *J Process Control* 79:41–55. <https://doi.org/10.1016/j.jprocont.2019.04.009>
4. Ghaderi A, Athitsos V (2016) Selective unsupervised feature learning with Convolutional Neural Network (S-CNN). In: 2016 23rd international conference on pattern recognition (ICPR). <https://doi.org/10.1109/icpr.2016.7900009>
5. Xiang L, Zhao G, Li Q, Hao W, Li F (2018) TUMK-ELM: a fast unsupervised heterogeneous data learning approach. *IEEE Access* 6:35305–35315. <https://doi.org/10.1109/ACCESS.2018.2847037>
6. Christou V, Tsiouras MG, Giannakeas N, Tzallas AT, Brown G (2019) Hybrid extreme learning machine approach for heterogeneous neural networks. *Neurocomputing*. <https://doi.org/10.1016/j.neucom.2019.04.092>
7. Zhou L, Ma L (2019) Extreme learning machine-based heterogeneous domain adaptation for classification of hyperspectral images. *IEEE Geosci Remote Sens Lett* 16(11):1781–1785. <https://doi.org/10.1109/LGRS.2019.2909543>
8. Kong Y, Sun G, Zhang A, Huang H (2018) Synergistic use of optical and SAR data with multiple kernel learning for impervious surface mapping. In: Fifth international workshop on earth observation and remote sensing applications (EORSAs). Xi'an 2018:1–4. <https://doi.org/10.1109/EORSAs.2018.8598552>
9. Liu F, Zhang G, Lu J (2017) Heterogeneous unsupervised domain adaptation based on fuzzy feature fusion. In: 2017 IEEE international conference on fuzzy systems (FUZZ-IEEE), Naples, 2017, pp 1–6. <https://doi.org/10.1109/FUZZ-IEEE.2017.8015569>
10. Yang A, Zhou J (2006) HClustream: A novel approach for clustering evolving heterogeneous data stream. In: Sixth IEEE international conference on data mining - workshops (ICDMW'06), Hong Kong, 2006, pp 682–688. <https://doi.org/10.1109/ICDMW.2006.89>
11. Salama M, Kader HA, Abdelwahab A (2021) An analytic framework for enhancing the performance of big heterogeneous data analysis. *Int J Eng Bus Manag*
12. Bache MK (2017) Uci machine learning repository. <http://archive.ics.uci.edu/ml/index.php>

# Artificial Intelligences-Based Approaches for Generating Image Caption



S. P. Ingale and G. R. Bamnote

**Abstract** Most of the visual information presented to humans doesn't have a description associated with it, but humans can generally understand the visual information without any detailed description accompanying them. But for a machine to perform the same task of interpreting the same images and generate description in a Natural language form is a hard task. For a machine, the process involves two basic human tasks of seeing things and understanding them. In computing, these problems come under the domains of Natural Language Processing and Computer Vision. Therefore creating descriptions of the Image we need a combination of both tasks to be performed. The image description process consists of two parts: first part is to detect the important objects, features attributes, and the relationships between them in an image that comes under the field of Computer Vision, and then it should also generate semantically and syntactically accurate phrases, sentences which the area of Natural Language Processing. As this task of Describing images with human-readable language is a cognitive task the recent progress in area of Machine and Deep learning which are the subset of Artificial Intelligence in which computer algorithms try to copy and adapt the workings of the human brain in processing data and has accelerated the study of this challenging problem of image description. In this paper, the basic techniques and ways to do image captioning are discussed.

**Keywords** Artificial intelligence · Image description · Machine learning · Deep learning · Image captioning

## 1 Introduction

For a Human mind it is very simple to carry out a wide range of simple tasks or processes that require complex visual recognition and understanding of the activity, and also for us as a species it is also easy to communicate with each other through Language. Hence many times Humans are performing a combination of these two

---

S. P. Ingale (✉) · G. R. Bamnote  
Prof Ram Meghe Institute of Technology and Research, Badnera, India  
e-mail: [sumedh3003@gmail.com](mailto:sumedh3003@gmail.com)

tasks of Visual Recognition and Communication in Natural language without giving much thought about it. For example, a simple glance at an image/scene is enough for most people to describe a significant amount of details present in a given image/scene.

It is an old English proverb that a picture is worth of thousand words. In our daily life, people come across a lot of visual information from many sources like the Internet, document, diagrams, news articles, and advertisements. Even though most of this visual information contains no description information written/associated with them, the human mind is intelligent enough to grasp the meaning conveyed by the images in its own way [1]. For example, most images on the internet don't have any written description with it, but people can generally understand and interpret them without having any trouble.

Easy it may be for lots of people but for Computers to perform this process is a hard task. For a Machine to be able to describe an image in well-structured English sentences is a demanding task, but the result will be having a great impact in many fields, such as providing help to blind people to have a better understanding and interpretation of content in images [2].

The tasks mentioned above require an understanding of both image and language processing. One can view this problem as an area of research in the field of artificial intelligence as basically, and we are trying to perform a task that tries to replicate human intelligence [3].

## ***1.1 Literature Review***

Understanding an image and conceptualizing the language description for it has wide applications in Image captioning: a prominent field in Artificial Intelligence (AI). To detect an object by recognizing the scenery, its location, physical characteristics, and their connections are the prerequisite in understanding an Image, whereas developing full-phrased sentences involves a clear understanding of the basic language, both syntactically and semantically [3].

Traditional machine learning practices use features which are handcrafted such as the scale-invariant feature transform, histogram of oriented gradients, local binary patterns, and their various combinations. The technique involves extraction of object features from input data and output is then processed using a classifier mechanism like Support Vector Machines [1]. The task-specific orientation of handcrafted features makes extraction of different features unrealistic/impractical using diverse and extensive datasets. Moreover, the actual images and video contents tricky in nature with different semantic interpretations. Contrary to that in deep learning mechanism large and intensive dataset can be handled precisely by examining training data to automatically identify the features and their properties. For instance, convolutional neural network is commonly used feature-learning algorithm and 'Softmax' is frequently used as the classifier [4]. Recurrent Neural Networks (RNN) is used after the CNN to generate captions for images.

The extensive research on image captioning identifies that machine and deep machine learning-based algorithms are able to deal with complexities and challenges presented in the process of image captioning and hence proves a good choice. Many researchers have reviewed various image captioning process template-based mechanisms, retrieval-based mechanisms, and some deep learning mechanisms in [5, 6]. With the advancement in technology, computing power and availability of extensive data have shifted the paradigm towards deep learning-based image captioning. This paper gives a brief survey of image captioning processes with artificial intelligence-based deep learning approaches. The survey/review process is divided into 4 Sections. Section 2 provides an overview of some dataset used to model. Various image captioning models are discussed in Sect. 3 and Sect. 4 on deep learning-based image captioning techniques.

## 2 Datasets

This section introduces the open-source datasets. It is observed that a decent database can help make the model or algorithm more robust and efficient. The image description or captioning process is analogous to machine translation. As in a machine learning approach, we require datasets for building model and to train these models good datasets are very important these are some of the datasets given below (Table 1).

### 2.1 MSCOCO

This dataset was established by Microsoft team to understand targeted scenes from daily complex scene to perform tasks like image segmentation, detection, description, and recognition. In this dataset, 82,783 images are available for training, 40,775 images are available for testing, and 40,504 images are available for validation.

**Table 1** Description of a number of images in each dataset

Dataset name	Size		
	Train	Test	Valid
AIC	210,000	30,000	30,000
MSCOCO	82,783	40,775	40,504
STAIR	82,783	40,775	40,504
Filckr30k	28,000	1000	1000
Filckr8k	6000	1000	1000

## 2.2 *Flickr30k/Flickr8k*

The pictures available in this dataset were taken from Flickr website. The Flickr8k pictures dataset contains 1000 pictures for validation, 6000 pictures for training, and 8000 total number of pictures. The Flickr30k dataset was also taken from Flickr website and contains total of 31,783 pictures, 1000 each for testing and validation, and 28,000 pictures for training.

## 2.3 *AI Challenger*

AIC is first largest Chinese database obtained from the AI Challenger in the domain of image description generation. This database contains 30,000 images each for testing and validation and 210,000 images for training. AIC is analogous to MSCOCO and each image is conveyed by five descriptions, which describe image information, scenes, main characters, actions, and other contents.

## 2.4 *STAIR Captions*

This dataset is the largest Japanese image database, which is developed based on MSCOCO dataset. It consists of 820,310 Japanese descriptions (each image corresponds to five descriptions) and 164,062 images.

# 3 Image Caption Approaches

Image caption models are broadly classified into.

- (1) Template-based Models
- (2) Retrieval-based Models
- (3) Novel caption generation.

Template-based techniques use fixed templates which are associated with certain number of blank places and to produce captions these blanks are filled according to objects attributes or action that can be detected from the image.

In Retrieval-based technique's and models, descriptions are obtained from a pool of already existing description/caption dataset, in the retrieval approach first find images with caption which are visually similar to a given the training data set. Then these captions from these similar images are called candidate captions. Both Template-based and Retrieval-based approaches lag behind in creating image specific description for a particular image.



In novel captions generation, the general approach is to first analyze the content attributes features present in the image and then create image description from the information obtained from image by using a language model. These methods can create good captions for the images, most of the novel caption generation methods are using on deep learning neural networks.

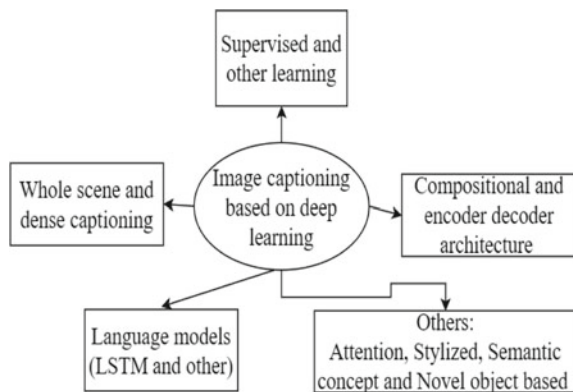
### 3.1 Deep Learning Features with Neural Network

Deep learning methods are good due to the recurrent neural networks (RNN) which have wide range of applications in deep learning domain. RNN method is implemented primarily to process natural language. The model uses Convolutional neural network attached to encode the features from various connected layers of image and generate the description for input image using language model based on RNN.

## 4 Methods

In this section, the methods are grouped according to their characteristics feature which are (i) Supervised and other deep learning-based captioning, (ii) Whole scene based and Dense captioning (iii) Compositional Architecture-based and Encoder-Decoder Architecture-based, also based on the language model used as (iv) LSTM (Long Short-Term Memory) language model based and others language model based. The captioning methods can be further classified as (v) Attention-Based methods, Stylized captions methods, Semantic concept-based methods, and Novel object-based methods. Figure 1 represents an overall classification of image captioning based on deep learning [1].

**Fig. 1** A classification of deep learning-based picture captioning techniques



## **4.1 Supervised Learning Models and Other Deep Learning**

The first way of grouping captioning method can be done on the basis of the data used for Training. The training outcome in supervised learning is based on labeled data. But due to vast amount of data generated every day, it's not always practical to annotate labels to data every time hence other unsupervised learning such as Generative Adversarial Networks (GAN) [7] has to work with unlabeled data. Another popular approach: Reinforcement learning is used to explore, identify, label, and allocate a signal. Many studies reported GAN-based approaches in reinforcement learning to caption an image and are largely categorized into "Other Deep Learning" mechanisms.

### **4.1.1 Supervised Learning-Based Image Captioning**

Many studies in past have productively used supervised learning-based networks for image captioning process that includes classification of images [8–11], object detection [12–14], and learning from the attributes. The higher efficiency in output deliverables have promoted supervised learning-based networks to employ in automatic image captioning process [15–18].

### **4.1.2 Other Deep Learning-Based Image Captioning**

As we know terabytes data is accumulating with each passing second, most of the data which is generated is unlabeled so annotating all the collected data will be impractical. Therefore, machine learning based on reinforcement and/or unsupervised techniques for image captioning are gaining more importance nowadays. The reinforcement learning is particular about choosing action, receiving reward values, and then processes to next state. Here, the agent decides the action with presumption of extensive and persistent obtainable rewards. Thus ascertaining the value function requires continuous state and action information which is a lacuna faced in traditional reinforcement learning approaches. The policy gradient approach type of reinforcement learning can adopt unambiguous policy for precise action [19]. The domain knowledge can be incorporated into policy for specific actions that come with guaranteed convergence.

Steps in these techniques are as follows:

1. A CNN and RNN-based combination network to generate captions/descriptions.
2. Another CNN-RNN network is used to evaluate the generated captions and then send feedback to the first network for creating good captions.

GAN-based approach is the most suitable method to examine unlabeled data and learn the features. The method employs a generator and a discriminator as a pair of networks for data analysis and learning. GAN has proved a powerful tool in

various domains of data learning such as image captioning [20, 21], image to image translation [22], text to image synthesis [23]. The GAN method of captioning as proposed by Dai et al. [20] involves only one set of image classification. However, GAN approach can deliver a wide classification range for image compared to conventional methods. Shetty et al. [21] used multi-captioning GAN approach and produced different captioning options to choose from.

## ***4.2 Dense Captioning and Whole Scene Captions***

Most of the time visual data such as images are rich with large amount of content present in their captioning methods and be grouped into the above two groups and they are explained below.

### **4.2.1 Dense Captioning**

Many available image captioning techniques can assign only single caption for the input image as a whole. But the different areas of image are also important which should be analyzed to gather the information on various objects. Hence, the main idea behind dense captioning is to describe various object attributes and actions present in image.

Steps in these techniques are as follows:

1. Region proposals are created for the different areas of the image.
2. CNN is used to get the area-based image features/attributes.
3. Outputs from second step are supplied to a language model for generating captions for every area of the image.

### **4.2.2 Captions for the Whole Scene**

In caption for whole scene method single or multiple captions are generated which tries to describe the whole scene rather than creating multiple captions for different parts of particular image. Many of the next architecture tried to address this approach.

### **4.2.3 Encoder-Decoder Architecture-Based Image Captioning**

Encoder decoder-based captioning techniques are same as that of the encoder-decoder framework-based in neural machine translation [24]. The architecture works in simple end-to-end manner to provide image captions.

In this process, neural network which is most of the time a CNN model is used to extract all features, attributes from image, then these extracted features and attributes are given to an LSTM model to produce a sequence of words for caption.

Steps in these techniques are as follows:

- (1) A CNN for obtaining the type of scene, to detect the object's attributes and features.
- (2) The output of first step is given to a language model to generate them into words, and phrases to generate image captions.

#### 4.2.4 Compositional Architecture-Based Image Captioning

Compositional architecture-based methods consist of more than one independent functional block. CNN is needed for extracting semantic concepts from the image. A language model is needed for generating a set of candidate captions. Then while generating the final caption, the candidate captions are rearranged to be re-ranked by deep multimodal similarity model.

Steps in these techniques are as follows:

1. Using a CNN image features are extracted.
2. Visual features are used to obtain Visual concepts.
3. Multiple captions are generated by using a language model using the output of the first and second steps.
4. Generated captions are re-ranked by using a deep multimodal similarity model to select good image caption [1].

### 4.3 LSTM and Others

Image captioning process bridges natural language processing (NLP) with computer vision. NLP process formulates learning with sequential manner. Many neural models of language learning follow sequence-to-sequence tasks, e.g., neural probabilistic language model [25], skip-gram models [26], recurrent neural networks (RNNs) [27], and log-bilinear models [28]. RNN have extensive scope in rigorous sequence learning. However, it has its own limitations in dealing with long-term temporal dependencies and liable to vanishing and exploding gradient problems. LSTM, on other hand, does not follow the structural hierarchy of sentence formation. Further, LSTM demands significant storage space to deal with long-term dependency over memory cell. CNN has advantage of learning the internal structural hierarchy of sentences with fast processing speed. Therefore, recently, many sequence-to-sequence tasks are being developed using convolutional architecture. Ass conditional image generation [29] and machine translation [30–32]. Examining the success ratio of CNN in sequence learning tasks, Gu et al. [33] put forth a CNN language model-based image captioning method that uses language-CNN as statistical language modeling tool. However, the model has limitation to process dynamic temporal behavior of the language using language-CNN and need to be coped with a recurrent network to address the temporal dependencies. Another model proposed by Wang et al. [34] is based on CNN + CNN processing-based image captioning which

connects the language-CNN to vision-CNN using a hierarchical attention module. The authors investigated several hyper-parameters, layers quantity, and kernel width of the language-CNN, and found that the hyper-parameters exhibit positive influence on the modeling process and improve the performance.

The image captioning methods can also be classified according to various approaches used like attention-based Image captioning which focuses on various areas of attention in image to generate image caption, other approach is semantic concept-based image captioning which selectively addresses to set of semantic concepts from and generate captions. Next method is Novel object-based image captioning methods. These methods depend upon image and caption database for results. Last approach can be stylized caption in which rather than creating flat factual descriptions of image an emotional-based description is created using a training data [1].

## 5 Conclusion

This paper presents a brief review of various image captioning methods. The classification of various techniques used for image captioning are mentioned and explained according to their approaches along with the generic steps involved in them, high lightening the pros and cons. We briefly outlined various research directions present in this area of image captioning. Although machine and deep learning-based image captioning methods have obtained a remarkable success in past years, a good robust image captioning model that will produce good quality captions for all types of images is yet to be achieved. Therefore, with the advancement of newer artificial Intelligence architectures and increase in data generation and processing power, Image captioning will be an active study subject for some time in the foreseeable future.

## References

1. Zakir Hossain MD, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. *ACM Comput Surv* 51(6). <https://doi.org/10.1145/3295748>
2. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: *Proceedings of the IEEE Computer Society Conference On Computer Vision And Pattern Recognition*. <https://doi.org/10.1109/CVPR.2015.7298935>
3. Vinyals O, Toshev A, Bengio S, Erhan D (2017) Show and tell: lessons learned from the 2015 MSCOCO image captioning challenge. *IEEE Trans Pattern Anal Mach Intell* 39(4):652–663. <https://doi.org/10.1109/TPAMI.2016.2587640>
4. LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324
5. Bai S, An S (2018) A survey on automatic image caption generation. *Neurocomputing*

6. Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B et al (2016) Automatic description generation from images: a survey of models, datasets, and evaluation measures. *J Artif Intell Res (JAIR)* 55:409–442
7. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: *Advances in neural information processing systems*, pp 2672–2680
8. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 770–778
9. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems*, pp 1097–1105
10. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: *International conference on learning representations (ICLR)*
11. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1–9
12. Girshick R (2015) Fast r-cnn. In: *Proceedings of the IEEE international conference on computer vision*, pp 1440–1448
13. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 580–587
14. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, pp 91–99
15. Chen X, Lawrence Zitnick C (2015) Mind’s eye: a recurrent visual representation for image caption generation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 2422–2431
16. Karpathy A, Fei-Fei L (2015) Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 3128–3137
17. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2015) Deep captioning with multimodal recurrent neural networks (m-rnn). In: *International conference on learning representations (ICLR)*
18. Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 3156–3164
19. Sutton RS, McAllester D, Singh S, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. In: *Advances in neural information processing systems*, pp 1057–1063
20. Dai B, Fidler S, Urtasun R, Lin D (2017) Towards diverse and natural image descriptions via a conditional GAN. In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp 2989–2998
21. Shetty R, Rohrbach M, Anne Hendricks L, Fritz M, Schiele B (2017) Speaking the same language: matching machine to human captions by adversarial training. In: *IEEE international conference on computer vision (ICCV)*, pp 4155–4164
22. Isola P, Zhu JY, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: *Proceedings of the IEEE international conference on computer vision (CVPR)*, pp 5967–5976
23. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis. In *Proc Mach Learn Res* 48:1060–1069
24. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: *Advances in neural information processing systems*, pp 3104–3112
25. Bengio Y, Ducharme R, Vincent P (2003) A neural probabilistic language model. *J Mach Learn Res* 3:1137–1155
26. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)

27. Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S (2010) Recurrent neural network based language model. In: Eleventh annual conference of the international speech communication association
28. Mnih A, Hinton G Three new graphical models for statistical language modelling. In: Proceedings of the 24th international conference on Machine learning. ACM, pp 641–648
29. Van den Oord A, Kalchbrenner N, Espeholt L, Vinyals O, Graves A et al (2016) Conditional image generation with pixelcnn decoders. In: Advances in neural information processing systems, pp 4790–4798
30. Gehring J, Auli M, Grangier D, Dauphin YN (2016) A convolutional encoder model for neural machine translation. [arXiv:1611.02344](https://arxiv.org/abs/1611.02344)
31. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence tosequence learning. [arXiv:1705.03122](https://arxiv.org/abs/1705.03122)
32. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
33. Gu J, Wang G, Cai J, Chen T (2017) An empirical study of language cnn for image captioning. In: Proceedings of the International Conference on Computer Vision (ICCV), pp 1231–1240
34. Wang Q, Chan AB (2018) CNN+ CNN: convolutional decoders for image captioning. [arXiv:1805.09019](https://arxiv.org/abs/1805.09019)

# Composite Reversible Data Hiding Scheme for Secure Image Reconstruction



Nandni Tandon and Abhishek Sharma

**Abstract** Intense growth of information technologies has ameliorated the means of access to digital information. Digital image processing manages the digital content to store, share more efficiently with lesser time and space complexity. However, these approaches beach the privacy of digital content. Recent research focuses on protecting digital content from illegal use and distribution by hiding reversible data scheme to handle the risk of privacy violations for digital content. In this paper, a Composite Reversible data hiding (CRDH) scheme is proposed. CRDH used the Integer Wavelet Transform (HAAR transform) with the HH band's Eigenvalue decomposition. The proposed CRDH initially applied the IWT transformation on the cover image (CI) and parsed it into four subsequent frequency sub-bands namely LL, HL, LH, and HH. Confidential data of the proposed scheme are embedded by combining the HH band of the cover image's individual values with the encrypted Eigenvalues of the confidential data. The selection of casing art is such a manner where values lie within a range. The Confidential data image and HH band's frequency band are approximately identical; therefore, changing the individual values significantly does not trigger the quality of the confidential data image and the HH band's content. The proposed scheme's main objective is to develop a data hiding scheme that prevents the verification of digital information by maintaining a high rate of PSNR. The PSNR of the current technology is less than 50% for the entire available data set. The PSNR value indicates the image's visual quality, where the PSNR improves the higher image quality, so hiding data is necessary for the scheme that prevents authentication and maintaining a high rate of PSNR. The proposed method achieves this goal, gains a PSNR rate of over 50%, and reaches 57% for Tiffany and high SSIM which represent the quality of image.

---

N. Tandon (✉)

Sagar Institute of Research & Technology, Bhopal, India

e-mail: [nandni456tandon@gmail.com](mailto:nandni456tandon@gmail.com)

A. Sharma

Oriental Institute of Science & Technology, Bhopal, India



**Keywords** Digital image processing · Integer wavelet transform · HAAR transform · Reversible data hiding · Secure image reconstruction · PSNR ratio · MSE

## 1 Introduction

With the global expansion of the Internet, the Internet has become a powerful and accessible platform for data transmission. Unfortunately, Sensitive information can be stealing and temper on the Internet. In recent times, information security has attracted the focus of the researcher as a data hiding scheme.

Data hiding [1] plays a vital role in Internet and multimedia-based secure communication. The primary purpose of data hiding is to secure the confidentiality of the message and share the sensitive news safely. If there is any splash in the confidential message or its media during data transmission, the secret message cannot be reorganized entirely after receiving it. Data hiding scheme for sharing confidential data in intelligence burro, para-military, military, and company financial report would have played a significant role in recent times and led to focus research in reversible data hiding scheme in recent years [1].

The reversible data hiding [2, 3] uses images as an input tool because of their easy access. Images can be downloaded with a scanner, digital camera, or directly online. Depending on the encryption method, the most recent algorithm data can be divided into three areas: locations, currents, and pressure points. Algorithms in the field section include encryption by changing the pixel value directly. However, algorithms in the network area begin to convert the image to coefficient [4–6]. Then the coefficient changes include individual messages. Pressure algorithms accept images generated by a series of compressed codes as their intervening medium. Inclusive details are achieved by modifying computer code [7–10].

This paper presents a Composite reversible data hiding (CRDH) scheme that employed Integer Wavelet Transform (HAAR transformation) with Eigen decomposition over both cover and confidential data image. The proposed data hiding scheme significantly improved the extracted hidden data measure as MSE and PSNR ratio and acquired higher PSNR and Lower MSE.

The rest of the paper is organized as follows: Sect. 2 focuses on recent reversible data hiding scheme. Section 3 presents description of the proposed framework for reversible data hiding scheme, successively with sender side and receiver side at Sects. 4.1, 4.2 and 5 includes implementation scenario detail information about image data set and performance evaluation. And finally, Sect. 6 concludes the paper, with outcomes of CRDH framework performance and research outcome.

## 2 Reversible Data Hiding

With the rapid advancement of communication through the Internet, the information exchanged could tamper intentionally or accidentally through unprivileged access. In recent years, reversible data hiding (RDH) has become an active research domain in data replication. In reversible data hiding, the bits hidden are embedded in the cover file (image) on the sender side. Confidential data and original cover media are extracted without distortion toward the receiver. RDH is also known as invertible or lossless data hiding. The data hiding technique is divided into immutable data hiding and reversible data hiding. The embedding ability to hide immutable data is high. However, while the original cover media is destroyed, the embedding capacity in RDH is low, but the original cover media can be recovered.

Some of the present data hiding schemes are not invertible; the presence of truncation error and round-off error in spread spectrum technique makes it non-reversible, because of the bit replacement without memory, and the least significant bit plane scheme is taken into account to be non-reversible and because of the quantization error quantization-index-modulation rendered as invertible. RDH also links two data sets, and a group belongs to the embedded information. Another set belongs to the cover media data such that the cover media will be losslessly recovered once the hidden information has been extracted out. Confidential data is embedded by modifying the frequency coefficient of the cover image by using some standard method like discrete Fourier Transform (DFT), Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT), and Integer Wavelet Transform, etc.

## 3 Literature Survey

Ahmad shaik, V. thanikaiselvan proposed the scheme of threshold-based histogram modification technique. Where the data is embedded in High-frequency sub-bands using IWT scheme. Secret data is hidden in the detail sub-band. Histogram of high-frequency sub-bands has been shifted with predefine threshold range to embed the secret data. Flag array method is used to overcome overflow and underflow condition. To obtain the input image inverse IWT is applied on approximation sub-band [1].

Amishi Mahesh Kapadia and Nithyanandam Pandian proposed the IWT-based lifting scheme. The secret data is converted to binary data before embedding. Once the embedding is done then the inverse IWT is applied to obtain the stego image. Two-level integer wavelet transform is performed in this paper. In extraction process, IWT applied on the stego image to extract the binary bits of secret data [2].

Mr. Nitin Laxman Shelake, Prof. Santosh R. Durugkar proposed an algorithm extracting hidden data from encrypted images using IWT. Apply IWT to divide the image into sub-bands. Calculate the average value of each pixel and then apply the reversible wavelet transform and generate the histogram of reversible wavelet transform image. Histogram is used to prevent the condition of overflow and underflow.

In this paper, compress the data for embedding. After embedding, inverse transform is applied to obtain marked image [3].

Guangyao Ma, Jianjun Wang proposed efficient reversible data hiding in encrypted images based on multi-stage integer wavelet transform. This paper presents the reversible data hiding for encrypted images using integer transform. The cover image is encrypted by encryption key. The IWT is applied on the encrypted image, which decomposed the image into four frequency sub-bands, i.e., LL, LH, HL, and HH. The preprocessing is applied on the high-frequency sub-bands. In the preprocessing, the location map is generated of high-frequency band for compression. The book-keeping data comprises with location map of high-frequency sub-bands. In this paper, the multilevel embedding is approached for high embedding capacity [4].

Bing Chena, Xiaotian Wub, c, d, Wei Lua, Honglin Rena proposed Reversible data hiding in encrypted images with additive and multiplicative public-key homomorphism. RDH-EI proposed scheme is implemented with two public key for encryption. Three parties are involved in this paper, i.e., content owner, data hider (provider), and receiver. Content owner utilized public key to the encrypt image of original image. Provider embedded the encrypt the secret data into the encrypt image of original image. Receiver generates the public and private keys by using homomorphism technique to extract the secret data and decrypt the secret data [5]. DNA cryptography is a method that encrypts the data using data encoding technique. DNA sequence consists of four different basic nucleotides, i.e., A, C, G, T. the pair is allowed between two A and T and C and G only. On the bases of these four, eight types of bits are presented in DNA sequence. XOR operation is performed on DNA sequence. DNA encoding is used to extract key for encryption of the secret data ad select the pixel in frequency band for data hiding. IWT is applied to get LL sub-band [6].

Block-DCT scheme proposed the prepressing technique that is used on cover image for preparing embedding process. Cover image is divided into  $8 \times 8$  non-overlapping block. Each block is transformed into frequency domain for quantization. Secret data pixel converted into base 8 for embedding in cover image pixel. After embedding IDCT applies to get the stego image [7].

Steganography technique has proposed two methods—encoding and decoding process. It also extracts the secret data and recovers the cover image from stego image. IWT-based steganography provides better visual quality than DWT-based digital steganography. Scrambling based on Arnold transformation is applied on secret data to generate private key for more security and robustness [8].

YCbCr format is better for hiding. Iwt is applied on Cb component to obtain four frequency sub-bands. Data has been hidden into the approximation sub-bands. Two bits of approximation coefficient of audio signal is XOR with the two bits of Cb component of cover image to get encrypted stego image. This encryption avoids the encryption key. In extraction process, RGB stego image is converted into YCbCr. Inverse IWT is applied to get approximation sub-bands of Cb component of stego image. Again apply the IWT to extract the audio file [9].

Data is hiding LSB of high-frequency band of CDF (Cohen-Daubechie- Faurae) (2, 2) proposed technique. Histogram modification is used to overcome overflow/underflow condition. Threshold embedding scheme is used to embed the

payload into high frequency using IWT coefficient with predefined threshold value and by replacing LSB bits. Stego image is generated from the inverse integer wavelet transform [10].

J K Mandal and S das proposed an information Hiding Scheme in wavelet Domain using Chaos Dynamic. The source image is converted into transform domain by using Daubechchies DWT algorithm. The secret data is embedded by using LSB bit stream after transform data is multiplied with value of pie. Secret data is encrypted using the same random bit stream generated by using key  $(\alpha, x_0, y_0)$ . After embedding inverse DWT transform is applied to generate stego image [11].

### 4 Proposed Method

This paper presents a Composite reversible data hiding (CRDH) scheme. CRDH competitively applied Integer Wavelet Transform (HAAR transformation) with Eigen decomposition over the cover image's cover image and brand of the cover image. CRDH decomposed the cover image (CI) into four sub-bands, namely LL, HL, LH, and HH, and evaluated the HH band's Eigen decomposition value for data hiding (Fig. 1).

In the proposed reversible data hiding scheme, Integer Wavelet transforms the cover image (CI) by decomposing, namely, into lower-lower (LL), higher-lower (HL), lower-higher (LH), and higher-higher (HH) frequency sub-band as shown in Fig. 2.

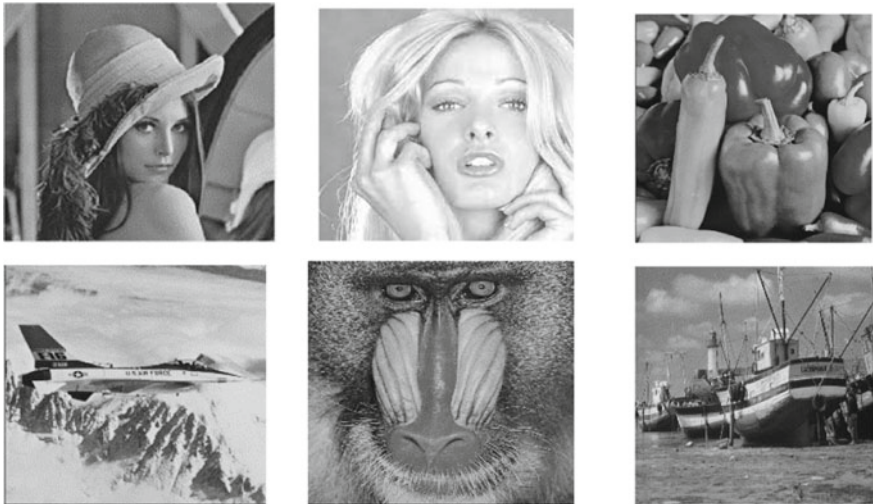
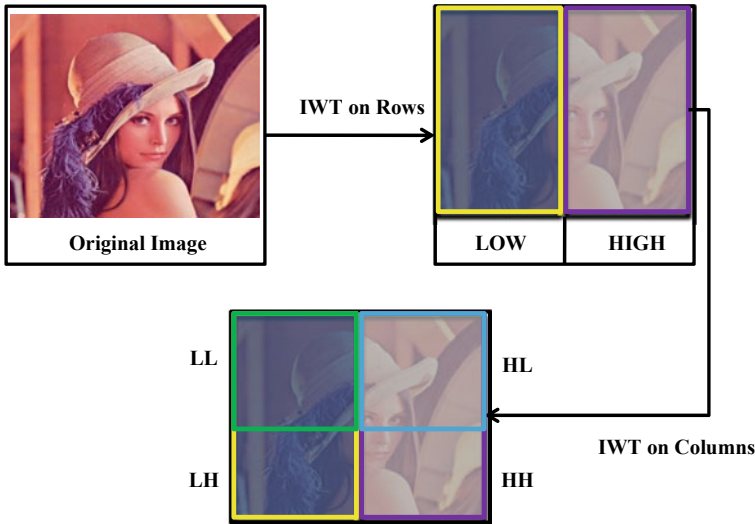


Fig. 1 Input images are Lena, Tiffney, fruit, plane, boat, and baboon



**Fig. 2** IWT transform of cover image

In CRDH algorithm, the hidden data is integrated into the host image by changing the band of high-frequency coefficients described in Sect. 3, that is, the HH sub-band. As shown in Figs. 3 and 4, the proposed reversible data hiding scheme is based on both Integer Wavelet Transform and Eigendecomposition with encryption. Initially, Integer Wavelet Transform divides the host image into four frequency subfields as LL, HL, LH, and HH bands. The LL tape works with rough details, the HL band deals with horizontal elements, the LH provides vertical information, and the HH band contains diagonal image information. The HH band uses the HH band to embed sensitive data because you lose data on the image energy. In this way, an embedded confidential data will not affect the perception accuracy of the cover image. The proposed reversible data hiding scheme has two steps for hiding and sending, and receiving as extraction steps.

## 5 Hiding Phase of Composite Reversible Data Hiding (CRDH) Scheme

Once the cover image decomposed into four subsequent frequency sub-band, CRDH uses the HH band to hide confidential data (CDI), as its accounted minimum noise level. The proposed data hiding scheme evaluates the Eigendecomposition value of HH band confidential data followed by encryption. Encrypted Eigendecomposition then superimposes over HH band of IWT (CI) after applying inverse Eigendecomposition operation. Finally, apply inverse Integer Wavelet Transform to generate an embedded cover image (ECI).

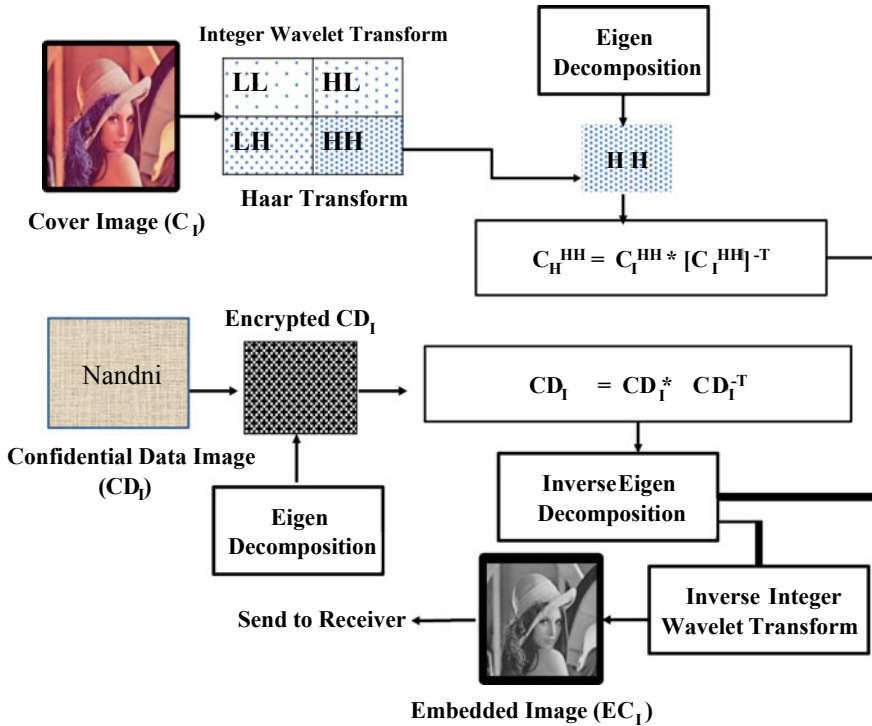


Fig. 3 Reversible data hiding embedding procedure

The extracting phase of the proposed scheme uses exactly the inverse operation of the hiding phase; in Extracting, the proposed scheme is used to apply Integer Wavelet Transform over the embedded image (ECI) to get the HH band confidential data (CDI) is hidden, as shown in Fig. 4.

CRHD scheme enforces Eigendecomposition on HH frequency band of received data. The retrieved Eigendecomposition value compositely contains Eigendecomposition on the cover image’s HH frequency band and confidential data.

As the image dimension of covered data is already shared among sender and receiver before established data communication, extracting phase compared shared cover image with received image. The received image contains both the cover image and confidential data. The extraction step’s subtracting returns the eigenvalue of encrypted confidential data and subsequently applied all inverse operation to get confidential data at the receiver side.

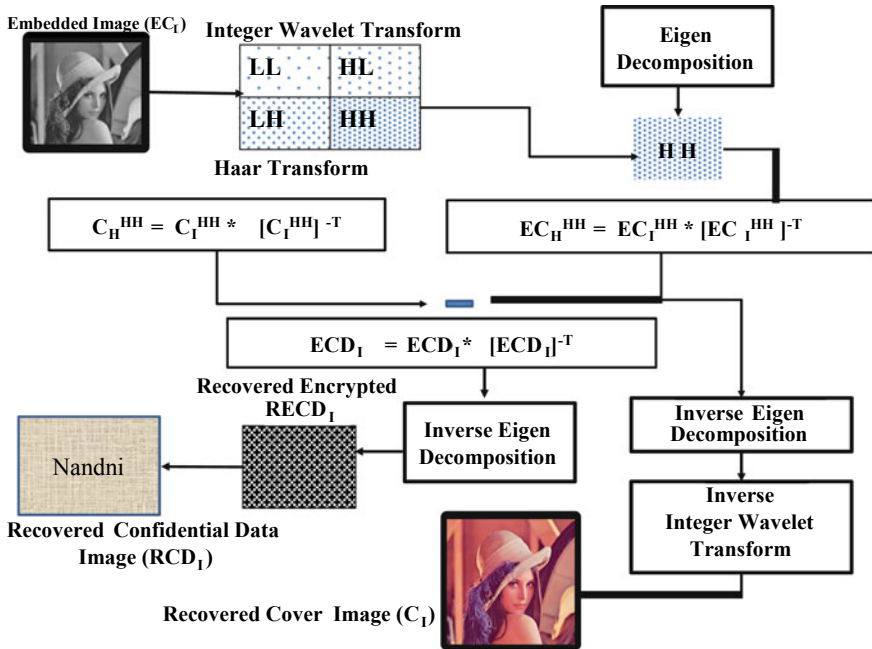


Fig. 4 Reversible data hiding extraction procedureextraction phase of composite reversible data hiding (CRDH) scheme

## 6 Result Analysis

The proposed works have tested on various data set image of size  $512 \times 512$ . These entire data set image are grayscale. Here data set image is used as Fruits, Elaine, LENA, and Tiffany. The size of the confidential image is also the same as the original image. To simulate the proposed work, the implementation is done in MATLAB. The i3 processor is executed with 4 GB RAM and 500 GB HDD.

## 7 Performance Parameter

PSNR is an image factor used to determine the quality of an image by comparing quality differences between the original image and the resulting image.

It is calculated using the mean square error (MSE). The following formula calculates both parameters.

$$PSNR = \log_{10} \left( \frac{MAX^2}{MSE} \right)$$

$$MSE = \frac{\sum_{i,j} [I_1(i, j) - I_2(i, j)]^2}{i * j}$$

SSIM (structural similarity index ratio) represents similarity between cover and embedded image. The value ranges of SSIM between [0, 1], where 0 represents least similarity and 1 represents high similarity.

$$SSIM = \frac{(2\mu_l\mu_r + k_1)(2\sigma_{l,r} + k_2)}{(\mu_l^2 + \mu_r^2 + k_1)(\sigma_l^2 + \sigma_r^2 + k_2)}$$

where  $\mu_l, \mu_r, \sigma_l^2, \sigma_r^2$  means ad variance of cover and embedded image are  $k_2, k_1$  are two constants. This calculation has been gathered by the computer program. There is various testing images that have been used. Some of them have been shown here with their results.

Figure 5 shows the PSNR ratio, as the graph shows that there are some images used as input. This input and the generated output image have been used for calculating the PSNR. Graph also shows the proposed approach which shows the better results (Table 1).

As shown in Table 2, PSNR of existing technique for all the available data set described in Fig. 5 is less than 50%. The PSNR value indicates the image’s visual quality, where a higher PSNR value leads to better image quality. It is needed to develop a data hiding scheme that prevents authentication of digital information with maintaining a higher PSNR ratio. The proposed scheme significantly achieves this goal and gains PSNR ratio greater than 50 and achieves up to 57% for Lena.

As shown in Table 3 SSIM of existing technique for all available data sets is less than the proposed method. High SSIM value shows maximum similarity.

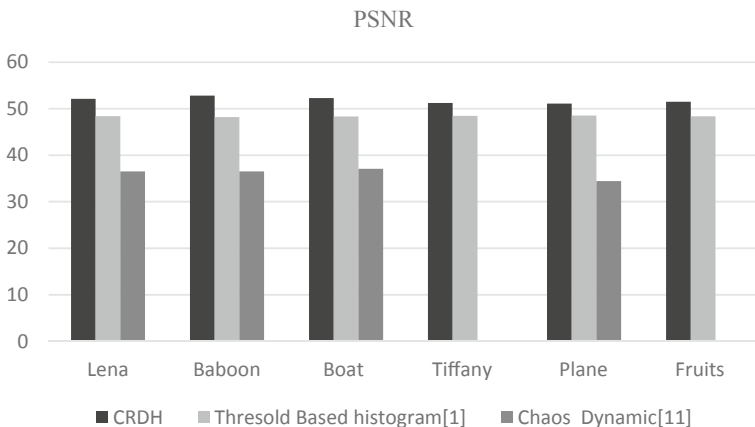


Fig. 5 PSNR proposed result is compared with Ahmad shaik et al. [7] and Mandal et al. [10]



**Table 1** Mean square error

Images	MSE (CRDH)
Fruits	0.36757
Lena	0.36765
Tiffany	0.243832
Boat	0.3823
Plane	0.41597
Baboon	0.32894

**Table 2** Comparative analysis PSNR with [7, 8]

Images	Chaos dynamics technique [11]	Threshold based histogram scheme [1]	Proposed scheme(CRDH)
Fruits	–	48.36	48.819
Lena	36.54	48.39	48.9611
Tiffany	–	48.45	57.8560
Boat	37.08	48.33	52.001
Plane	34.44	48.53	49.677
Baboon	36.54	48.21	55.797

**Table 3** Comparative analysis of SSIM with Ahmad shaik et al. [5] and Mandal et al. [6]

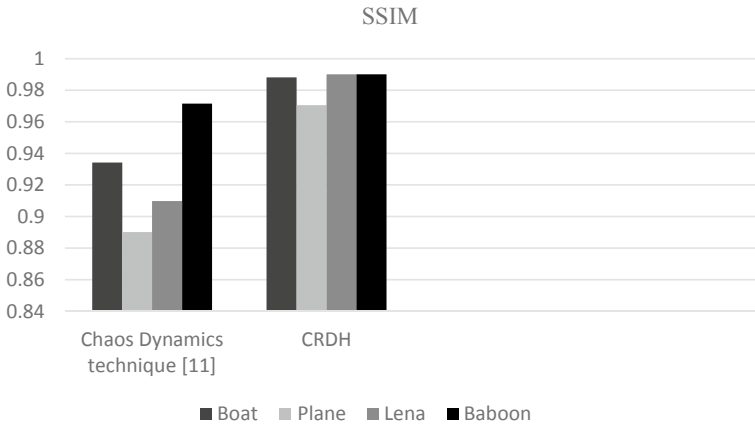
Images	Chaos dynamics technique [11]	Threshold based histogram scheme [1]	Proposed scheme(CRDH)
Boat	0.9342	0.9992	0.9889
Plane	0.8901	0.9984	0.9921
Lena	0.9098	0.9985	0.9900
Baboon	0.9715	0.9993	0.9977

Maximum similarity increases the imperceptibility of secret data. The proposed scheme achieved SSIM value of nearly to 1 which increases imperceptibility.

Figure 6 shows the SSIM, as the graph shows that there are some images which has been used as input. SSIM graph also shows that the proposed approach shows the better results.

## 8 Conclusion

Confidential data exchange is virtually insecure in this era of wireless communication. Intentional or unintentional changes to the transmitted data are possible. Recently researchers focused on hiding confidential data within the covered image



**Fig. 6** SSIM proposed result is compared with Mandal et al. [11]

via a reversible data hiding scheme. But data hiding scheme still faces challenges to extract distortion less and noise-free confidential data at the receiver side. This paper presents a Composite reversible data hiding (CRDH) scheme. CRDH competitively applied Integer Wavelet Transform (HAAR transformation) with Eigendecomposition over the cover image and confidential data image. CRDH decomposed the cover image (CI) into four sub-bands, namely LL, HL, LH, and HH, and evaluated the HH band’s Eigendecomposition value for data hiding. The proposed data hiding scheme significantly improved the extracted confidential data measure as MSE and PSNR ratio and acquired higher PSNR (up to 57%) and Lower MSE.

## References

1. Shaik A, Thanikaiselvan V (2018) Comparative analysis of integer wavelet transforms in reversible data hiding using threshold based histogram modification. *J King Saud Univ Comput Informat Sci* 2–12.
2. Kapadia AM, Pandian N (2020) Reversible data hiding methods in integer wavelet transform. *Int J Informat Comput Security* 12(1):70–89
3. Shelake MN, Durugkar SR (2015) An algorithm extracting hidden data from encrypted images using IWT. *Int J Eng Res Technol (IJERT)*, 4:1408–1412
4. Ma G, Wang J (2019) Efficient reversible data hiding in encrypted images based on multi-stage integer wavelet transform. *Signal Process Image Commun* 75:55–63
5. Chen B, Wu X, Lu W, Ren H (2019) Reversible data hiding in encrypted images with additive and multiplicative public-key homomorphism. *Signal Process* 164:48–57
6. Firas A, Abdullatif A, Alaa A, Abdullatif A, Namar A, Taha (2020) Data hiding using integer lifting wavelet transform and DNA computing. *Periodicals Eng Nat Sci* 8(1):58–66
7. Koikara R, Goswami M (2015) A data hiding technique using Block-DCT. *Int J Eng Res Technol (IJERT)*, 5:81–85

8. Sehgal P, Sharma VK (2013) Performance improvement in discrete wavelet transform based digital image steganography by the use of integer wavelet transform. *Int J Eng Res Technol (IJERT)* 2:972–977
9. Punidha R, Sivaram M (2017) Integer wavelet transform based approach for high robustness of audio signal transmission. *Int J Pure Appl Math* 4(23):295–304
10. Xuan G, Shi YQ, Yang C, Zheng Y, Zou D, Chai P (2005) Lossless data hiding using integer wavelet transform and threshold embedding technique
11. Mandal JK, das S (2018) An information hiding scheme in wavelet domain using chaos dynamic, vol 77, pp 264–267

# Semantic Relation-Based Modularity-Optimized Community Detection Algorithm for Heterogeneous Networks



Rishank Rathore and Ravi Kumar Singh Pippal

**Abstract** This paper presents the semantic relation-based modularity-optimized community detection algorithm for Heterogeneous Networks. This paper aims to try to increase the modularity value of the network by using the content analysis of the network and the link analysis together. Therefore, similarity values between people's shares were calculated and included in the network as indirect links. The proposed content- and link-based methods are a greedy hierarchical clustering algorithm that uses indirect connections with the network structure and ensures that the nodes most relative to each other are topologically and semantically grouped with priority. This paper presents a comparative analysis to analyze the impact of semantic relation over the optimization algorithm, i.e. Parliamentary Optimization Algorithm (POA) and Modularity Optimization Algorithm (MOA) for community detection. Finally, the modularity and NMI of the proposed work were evaluated over six real network-based heterogeneous network data sets and gained a satisfactory modularity rate over the resultant informative community.

**Keywords** Social media · Community detection · Clustering · Optimization algorithm · Modularity · Semantic relation · Hierarchical clustering algorithm · Seed community

## 1 Introduction

Community detection in social networks is one of the most studied problems recently. Analyzing the relationships, hierarchy and common interests among people is vital for many fields such as security, marketing and sociology [1, 2].

Existing studies on the community recognition problem generally use link analysis approaches [3–6]. Since the community is defined as a set of similar entities in terms of specific characteristics, the said similarity measure varies according to the network. Classical methods generally use the topological features of the network

---

R. Rathore (✉) · R. K. S. Pippal  
Department of Computer Science, RKDF University, Bhopal, India  
e-mail: [rishank1989@gmail.com](mailto:rishank1989@gmail.com)

while separating the network into communities. The most important criterion on which the developed methods are based is the best possible fragmentation of the network into communities. For this reason, the modularity parameter is used to compare the advantages of the ways [7]. Different optimization-based algorithms are being developed to increase the modularity value.

Another approach used in community detection is grouping individuals according to their common interests with content analysis [8–11]. Individuals using social networks share different, meaningful and actionable data such as media shares, blogs and e-mails. It is thought that individuals who share similar data have common interests. Community detection of the network can be made using clustering methods according to the attributes obtained from the shares.

When both approaches are examined, it is seen that the methods that use the topological features of the network ignore the interests of the users. In contrast, the methods that focus on content analysis ignore the topological structure of the network [12]. By addressing this problem, two-stage community detection approaches are being developed [13]. After breaking down the network according to content analysis, two-stage methods apply link analysis for each obtained community.

The research objective of this paper is to try to increase the modularity value of the network by using the content analysis of the network and the link analysis together. Therefore, similarity values between people's shares were calculated and included in the network as indirect links. The developed method is a greedy hierarchical clustering algorithm that uses indirect connections with the network structure. The modularity optimization [7] algorithm, a modularity-based approach, has been reinterpreted based on the semantic similarities between the people in the network. The modularity optimization algorithm groups the nodes that are closest to each other topologically. On the other hand, our approach ensures that the nodes most relative to each other both topologically and semantically are grouped with priority. When the developed method is compared with the current work, it is seen that it produces communities with higher modularity. Experimental results have confirmed this.

In the second part of the article is present recent research on studies on community detection. In the third section, the semantic similarity calculation of the people in the network and the developed algorithm are given. In the fourth section, the test results will be presented. Finally, in the fifth section, the result will be explained.

## 2 Related Work

Social network nodes can be modeled with graph structures representing people and edges representing relationships between people. In topology-based community detection, networks are considered as graphs. However, the essential elements in the problem of dividing the charts into certain groups differ according to the applied methods. For this reason, first of all, the structure of the communities to be formed belonging to the graph should be defined. The most general approach is to ensure that the community's connection density in the community to be created is higher than the

connection density outside the community. Another method is to group close nodes by calculating the distance between pairs of nodes according to specific similarity criteria.

For this reason, some classical data clustering algorithms are used. But, unfortunately, many data clustering problems are NP-hard problems. For this reason, methods trying to approach the optimum solution are developed by using various approach algorithms with low complexity.

For the ensemble detection problem, grouping methods such as graph segmentation, spectral clustering and divisive algorithms such as Girvan Newman, modularity-based optimization algorithms and statistics-based methods are used.

The Kernighan-Lin algorithm [5] is one of the first graph segmentation methods, which detects the community by randomly dividing the graph into two equal parts and replacing the node pairs in different parts consecutively. The spectral bisection approach is one of the most used methods according to the boundaries of the Fiedler vector coordinates [6]. An example of discriminating algorithms is the Girvan-Newman [4] algorithm, which splits according to the middle link value (edge betweenness). Since the ensemble detection problem in complex networks is an NP-hard problem, many heuristic optimization algorithms have also been developed. These methods use the modularity function as the objective function to ensure that the network is best divided into communities. For example, while a node is allocated from one community to another for local optimization, a simulated backgammon approach has been developed for global optimization where communities are divided or combined with increasing modularity [13]. In addition, many heuristic algorithms, such as the tabu search algorithm [14], are used for community detection problems [15, 16]. Another method, one of the modularity-based approaches and frequently used in studies, is the modularity optimization algorithm. A greedy hierarchical clustering approach uses the concept of modularity [7] as a stopping criterion.

Grouping by analyzing the content of social objects in the network is another community detection approach. Topic models have been developed that summarize people's interests and allocate people interested in similar topics to the same community. The best-known subject model is the Latent Dirichlet Assignment (LDA) [11]. Topic models are developed and interpreted for social networks. CUT [8], CART [9] and CT [10] models are used to group network members with common interests. Since the subject models ignore the topological structure of the network, they produce a low modularity value.

The methods based on the topology of the network and the content shared by the people in the network in the community detection problem are relatively new studies. Considering the algorithms mentioned above, in this study, an improved modularity optimization algorithm, which uses the connection structure of the network and the sharing similarities between the people in the network, has been studied to increase the modularity value.

### 3 Parliamentary Optimization Algorithm (POA)

The parliamentary system, a system of government making and regulating laws, is also known as parliamentarism. The people elect members of Parliament in general elections. People often vote for their favorite party. Members of parliament who are members of political parties support their parties in parliamentary elections. Parliamentary groups of members based on the party they belong to strive to gain superiority over other parties in the competition between parties. In almost all democratic countries, the parliamentary population is formed by political parties.

There are two systems in parliamentary elections, the majority election system and the proportional representation system. While only one member is elected from each constituency in the majority electoral system, several members may be selected from one constituency in the balanced representation system. Generally, each political party presents its list of candidates, and voters can choose the political parties to vote for. Parties are given seats in the parliament in proportion to their votes [17].

Members of political parties within or outside parliament have different power values. These members of the party strive with little power to make a good impression on other noble members. They make this effort to get their support and votes during the elections. Essential members of the party get involved in races and try to find support among the noble members. On the other hand, Noble members tend to be more resourceful and often vote for those they trust. In this process, high-capacity general members are replaced with previous candidates. This part of the competition takes place between individuals within the party. Another race of the algorithm takes place between parties. Parties compete to increase their power. Parties have two main objectives for success: having the highest number of seats in the parliament and taking control of the government [17].

In the Parliamentary Optimization Algorithm (POA), optimization steps begin by creating the initial population of individuals. The individuals created are considered members of the parliament. In the next step, the population is divided into political groups (parties), and the candidate for the fixed number of member groups with the highest fitness is considered. After this step, the in-group competition starts. In the in-group competition step, the leading members turn to the candidate members suitable for them. This situation is modeled as the weighted average of vectors of principal candidates [17], as shown in the flowchart of POA, i.e. given in Fig. 1.

At the end of the in-group competition step, several candidates with the highest qualification are determined as the final candidates for each group. In the next step, the final candidates compete with the candidates of other groups. Principal and candidate members of the group are essential in determining the total power of the group. After the intra-group competition step, the competition between groups starts. Political groups within parliament compete with other groups to strengthen their candidates. Strong groups sometimes unite and become one group to increase their chances of winning. Algorithm 1 shows the process steps of POA.

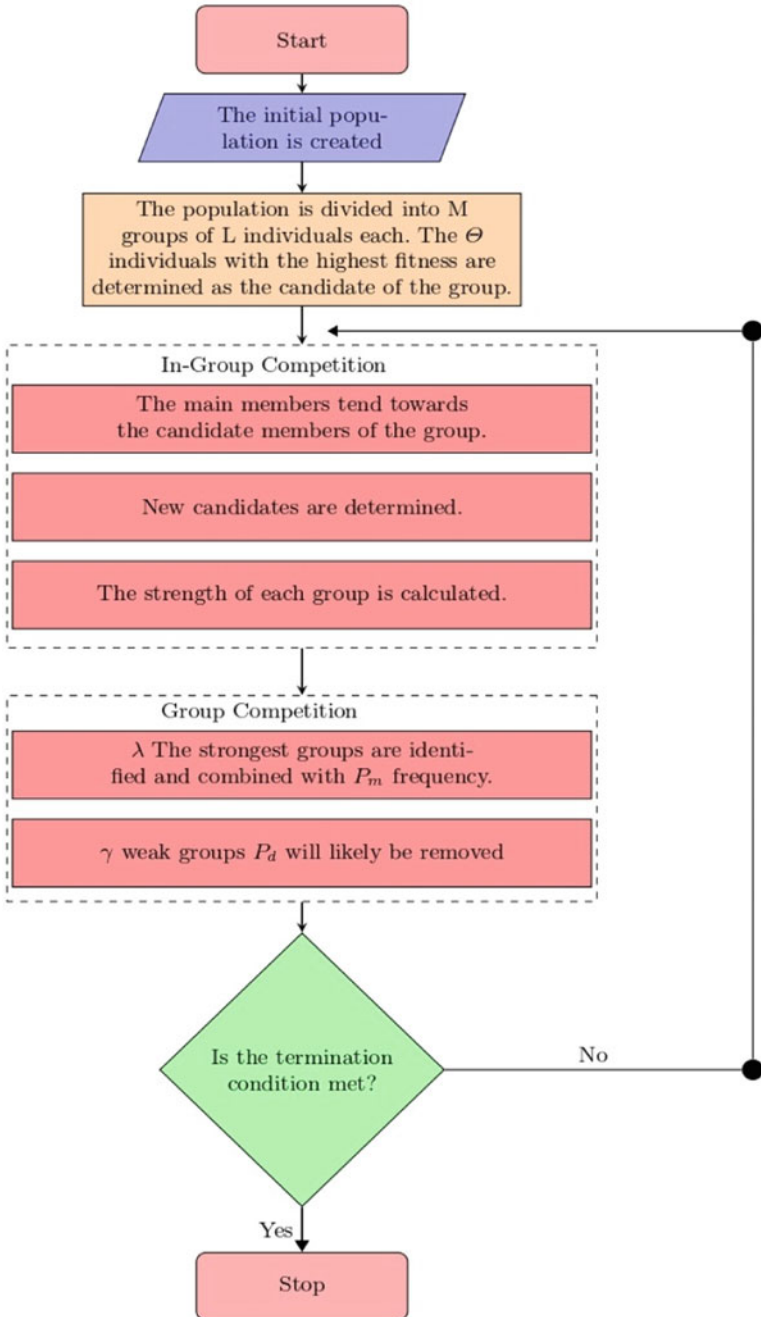


Fig. 1 Flowchart of parliamentary optimization algorithm



**Algorithm 1 [Process steps of POA]**

1. Start
2. The initial population is created.
  - (a) The population is divided into  $M$  groups consisting of  $L$  individuals.
  - (b) The highly fit individual is selected as the candidate for each group.
  - (c) In-group competition.
    - (a) The prominent members head toward the candidate members of each group.
    - (b) New candidates are appointed.
    - (c) Calculate the power of each group.
3. Competition between groups
  - (a) The most influential group is determined, and these groups are combined with  $P_m$  probability.
  - (b) The weakest group  $P_d$  will likely be deleted.
4. If the termination condition is not met, step 3 is repeated.
5. The best candidate is considered the solution to the optimization problem.
6. End.

**4 Modularity Optimization Algorithm (MOA)**

The algorithm is a hierarchical clustering method that combines pairs of nodes to increase the modularity value. The union giving the highest modularity value shows the number of ensembles. The algorithm defines the network as a graph and decides which pairs of nodes will have the maximum modularity value by using the network's link structure. The method is given in Algorithm 2.

A social network is defined as a graph of the shape  $G = (V, E)$ , where  $V$  is nodes and  $E$  is edges. The set of nodes  $V$  represents all users in the social network, and the collection of edges  $E$  represents the connections between users. Equation (1) gives the modularity value  $Q$  for an un-weighted and undirected graph  $G$  containing  $m$  number of edges.

$$Q = \frac{1}{2m} \sum_{uv} \left[ A_{uv} - \frac{d_u d_v}{2m} \right] \delta[r(u)r(v)] \quad (1)$$

$r(u)$  and  $r(v)$  indicate whether nodes  $u$  and  $v$  are included in the ensemble  $r$ . The function  $\delta$  produces  $\delta_{ij} = 1$  in the case of  $i = j$  and 0 in the other case.  $d_u$  and  $d_v$  represent the degree of nodes of  $u$  and  $v$  nodes, respectively. For the neighborhood matrix  $A$ , the case where  $A_{uv} = 1$  is where the nodes  $u$  and  $v$  are adjacent. The connection between nodes is defined as a direct connection. According to the modularity optimization algorithm, each node is defined as a separate class, and as many classes as the number of nodes are created. Significant groups are obtained by combining the

groups that increase the partial modularity value ( $\delta Q$ ) at each stage. The  $\delta Q$  value gives the effect of a merger on modularity. By combining two communities with a high  $\delta Q$  value, the modularity value ( $Q$ ) is increased. The  $\delta Q$  value is given in Eq. (2).  $e_{ij}$  is the ratio of the number of edges in the group to the number of all edges,  $a_i$  is  $i$ . class  $j$ . is equal to the side fractions  $\sum_j e_{ij}$  with the class.

$$\delta Q = 2(e_{ij} - a_i a_j) \tag{2}$$

**Algorithm 2 [Modularity Optimization Algorithm]**

Input:  $V, E$

Output: Communities.

1. Start.
2. Create  $k$  classes and assign nodes to each class.
3. Calculate the initial modularity value  $Q$ .
4. Set  $s$  to 1.
5. Repeat (until  $s$  equals  $k$ ).
6.  $i$  and  $j$ . Calculate  $\delta Q_{ij}$  values for class.
7. Combine the classes with the highest  $\delta Q$ .
8. End.

**5 Semantic Relation-Based Optimization Algorithm**

In this section, the modularity and Parliamentary optimization algorithm developed by utilizing semantic similarity are explained. The developed approach provides priority grouping of nodes that are closest to each other both topologically and semantically.

**6 Semantic Relation-Based Modularity Optimization Algorithm (SR-MOA)**

Semantic proximity between network members is the similarity value between documents shared by individuals (e-mail, blog, etc.). Word distributions represent each member of the network. The similarity  $C_{uv}$  of the word distributions of  $u$  and  $v$  nodes are added to the network as an indirect link.

Let  $d_u = d_{u1}, d_{u2}, \dots, d_{uz}$  be the set of  $z$  documents belonging to node  $u$ . To find the word distributions for each individual in the network,  $u$ , all documents belonging to the node are considered as a single document.  $d_u = w_1, w_2, \dots, w_{ku}$  represents  $k$  sets of words belonging to the node. The indirect connection value (semantic

similarity) between nodes  $u$  and  $v$  is calculated using the Cosine similarity method [18] (Eq. (3)).

$$C(d_u, d_v) = \frac{d_u * d_v}{|d_u| * |d_v|} \quad (3)$$

Pattern detection and vector representation of words are required to detect the best similarities and differences between the word distributions of the members. In the study, meaningless words in the documents belonging to the members were extracted, and the root of each word used in the network was obtained. Using the TF-IDF method [19], the terms were weighted according to the frequency of use of the members.

With the cosine similarity method, an  $n*n$  size  $C$  similarity matrix is created for an  $n$ -node network  $C_{ij}$ ,  $i$  with class  $j$ . It represents the semantic similarity between the classes. The developed algorithm initially has  $\delta Q$  and  $C$  matrix with  $n*n$  dimensions. The combined strength  $\delta QC_{ij}$  of the  $i$  and  $j$  classes is obtained by multiplying the  $\delta Q_{ij}$  and  $C_{ij}$  values. The two types with the highest  $\delta QC$  value are combined. As seen in algorithms 3 and 4, the  $C$  matrix is updated after each merge operation. Similarity values for the combined class are calculated by averaging the similarities of the two types. If the computed sharing similarity  $C_{ij}$  of the two members is equal to 0 initially,  $\delta QC_{ij}$  is assigned the lowest similarity value in the Cosine matrix.

#### Algorithm 3 [SR-MOA Algorithm]

Input:  $V, E, C_{n*n}$ .

Output: Communities.

1. Start.
2. Create  $k$  classes and assign nodes to each class.
3. Calculate the initial modularity value  $Q$ .
4. Set  $s$  to 1.
5. Repeat (until  $s$  equals  $k$ ).
6.  $I$  and  $j$ . Calculate  $\Delta Q_{ij}$  values for class.
7. Calculate  $\Delta QC_{ij} = \Delta Q_{ij} * C_{ij}$ .
8. Combine classes with the highest  $\Delta QC$ .
9. Update  $C$  matrix.
10. End.

#### Algorithm 4 [SR-POA Algorithm]

Input:  $V, E, C_{n*n}$ .

Output: Communities.

1. Start.
2. Create  $k$  classes and assign nodes to each class.
3. Calculate the initial modularity value  $Q$ .

4. Set  $S$  to 1.
5. Repeat (until  $s$  equals  $k$ ).
6. The initial population is created.
  - a. The population is divided into  $M$  groups consisting of  $L$  individuals.
  - b. The highly fit individual is selected as the candidate for each group.
  - c. In-group competition.
  - d. The prominent members head toward the candidate members of each group.
  - e. New candidates are appointed.
  - f. Calculate the power of each group.
7. Competition between groups
  - a. The most influential group is determined, and these groups are combined with  $P_m$  probability.
  - b. The weakest group  $P_d$  will likely be deleted.
8. If the termination condition is not met, step 3 is repeated.
9. The best candidate is considered the solution to the optimization problem.
10. End.

## 7 Experimental Setup and Result analysis

Performance evaluation to detect the impact of single- and multi-purpose-based heuristic community detection algorithm has been carried out over six different graphical social media data sets, namely Word adjacencies, Zachary karate club [20], Dolphin social network [21], Les Miserable's Books about US politics and American College football [22] over the evaluation parameter modularity and normalized mutual information.

Modularity is a network structural measurement that evaluates the strength of subgraph (groups, clusters or communities) in a network for extracting community structure [23]. In a network, a group of nodes having higher modularity are relatively dense to each other and leads to the appearance of communities in a given network as

$$M = \frac{1}{|2E|} \sum_{xy} \left[ e_{xy} - \frac{w_x w_y}{2|E|} \right] \delta(c_x, c_y) = \sum_{i=1}^n (f_{ii} - f'_i)^2 \quad (4)$$

where  $e_{xy}$  represents the edge from node  $x$  to node  $y$ ,  $W_x$  represents the summation of the weights of the edges linked to node  $x$ ,  $c_x$  is the belonging community structure of node  $x$  and  $(c_x, c_y)$  is a probabilistic function that equals 1 if both the respective node  $x$  and  $y$  belong to same community structure, otherwise 0.  $f_{ii}$  represents the edge in community  $i$ , and  $F'_i$  is the belonging probability of random edge to community  $i$  that is attached to vertices in community  $i$ .

Normalized mutual information is a normalization of intra-community mutual information score to scale the similarity between intra-community nodes as

$$nmi(x, c) = \begin{cases} 0 & \text{Node are totally dissimilar} \\ 1 & \text{node are totally similar} \end{cases} \tag{5}$$

And mutual information is calculated as

$$nmi(x, c) = \frac{2 * i(x, c_i)}{e(x) + e(c)} \tag{6}$$

where x is the class label, c is the community structure, e is the Entropy and i(x;c) is the information gain for element c<sub>i</sub> for class label x.

Performance evaluation of community detection algorithm with and without semantic relation is shown in Tables 1 and 2 as Modularity and Normalized Mutual information, respectively. Both the evaluation parameters are significantly improved after incorporating social theories with a community detection algorithm.

The community detection algorithms MOP, POA gain 45.12%, 60.45% modularity and 72.45%, 74.45% NMI over ZKC data sets, respectively, as shown in Figs. s2 and 3, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the modularity and NMI information are significantly increased and acquired 52.45%, 70.14% modularity and 81.78%, 84.56% NMI, respectively, over ZKC data set.

Over the AFC data set, community detection algorithms MOP, POA gain 55.68%, 68.56% modularity and 64.45%, 70.56% NMI, respectively, as shown in Figs. s4 and 5, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the

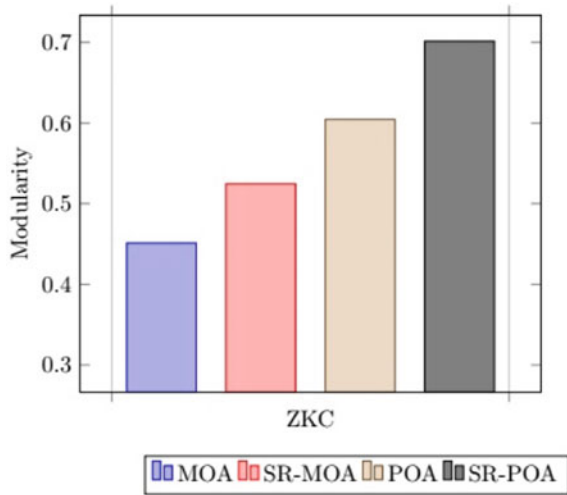
**Table 1** Comparative analysis of impact of semantic relation on modularity

Classification technique	Modularity					
	ZKC	ACF	DCN	BUP	LM	WA
MOA	0.4512	0.5568	0.5684	0.6575	0.5423	0.3956
POA	0.6045	0.6856	0.6423	0.6953	0.6845	0.4956
SR-MOA	0.5245	0.6435	0.6462	0.7512	0.6075	0.5056
SR-POA	0.7014	0.7856	0.7126	0.7856	0.7456	0.6586

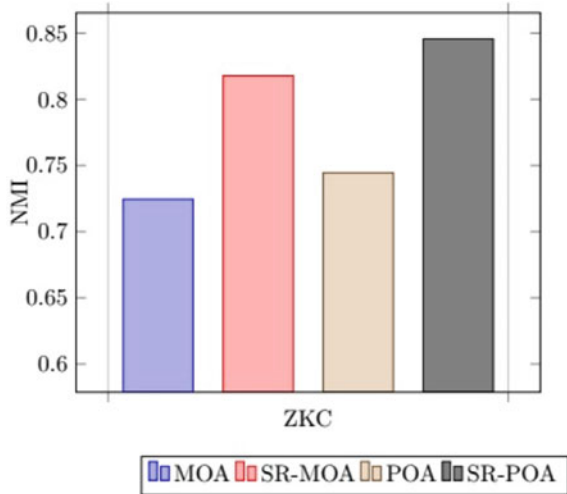
**Table 2** Comparative analysis of impact of semantic relation on NMI

Classification technique	Normalized Mutual Information					
	ZKC	ACF	DCN	BUP	LM	WA
MOA	0.7245	0.6845	0.6256	0.4567	0.3845	0.4125
POA	0.7445	0.7056	0.6645	0.5124	0.4756	0.4856
SR-MOA	0.8178	0.7456	0.6856	0.5042	0.4123	0.4986
SR-POA	0.8456	0.7635	0.7456	0.6845	0.5945	0.5896

**Fig. 2** Modularity of community detection over ZKC data set



**Fig. 3** Normalized mutual information of community detection over ZKC data set

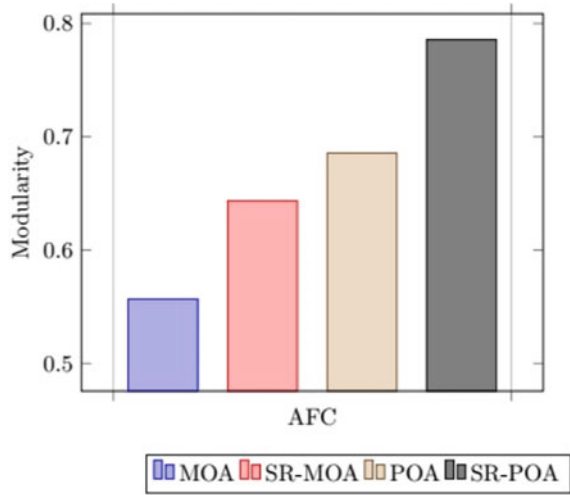


modularity and NMI information are significantly increased and acquired 64.35%, 78.56% modularity and 74.56%, 76.35% NMI, respectively, over the AFC data set.

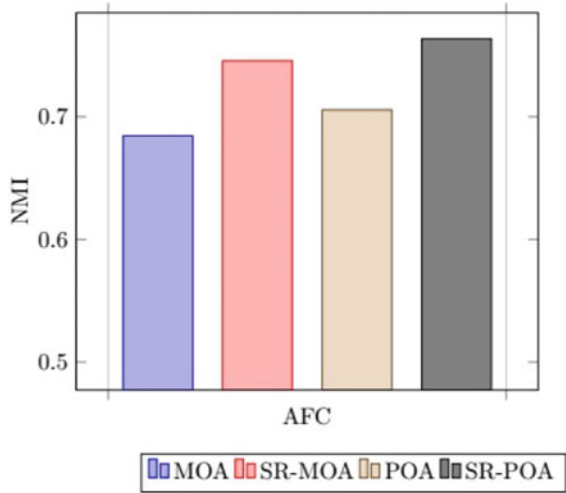
Over the DCN data set, community detection algorithms MOP, POA gain 56.84%, 64.23% modularity and 62.56%, 66.45% NMI, respectively, as shown in Figs. s6 and 7, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the modularity and NMI information are significantly increased and acquired 64.62%, 71.26% modularity and 68.56%, 74.56% NMI, respectively, over the DCN data set.

SEOA algorithm leads the modularity, whereas SBA and HSA algorithms achieve the highest NMI information.

**Fig. 4** Modularity of community detection normalized mutual information of community detection over WA Data Set over AFC Data Set



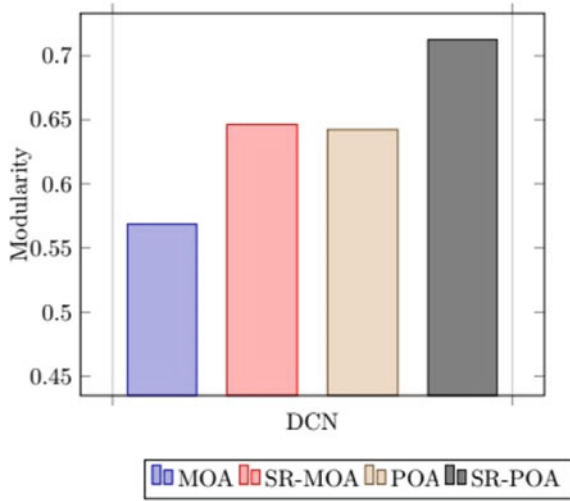
**Fig. 5** Normalized mutual information of community detection over AFC Data Set



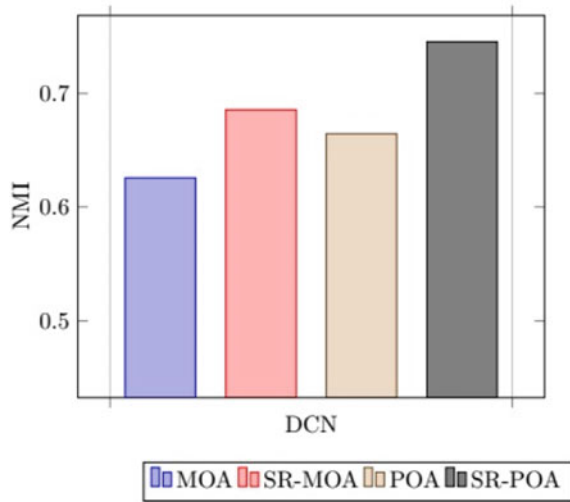
Over the BUP data set, community detection algorithms MOP, POA gain 65.75%, 69.53% modularity and 45.67%, 51.24% NMI, respectively, as shown in Figs. s8 and 9, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the modularity and NMI information are significantly increased and acquired 75.12%, 78.56% modularity and 50.42%, 68.45% NMI, respectively, over the BUP data set.

Over the LM data set, community detection algorithms MOP, POA gain 54.23%, 68.45% modularity and 38.45%, 47.56% NMI, respectively, as shown in Figs. s10 and 11, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the modularity and NMI information are significantly increased and acquired 60.75%, 74.56% modularity and 41.23%, 59.45% NMI, respectively, over the LM data set.

**Fig. 6** Modularity of community detection over DCN Data Set



**Fig. 7** Normalized mutual information of community detection over DCN Data Set

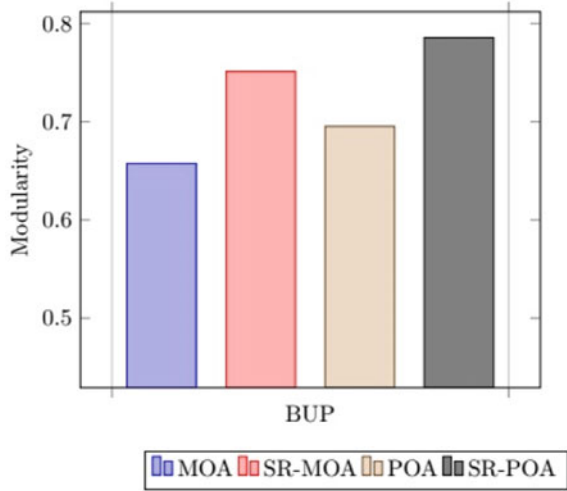


Over the WA data set, community detection algorithms MOP, POA gain 39.56%, 49.56% modularity and 41.25%, 48.56% NMI, respectively, as shown in Figs. s12 and 13, whereas after incorporating semantic relation, i.e. SR-POA and SR-MOA, the modularity and NMI information are significantly increased and acquired 50.56%, 65.86% modularity and 49.86%, 58.96% NMI, respectively, over the WA data set.

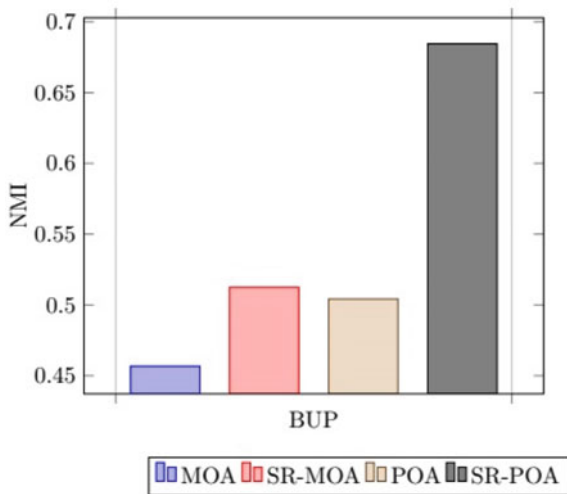
The performances of the POA and MOA algorithms with semantic relation are significantly increased and achieve a higher performance rate over higher dense ACF and ZKC networks and relatively lower over lightly dense WA data set.



**Fig. 8** Modularity of community detection over BUP Data Set



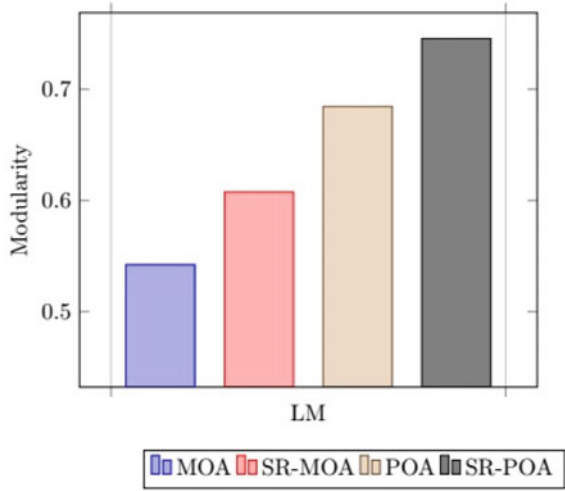
**Fig. 9** Normalized mutual information of community detection over BUP Data Set



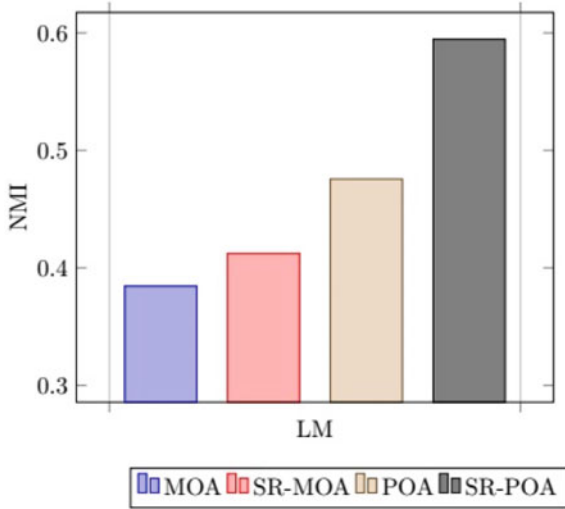
## 8 Conclusion and Future Studies

Recently, researchers have been focusing on community detection problems by analyzing the connection structure of the network. In this case, the obtained communities reflect only the topological feature of the network, while the documents shared among the people in the network are ignored. In this paper, the semantic similarities between people were used together with the topological structure of the network by making content analysis of the network. The results show that the use of semantic closeness between individuals and the link structure increases the performance of

**Fig. 10** Modularity of community detection over LM Data Set

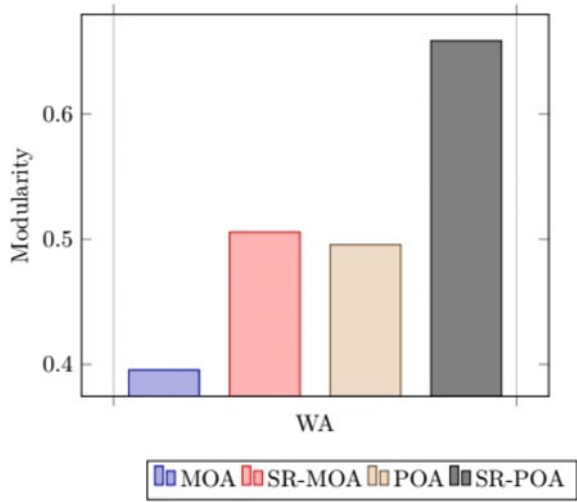


**Fig. 11** Normalized mutual information of community detection over LM Data Set

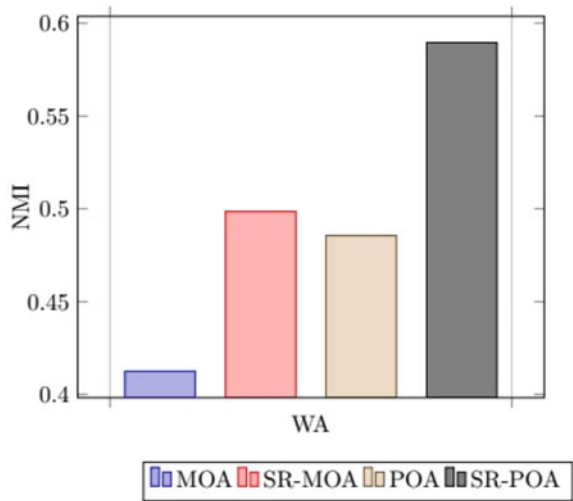


the community detection problem. In the future work, the variation of modularity performance will be investigated using the semantic similarity approach in different optimization algorithms.

**Fig. 10** Modularity of community detection over WA Data Set



**Fig. 11** Normalized mutual information of community detection over WA Data Set



## References

1. Morichetta A, Mellia M (2019) Clustering and evolutionary approach for longitudinal web traffic analysis. *Performance Evaluat* 135:102033
2. Li M, Wen L, Chen F (2021) A novel collaborative filtering recommendation approach based on soft co-clustering. *Phys A* 561:125140
3. Yang B, He H, Hu X (2017) Detecting community structure in networks via consensus dynamics and spatial transformation. *Phys A* 483:156–170
4. Moscato V, Sperla G (2021) A survey about community detection over on-line social and heterogeneous information networks. *Knowl-Based Syst* 224:107112

5. Smith NR, Zivich PN, Frerichs LM, Moody J, Aiello AE (2020) A guide for choosing community detection algorithms in social network studies: the question alignment approach. *Am J Prev Med* 59(4):597–605
6. Bruglieri M, Cordone R (2021) Metaheuristics for the minimum gap graph partitioning problem. *Comput Oper Res* 132:105301
7. Abduljabbar DA, Hashim SZM, Sallehuddin R (2020) An improved multi-objective evolutionary algorithm for detecting communities in complex networks with graphlet measure. *Comput Netw* 169:107070
8. Costa G, Ortale R (2020) Integrating overlapping community discovery and role analysis: Bayesian probabilistic generative modeling and mean-field variational inference. *Eng Appl Artif Intell* 89:103437
9. Zhou Q, Zhang C (2021) Breaking community boundary: comparing academic and social communication preferences regarding global pandemics. *J Informat* 15(3):101162
10. Bahadori S, Zare H, Moradi P (2021) Pod cd: Probabilistic overlapping dynamic community detection. *Expert Syst Appl* 174:114650
11. Yoon J, Jeong B, Kim M, Lee C (2021) An information entropy and latent dirichlet allocation approach to noise patent filtering. *Adv Eng Inform* 47:101243
12. Reihanian A, Feizi-Derakhshi M-R, Aghdasi HS (2018) Overlapping community detection in rating-based social networks through analyzing topics, ratings and links. *Pattern Recogn* 81:370–387
13. Hafiene N, Karoui W, Romdhane LB (2020) Influential nodes detection in dynamic social networks: a survey. *Exp Syst Appl* 159:113642
14. Chen S, Wang Z-Z, Bao MH, Tang L, Zhou J, Xiang J, Li J-M, Yi C-H (2018) Adaptive multi-resolution modularity for detecting communities in networks. *Phys A* 491:591–603
15. Cheikh-Graiet SB, Dotoli M, Hammadi S (2020) A tabu search based meta heuristic for dynamic carpooling optimization. *Comput Industr Eng* 140:106217
16. Pourasghar B, Izadkhah H, Isazadeh A, Lotfi S (2021) A graph-based clustering algorithm for software systems modularization. *Inf Softw Technol* 133:106469
17. Osaba E, Del Ser J, Camacho D, Bilbao MN, Yang X-S (2020) Community detection in networks using bio-inspired optimization: Latest developments, new results and perspectives with a selection of recent meta-heuristics. *Appl Soft Comput* 87:106010
18. Kim D, Seo D, Cho S, Kang P (2019) Multi-co-training for document classification using various document representations: Tf-IDF and doc2vec. *Inf Sci* 477:15–29
19. Bozkir AS, Akcapinar Sezer E (2018) Layout-based computation of web page similarity ranks. *Int J Human-Comput Stud* 110:95–114
20. Zachary WW (1977) An information flow model for conflict and fission in small groups. *J Anthropol Res* 33(4):452–473
21. Lusseau D (2003) The emergent properties of a dolphin social network. In: *Proceedings of the royal society of London. Series B: Biological Sciences*, vol 270, no. suppl 2, pp S186–S188
22. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci* 99:7821–7826
23. Newman MEJ (2006) Modularity and community structure in networks. *Proc Natl Acad Sci* 103(23):8577–8582

# Sentimental Analysis with Emojis by Using Machine Learning



Balajee Maram, B. Srinivas Kumar, and P. Swaroopni

**Abstract** In the current situation, it is very common for people to share their reviews or feedback on the products they buy or the services they use or share their ideas on any event. A person's emotional impact plays an important role in their daily lives. Sentiment analysis is the default process by which text information is assessed and comments are commented on as bad, good or neutral. An existing system involved in dynamic energy construction of Dictionaries and classification techniques, but the advent of electronic learning processes in him included when the models work well because of their automatic reading ability. It's a way of a subset of artificial intelligence and is also involved in the scientific study of algorithms and mathematical models used by computer programs to perform a specific task without explicit user commands, relying on patterns and problems instead. Involved in emotion analysis or an emotional or customer document in which emotions can be represented by emoji. Various forms of human emotions can be positive or negative.

**Keywords** Machine learning · Emoji · Sentiment · SVM · Random forest classifier

## 1 Introduction

The main motto is to simplify the customer return process effectively. This can only be done if the real purpose of the customer is known and what they actually feel about the services and products offered by the companies. Analysis of traditional emotions and existence will only provide an indication of the pros and cons. This will not help companies to know exactly how the customer is feeling. The real intention of the customers is well known if the companies were able to translate the text embedded

---

B. Maram (✉) · B. S. Kumar · P. Swaroopni  
Department of CSE, GMR Institute of Technology (Autonomous), Rajam, India  
e-mail: [balajee.m@gmrit.edu.in](mailto:balajee.m@gmrit.edu.in)

B. S. Kumar  
e-mail: [17341A05F7@gmrit.edu.in](mailto:17341A05F7@gmrit.edu.in)

P. Swaroopni  
e-mail: [17341A05D1@gmrit.edu.in](mailto:17341A05D1@gmrit.edu.in)

in them into related emojis. Emojis are used to express many emotions in a better way. Every word has one related emoji. Emojis are used to record data in training and testing databases. In-depth learning models are used to know the exact feelings of the customer through emojis. In-depth learning is part of a family of mechanical learning approaches based on the learning representation of neural networks. Reading can be monitored, supervised or partially monitored. Supervised reading is a type of machine learning that maps input to both input and output. In-depth learning is a category of machine learning algorithms that use multiple layers to gradually extract high-level features of raw inputs. The glove database is used to train our in-depth learning model. It is used to train words that occur more often. RNN is a type of neural network, in which output from the previous step is given to the current step as input. Therefore, RNN is used to remember previous words. A single reversal in standard RNNs is a problem of gradient disappearance. The solution to the problem is LSTM, an artificial Recurrent Neural Network with its own structures used in the field of in-depth learning. Long-term memory (LSTM) works in tasks such as non-fragmented and connected handwriting or speech recognition. LSTM networks are well suited to classify, process and predict data based on time series data, as there may be unknown downtime between key events in the timeline. By displaying predicted output and test data results, using an integrated Flask, the output has two fields—one to take sentences and the other to display emojis.

## 2 Literature Survey

Dipak R. Kawade, Dr. Kavita S. Oza, “Sentiment Analysis: Machine Learning Approach”, 2017 [1].

Twitter is the well-known social networking site where people can express freely their thoughts, ideas and feelings. These tweets are recorded and analyzed human emotions related to a terrorist attack (Uri attack). The study returns tweets about the attack on Ur and found the feelings and intensity of the tweets. For studying emotions and polarity in tweets, text mining techniques are used. About 5000 tweets were rearranged and processed in advance to create a database of frequently repeated words. R is used for mining sensations and polarity. The results of the survey showed that 94.3% of the people were disgusted with the attack on Ur. With the increase in social awareness, the need to connect with people like Twitter is increasing. Twitter is the most important and popular social media where people can post tweets for any event. This is an open platform where people can share their thoughts / ideas or feelings freely. Due to low Internet prices, portable devices are less expensive and increase public value, people have a Twitter account. Most of them use tweets for various events. In the years of social media, people express their choices and their feelings via Twitter. So Twitter contains a huge amount of information. We know that the length of a tweet does not exceed more than 140 characters so that people can write tweets with the appropriate feelings/feelings for each word. Emotional analysis determining ideas is nothing but an analysis of ideas or feelings from textual details.

It identifies each person's opinion or feelings about a particular event. In emotional analysis, we need to convey a text or text that cannot be analyzed and create a program or model that represents a concise approach to a given document perspective.

Twitter Emotional Analysis is one of the newest and most challenging research areas. Since social media such as Twitter contains a large amount of information to hear the text in the form of tweets, it helps to identify people's feelings or opinions about a particular event. Emotional analysis or solicitation of ideas helps in reviewing movies, products, customer services, opinions on any event, etc. This helps us determine whether a particular item or service is good / bad or popular or not. It is also helpful to identify people's opinions about any event or people and to find out whether the text is right, wrong or neutral. Emotional analysis is a type of text separation that can separate text into different emotions.

Varsha Sahayak et al., "Sentiment Analysis on Twitter Data", International Journal of Innovative Research in Advanced Engineering (IJIRAE), 2015 Aspect based sentiment classification using BERT [2].

Emotional analysis is important in understanding natural language and there are many types of real-world applications. Normal emotional analysis focuses on determining the good or bad coherence of a given sentence. A common and complex task can be to predict the features mentioned in a sentence and the feelings associated with each of them. Normal emotional analysis is concerned with distinguishing the general feeling of the text, but this does not include other important details such as the business, topic or aspect within the text the target feeling. Aspect-based emotional analysis (ABSA) is a very complex task consisting of identifying emotions and characteristics. Aimed at identifying a clearer perspective on a particular aspect, it is a less challenging task of emotional analysis (SA). We have a well-prepared pre-trained model from BERT and have achieved new technological results in the database.

Felbo, B., Mislove et al., Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm, 2017 [3].

This paper shows the extension of long surveillance to a variety of sound labels and models can read rich presentations. Often NLP activities are limited by a lack of manually defined data. Emotional expression has been used to monitor long-term analysis of communication emotions and activities related to making models learn useful text representation. Also, hashtags from previous research are designed for the emotional stages of emotional analysis. Readings of one previously trained model were made in five domains. Emoji can be used to classify the emotional content of a text in an accurate way. Here, it shows how millions of instantaneous emojis on Twitter can be used to prepare training models to learn rich emotional representation. Here, the information is transferred to the identified tasks using a new system of good organization. It acquires a state of the art in many works such as emotions, sarcasm and emotional discovery. The modeled version of the emoji model is created in such a way that some can be used in various NLP activities. Emotional expressions are used as sound labels in texts. Manual separation required an understandable emotional content of every statement. It's hard when we take a job. This type of activity can capture the various uses of 64 types of emoji. This type of route has two restrictions. First, it requires emojis during testing while there are many domains with limited or

unused use of emojis. One limitation is that the tables do not include the power of the emojis used as that draws on the intended meaning of the emoji over time. Much learning has been done on this project because of its positive results. This requires an emoji database whenever it needs it. Due to a storage problem transfer reading, for the reading type, you do not need access to the actual database. It only needs a pre-made split. A large number of short texts with emojis can be used as sound labels for pre-training. To make common words, the correct methods of making tokens are followed. This data processing allows pre-training work to take on more variety.

HariPriya, V., & Patil, P. G A survey of sarcasm detection in social media, 2017 [4].

In this paper, a mock survey was conducted on social media. It depends entirely on the emotional analysis. Sarcasm word refers to the indirect expression of words in any context. This is the hardest job a person can get. Emotional analysis is a process of different perspectives on the good, the bad and the neutral. Indigenous language processing and data extraction help to classify data into positive, negative and neutral. The discovery of Sarcasm is part of NLP, and it also deals with the discovery of a person. In sarcastic perception, there will be three modes called lexical, hyperbolic and pragmatic. Each of these will have individual elements associated with it. In the lexical, it has three elements namely unigram, bigram N-gram and hyperbole, and it has four elements which are interjection, punctuation, quotes and reinforcement. Pragmatic, it has three elements namely, emotions, smileys and responses. With complete precision of sensory separation, mechanical learning methodology and semantic orientation provide the best generalization. Another method of emotional separation is a dictionary-based approach. ML can be split into two, called supervised reading and unchecked reading. Supervised learning requires training database and assessment database. types of planning used for this type of decision-learning study, SVM, neural network, bay bay naive high entropy. In a dictionary-based approach, it is based on a dictionary of emotions, a collection of ideas for emotional awareness and pre-construction. At this point, the clever details of the interaction are used by asking questions in the search engine considering a certain number of limits of good and evil whereas, in corpus and in the dictionary, it is as bad as a general dictionary concept, some words of ideas are handwritten. That list will be extended by searching on well-known word networks. The polarity strength of each word is also listed in the dictionary. Both dictionary-based methods and corpus and dictionaries are subject to unsupervised reading. Some of the ML methods are hybrid methods and in-depth learning methods. In a hybrid manner, it improved data extraction from informal documents by combining high entropy modeling and was categorized based on data-driven experiments. In an in-depth learning approach, it uses many different layers of processing units that are not compatible with feature extraction and modification. Three different databases were used and obtained the F1 school result. In this case, the authors used different data sets in a variety of ways, showing the correctness of each method as a result. In each case where it has individual SVM results, if the data sets are too large, they may not be able to show the correct result. Hinge losses are often dispersed affecting the accuracy of the results. In cutting trees, it is simple and easy



to use but it has some drawbacks such that it only works with well-known groups. In-depth learning process gives the best results of big data with pre-trained data whereas in the random forest, it can work with small details and many symbols does not work well. In-depth learning works with a large number of data sets. The discovery of sarcasm is a challenging task as there is a lot of ambiguity and facial expressions. Many ways are available to perform emotional analysis on this discovery of humor. The problem of finding jokes has not been resolved because of the increase in data day by day. This paper provides research on the various methods used to find sarcasm on various social media platforms. Analysis of the various dividers is performed on this. Here most methods are performed on a small data only. For the management of large amount of databases, in-depth learning is the best way to get jokes on big databases.

Meishan Zhang et al., “Tweet Sarcasm Detection Using Deep NeuralNetwork”, Proceedings of COLING 2016, the 26th International Conference on ComputationalLinguistics: Technical Papers, pages 2449–2460, Osaka, Japan, 2016 [5].

This paper is defined by the discovery of humor that is used automatically in the management of ideas and reputation management. By finding any sarcasm, details are taken from social media Twitter. It’s all focused on the Twitter details. The discovery of Twitter mockery can be made as a function of separating binary texts. There are two sources used. Firstly, much of the past work removes a rich feature that is different from the content of the tweet itself and secondly the recent work has exhausted the tweet feature of the status quo. Some of these activities include tags such as POS tags, dependent drug structures, brown clusters and emotional indicators. Relying on external sources, many ways to detect sarcasm are found in textbooks that use different methods. Neural network models benefit greatly from this work such as emotional analysis and perception. The advantages of using a neural network are that they are used to add features automatically and also make features unnecessary. Another advantage is neural models that use word input with real value. This paper provides a clear description of the deep neural network for the detection of sarcasm. The work can be done for the first time, with a different basic model built. After the neural model is created, it will have a sub-neural network and capture tweet content and information details. Tweet content is modeled through a normal internal neural network. Further analysis shows features from historical tweets and is useful for the neural model as a separate model. The diagnosis of Sarcasm is a type of separation problem. Different models have been used and many of the existing research efforts have focused on finding a functional feature. Some semi-supervised patterns have been extracted from the sarcastic sentences of amazon product reviews. Here, traditional art features are drawn from historical tweets and demonstrated good performance in detecting sarcasm. Various neural network structures have been used for sensory analysis. It ncludes repetitive autoencoders, dynamic integration networks, deep belief networks, deep convolution networks and neural CRF. This function provides great neural networks in sensory analysis. Due to the handmade features of semantic patterns, those are very difficult to capture. The main part of the work you put in, first, is to release the most recent edited historical tweets. Here, a limited number of tweet words have very high standards of contextual elements.

Neural pathways provide a dimensional dense bearing as input. The proposed neural model will consist of two objects that are considered to be local objects and themes. The repetitive neural network is used to capture sequence features automatically because it provides semantic information through input tweets. Repetitive neural networks are like short-term memory that will effectively reduce explosive issues and reduce gradients. It has been used as a more dynamic neural network. Supervised learning for training purposes has been used to reduce lost entropy crossing over a set of training examples. These types are tested on a balanced and non-balanced database. Experiments here are ten times the cross-verification test and overall's description of the discovery of satire as a major test metric. For modification of model parameters, here they selected 10% of the training data sets. Final results are displayed in balanced and unequal data sets on board. Neural models also show better results. It also offers a very high accuracy of 79.29% with a different model. Here a neural network model is developed for the model of tweet sarcasm detection and is also compared to small independent features. This neural network has two main advantages discussed before and it also provides the positive effects of a different art form.

Imane El Alaoui et al., "A novel adaptable approach for sentiment analysis on big social data", El Alaoui et al., 2018 [6].

Emotional analysis, also called opinion polls, aims to clarify people's feelings about a topic by analyzing their posts and different actions on social media. Tenth, it contains the division of posts into different emotions like good, bad and so on.

V.Subramaniaswamy et al., "Sentiment analysis of tweets for estimating criticality and security of events", 2017 [7].

Social Media has become a huge part of the world. It has been observed that about three-quarters of the world's population use social media. This has encouraged a lot of research on social media. One of the most useful applications is the real-time emotional analysis of contact information for security purposes. Here, the authors suggest a complete software program that will add deletion data to Twitter and analyze it using dictionary-based analysis to detect potential threats. They suggest a method of getting rating results called critical to assess the level of threat of a public event. The proposed program includes this heart-based dictionary analysis as well as in-depth data collection and emotion analysis at a different level to process event threats.

Wataru Souma et al., "Enhanced news sentiment analysis using deep learning methods", 2019 [8].

In this paper, we explore the power of predicting historical news sentiment based on the previous new financial market. of TensorFlow. After taking the word input as an input, then we analyze the long history of the Thompson Reuters stocks and the history of the high price values of the Dow Jones Industrial A average. This approach is advanced in analyzing stories with high scores as positive and high-value stories as one. Predicting the financial market is probably the most difficult task based on past financial news. Researchers should analyze a wide range of data for financial market transactions and use the most advanced technology tools to detect hidden patterns. The amount of information generated in the news and financial information.

### 3 Methodology

Classifiers.

The following classifiers are used for this project:

- SVC
- Linear SVC
- Random Forest Classifier
- Decision Tree Classifier

#### SVC

By combining the same and different things from other collections. Integration is the process of extracting data from unlabeled data. This is the operator of the implementation of Support Support Vector Clustering in Ben-Hur et al. (2001). In this field, the Support Vector Clustering (SVC) algorithm points to a map from the data space to the advanced feature space using the Gaussian kernel. Non-importing field The smallest field containing the image data is searchable. This field. The map returns to the data space, where it forms a set of parameters that include data points. These lines are interpreted as merged parameters. Points are placed separately for each item the line is associated with the same collection. As the parameter the width of the Gaussian character decreases, the number of limits limited to the data space increases, leading to a growing number of collections. Since observations can be defined as defining support for the allocation of basic opportunities, this algorithm can be viewed as the one that identifies regions in this distribution of opportunities. Integration is combined with the elements of group collections together in the same way and different from the objects of other collections. Icon the process of extracting data from unlabeled data and can be very helpful in many different situations, e.g., in the marketing app, we may be interested in finding customer collections with similar purchasing behavior.

#### LINEAR SVC

Implementation is based on libsvm Minimum time scale scales quarterly on average samples and may not work more than tens of thousands of samples. For large databases, consider using a linear SVC or SGD classifier instead, perhaps behind the Nystrom transformer.

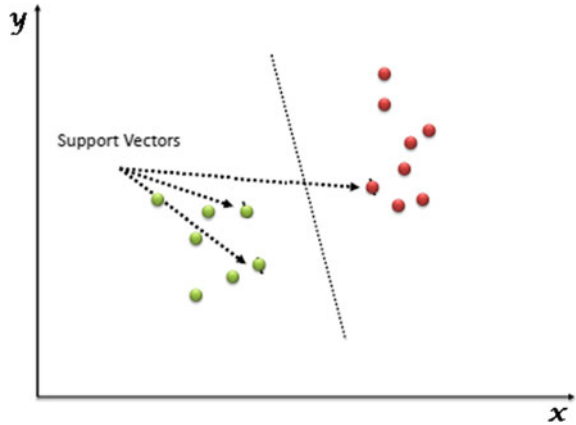
#### Support Vector Machines (SVMs)

In the SVM algorithm, we set each data item as a point in space  $n$  (where  $n$  is the number of attributes you have) and the value of each item is the value of a specific link. After that, we divide by finding a hyper-plane file that separates the two classes very well (see Fig. 1).

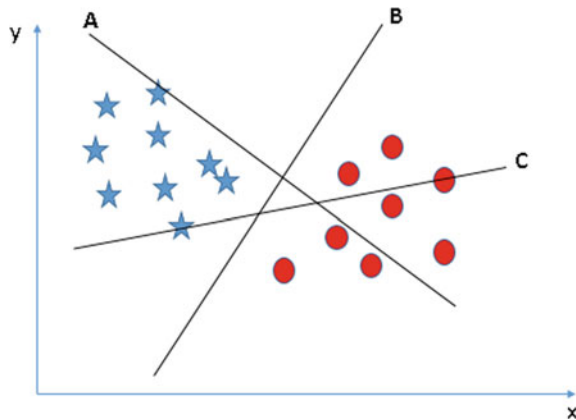
Identify the right hyper-plane (Scenario-1).

Select the hyper-plane which segregates the two classes better. In this scenario, the hyper-plane “B” has excellently performed this job. Figure 2 gives the explanation.

**Fig. 1** Support vector machine



**Fig. 2** Right hyper-plane (scenario-1)



Identify the right hyper-plane (Scenario-2).

Here, maximizing the distances between the nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as Margin. The right hyper-plane with scenario-2 can be explained with Fig. 3.

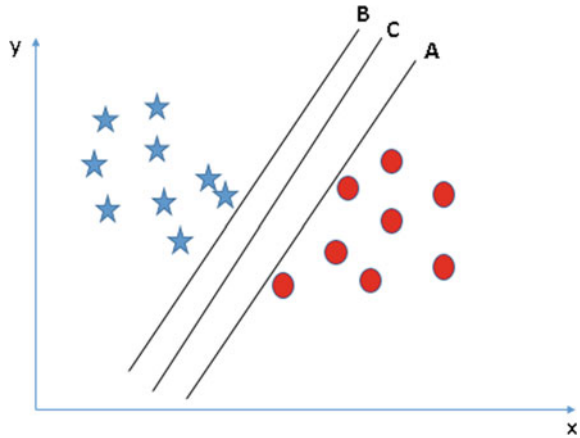
Identify the right hyper-plane (Scenario-3).

Here SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin. Here, hyper-plane B has a classification error and A has classified all correctly. Therefore, the right hyper-plane is A. In Fig. 4, it was clearly explained.

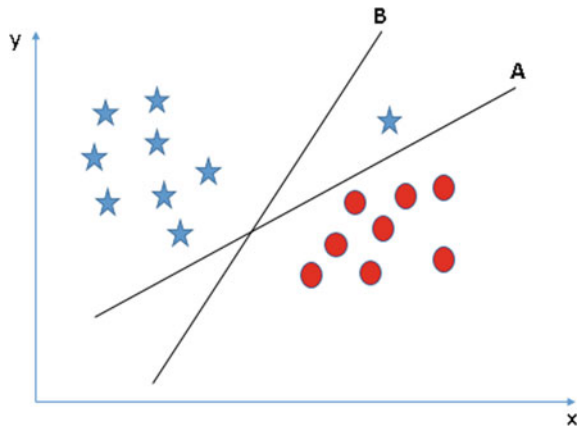
### Random Forest Classifier

Random forest is a machine learning method used to solve division problems. It uses collective bargaining, which is integrated into many chapters to provide solutions

**Fig. 3** Right hyper-plane (scenario-2)



**Fig. 4** Right hyper-plane (scenario-3)



to complex problems. The algorithm in the random forest contains many decision trees. “Forest” is made up of a random forest algorithm that is trained by combining combining bootstrap. The (random forest) algorithm determines the outcome based on predictive tree predictions. It predicts taking a measure or rate of extraction of various trees. Extending the number of trees increases the accuracy of the result. The random forest removes the algorithm limitations of the decision tree. Reduce data overload and overload accurately. It produces predictions without the need for multiple configurations in packages Random jungle algorithm of targeted machine built on the decision of the art of the tree.

### Classifier Decision

These are supervised non-formal learning methods used for classification as well as retreat. The purpose is to create a model that predicts the amount of variance. The tree can be seen as a clip smart measurement always. For example, decision trees learned

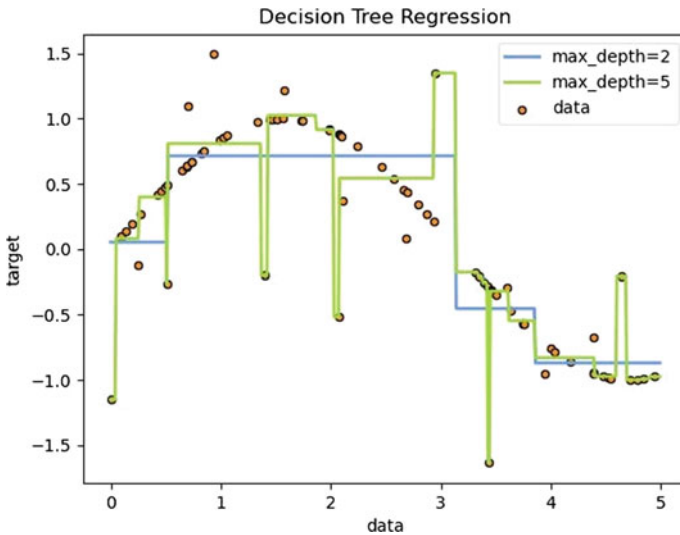


Fig. 5 Decision tree regression

from measurement data curve sine and set of rules of decision if then another. Depth of a tree, in which the process controls the decision and calibration of the model.

### Classifier Tree Decision

These are supervised non-formal learning methods used for classification as well as retreat. The purpose is to create a model that predicts the amount of variance targeted. The tree can be seen as a clip smart measurement always. For example, decision trees learned from measurement data curve sine and set of rules of decision if then another. Depth of a tree, in which the process controls the decision and calibration of the model.

### Decision Tree Regression

Decision trees can also be applied to regression problems, using the Decision tree regressor class where the fit method will take as argument arrays x and y. Figure 5 gives a clear explanation.

## 4 Results

The training and test accuracies for each classifier are summarized in Table1 which is given.

By looking at the numbers, the support vector machine works well, which means it is not overly constrained and unsuitable. The test accuracy is slightly higher than

**Table 1** Accuracies for each classifier

Classifier	Training accuracy	Test accuracy
SVC	0.1458890	0.1410428
LinearSVC	0.8899302	0.5768717
RandomForestClassifier	0.9911430	0.4304813
DecisionTreeClassifier	0.9988302	0.4585561

other classifiers. Some classifiers are overcrowded because training accuracy is much higher than accuracy tests.

Final model:

The final model of this project performed a test accuracy of 0.576872. As expected, the accuracy is low.

Confusion matrix of LSVMs are shown in Figs. 6, 7, 8, 9, 10, 11 and 12:

Output for emoji prediction when blank sentences were given.

Testing

Output for correct emoji prediction.

Output for incorrect emoji prediction.



**Fig. 6** Confusion matrix of LSVM

```
[37] emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = "DON'T IRRITATE ME"
t2 = "I LOVE FOOD"
t3 = "I WAS AFRAID OF DOGS"
t4 = "I LOST MY BIKE"

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))
```

😊 DON'T IRRITATE ME  
😊 I LOVE FOOD  
😨 I WAS AFRAID OF DOGS  
😊 I LOST MY BIKE

Fig. 7 Screenshot 1

```
emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = "Thank you for dinner!"
t2 = "I don't like it"
t3 = "My car skidded on the wet street"
t4 = "My cat died"

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))
```

😊 Thank you for dinner!  
😊 I don't like it  
😨 My car skidded on the wet street  
😊 My cat died

Fig. 8 Screenshot 2

```
emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = ""
t2 = ""
t3 = ""
t4 = ""

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))
```

😊  
😊  
😊  
😊

Fig. 9 Screenshot 3



```
emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = ""
t2 = "I didn't write the exam"
t3 = "he bought a car"
t4 = "My trip was cancelled"

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))

😊
😞 I didn't write the exam
😊 he bought a car
😞 My trip was cancelled
```

Fig. 10 Screenshot 4

```
emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = "we went to the party"
t2 = "she don't have any idea"
t3 = "he lost his pet dog"
t4 = "I was preparing for exam"

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))

😊 we went to the party
😞 she don't have any idea
😊 he lost his pet dog
😨 I was preparing for exam
```

Fig. 11 Screenshot 5

## 5 Conclusion

Each machine learning model has two important steps which are to train the model and test the model. This project has four stages and it trains each classifier to check the performance of each. Here the data set is divided into two parts like 80 and 20%. 80% of data is used to train the model and 20% data is used to test the model.

```
emoji_dict = {"joy": "😊", "fear": "😨", "anger": "😡", "sadness": "😞", "disgust": "😤", "shame": "😳", "guilt": "😓"}

t1 = "I was very happy with the results but when i saw my physics and maths i was disappointed"
t2 = "I was not happy when i saw my physics and maths but i was happy with the overall result"
t3 = "he lost his pet dog"
t4 = "I was preparing for exam"

texts = [t1, t2, t3, t4]
for text in texts:
    features = create_feature(text, nrange=(1, 4))
    features = vectorizer.transform(features)
    prediction = grid_obj.predict(features)[0]
    print("{} {}".format(emoji_dict[prediction], text))

😊 I was very happy with the results but when i saw my physics and maths i was disappointed
😞 I was not happy when i saw my physics and maths but i was happy with the overall result
😞 he lost his pet dog
😨 I was preparing for exam
```

Fig. 12 Screenshot 6

This paper checks the performance of each model with the help of Linear SVC and confusion matrix. The confusion matrix is a table that summarizes the performance of the division algorithm and reveals the type of misalignment that occurs. In other words, it shows the confusion of the distinction between classes. The row in matrix labels true labels and columns are predicted labels. By looking at the distribution of labels, the details are still well distributed. Errors are likely to occur because there is insufficient data to differentiate training. Also, a ngram with a width of 1 to 4 may not fit or add sound to the separators.

## References

1. Kawade DR, Oza KS (2017) Sentiment analysis: machine learning approach. Int J Eng Technol (IJET), 9(3):2183–2186
2. Agarwal A, Xie B, Vovsha I, Rambow O, Passonneau RJ (2015) Sentiment analysis on Twitter data. Int J Innovative Res Advanc Eng (IJRAE), ISSN 2349–2163 2(1):178–183
3. Felbo B, Mislove A, Søgaard A, Rahwan I, Lehmann S (2017) Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proceedings of the 2017 Conference on empirical methods in natural language processing, pp 1615–1625. Copenhagen, Denmark
4. Haripriya V, Patil PG (2017) A survey of sarcasm detection in social media. Int J Res Appl Sci Eng Technol (IJRASET) ISSN 2321–9653; IC Value: 45.98; SJ Impact Factor :6.887 Volume 5 Issue XII December 2017, pp 1748–1753

5. Zhang M, Zhang Y, Fu G (2016) Tweet sarcasm detection using deep neural network. In: Proceedings of COLING 2016, the 26th international conference on computational linguistics: technical papers, pp 2449–2460. Osaka, Japan
6. El Alaoui I, Gahi Y, Messoussi R, Chaabi Y, Todoskof A, Kobi A (2018) A novel adaptable approach for sentiment analysis on big social data. *J Big Data* 5:12
7. Subramaniaswamy V, Logesh R, Abejith M, Umasankar S, Umamakeswari A (2017) Sentiment analysis of tweets for estimating criticality and security of events. *J Organizat End User Comput (JOEUC)*
8. Souma W, Vodenska I, Aoyama H (2019) Enhanced news sentiment analysis using deep learning methods. *J Comput Soc Sci SpringerOpen*

# Security Risk Analysis and Design Reengineering for Smart Healthcare



**Madhu Sharma Gaur, Navneet K. Gaur, Sanjeev Kumar,  
and Prem Sagar Sharma**

**Abstract** In view of evolving the smart cities, emerging technologies interventions and the adoption of embedded sensor devices increasing rapidly in every domain. From the smart home to the smart workplace, smart education to smart healthcare, routine activities are technology-enabled and always connected through the internet. A large number of small, seamless but powerful sensor devices connected through wireless sensor networks over underlying architecture and heterogeneous environment. Adoption of IoT in healthcare as an Internet of Medical Things (IoMT) provides healthcare-assisted services with remote monitoring and treatment. IoT-enabled healthcare applications where human-centric information is ubiquitous and security risks rising irrationally. With the aim of enabling, secure connected healthcare applications and services anytime, anywhere in the highly dynamic and heterogeneous environment need for continuous improvement in the existing security solutions. For smart healthcare security assurance, privacy and interoperability are the most common thrust areas for research. Emerging technologies like artificial intelligence, machine learning, and deep learning are the widespread technologies offering risk prediction, data analyzing, and process automation. In this paper, we explore emerging security risks and propose an adaptive security reengineering strategy using artificial intelligence with risk requirement alignments for design re-engineering of a typical smart healthcare system development where security is critical in adopting this fast digital transformation.

**Keywords** Smart healthcare system · Internet of Things IoT · Internet of Medical Things IoMT · Security requirement engineering · Machine learning

---

M. Sharma Gaur (✉) · S. Kumar · P. S. Sharma  
G. L. Bajaj Institute of Technology and Management, Greater Noida, India  
e-mail: [madhu14nov@gmail.com](mailto:madhu14nov@gmail.com)

N. K. Gaur  
Telus International, Noida, India

# 1 Introduction

Emerging wireless sensor networks and communication technologies are fascinating rapid digital transformation and automation to deploy the smart systems in every domain. From smart home to smart workplace, smart healthcare fabricates smart cities where everything is connected over the internet through several embedded sensor devices known as Internet of Things (IoT) [1]. Such small but powerful sensor devices are seamlessly communicating in a virtual world from anywhere, any time, and bringing the smart world in our mist. Adoption of IoT in healthcare to create an Internet of Medical Things (IoMT) where medical sensor devices and equipment are connected through wireless or mobile networks which provide virtual and smart healthcare assistants and continuous monitoring with smart or mobile healthcare applications systems shown in Fig. 1.

IoT and IoMT healthcare data are voluminous and collected from dissimilar sources, ensuring security of this highly sensitive data is equally critical as ensuring all the functional requirements and meeting the objectives of the smart healthcare system. In view, the smart grid and rapid development of IoT or IoMT-based applications security assurance and trustworthy services are being prime concerns for solution providers as well as community researchers.

This paper aims to reduce smart healthcare adoption concerns by security risk analysis and requirement alignment for design reengineering. We explore emerging security risks and propose adaptive security reengineering strategy using artificial intelligence with risk requirement alignments for design re-engineering of a typical smart healthcare system development where security is critical in adopting this fast digital transformation. The idea of design by security risks analytics to endorse adaptive security requirement engineering strategy as an essential system development process for smart healthcare system where managing security is critical. Rest of the paper organized as related work presented; later to it, IoT security risk and requirement Alignment discussed in the third section. In the fourth section, proposed

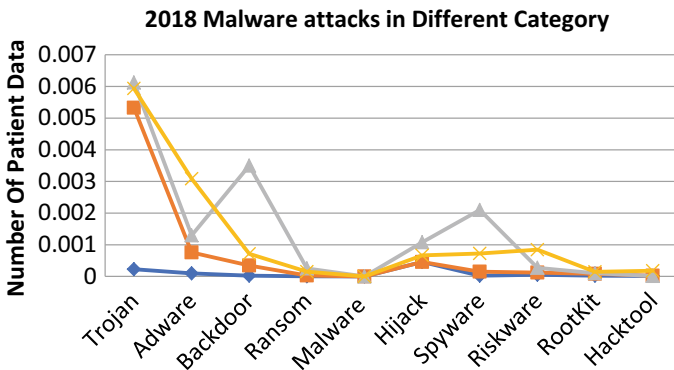


Fig. 1 2018 Malware attacks in different category

adaptive security reengineering strategy discussed and finally conclusion is given to summarize the paper.

## 2 Related Work

IoT devices deployed in an interconnected and heterogeneous network where embedded sensor devices communicating seamlessly. In the IoMT various medical devices and equipment are connected through wireless or mobile networks to form a smart healthcare system. The big data generated by IoT devices is massive and collected from dissimilar sources for patient diagnosis and remote monitoring. Furthermore, this big data used for various analytics, machine learning, risk predictions and pattern or behavior assessments. the heterogeneity of data sources, collection and access raise security requirements Security in ICT-enabled Healthcare is a sensitive Concern.

IoT-based Healthcare community researchers have realized the security requirement [2–8]. With rapidly growing technology and emerging security risks, for successful adoption of IoT in healthcare Security assurance is one of the key challenging issues. As IoT, technology is not new now, popularly being used for automation and smart applications [9]. A fog-based novel approach as middleware has been presented in [10] for security compliance regarding security vulnerability and patient privacy in IOT healthcare. In IOT open security issues and challenges explored and discussed in [11, 12]. Extensive research on Machine learning, deep learning-based intrusion detections model, malware detection in IOT backbone network discussed in [13–17]. In 2015, K. Zhang [18] has presented a structural design for Mobile Healthcare Networks (MHN) with security and privacy limitations and countermeasures in healthcare data aggregation, misbehavior recognition and protected health data processing. Hossain and Muhammad in 2016 [19] has recommended a health IoT-enabled monitoring framework in which ECG and other patient vital data collected by sensor devices over MHN with safe and flawless transmission. Watermarking and Signal improvement-based analytics also done to identify theft and avoid clinical errors with correctness assessment. Current solutions and future Challenges of machine learning in IoT security are shown in [20]. This author describes existing machine learning-based solutions, IoT network characteristics, identify gaps in IoT networks, and different types of attacks.

## 3 Security Risk Analytics and Requirement Alignment

E-Healthcare and Security both are equally critical concerns. Consideration of trustworthiness and security as one of the greatest need while designing the system with security risk predictive modeling. There is a need for requirement engineering change for refining the application development lifecycle processes with keeping security

objective parallel to functional objectives for an application design. Security assurance modelling is equally important to align with the design process of healthcare application development and modelling data analytics. It will include following.

- (a) Security Prototyping aligned with Application Prototyping: It refers to the rigorous security requirement specification with application requirement specification and analysis by augmenting any modeling language like UML or software modeling in order to better conceptualize the security requirement, Analysis, and outcome in association with the overall design of the application.
- (b) Security Declaration: Every security requirement and outcome chronicled with the application design, development, and deployment modelling through security risk analytics
- (c) Security Policy Enforcement: Throughout the AI modeling, specific security requirements with well-defined policies and structured evaluation process positioning are needed for data access, cleansing, computation, and classification for risk predictive analysis.
- (d) Security Evaluation and Assurance Policy Enactment: With security Policy, modularizing security evaluation and assurance strategy specification and endorsement in a well-structured manner must be included.

In the smart healthcare system, Artificial Intelligence (AI) provides effective algorithms to learn classifies the security risks and techniques to implement a rule-based or learning-based automation for managing attack surface and black spots as well as at both the sides; at defense and offence. Through AI security, the security risks analyzed to-

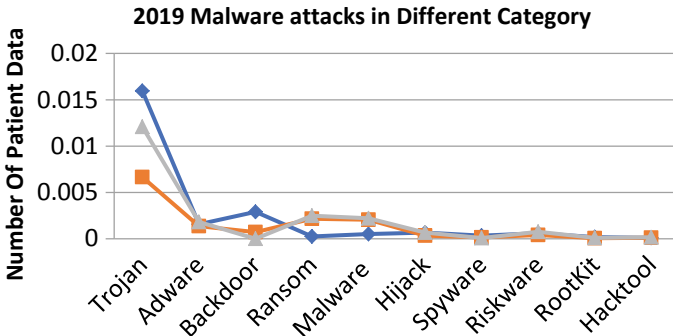
- Identify and learn new threats/ attack patterns and risk prediction
- Pattern recognition for user behavior learning and isolating compromised medical devices
- AI-based machine learning and deep learning to respond the attacks and data breaches
- Cyber-Physical resources monitoring and decision support system to protect the devices
- Well understand the security and performance tradeoffs in terms of interpretability, scalability, and extensibility.

### ***3.1 Healthcare Data Breaches***

Continuously healthcare data breaches are reported. As per Digital health news only in the month of July 2018 through various insecure data storage and access patient data affected [26]. This July month was bad for patients as significant number incidents were recorded and publicized by the US Department of Health & Human Services' office for civil rights where 858,411 patients are at risk. Although it's not only a case where each case must be reported (Table 1).

**Table 1** Summarizes of the patients' data affected in different ways

Patients affected category	No of patients affected
Improper disposal of data	317,154
Unauthorized access and disclosure	245,597
Hacking IT Incidents	291,465
Data theft and loss	4195



**Fig. 2** 2019 Malware attacks in different category

Network and Internet enabled medical devices are being used to patient vital parameters for mobile or remote healthcare services. However, most of the IoT devices fail to ensure are design and deployment with security certificates, primarily focused to provide automation and speedy digital transformation, which makes them highly vulnerable to compromise. A study done by Health IT Security [27] reveals that there is an average of 164 cyber threats detected/1000 connected host devices. Figures 1 and 2 shows 2018 and 2019 Malware attacks in a different category.

### 3.2 Security Risk Factors in Healthcare Internet of Things

- Some security measures often not carefully deployed on underlying network and if any how a threat remains inside the network slow down for and other threats may lead to security breaches
- IoMT devices operate on custom operating systems. In addition, being difficult to patch, maintain or regular exchange in view of security upgradation and become vulnerable to emerging attacks and threats.
- Along with IoMT devices functionality these devices being popular for attack vendors. Such devices allowed to so many critical but obvious and unpredictable accesses to much of the data stored on the network, which becomes an ideal point of cyber data theft.



- IoMT devices are disgracefully insecure as many of such devices are being designed by focusing only on healthcare functionality with little or no protection from cyber-attacks and devices are not designed with inherent security assurance
- IoMT devices deployed in a medical connected system and security policies enforced locally with flexibility in compliance.
- Containment reducing rather than enforcing end-to-end expel of attacks rather than only at network segment perimeter
- Misuse of good things offered by IoMT is not restricted to unwanted browsing. Lack of proactive identification of potential vulnerabilities to avoid malicious and compromised activity.

**Common IoT Security threats** As per the most recent research [28], IoT security gaps were identified. Although continuous research for improving security, trust, privacy, detecting malware attacks, and using enhanced solutions are always the open research [29–35]. Particular security requirements like measure Authentication, Device-to-Device secure communication, layered security and protection against specific risks, ensured access control need to be determined for an IoT/IoMT-based solutions. Common security threats while data collection and exchange from and through different sources in different formats are:

- Spoofing: False or multiple identities gaining access.
- Repudiation: Denial of specific action on user behalf.
- Tampering: Unauthorized access and alteration in data
- Denial of Services (DDoS): making the system unavailable
- Disclosure of Information: Exposing personal and sensitive data.

## 4 Proposed IOT Security Risk Assessment

An IoT-based solutions need novel self-adaptive security mechanisms to manage dynamic vulnerabilities and contextual change management. Accompanying aspects of IoT-based mobile/electronic healthcare solutions and data analytics with security Requirement Engineering must offers to make it inclined towards no misuse of personal data and no vulnerabilities with more disclosures for ultimately achieved security evaluation. In the highly connected world IoT devices also designed by incorporating Inherent contextual security. In view of Healthcare IoT, It is difficult for a team of people to manage embedded connected devices over wireless or mobile network and high volume of information with security assurance. This is where Machine learning can play an important role to recognize the compromised behavior/patterns and threats predicted on massive data sets at all on the machine speed.

All the stakeholders system, need to realize that security is not optional rather it is inherent requirement and requirement of strictly enforcing each and every data disposal and access point security critical A big caution regarding the security assurance needs to integrally upstretched for all the IoMT vendors, deployment IT teams

and services providers. Large number IoMT devices are connected on an open and heterogeneous environment and as large number of devices means larger attacks and security breaches are significant challenges.

**Challenges for Security Risk Analytics:** In view of implementing rule-based automated algorithms using AI, several challenges also arise like lack of training data, unstructured, massive data regarding device profiling, logs, binaries, network underlying infrastructure, protocols, and healthcare stakeholders. Thus, semi-supervised learning algorithms are applied to learn about new attacks, threats and classify the vulnerabilities. Similarly, intruders may use to find the path to identify the security loopholes. System and domain study including state-of-the-art survey to understand the comprehensive security requirements.

- Risk Analytics plan: Vulnerabilities path finding
- Formation/correlation of attacks and designing threat vectors from internal/external impetus factors.
- Eliminating un-essential, less essential, and most essential tasks to reduce false reduction while rule-based or learning-based automation.
- Vulnerability Ranking to set the priority as per system need.

#### ***4.1 Requirement Engineering in Smart Healthcare***

Smart healthcare solutions are driven by the use of smart sensor devices, Internet of Things (IOT), Artificial Intelligence, and predictive modelling providing significant advantages for digitized disease care. Emerging technology advancement also has another potential distressing counterparts associated with trust and security. Figure 3 represents the machine learning-based requirement engineering in an IOT-based smart healthcare applications.

Every day new solutions or services proposed and tossed in the market by the healthcare industry and technology experts. Worldwide number of diabetes patient is rapidly growing and smart healthcare stakeholder has highly influenced digitized solutions for disease care. Diabetes care solutions and services are data driven where voluminous past, current and environmental big data need to analyze, Medical firms are adopting IoT-enabled smart healthcare architecture and more robust AI/ML based such big data handling and risk analytics strategies. For the highly data-driven chronic disease like diabetes where continuous data collection is needed, security and trust implications are gaining importance as continuously more security breaches are reported. Trusted security policies must ensure the best security posture with stringent security assurance and compliance measures. Figure 4 represents a complete view of the Healthcare system/application development life cycle process alignment with the Intelligent Security Requirement engineering process.

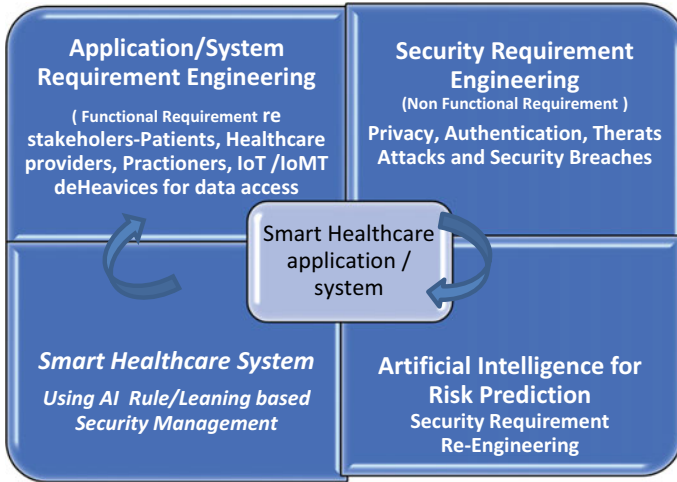


Fig. 3 Security requirement engineering in smart healthcare system

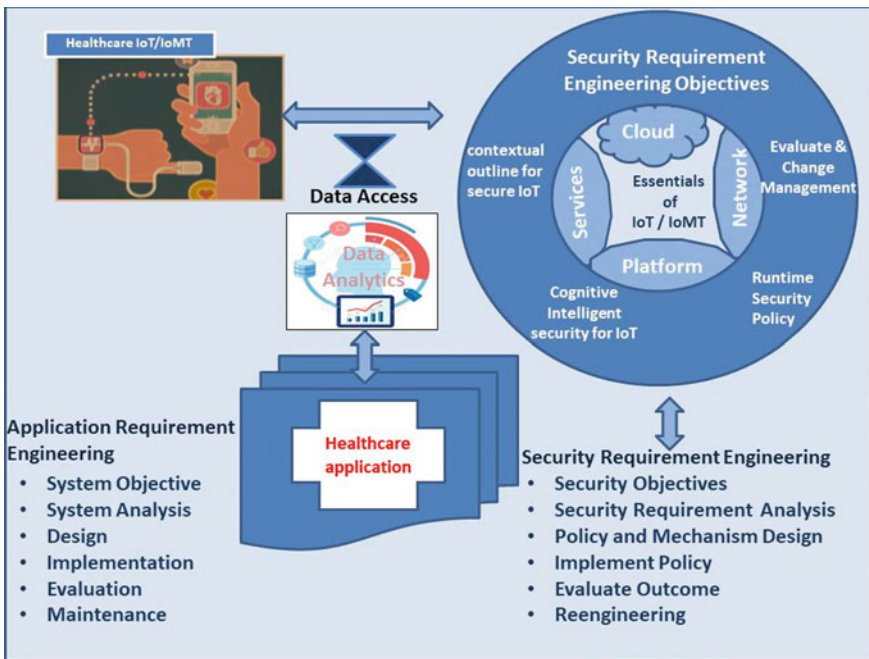


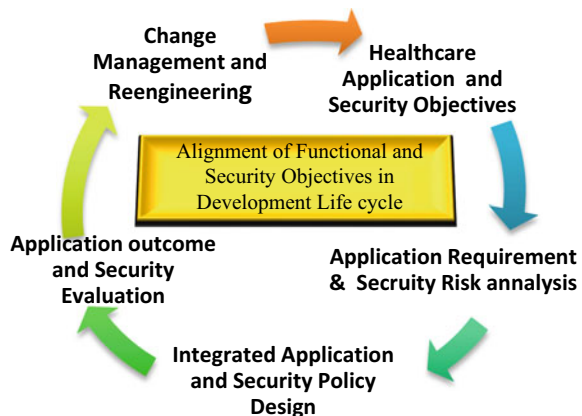
Fig. 4 Smart healthcare development life cycle with security requirement engineering

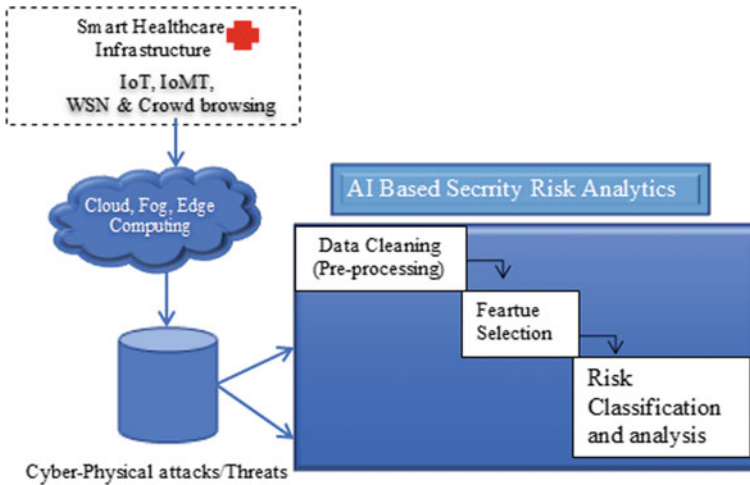
### 4.2 Restructuring Security Requirement Engineering Lifecycle

Ensuring the highly critical and sensitive healthcare data in safe hands is an important goal of the smart healthcare system as its functional objective. Predictive analytics is used to predict security requirements while designing the system from end to end at all the layers of IOT-enabled systems with inherent considerations for secure and trustworthy emerging concerns. Healthcare providers are always seeking data integrity, privacy, and security assurance with trustworthiness, in spite of that day by day breaches have increased in prevalence and destruction. Along with data, analytics, and predictive modeling security engineering techniques and strategies are expected to provide secure and trustworthy usage rather than only increasing global healthcare applications and services. According to the PwC Global State of Information Security, 2018 need for increased data. Inherent Security Requirement Engineering Lifecycle has been shown in Fig. 5, which needs to include the following process while application development.

- Designing an Organizational Contextual outline for secure IoT application: We design method is used to create a channel between system engineering and Security requirement Engineering which can address the gaps
- Formal and applied security requirement engineering in the development and deployment of human-centric security-critical systems.
- Adoptable change management for continuously evolved, adhoc vulnerabilities with the healthcare IoT devices including usage classification and guidelines of contextual environment, standards, and platform.
- Open Cognitive IoT Healthcare Architecture development for Intelligent Security: In Healthcare data analytics data is collected from diverse IoT devices.

Fig. 5 Security requirement engineering lifecycle





**Fig. 6** Using AI for security risk analysis for smart healthcare system

### 4.3 Artificial Intelligence in Security Risk Analysis for Requirement Engineering

In a typical smart healthcare system, as shown in Fig. 6, risk analysis can be automated using AI/ML including.

How AI/ML can strengthen security management? The existing work presented in [36–39], to identify network, IoT attacks, and threats using deep learning, machine learning with predictive framework. A machine learning-based classification technique [40]; training and testing done for managing security as an integral part of application for security risk prediction and analysis, machine learning can play an important role for security risk prediction and analytics. Inspiring from such scenarios, here we propose a machine learning-based security design by dynamic analytics as per dynamic attacks and threats. An integral security reengineering by security risk analytics can be a better approach to deal with emerging attacks and the corresponding solution. To define a roadmap for an intelligent security strategy following steps recommended.

Step: 1 Recognizing data sources and platforms by evaluating rubrics for using AI/ML.

Step: 2 Define use cases for man–machine behavior pattern learning and analysis for the IoT or IoMT devices with the lowest implementation complexity and highest risk prediction like an intrusion, malware, fraud detection, and network attacks/threat.

Step 3: Threat intelligence for implementing rule-based automation using AI algorithm/ML techniques to detect threats/attacks.

Step 4: Training and Learning based automation for repetitive security concerns detection and management.

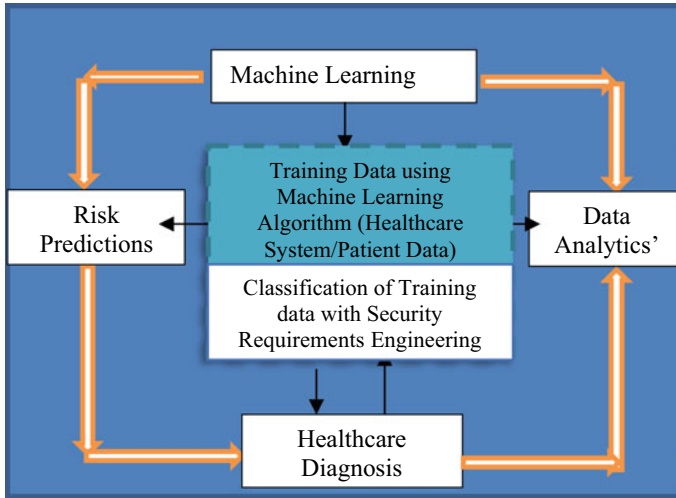


Fig. 7 Machine learning based data classification with security requirement prediction

For a smart healthcare system, managing voluminous unstructured data security is major area of concern than structured data. In Fig. 7, machine learning-based training and classification based on security requirement reengineering flow for an IOT enabled smart healthcare data Analytics shown. In the supervised learning calls on sets of training data in the form of correct question-and-answer pairs framed and freezes as a ground truth of an intelligent system with security requirement prediction. This training helps to classify the mainstays of machine learning-based security requirement engineering process, analysis, and categorize the observations. There are various supervised and unsupervised learning algorithms and methods used for classification though Security risk prediction and analysis of healthcare IoT devices data breaches is our work in progress for further realization of the proposed intelligent approach using the following supervised machine learning methods.

- Support Vector Machine (SVM): An algorithm based on statistical learning theory and look ahead to data points close to the opposing class. The data points called support vectors, used for the important classification task. The best separation line known as decision boundary is defined to segregate different classes from data sets.
- Naive Bays (NB): A statistical method based on Bayes’ theorem (Lewis, 1998) with the strong assumption of independence between features. It calculates the probability of input data that is relevant to a specific pre-defined class with the certain statistical functions and results the class with highest probability.
- Decision Tree (DT): A logic-based algorithm where data sets are modeled in hierarchical structures.

**AI-Driven Model for Security risk prediction in Healthcare IoT:** For security by design reengineering a continuous threat modeling is recommended to include in an application design development and deployment life cycle for continuous threat assessment and risk reduction. Threat Modeling includes the following set of activities.

- i. Analyzing Security Requirements
- ii. Application Functional Design
- iii. Identification of threats/attacks
- iv. Qualifying threats/attacks
- v. Evaluation and validation of qualifying threats/attacks.

For an Intelligent Security Requirements Engineering using Machine Learning techniques, an IoT-enabled healthcare system will make inherent intelligent detection of compromised behavior and pattern. This can rapidly ascertain an isolation state of the user, device, or network access points and security threats prediction for where deeper human analysis is required.

## 5 Conclusion

In this paper, we explored the smart healthcare system with security apprehensions. Ensuring Security is equally necessary as meeting any system's functional objectives. To enhance the confidence and adoption of the IoT-enabled smart healthcare system development, security requirement engineering needs to be defined and included at the early stage of the system development life cycle as an equally important process. Human-centric systems like e-healthcare where data security is critical a system must be bided on some specific evidence to meet the security requirements. Existing IoT-enabled healthcare solutions and underlying approaches focused on developing and deploying security techniques to detect attacks, threats, and other vulnerabilities regardless of much consideration of security requirements alignment with solution objectives. Despite increasing IoT-based smart healthcare system deployments still ensuring security is challenging. In this paper, we proposed machine learning based on an Inherent Intelligent Security Requirement Engineering strategy for smart healthcare system relying on IoT devices and underlying architecture. A survey-based data collection and analysis of healthcare IoT devices' data security and breaches is our work in progress for further realization of the proposed intelligent approach using the following supervised machine learning methods. In an IoT-enabled smart healthcare system predictive analytics are used to predict security requirement relevance and vulnerability impact and need to be addressed by all the stakeholders equally to freeze solution objectives with security requirement objective.

## References

1. Novo O, Beijar N, Ocak M (2015) Capillary networks bridging the cellular and IoT worlds. *IEEE World Forum Internet Things (WF-IoT)* 1:571–578
2. Binti Mohamad Noor M, Hassan WH (2019) Current research on internet of things (IoT) security: a survey. *Comput Netw* 148:283–294
3. Rault T, Bouabdallah A, Challal Y (2018) Internet of things security: a top down survey. *Comput Netw* 141:199–221
4. Alaba A, Othman M, Hashem IA, Ajotaibi F (2017) Internet of things security: a survey. *J Netw Comput Appl* 88:10–28
5. Porabage P, Braeken A, Schmitt C, Gurtoy A, Ylianttila M, Stiller B (2015) Group Key Establishment for enabling secure multicast communication in wireless sensor network deployed for IoT applications. *IEEE Access* 3:1503–1511
6. Neisse R, Steri G, Fovino IN, Baldini G (2015) SecKit: a model-based security tool kit for internet of things
7. Sicari S, Rizzardi A, Grieco L, Coen-Porisini S (2015) Privacy and trust in internet of things : the road ahead. *Comput Netw* 76:146–164
8. Kai KA, Pang ZB, Cong W (2013) Security and privacy mechanism for health internet of things. *20 (Suppl. 2)* 64–68
9. Mohammad D, Ahmed M (2019) IoT service utilization in healthcare, internet of things (IoT) for automated and smart applications. Ismail Y (ed) IntechOpen, June 2019. <https://www.intechopen.com/books/internet-of-things-iot-for-automated-and-smart-applications/iot-service-utilisation-in-healthcare>
10. Gope P, Hwang T (2016) A fog based middleware for automated compliance with OECD privacy principles in Internet of healthcare things. *IEEE Access* 4:8418–8441
11. Mosenia A, Jha NK (2017) A comprehensive study of security of Internet of Things. *IEEE Trans Emerg Topics Comput* 5:586–602
12. Sha K, Wei W, Yang TA, Wang Z, Shi W (2018) On security challenges and open issues in internet of things. *Fut Generat Comput Syst* 83:326–337
13. Sen DBJ (2011) Internet of things-applications and challenges in technology standardization. *IEEE Trans Wireless Personal Commun*
14. Restuccia F, Doro S, Melodia T (2018) Securing Internet of Things in the age of machine learning and software- defined networking. *IEEE Internet Things J* 5:4829–4842
15. M et al (2018) Machine learning for internet of Things data analytics: a survey. *J Digital Commun Netw Elsevier*, 1:1–56.
16. Azmoodeh A, Dehghantanha A, Choo KR (2018) Robust malware detection for internet of (Battlefield) Things device using deep Eigen space learning. *IEEE Trans Sustain Comput* 1–1
17. Pajouh HH, Javidan R, Khayami R, Dehghantanha A, Choo KK (2018) A two layer dimension reduction model for anomaly-based intrusion detection in IOT backbone networks. *IEEE Trans Emerg Topics Comput*:1–1
18. Zhang K, Yang K, Liang X, Su Z, Shen X, Luo HH (2015) Security and privacy for mobile healthcare networks: from a quality of protection perspective. *IEEE Wireless Commun* 22(4):104–112
19. Hossain MS, Muhammad G (2016) Cloud based Industrial Internet of Things (IoT)- Enabled Framework for Health Monitoring. *Comput Netw* 101:192–202
20. Hussain F, Hussain R, Hassan SA, Hossain E (2019) Machine learning in IoT security: current solutions and future challenges. *ArXrv*:1904.05735
21. <https://appinventiv.com/blog/iot-in-healthcare/>
22. [https://idc-cema.com/dwn/SF\\_177701/driving\\_the\\_digital\\_agenda\\_requires\\_strategic\\_architecture\\_rosen\\_idc.pdf](https://idc-cema.com/dwn/SF_177701/driving_the_digital_agenda_requires_strategic_architecture_rosen_idc.pdf)
23. Madhu S (2015) Trusted and secure clustering in mobile pervasive environment. Springer Open *J Human-Centric Comput Informat Sci (HCIS)*. <https://doi.org/10.1186/s13673-015-0050-1>
24. Gaur MS, Pant B (2015) Impact of signal-strength on trusted and secure clustering in mobile pervasive environment. *Elsevier Procedia Computer Science* 57:178–188



25. Gaur MS, Pant B (2014) A bio-inspired trusted clustering for mobile pervasive environment. Published by Springer series Advances in Intelligent Systems and Computing AISC, ISSN-2194-5357, Proceedings of the Third International Conference on Soft Computing for Problem Solving (SocPros2013), 259:553-564
26. <https://www.idigitalhealth.com/news/with-860k-affected-patients-july-among-worst-data-breach-months-of-year?rel=0>
27. <https://healthitsecurity.com/news/medical-device-security-critical-with-fda-interoperability-guide>
28. Aufner P (2020) The IoT security gap: a look down into the valley between threat models and their implementation. *Int J Informat Security* 19:3-14. <https://doi.org/10.1007/s10207-019-00445-y>
29. Shayesteh B, Hakami V, Akbari A (2020) A trust management scheme for IoT-enabled environmental health/accessibility monitoring services. *Int J Informat Security* 19:93-110. <https://doi.org/10.1007/s10207-019-00446-x>
30. Ozawa S, Ban T, Hashimoto N et al (2022) A study of IoT malware activities using association rule learning for darknet sensor data. *Int J Informat Security* 19:83-92. <https://doi.org/10.1007/s10207-019-00439-w>
31. Jing Q, Vasilakos AV, Wan J, Lu J, Qiu D (2018) Security of the Internet of Things: perspectives and challenges. *Wireless Netw* 20(8):248
32. <https://doi.org/10.1007/s11276-014-0761-7>
33. Federal Trade Commission (2016) Internet of Things: privacy and security in a connected world. FTC Staff Report
34. Brian Mosley (2017) NSF Funded IoT Security Research Excites at the 2017 CNSF Exhibition. <http://cra.org/govaffairs/blog/2017/05/2017-cnsf-exhibition/>.
35. United States Department of Health and Human Services Office for Civil Rights. n.d.. Summary of the HIPAA Security Rule. (n.d.). <https://www.hhs.gov/hipaa/forprofessionals/security/index.html>
36. Alshammari M, Simpson A (2017) Towards a principled approach for engineering privacy by design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, LNCS, vol 10518, p 161. [https://doi.org/10.1007/978-3-319-67280-9\\_9](https://doi.org/10.1007/978-3-319-67280-9_9)
37. Ko I, Chambers D, Barrett E (2020) Feature dynamic deep learning approach for DDoS mitigation within the ISP domain. *Int J Informat Security* 19:53-70
38. Jayasinghe U, Otebolaku A, Um TW, Lee GM (2017) Data centric trust evaluation and prediction framework for IoT. In: *ITU Kaleidoscope: Challenges for a Data-Driven Society (ITU K)*, pp 1-7. IEEE
39. Dao NN et al (2017) Securing heterogeneous IoT with intelligent DDoS attack behavior learning. [arXiv:1711.06041](https://arxiv.org/abs/1711.06041)
40. Sultana N, Chilamkurti N, Peng W, Alhadad R (2018) Survey on sdn based network intrusion detection system using machine learning approaches. *Peer-to-Peer Netw Appl*. <https://doi.org/10.1007/s12083-017-0630-0>
41. Chourasiya R et al (2018) Classification of cyber attack using machine learning technique at microsoft azure cloud. *Int Res J Eng Appl Sci*

# Prediction of the Risk of Heart Attack Using Machine Learning Techniques



Pinaki Ghosh, Umesh Kumar Lilhore, Sarita Simaiya, Atul Garg, Devendra Prasad, and Ajay Kumar

**Abstract** Heart attack is the number one cause of global mortality. As per World Health Organization (WHO), 17.9 million lives are lost each year from heart-related diseases, representing 32% of all deaths worldwide. The cases of heart attacks are increasing day by day at an alarming rate. It is essential to predict any such risks. This paper is focused on which patients are more likely to have heart disease. In this paper, numerous machine learning models for anticipating the risk of heart attack are addressed. Such algorithms are implemented on a wide range of datasets and the results are compared.

---

P. Ghosh · U. K. Lilhore (✉) · S. Simaiya · A. Garg · D. Prasad · A. Kumar  
Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India  
e-mail: [umeshlilhore@gmail.com](mailto:umeshlilhore@gmail.com)

P. Ghosh  
e-mail: [pinaki.ghosh@chitkara.edu.in](mailto:pinaki.ghosh@chitkara.edu.in)

S. Simaiya  
e-mail: [sarita.simaiya@chitkara.edu.in](mailto:sarita.simaiya@chitkara.edu.in)

A. Garg  
e-mail: [atul.garg@chitkara.edu.in](mailto:atul.garg@chitkara.edu.in)

D. Prasad  
e-mail: [devendra.prasad@chitkara.edu.in](mailto:devendra.prasad@chitkara.edu.in)

A. Kumar  
e-mail: [akumar@chitkara.edu.in](mailto:akumar@chitkara.edu.in)

P. Ghosh  
Institute of Advance Computing, SAGE University, Bhopal, India

U. K. Lilhore  
KIET Group of Institutions, Delhi-NCR, Ghaziabad, India

D. Prasad  
Department of Computer Science and Engineering, Panipat Institute of Engineering & Technology, Panipat, India

U. K. Lilhore  
KIET Group of Institutions, Delhi-NCR Ghaziabad, (UP), India

**Keywords** Machine learning · Prediction model · Heart attack prediction · Cardiovascular diseases · Decision tree · Random forest · Naive Bayes · Multilayer perceptron

## 1 Introduction

Approximately 18 million people are killed each year globally due to cardiovascular diseases, around 32% of all deaths worldwide. Heart attack is the primary cause of 85% of these deaths [1]. The rate of heart-related diseases is increasing rapidly in low- and middle-income countries because of the lack of primary healthcare facilities for early detection and treatment. Heart failure is usually caused by high blood pressure, high blood sugar, irregular heartbeats, and other conditions [2]. There is a need for a system that can predict the risk of heart attack based on any person who has given biological and medical data.

The evolution of the Internet of Things [3, 4] makes it very easy to collect medical data in real time. With various sensors and actuators, medical data such as heart rate, blood pressure, ECG, and oxygen saturation level can be collected as and when necessary. This data can be transmitted and processed at remote data centers using IoT devices. Also, it can trigger alarms and inform the nearest medical facility in case there some anomaly has been detected [5].

Machine learning is broadly used in various business and automation applications, including market prediction analysis, fraud detection, e-commerce, customer recommendation systems, behavior detection, etc. [6, 7]. Machine learning (ML) techniques and algorithms are primarily used to predict problems. In this paper, we are using comparisons of various machine learning methods to predict a heart attack. This complete article is divided into several sections that cover related work, methodology, result, discussion, and conclusion.

## 2 Related Work

Various authors have carried out a significant amount of work related to the prediction of heart disease using machine learning. Various machine learning techniques, including logistic regression, KNN, and random forest, have been used. In this section, some of the works are discussed in brief.

Mohammed Abdul Khaleel [8] has given a survey of various mining techniques on medical data for finding frequent diseases. It shows various mining techniques to find diseases like lung malignancy, heart weakness, bosom disease, etc. The author used the Naïve Bayes algorithm and the Weka tool for classification. The dataset has been taken from a diabetic research institute in Chennai, India, and contains medical data of 500 + patients.

Nabil Alshurafa et al. [9] have suggested a remote health monitoring system, Wanda-CVD. It is a smartphone-based system that provides wireless coaching and support to users. It provides a cost-consciousness solution for medical outcome success prediction. This system has been tested on a group of women aging between 25 and 45 years and achieved an F-measure of 75.4%.

Sathish Kumar and Padmapriya [10] proposed an idea for predicting common diseases through traditional media like television and mobile phone. This framework uses the ID3 algorithm to help people know about diseases and helps reduce the death rate.

Nishara et al. [11] developed a prediction system to find the accuracy of the risk level of the heart patients. The authors used classification and clustering algorithms for the prediction of heart diseases. This system uses different data mining techniques for predicting the risk of heart disease.

An intelligence system has been proposed by Wiharto et al. [12] to diagnose coronary heart disease considering the data imbalance problem. This system uses random sampling, clean data out of range, synthetic minority oversampling, and duplicate removal techniques. It uses a k-fold cross-verification and k-star multiclass classification algorithm. This system achieved an 80.1% F-measure value.

Jayshril and Patil [13] suggested a framework for heart infection prediction using learning vector quantization. A neural-based system uses 13 clinical information-like features and predicts the presence of heart disease with different performance measures. The research article [14] defined three procedures for conducting a comprehensive evaluation, and exciting results were obtained. Researchers discovered that ML algorithms significantly improved inside this evaluation.

In the research article [15], numerous studies have extensively suggested that perhaps motion will remain in motion and really ought to use ML whenever the data source is small, whereby this paper proves. The processing complexity had also been lowered, which is useful when dispatching a framework.

### 3 Methodology

The main focus of this research is to identify a machine learning technique that can correctly predict the risk of heart attack based upon the given data. The following machine learning algorithms are used for this research.

- **Naïve Bayes:** This approach is based on Bayes' theorem. This is a relatively simple classification algorithm [16]. Amidst its architecture and simplifications, this classifier has continued to work so well in numerous operating practical applications.
- **Logistic Regression** is a logistic function-based statistical model for a binary variable. It uses a logistic approach to calculate a type of binary dependent variable [17].

- **Random Forest:** These are collaborative methods that use a large decision tree to perform Regression, classification, and other activities during training. The classification outcome is just the class chosen by the majority of the tree branches. This is the measure or median of the specific tree's forecasting in Regression [19].
- **Multilayer Perceptron:** The multilayer perceptron, or MLP, is a type of feed-forward artificial neural network (ANN). A multilayer perceptron consists of a minimum of three layers of nodes; the first layer is an input layer, followed by one or more layers of nodes known as a hidden layer, and the last layer is the output layer. Each node except the nodes in the input layer uses a non-linear activation function. MLP uses a supervised learning technique known as a back-propagation training algorithm [20].

## 4 Dataset

The dataset is collected from the University of California Irvine (UCI) Machine Learning Repository [18]. The data in this dataset has been initially collected from the Faisalabad Institute of Cardiology, Faisalabad, Pakistan. It contains records of 299 heart patients consisting of 194 males and 105 females ranging between 40 and 95 years. There is a total of 13 attributes or features showing a report of clinical, lifestyle, and physical information of a patient (Table 1). Most features are numerical data (age, creatinine phosphokinase, ejection fraction, platelets, serum creatinine, serum sodium, and time). The quantitative statistical description of the category features and numerical features are shown in Tables 2 and 3.

**Table 1** Data dictionary

Feature	Description
Sex	Gender of patient
Age	Patient's age
Time	Observation period
Smoking	The patient is a smoker or not
Diabetes	If the patient has diabetes or not
Platelets	Blood platelets
Blood pressure	If the patient has high blood pressure or not
Anemia	Deficiency of RBC in blood
Creatinine phosphokinase	Level of CPK in blood
Ejection fraction	Percentage of blood ejected at each contraction cycle
Serum sodium	Blood sodium level
Serum creatinine	Blood creatinine level
Death event	If the patient died due to a heart attack during the observation period

**Table 2** Statistical quantitative description of category features

Feature	Samples		Dead		Survived	
	#	%	#	%	#	%
Anemia (yes)	129	43.14	46	35.66	83	64.34
Anemia (no)	170	56.86	50	29.41	120	70.59
Diabetes (yes)	125	41.81	40	32.00	85	68.00
Diabetes (no)	174	58.19	56	32.18	118	67.82
High BP (yes)	105	35.12	39	37.14	66	62.86
High BP (no)	194	64.88	57	29.38	137	70.62
Sex (male)	194	64.88	62	31.96	132	68.04
Sex (female)	105	35.12	34	32.38	71	67.62
Smoking (yes)	96	32.11	30	31.25	66	68.75
Smoking (no)	203	67.89	66	32.51	137	67.49

Sample size: 299 patients, Survived patients: 203, Dead patients: 96

**Table 3** Statistical quantitative description of numerical features

Feature	Samples		Dead		Survived	
	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
Age	60.834	11.895	65.250	13.132	58.793	10.658
Creatinine phosphokinase	581.839	970.288	666.577	1309.798	539.433	753.347
Ejection fraction	38.084	11.835	33.215	12.551	40.031	11.625
Platelets (K)	263.358	97.804	256.406	98.139	266.662	97.329
Serum creatinine	1.394	1.035	1.821	1.457	1.185	0.656
Serum sodium	136.625	4.412	135.471	4.977	137.341	3.924
Time	130.261	77.614	70.847	62.070	158.264	67.563

Sample size: 299 patients, Survived patients: 203, Dead patients: 96

WEKA (Waikato Environment for Knowledge Analysis) version 3.8.5 is used to implement the machine learning algorithms [22–25]. Figure 1 shows the class view of all the attributes (features). The risk of a heart attack upon patients' age has been shown in Fig. 2.

The dataset is divided into a ratio of 60:40 for training and testing. The dataset is then compared with Naïve Bayes, Random Forest, multilayer perceptron, Logistic Regression, and K\* algorithms.

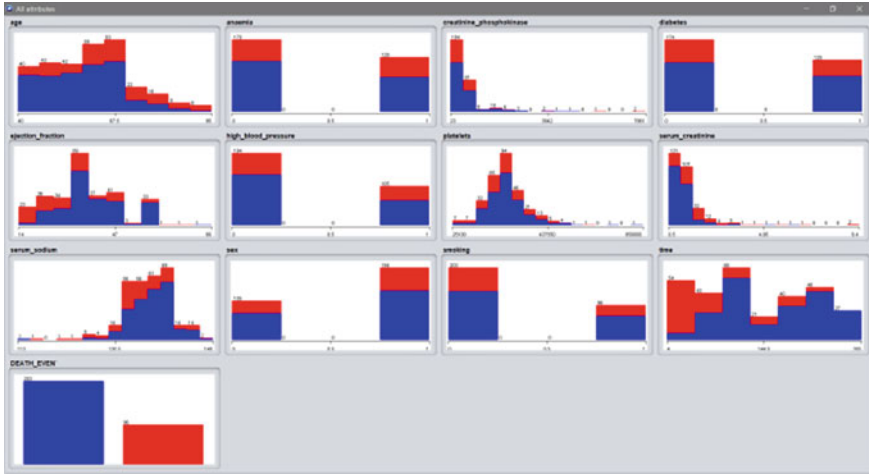


Fig. 1 Visualization of attribute class view

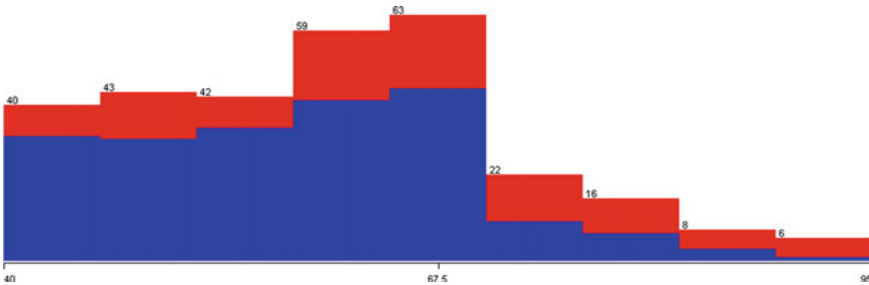


Fig. 2 Risk of heart attack based upon age

### 5 Results and Discussions

The given dataset has been trained and tested on different algorithms; Naïve Bayes, random forest, multilayer perceptron, Logistic Regression, and K\*. The confusion matrix of these algorithms is shown in Fig. 3.

The comparative result of various machine learning algorithms for predicting heart attack has been shown in Table 4 and Fig. 4.

Based on the experimental results, it is clear that the logistic regression algorithm has better precision, recall, and f-measure among these five machine learning techniques for the given dataset. Logistic Regression has classified 104 instances correctly out of 120 test instances, giving an accuracy of 86.7%. The random forest has also given very close prediction by correctly classifying 101 instances out of 120 with an accuracy of 84.2%.

```

a b <-- classified as
83 5 | a = no
17 15 | b = yes
a. Naïve Bayes

a b <-- classified as
82 6 | a = no
10 22 | b = yes
b. Logistic Regression

a b <-- classified as
72 16 | a = no
16 16 | b = yes
c. Multilayer Perceptron

a b <-- classified as
81 7 | a = no
12 20 | b = yes
d. Random Forest

a b <-- classified as
66 22 | a = no
20 12 | b = yes
e. K*
    
```

**Fig. 3** Confusion matrix

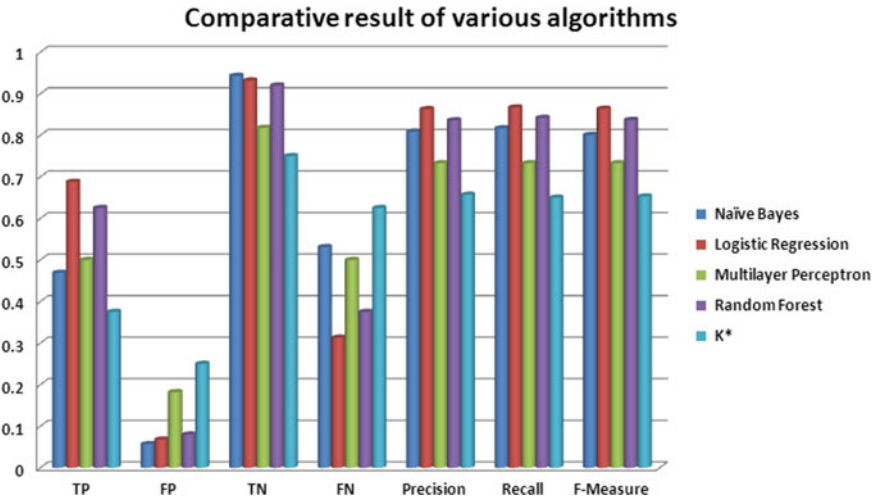
**Table 4** Comparative result of various algorithms

Classifier	TP	FP	TN	FN	Precision	Recall	F-Measure
Naïve Bayes	0.469	0.057	0.943	0.531	0.809	0.817	0.801
Logistic Regression	0.688	0.068	0.932	0.313	0.863	0.867	0.864
Multilayer Perceptron	0.500	0.182	0.818	0.500	0.733	0.733	0.733
Random Forest	0.625	0.080	0.920	0.375	0.836	0.842	0.837
K*	0.375	0.250	0.750	0.625	0.657	0.650	0.653

## 6 Conclusion

Predicting medical conditions and their outcomes is always a challenge for scientists and medical practitioners. In this paper, the main focus is solving a challenging heart attack prediction problem using various machine learning techniques. In this work, different ML techniques are applied to the same dataset, and found logistic regression algorithm gives an accuracy of 86.7%. Using a large dataset ensures higher chances of accuracy. Cleaning the dataset and applying multiple machine learning techniques can improve prediction accuracy.





**Fig. 4** Comparative result of various algorithms

## References

1. World Health Organization (WHO), Cardiovascular Diseases (CVDs), Fact sheets, 11 June 2021. <https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-cvds>. Accessed 28 June 2021
2. National Heart, Lung, and Blood Institute (NHLBI), “Heart Failure,” <https://www.nhlbi.nih.gov/health-topics/heart-failure>. Accessed 28 June 2021
3. Datta P, Sharma B (2017) A survey on IoT architectures, protocols, security, and smart city-based applications. In: 8th international conference on computing, communication and networking technologies (ICCCNT), Proceedings, pp 1–5. <https://doi.org/10.1109/ICCCNT.2017.8203943>
4. Ghosh P, Mahesh TR (2015) Smart City: Concept and Challenges. *Int J Advanc Eng Technol Sci (IJAETS)*, **1**(1): 25–27
5. Ghosh P, Prasad D, Guleria K (2020) An m-IoT framework for remote monitoring of ECG signals. *J Advanc Res Dynam Control Syst (JARDCS)* **12**(8):296–300. <https://doi.org/10.5373/JARDCS/V12I8/20202477>
6. Sarker IH (2021) Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci* **2**(160). <https://doi.org/10.1007/s42979-021-00592-x>
7. R Chhabra, S Verma, and C R Krishna, “A survey on driver behavior detection techniques for intelligent transportation systems,” 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence, 2017, pp. 36–41. <https://doi.org/10.1109/CONFLUENCE.2017.7943120>
8. Mohammed Abdul Khaleel (2013) Sateesh Kumar Pradhman and G N Dash, “A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases.” *International Journal of Advanced Research in Computer Science and Software Engineering* **3**(8):149–153
9. Alshurafa N, Sideris C, Pourhomayoun M (2017) HaikKalantarian, Majid Sarrafzadeh, and Jo-Ann Eastwood, “Remote Health Monitoring Outcome Success Prediction using Baseline and First Month Intervention Data.” *IEEE J Biomed Health Inform* **21**(2):507–514. <https://doi.org/10.1109/JBHI.2016.2518673>

10. L Sathish Kumar and A Padmapriya, "Prediction for Common Disease using ID3 Algorithm in Mobile Phone and Television", *International Journal of Computer Applications*, 50(4), 2012, pp. 30–33
11. M A Nishara Banu and B Gomathy, "Disease Forecasting System using Data Mining Methods", *International Conference on Intelligent Computing Applications (ICICA)*, 2014, doi: <https://doi.org/10.1109/ICICA.2014.36>
12. Wiharto, Hari Kusnanto, and Herianto, "Intelligence System for Diagnosis Level of Coronary Heart Disease with K-Star Algorithm", *Healthcare Informatics Research*, 22 (1), 2016, DOI: <https://doi.org/10.4258/hir.2016.22.1.30>
13. Jayshril S Sonawane and D R Patil, "Prediction of Heart Disease using Linear Vector Quantization Algorithm", *Conference on IT in Business, Industry, and Government (CSIBIG)*, 2014, doi: <https://doi.org/10.1109/CSIBIG.2014.7056973>
14. S Madeh Piryonesi and Tamer E El-Diraby, "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems", *Journal of Transportation Engineering, Part B: Pavements*, 146 (2), 2020, doi:<https://doi.org/10.1061/JPEODX.0000175>.
15. Juliana Tolles and William J Meurer, "Logistic Regression Relating Patient Characteristics to Outcomes", *The Journal of the American Medical Association (JAMA)*, 316 (5), 2016, pp. 533–4, doi:<https://doi.org/10.1001/jama.2016.7653>.
16. Trivedi NK, Simaiya S, Lilhore UK, Sharma SK (2020) An efficient credit card fraud detection model based on machine learning methods. *International Journal of Advanced Science and Technology* 29(5):3414–3424
17. Patil V, Lilhore UK (2018) A survey on different data mining & machine learning methods for credit card fraud detection. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology* 3(5):320–325
18. UCI Machine Learning Repository, "Heart Failure Clinical Records Dataset", 5 February 2020, Available at: <https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records>, (Accessed on 27 June 2021).
19. Pawar N, Lilhore UK, Agrawal N (2017) A hybrid ACHBDF load balancing method for optimum resource utilization in cloud computing. *Int J Scientif Res Comput Sci Eng Informat Technol (IJSRCSEIT)*, ISSN 2456, 3307:367–373
20. Guleria K, Sharma A, Lilhore UK, Prasad D (2020) Breast cancer prediction and classification using supervised learning techniques. *J Comput Theor Nanosci* 17(6):2519–2522
21. Lilhore UK, Simaiya S, Guleria K, Prasad D (2020) An efficient load balancing method by using machine learning-based VM Distribution and dynamic resource mapping. *J Comput Theor Nanosci* 17(6):2545–2551
22. Sharma SK, Lilhore UK, Simaiya S, Trivedi NK (2021) An improved random forest algorithm for predicting the COVID-19 pandemic patient health. *Ann Romanian Soc Cell Biol*:67–75
23. Lilhore UK, Simaiya S, Prasad D, Verma DK (2021) Hybrid weighted random forests method for prediction & classification of online buying customers. *J Informat Technol Manag* 13(2):245–259
24. Trivedi NK, Simaiya S, Lilhore UK, Sharma SK (2021) COVID-19 Pandemic: role of machine learning & deep learning methods in diagnosis. *Int J Cur Res Revl* 13(06):150
25. Simaiya S, Lilhore UK, Prasad D, Verma DK (2021) MRI brain tumour detection & image segmentation by hybrid hierarchical K-means clustering with FCM based machine learning model. *Ann Romanian Soc Cell Biol*:88–94

# An Integrated CBIR Approach for Medical Image Retrieval System



Anubhav Sharma and Shiv Shakti Shrivastava

**Abstract** This paper introduces an information retrieval model to solve the semantic gap problem in content-based image retrieval. The proposed model aims to eliminate the semantic gap by adding image descriptors to the vector space model approach, often used in text information retrieval. In this way, semantically inadequate image descriptors are expressed in the same place as textual concepts. As a result, a relationship between textual concepts and image descriptors can be established automatically. The proposed model was evaluated on the *Image CLEFmed* dataset. The evaluation results are promising and outperform existing techniques.

**Keywords** Image retrieval · Content-based image retrieval · Deep learning · Information retrieval · Medical image archives

## 1 Introduction

Information Retrieval Systems (IR) aim to provide access to undefined data. On the other hand, information systems are designed and implemented on defined data, for example, in a hospital information system, such as patient number, patient name, test date, etc. Such data with a defined structure are usually stored in Database Management Systems (DBMS). Data with the undefined design usually appear as reports containing documents written in plain text, and DBMS is insufficient to search on such data. Although such documents can be stored in database systems, database systems only support standard SQL queries to access this type of information. With this approach, it is not possible to meet the need for information. Information retrieval, the sub-branch of computer science, which is the subject of access to undefined or semi-defined data, deals with the systems developed for this purpose, and the techniques developed for this purpose are called the Information Retrieval System [1–3].

---

A. Sharma (✉) · S. S. Shrivastava  
Department of Computer Science, Rabindra Nath Tagore University, Bhopal, India  
e-mail: [anubhav.sharma0025@gmail.com](mailto:anubhav.sharma0025@gmail.com)

Today, there are many image sources for diagnosis and treatment in hospitals. These may originate from X-ray [4], CT, MRI [5], ultrasound, nuclear medicine, ophthalmology, cardiology, pathology and gastroenterology. Regardless of their source, the data in a Medical Image Archive (MIA) application can be grouped into three main groups: film, still image, and video. The purpose of MIAs is to store these three types of data groups in computer-based environments, run queries on these data, retrieve them when desired, and present them to the user as they wish [6, 7].

Today, access to images stored in medical image archives (MIA) is provided by external attributes such as patient names and file number associated with these images. However, Content-Based Image Access (CBIA) requires the file number, patient name, surname, etc., associated with the images. In addition to external features such as shape, texture, and color distribution also enable access according to internal components such as color distribution. This technique is based on the basic principle of using the contents of the images directly in the archive in querying and retrieval. In this way, images in the library are not just passive objects stored in the database but become active objects directly participating in the query and retrieval process [8, 9]. A typical SQL query would be the user retrieving other images from the archive that include the texture or shape of a tumor in any image. The use of CBIA in MIA applications is a relatively new concept, and it is likely to find a place in computer-aided diagnosis and evidence-based medicine [10–12].

Although the success of the CBIA is sufficient for some specialized restricted tasks, there is still a semantic gap between the high-level concepts (such as a tumor, abnormal tissue) that the user naturally wants to search for and low-level visual features (texture, color). Successful access results are expected from a combination of textual and visual attributes for medical image access. In addition to the automatically extracted visual features, descriptions that describe the image are a possible way to close the semantic gap. However, it is also necessary to determine which pairs of visual and textual attributes will best optimize the search results [13–15].

More importantly, there is a strong need for an integrated CBIA model that combines both low-level visual descriptors and high-level textual descriptions at the same time. Because, with the help of a new model designed in this way, low-level image descriptors and textual terms can be used together. In this way, the success of the CBIA improves with the help of textual representations, and the semantic space problem of the CBIA can be overcome by representing visual terms and textual terms in the same space [16–18]. In addition, with the help of an integrated CBIA model, the way is cleared for the automatic assignment of semantic concepts corresponding to lower level descriptors.

This study proposes a new model called Integrated Information Retrieval Model (IIRM), which aims to model high-level textual terms and low-level image descriptors in the same mathematical space. The proposed model can also be considered as a solution to the CBIA semantic space problem. The model tries to overcome the inadequacy of image descriptors in identifying semantic expressions using textual information associated with the image. The performance of the IIRN has been tested on a dataset Image CLEFMed and found to be better than known existing methods.

The remainder of this study is organized as follows: In the second chapter, the Integrated Information Retrieval Model (IIRM) is explained in the light of existing information access models. In the third part, the results obtained from the experiments conducted to test the proposed model are presented. In the fourth section, the results obtained from the study and the discussion about these results are presented. In addition, efforts have been made to shed light on future studies in this area.

## 2 Content-Based Medical Image Access System

Content-based image access systems aim to search for images in an image database closest to a given query image. Here, the query is made entirely with picture content, without using any textual information. In these systems, the query process is given to the design and the system is asked to bring pictures similar to this sample image. This approach is called questioning by example and was first proposed by Nib Lack and his friends in the early 1990s when they introduced the query system based on picture content (QPC). As a result, content-based image access systems and query-by-sample concepts aim to extract similar data from extensive data rather than summarize existing data. In addition, the methods used in content-based image access systems are similar to data mining in image databases [19–21].

Today, the ancestor of content-based video access systems is the IBM-QPC system. As mentioned above, this system has brought the concept of questioning according to the example. Another commercial system is Virago, which CNN uses. Despite all this, most content-based video access systems are being developed academically. Examples of these systems are Candid, Photo book, and Nitra, all of which use basic texture and color characteristics to describe image content. Using high-level information such as using segmented parts for querying has been suggested by the blob world system. In this system, the picture is divided into ellipsoids, and an interpretation layer has been developed by using the color or texture properties of these ellipsoids [22, 23].

Most of these systems have similar architecture, and all systems try to describe the content of the images using visual descriptors. Visual descriptors can be examined in three groups: color, texture, and shape. The other parts that make up the system are the storage access manager, which ensures efficient storage of the extracted features of the images and quick access to this information when necessary, and the Graphical User Interface which enables the query results and the query to be transmitted to the system, and the visual descriptor vector of the query image and the visual descriptor of the images in the image archive. It is the measurement metric that measures the distance between vectors. All systems use one of the available algorithms for these components [3].

## **2.1 Visual Descriptors**

As mentioned above, content-based image access systems use visual descriptors to describe the content of images. These descriptors can be examined in three primary groups: color, texture, and shape. Also, besides these primary groups, logical identifiers such as the identification of objects in the image and summary descriptors such as the importance of the image are available [24]. None of the image access systems can provide intermediate information obtained by searching through textual information using this information. This problem is called semantic space. Although the semantic gap causes the rejection of this whole structure, image access systems can be maintained as long as the user is aware of the benefits and problems provided by the system.

### **2.1.1 Color Identifiers**

Color descriptors are one of the most preferred and most compelling visual descriptors in content-based image access systems. Because almost all systems contain color. Today, all images are encoded using RGB color space for storage. However, in image access systems, this space is not preferred to index the images using color descriptors. RGB color space is a color space far from human perception [5]. Here, the proximity or distance to human perception means that humans perceive color changes in the color space in the same way. This is why image access systems convert the images they contain into CIE Lab or HSV color spaces closer to human perception to index. Among these descriptors, the most preferred is to compare the color distribution of the images. They use the approach “Images with close color distribution can be similar images”. However, the biggest problem of this approach is that, because the color distribution of an image is independent of the location information, the images that are positionally very different from each other can be evaluated as similar to each other [4].

There is no color information in the images produced in the medical field. Still, the images use a grayscale well above the grayscale level accepted by the existing color spaces (the number of gray shades accepted by color spaces is 256 and the number of gray shades used is 4096). Therefore, color descriptors cannot represent images produced in the medical field alone [15].

### **2.1.2 Texture Descriptors**

Although the exact visual texture cannot be defined, texture descriptors have more variety than color descriptors. The most important of these are wavelet and Gabor filters. Gabor filters are the most preferred texture descriptors because they are more successful than wavelets and define edge detection better than human perception. Texture descriptors are descriptors that try to identify the characteristics of the

changes in specific directions in the image and the dimensions of these changes and describe the image in the light of this information [26].

### 2.1.3 Shape Descriptors

To use these descriptors, the objects in the image must be extracted entirely. However, extracting the things contained in the image from the image is a process that has not been solved yet. In image retrieval, many systems try to extract the objects of the image they contain to extract the shape descriptors and use the resulting parts' color and texture properties to perform a more efficient extraction process [2].

Many people have suggested methods to remove medical images. Therefore, after the obtained medical image is extracted, the image can be defined with shape descriptors independent of shift, rotation, and size [25].

## 3 Application of Content-Based Image Retrieval Systems in the Medical Field

Images obtained in medical clinics in our country cannot be stored digitally, except in large hospitals or university hospitals. However, the storage of digital images obtained in clinics in most countries is identified as an essential economic and clinical factor for the hospital environment. The current image archives used today use DICOM images and the title information contained in these images. These systems use some of the DICOM header information for image access, such as patient number, patient location, and image type. However, this information is not sufficient for an effective search. In addition, despite the increase in DICOM-compatible imaging devices, an error rate of up to 16% was found in this information in the research [6]. Therefore, the need for new image access techniques for existing image archives has increased.

Medical images have often been used for image access systems. They are usually recommended as base images for content-based image access technologies because of their effects on medical images. However, there is still not enough work on the performance and definition of these systems. Assert and IRMA are medical image access systems that are currently being studied. Assert is designed to classify and access high-resolution lung tomography images.

On the other hand, IRMA classifies medical images according to the anatomical region, type, and patient's location without discrimination and compares similarities within the classes it has created [4]. Digital images are produced in many medical fields. Studies on the classification of dermatological images are carried out. Content-based image access methods have been proposed because the color and texture descriptors of pathological images can be distinguished easily. In this way, the task of a pathologist is to use a content-based image access system rather than mixing reference books and finding similar events. Mammon graphics are among the most

frequently used image types in the radiology department for content-based image access systems. The aim here is to reduce the postoperative negative psychological impact in patients with incorrect diagnosis [6].

### **3.1 MPEG-7**

Nowadays, multimedia data is increasing at a rapid rate in both personal and commercial environments. Therefore, access to this information is getting more and more difficult. As mentioned above, content-based access systems have been proposed for this purpose, and some visual descriptors that these systems can use have been determined. However, there is still no standard for these systems, and each system uses the visual identifier it chooses, with the storage type it chooses. MPEG-7 is a standard proposed by the MPEG group to overcome this deficiency. The goal of MPEG-7 is to create shareable multimedia databases on the Internet and facilitate access to all multimedia data in the world through these databases [26].

MPEG-7 is a standard developed for describing multimedia content. MPEG-7 is not intended to restrict applications using this standard to a specific area. That is, MPEG-7 is not a standard based on a particular application, but a standard that makes it possible to develop applications in as many fields as possible [5]. MPEG-7 has to be a flexible and extensible standard since it supports many applications in many areas. Therefore, MPEG-7 is not a fixed set of components but rather a set of methods and tools for defining visual and audio content. Therefore, MPEG-7 is designed as general as possible. XML technology has been used to display the descriptors with MPEG-7 standard in the text to ensure independence from the platform.

The extraction and querying of descriptors, which are parts of content-based multimedia access systems, are closely related to MPEG-7. However, MPEG-7 does not set a standard for neither sorting nor querying descriptors. MPEG-7 aims to keep the identifiers obtained independently from the platform and standardize how they are stored. Elements contained in MPEG-7 address a wide range of applications. In addition, MPEG-7 enables the search systems, which can be performed textually on the Internet, to be performed on multimedia data. Other usage areas of MPEG-7 are Interior Architecture, TV Broadcast Selection, Medical Applications, Educational Applications, Cultural Services, and Digital Libraries.

## **4 Related Works**

It has recently become essential to use textual properties to increase the performance of the CBIA systems. In this study too, we propose a new integrated information access model. The method presents the use of low-level image descriptors and the features extracted from the texts attached to the images, and in this way, it tries to solve the CBIA's semantic space problem.



There are several studies on this subject in the literature. Ahmed and Malebary [10] conducted a survey proposing to associate image descriptors with medical concepts with the help of the Unified Medical Language System (UMLS) [4] dictionary. In the study, textual information was indexed using the vector space model (VSM) [13]; images have also tried to match the UMLS terms both in general and locally with the help of support vector machines (SVM).

Dai et al. [16] created visual dictionaries and aimed to create an information retrieval system with the help of these dictionaries. In the proposed method, images are first divided into small pieces using a segmentation algorithm and image descriptors extracted from each sample. The extracted identifiers are clustered with a clustering algorithm, and the representative vector of each cluster is called a visual term. In this way, any image is expressed with clusters of visual descriptors instead of visual descriptors in the proposed system. However, since textual or semantic concepts are not used in this approach, it cannot be said precisely that the semantic gap is eliminated. In addition, the success of the system is directly proportional to the success of the clustering algorithm used.

Peng et al. [23], on the other hand, suggested using a probabilistic model to describe visual terms in their work. With the help of this model, the descriptors extracted from the images are grouped, and each group is tried to be expressed with a probability distribution. The work is similar to the word bag method in the textual field, as the technique breaks down the images into a grid. However, the proposed method requires a lot of computational loads limiting the applicability of the system.

## 5 Technique

The IIRM proposed in this study is suitable for archives that contain both image and textual descriptions. For this reason, firstly, how the textual cerise are indexed with the classical vector space model, then how the descriptors extracted from the images are integrated with the textual vector space model are presented.

### 5.1 Vector Space Model

Textual data are indexed with VSM. In VSM, the report data of each case is expressed with a term vector. A report archive becomes a matrix made up of report (document) vectors. This matrix is called the document matrix,  $D$ , and the rows of this matrix are called document vectors and its columns are term vectors.

$$D = \begin{bmatrix} W_{11} & \cdots & W_{1n} \\ \vdots & \ddots & \vdots \\ W_{n1} & \cdots & W_{nn} \end{bmatrix} \quad (1)$$

Here,  $w_{ij}$  indicates the weight of the  $j$  term in the  $i$ th document,  $n$  indicates the total number of terms in the system and the total number of documents. In the literature, many weighting models have been proposed to express the weighting of textual terms in documents. In our study, the axial document length normalization model was used [8, 9]. According to this model, how much a term will weigh for the document is Eq. 2.

$$W_{ij} = \frac{\log(dtf) + 1}{\text{Sum}dtf} * \log \frac{N - n_f}{n_f} * \frac{u}{1 + 0.115u} \tag{2}$$

Here,  $dtf$  refers to the number of times the term occurs in the document,  $\text{sum}dtf$  refers to the sum of the expression for all terms in  $\log(dtf) + 1$  the document,  $N$  is the total number of documents, the total number of documents containing  $n_f$  terms, and the number of unique terms in the document in  $u$ .

## 6 Integrated Information Retrieval Model (IIRM)

IIRM is based on the integration of lower level descriptors of images into VSM. It aims to process image descriptors and textual terms in the same space and establish automatic relationships between image descriptors and textual representations. Ultimately, the  $D$  matrix transforms into Eq. 3, including the image properties.

$$D = \begin{bmatrix} w_{11} & \dots & w_{1n} & i_{1,n+1} & \dots & i_{1,n+k} \\ \vdots & \ddots & \vdots & | & \vdots & \ddots & \vdots \\ w_{1n} & \dots & w_{nn} & i_{m,n+1} & \dots & i_{m,n+k} \end{bmatrix} \tag{3}$$

Here,  $i_{pq}$  is the weight of the  $q$ th image term for the  $p$ th document and  $k$  means the total number of visual terms. In the study, visual terms and textual terms are normalized independently from each other.

Here,  $\theta$  shows the angle between the query vector and the  $i$ th document vector,  $n$  the total number of terms in the  $D$  matrix,  $q_j$  the weight of the  $j$ th term in the query vector, and  $b_{ij}$  the weight of the  $j$ th term in the  $i$ th document vector. Since the term weighting technique we use in this study is also a document vector normalization technique, textual term vectors can be considered normalized.

In this paper, two visual terms were used for experimental purposes. The first of these terms is the probability that dots in an image are gray. The other is the complement of this term: the possibility that the points in the image are colored. We took these two terms as two components of the point universe and added them directly to the  $D$  matrix. As a result, we calculated the similarity between the query and a document as follows:

$$Sim(\vec{q}, \vec{b}) = \sum_{j=1}^n q_j b_{ij} + \sum_{j=1}^k q_j b_{ij} \quad (4)$$

Here the value of  $k$  represents the total number of visual terms in the  $D$  matrix. Using the above equation, similarity values between the query vector and all documents in the archive are calculated. The resultant set was created by ordering all the similarity values other than zero calculated in descending order.

## 7 Experiments and Evaluations

In this paper, the proposed IIRM was tested on the Image CLEFmed dataset. The Image CLEFmed dataset is a data set that is regularly held every year and used within the Image CLEF competition, whose purpose is to measure the CBIA model and IR performance. This dataset includes images in articles published in a radiology journal, figure titles of these images, and article titles. The data set contains a total of 74,902 images and textual descriptions of these images. In addition, to evaluate the performance of the developed information retrieval systems, there are 25 queries grouped under three categories and an accuracy set containing the answers to these queries. Each query consists of a textual description and at least one image. As a result, the Image CLEFmed dataset is one of the most appropriate and up-to-date datasets for evaluating the performance of a CBIAS.

Before indexing the textual information in the data set, we standardized it through some processing. First, we converted the text in all documents to lowercase so that all textual expressions have the same letter mode. Since the dataset consists of medical records, we did not perform any word or number group elimination from the newly obtained documents. However, characters such as dash (-) and apostrophe (') have been removed in texts, and we have replaced the comma (,) and slash (/) characters with a space character. In this way, we tried to prevent problems that may occur due to possible spelling errors in the texts. For example, although x-ray and X-ray words mean the same term, they will be interpreted as two different terms because they are spelled differently. The above procedure will eliminate this difference, making the x-ray term the same as the X-ray term. After these operations were carried out, we separated the remaining textual expressions from the space character. We obtained words and created a document-term matrix so that each word was the term that make up the document. The number of terms in the document term matrix created is 33,613.

After indexing the textual expressions, we generated image descriptors for each image in the dataset, defining the probability that the points in the image are grayscale or color. We took the generated descriptors as terms for the related images and added them to the document-term matrix. This paper integrated only two simple visual descriptors with the textual information retrieval model [27].

After the indexing process was completed and we created the document-term matrix containing both textual and visual terms, we created the answer set by automatically running each query in the experimental set and sorting the result set according to the similarity with the query.

$$\text{Precision (P)} = \frac{\text{No. of documents suitable for the query in the result set}}{\text{No. of eligible documents in the result set}} \quad (5)$$

Average precision is the average of the system's sensitivity values for multiple queries, and  $n$  is expressed as follows to show the total number of queries:

$$\text{Average Precision (AvgP)} = \frac{1}{n} \sum_{i=1}^n P_i \quad (6)$$

It is essential to evaluate a system that the documents in the first order of the answer set are suitable for the query. For this reason, another performance criterion is the sensitivity of the documents brought in the first place. This criterion is called precision in the first  $n$  documents and is calculated as follows:

$$P@n = \frac{\text{No. of documents suitable for the query in the first } n \text{ documents}}{n} \quad (7)$$

The recall measure, on the other hand, shows the probability that the documents suitable for the query will be found in the answer set and is calculated as follows:

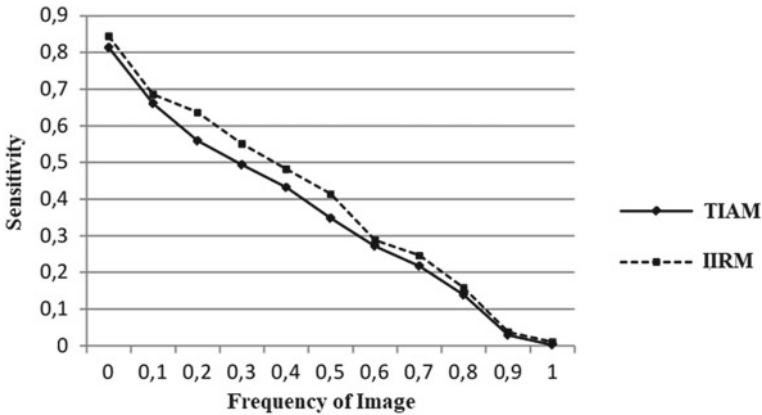
$$\text{Recall} = \frac{\text{No. of documents in the answer set suitable for the query}}{\text{No. of documents suitable for the query}} \quad (8)$$

Table 1 shows the comparison of IIRM and the textual information access model (TIAM). As can be seen from the table, while the Average Precision value for TIAM was 33%, the Average Precision value of the proposed model was 36%. Therefore, it is promising that only one visual descriptor, unlike classical textual information retrieval, positively affects the system's performance. Also, the table shows the sensitivity values in the first 5, 10, 30, and 100 documents of the answer set. The Precision@5 value of IIRM was measured as 0.632. In other words, 63% of the first five documents in IIRM were compatible with the given query, while this value was estimated as 58% in TIAM. When the table is examined, in general, it can be seen that the document sensitivity of IIRM in the first place is better than TIAM.

Figure 1 shows the sensitivity-remembrance graph of IIRM and TIAM. In the figure, the sensitivity rates of the TIAM at different rated levels are indicated with the straight line and the IIRM with the dashed line. As shown in the figure, IIRM showed better sensitivity values than TIAM at all nominal levels.

**Table 1** Comparison of TIAM and IIRM

Evaluation parameter	TIAM	IIRM
Mean average precision	0.339	0.368
Precision @ 5	0.584	0.632
Precision @ 10	0.520	0.544
Precision @ 30	0.448	0.483
Precision @ 100	0.303	0.324



**Fig. 1** Sensitivity-remembrance graph of IIRM and TIAM

## 8 Conclusion and Future Studies

In this paper, the integrated information access model (IIRM), which we proposed as a solution to the semantic space problem of the content-based image access (CBIA) model, was introduced. The proposed model defines a new integrated access method by combining the textual information access model with image descriptors. The process was tested with the trial set of the Image CLEFmed 2009 competition, which aims to measure the performance of the CBIA systems. The results obtained achieved the best rank in the competition area where textual and visual elements were used together.

In the experiments we have done, we have observed that IIRM’s success in the medical field is better than classical textual information access methods. The model improved the performance of textual information retrieval using simple image descriptors. The results support the accuracy of the method, and the future of the model is promising. We think we can get better results by adding new image descriptors to the technique.

One of the main goals of the IIRM, which is still in its infancy, is to close the semantic gap problem in the CBIA. In the proposed model, since image descriptors are located in the same space as semantic concepts, relationships between these

descriptors and semantic concepts can be established automatically with advanced information retrieval methods.

In today's hospitals, access to images is made only with patient information in the MIAS systems used in today's hospitals. On the other hand, the CBIA aims to find and retrieve images resembling a given image from the archive. In this way, medical image archives will cease to be systems where images are stored. In addition, archives that are accessed according to their image content can be used as an effective decision support system in evidence-based medicine and computer-aided diagnosis.

## References

1. Pengfei C (2019) Interactive image contents search based on high dimensional information theory. *IEEE Access* 7:141941–141946
2. Yang Y, Jiao S, He J, Xia B, Li J, Xiao R (2020) Image retrieval via learning content-based deep quality model towards big data. *Futur Gener Comput Syst* 112:243–249
3. Li X, Yang J, Ma J (2021) Recent developments of content-based image retrieval (CBIR). *Neurocomputing*
4. Haq NF, Moradi M, Wang ZJ (2021) A deep community based approach for large scale content based x-ray image retrieval. *Med Image Anal* 68:101847
5. Shijin Kumar P, Udaya Kumar N, Ushasree A, Sumalata G (2020) Key point oriented shape features and SVM classifier for content based image retrieval. *Mater Today: Proc*
6. Preethy Byju A, Demir B, Bruzzone L (2020) A progressive content-based image retrieval in jpeg 2000 compressed remote sensing archives. *IEEE Trans Geosci Remote Sens* 58(8):5739–5751
7. Yang X, Wang N, Song B, Gao X (2019) Bosr: a CNN-based aurora image retrieval method. *Neural Netw* 116:188–197
8. Shen M, Cheng G, Zhu L, Du X, Hu J (2020) Content-based multi-source encrypted image retrieval in clouds with privacy preservation. *Futur Gener Comput Syst* 109:621–632
9. Wang Y, Liu F, Pang Z, Hassan A, Lu W (2019) Privacy-preserving content-based image retrieval for mobile computing. *J Inf Secur Appl* 49:102399
10. Ahmed A, Malebary SJ (2020) Query expansion based on top-ranked images for content-based medical image retrieval. *IEEE Access* 8:194541–194550
11. Zhang C, Lin Y, Zhu L, Liu A, Zhang Z, Huang F (2019) CNN-VWII: an efficient approach for large-scale video retrieval by image queries. *Pattern Recogn Lett* 123:82–88
12. Hassan A, Liu F, Wang F, Wang Y (2021) Secure content based image retrieval for mobile users with deep neural networks in the cloud. *J Syst Arch* 116:102043
13. Ramos J, Kockelkorn TTJP, Ramos I, Ramos R, Grutters J, Viergever MA, van Ginneken B, Campilho A (2016) Content-based image retrieval by metric learning from radiology reports: application to interstitial lung diseases. *IEEE J Biomed Health Inform* 20(1):281–292
14. Yang K, Hua X, Wang M, Zhang H (2011) Tag tagging: towards more descriptive keywords of image content. *IEEE Trans Multimed* 13(4):662–673
15. Şaban Öztürk (2020) Stacked auto-encoder based tagging with deep features for content-based medical image retrieval. *Expert Syst Appl* 161:113693
16. Dai OE, Demir B, Sankur B, Bruzzone L (2018) A novel system for content-based retrieval of single and multi-label high-dimensional remote sensing images. *IEEE J Sel Top Appl Earth Obs Remote Sens* 11(7):2473–2490
17. Duan F, Zhang Q (2020) Stereoscopic image feature indexing based on hybrid grid multiple suffix tree and hierarchical clustering. *IEEE Access* 8:23531–23541
18. Tzelepi M, Tefas A (2018) Deep convolutional learning for content based image retrieval. *Neurocomputing* 275:2467–2478

19. Ahmed A (2020) Implementing relevance feedback for content-based medical image retrieval. *IEEE Access* 8:79969–79976
20. Zheng Y, Jiang Z, Zhang H, Xie F, Ma Y, Shi H, Zhao Y (2018) Size-scalable content-based histopathological image retrieval from database that consists of WSIs. *IEEE J Biomed Health Inform* 22(4):1278–1287
21. Ghrabat MJJ, Ma G, Abduljabbar ZA, Al Sibahee MA, Jassim SJ (2019) Greedy learning of deep Boltzmann machine (GDBM)’s variance and search algorithm for efficient image retrieval. *IEEE Access* 7:169142–169159
22. Xie G, Guo B, Huang Z, Zheng Y, Yan Y (2020) Combination of dominant color descriptor and Hu moments in consistent zone for content based image retrieval. *IEEE Access* 8:146284–146299
23. Peng X, Zhang X, Li Y, Liu B (2020) Research on image feature extraction and retrieval algorithms based on convolutional neural network. *J Vis Commun Image Represent* 69:102705
24. Zhao M, Liu J, Zhang Z, Fan J (2021) A scalable sub-graph regularization for efficient content based image retrieval with long-term relevance feedback enhancement. *Knowl-Based Syst* 212:106505
25. Sukhia KN, Riaz MM, Ghafoor A, Ali SS (2020) Content-based remote sensing image retrieval using multi-scale local ternary pattern. *Digital Signal Process* 104:102765
26. Pavithra L, Sharmila TS (2019) Optimized feature integration and minimized search space in content based image retrieval. *Procedia Comput Sci* 165:691–700. In: 2nd international conference on recent trends in advanced computing ICRTAC disruptive innovation, 11–12 Nov 2019
27. Piras L, Giacinto G (2017) Information fusion in content based image retrieval: a comprehensive overview. *Inf Fusion* 37:50–60

# Precise Forecasting of Stock Market Pricing Using Weighted Ensemble Machine Learning Method



Umesh Kumar Lilhore, Sarita Simaiya, Advin Manhar, Shilpi Harnal, Pinaki Ghosh, and Atul Garg

**Abstract** Forecasting stock market pricing is essential in the current marketplace. As a result, academics' interest in innovative ways to forecast the stock exchange has surged. Throughout this investigation, previous studies had insufficient knowledge, such as sociological climate. In this research, we are presenting an ensemble machine learning method (WEML). A long-short term memory (LSTM), weighted support vector regression (WSVR), and multiple regression method have been used to create the proposed method "weighted ensemble model." We also describe a practical approach for evaluating various data resources to bridge the disparity and forecast an accurate closing price. The primary aim of the investigation is to determine the most acceptable method for forecasting future stock prices. The quality of several machine learning methods, Multiple Regression, Support Vector Regression and Long Short-Term Memory Network (LSTM), ensemble machine learning method has been evaluated on NIFTY-50 Kaggle dataset (2000–2017), and various performance measuring parameters like as accuracy, precision, F-measure, and recall

---

U. K. Lilhore

KIET Group of Institutions, Delho-NCR, Ghaziabad (UP), India

e-mail: [umesh.lilhore@chitkara.edu.in](mailto:umesh.lilhore@chitkara.edu.in)

S. Simaiya (✉) · S. Harnal · P. Ghosh · A. Garg

Chitkara University Institute of Engineering and Technology, Chitkara University, Rajpura, Punjab, India

e-mail: [sarita.simaiya@chitkara.edu.in](mailto:sarita.simaiya@chitkara.edu.in)

S. Harnal

e-mail: [shilpi.harnal@chitkara.edu.in](mailto:shilpi.harnal@chitkara.edu.in)

P. Ghosh

e-mail: [pinaki.ghosh@chitkara.edu.in](mailto:pinaki.ghosh@chitkara.edu.in)

A. Garg

e-mail: [atul.garg@chitkara.edu.in](mailto:atul.garg@chitkara.edu.in)

A. Manhar

Amity University, Raipur, Chhattisgarh, India

e-mail: [amanhar@rpr.amity.edu](mailto:amanhar@rpr.amity.edu)



have been analyzed to measure the efficiency of the proposed algorithm. The experimental result demonstrates that the proposed method achieves 5% better results than existing methods.

**Keywords** Weighted ensemble learning · Stock market · Price forecasting · Machine learning · LSTM · Support vector regression · Multiple regression

## 1 Introduction

There has been much intriguing research conducted in this area of using Machine Learning methods to detect market behavior, take stock prices, and describe information. Many stockbrokers nowadays rely on Ai Stock Exchanges to anticipate prices due to numerous scenarios and variables, allowing them to make direct investments [1]. Predictive analysis of the stock market is a complex problem that has piqued the interest of many experts and individuals. Traders from all over the world have expressed a strong interest in stock forecasting models. Shareholders begin to rely on forecasting ideas and making critical decisions [2]. Much research has been done in this area, but no complete solution has been found. Is it feasible to fully foresee the financial markets? It is still a hotly debated topic due to the difficulty in accurately anticipating the stock market due to the significant effect of other elements (administrative, geopolitical, emotional, and financial) [3].

Numerous machine learning algorithms have been used to forecast the stock market, and there is no standard that we may use to select the best answer for forecasting. This manuscript will look at several machine learning methods used for forecasting. The investigation aims to perform a comparative approach to determine the best methodology for stock price prediction. Their level of performance will be used to compare them. Every strategy has its benefits and drawbacks. This study will look at the benefits and drawbacks of various strategies. Moreover, determine which strategy is superior in stock market forecast [4].

The research article is divided into various sections that cover introduction work, related work, machine learning methods, types, stock prediction dataset, simulation results, discussion, and conclusion.

## 2 Related Work

There are three main approaches to stock market prediction: technical analysis, conventional estimation, and machine learning techniques.

1. **Fundamental Analysis:** This technique is conducted by a Structural Analyst; this technique seems to focus more on the business than the stock. Experts make their conclusions on the business's prior performance, ability to accumulate, etc. Fundamental analysis requires reliable data from a corporation's financial

statement, competitiveness, and market conditions to determine the product's actual price and what they are interested in [5]. That quality is utilized to make financial decisions. It is based on the notion that "if a firm's value exceeds market rate, invest; otherwise, considering this a stupid investment and ignore them."

2. **Technically Analysis:** The technique is used by Technical Experts to determine the market price depending on a stock's past trends (it mainly takes place by a time-series past data analysis) [6]. Time series are sorted collections of data for a linear model or factor presented in equal intervals of time. The forecast continues a tendency through time, such as revenue growth, share price assessment, and gross national income. The forecast continues a tendency through time, such as revenue growth, share price assessment, and gross national income. The two most popular time series used for stock exchange are ARMA (Autoregressive Move Averaged) and ARIMA (Autoregressive Integrate Movement Average).

This section mainly covers the various existing research in the field of stock market prediction.

In the research article [5], researchers mainly discussed "Investor confidence and forecastability of U.S. stock market realized volatility: Evidence from machine learning." Experts calculated random forest data range 2001 to 2020 and then investigated the perspectives of maximum one year after computing predictions for recursion and a continuous estimation period. Experimental results showed the strength of the proposed method. In the research article [6], researchers mainly discussed "A time series analysis-based stock price prediction using machine learning and deep learning models." They demonstrate that business investment, particularly business investment uncertainties, provides others with the predictive ability for total realized volatility and variations. The findings have far-reaching ramifications for both investment and authorities.

In the research article [7], the researchers mainly discussed a time series analysis based stock price prediction using machine learning and deep learning models. In this paper, convolutional neural network and decision tree approaches have been used to forecast the day's current value for four companies from various industries. Primary business data: opening, median, lowest, and closing stock markets have been used to create new factors used as change is known.

The research article [8] discussed the "Effectiveness of artificial intelligence in stock market prediction based on machine learning." Researchers mainly used iterative and prescriptive techniques such as the ANN model to investigate multi-periodic stock price forecasting. In the research article [9], researchers mainly discussed "Technical analysis strategy optimization using a machine learning approach in stock market indices." Additional deep learning algorithms, such as "Hybrid Attention Networks" (HAN), self-paced achieving (NLP), interneuron network (MFN), and Wave net, anticipate stock and Forex movements.

In the research article [10], researchers mainly discussed "Effective forecasting of stock market price by using extreme learning machine optimized by PSO-based group-oriented crow search algorithm."

In the research article [11], researchers mainly discussed the “Machine learning model for stock Market Prediction.” Modifications of ANN were employed to forecast the stock market. However, the neural network used to develop the network determines the effectiveness of ANN prediction.

Table 1 shows the comparative analyses of various machine learning methods applied to the stock prediction by various researchers.

**Table 1** Comparative analysis of machine learning research in stock markets prediction

Reference	Methods	Key parameters	Data set
[12]	Clustering method for stock predictions, SVM, KNN	Opening price, higher cost, reasonable price, adjusted closing price, quantity, and adjusted market close	NSE listed companies
[13]	Machine learning for future market analysis, random forest	Sentiment (including good, impartial, and negative emotions)	Thomson Reuters
[14]	Indian stock market prediction by machine learning methods, SVM, CNN	Membership function computed by using linear relationship and item occurrences	The S&P 500 index is a stock market index that measures the performance of 500 companies (S&P 500)
[15]	Machine learning repressors, regression	Quantity, CCI, RSI, K.D., ROC, W.R., maximum, lowest, and closing price; quantity	30 equities from Fortune Global 500 businesses, including Walmart, XOM, and Mac
[16]	Machine learning for future stock market analysis, J-48	Volume; probabilistic analyzer (percent K); Lawrence Williams (L.W.) percent R indication; comparative intensity	The S&P 500 index company’s dataset
[17]	Moving averages methods and machine learning method, SVM	The closing prices, the RSI, the SMA, the MACD, the MFI, the Williams percent R, the sequential oscillation, and the ultimate optocoupler are all indicators	Digging through economic markets and the Baidu Rating
[18]	Market crashes using machine learning kNN, random forest	The opener, peak, lower, closing, revised close, and quantity	NSE and BSE’s top four outperforming firms

### 3 Machine Learning Methods

This report discusses the Machine Learning (ML) methodologies used here in this research and how they were implemented to forecast the stock marketplace.

1. **SVM:** It is a supervised method commonly used in classification and regression assignments. To discover two distinct classes inside a multi-dimensional environment, the support vector machine acts as just a sequential divider placed between two data blocks. The stages below demonstrate how support vector machines are implemented. Assume  $T_s$  be the training set of data, with  $T_s = (x_j, y_j), \dots, (x_n, y_n)$ , where  $j = (1, 2, 3, \dots, n)$ . A support vector machine denotes  $T_s$  as locations inside an  $N$ -dimensional area and afterward efforts to find a higher dimensional space that divides the area into the pre-specified labels with such a reasonable standard deviation. Formulas (2) and (3) demonstrate the method used throughout the Support vector machine optimization technique [19].

$$\text{Min}_d \left[ \frac{1}{2} W^{T*} W + C \sum_{j=1}^n W_j \right] \tag{1}$$

2. **Decision Tree:** It is a tree data structure that looks like a process flow and utilizes a segmented method to explain every possible decision outcome. The usability and easiness of the decision tree and its small processing expense and power to make decisions visually have made a significant contribution to its growing use for categorization problems. An attribute selection method has been used to find the correct possessions for every node of a produced tree. So every existing single node test characteristic is selected based on the character with the most data [20]. The entropy can calculate as

$$E(T) = \sum_{k=0}^n \binom{n}{k} \log_2 p_j \tag{2}$$

where  $E(T)$  = entropy, = the number of classes,  $P_j$  = total no. of instances.

3. **Random Forest:** It is also known as random decision forests, which are ensemble classifiers, regression, and other activities that work by building a large decision tree based on instruction. For classifiers, random forest production has been the class chosen mainly through a majority of tree trunks. An average mean forecast of the individual plants is decided to return for regression problems. Random forests recompense for judgment trees' proclivity to overfit one training dataset. Random forests perform better classification trees overall, but one's precision is lesser than among gradient boosted forests. After all, the number of observations can impact the outcomes [21].

## 4 Proposed Weighted Ensemble Learning Method

A long-short term memory (LSTM) [22], weighted support vector regression (WSVR) [23], and multiple regression method [24] have been used to create the proposed method “weighted ensemble model.” An ensemble strategy proposed in the research mainly utilizes a weighted estimation technique, which yields more impressive outcomes than median and large majority voting processes. Initially, the stock prices have been projected using classification models such as linear regression, SVM, and multiple regression methods. Weight values are then allocated to base learners within the second phase step based on their precision. As such ensemble methods’ finished product, the third stage estimate of weighted accuracy and precision of such base outcomes have been acquired.

First, let  $(B1, W1)$  be now the weight and accuracy for multiple regressions,  $(B2, W2)$  is just the weight and accuracy for the support vector regression method, and  $(B3, W3)$  is just the weight and accuracy of LST method, whereupon an accuracy of ensemble method is given using equation.

$$AM = [(B1, W1) + (B2, W2) + (B3, W3)] / (W1 + W2 + W3) \quad (3)$$

where  $AM =$  Accuracy of model.

1. **Long Short-Term Memory Network (LSTM):** Since there can be slowdowns of the source added among both significant events in a data series, LSTMS are very well for categorizing, storing, and predicting the future time series analysis. It is a “deep learning artificial recurrent neural network architecture.” Except for traditional feed-forward algorithms, LSTM includes activation functions. It is capable of processing not only singular data points (including such image data) and also complete data scenes (from a speech or a video). LSTM, for instance, is responsible for functions like laterally compressed, linked handwritten text [22].
2. **Multiple Regression:** A statistical technique can be utilized to examine the linear relationship between a completely reliant parameter and several exogenous variables. Multiple linear regression analysis aims to estimate the significance of a single utterly reliant variable using known significant factors [23].
3. **Support Vector Regression:** It can also be utilized as a regression model while retaining all of the system’s main characteristics (maximal margin). The Support Vector Regression (SVR) utilizes the same categorization ethics as the support vector machine [24, 25].

## 5 Experimental Results, Analysis, and Discussion

The accuracy of the proposed ensemble model is used to analyze the quality of machine learning techniques, i.e., base learning like Multiple Regression, Support Vector Regression, Long Short-Term Memory Network (LSTM), and Multiple Regression.

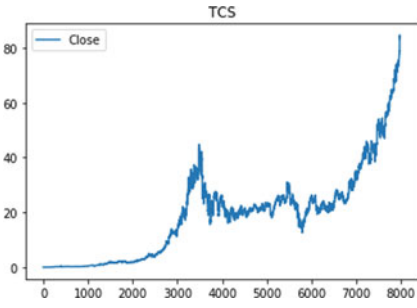
1. **Dataset:** The statistics is the market value background and share prices of the 50 stocks inside the NIFTY 50 from India's National Stock Exchange. All other data sets seem to be daily, as sales prices and exchanging values spread across. CVS documents for every share and a metadata document containing some macro information well about inventory on its own. The statistics range from January 1, 2000 to June 30, 2020 [26]. We use the top two companies' data for experimental analysis, TCS and Microsoft.
2. **Experimental Results:** The proposed method and existing methods (Long-short term memory (LSTM), weighted support vector regression (WSVR), Multiple Regression method, and SVM, Random Forest, and Decision tree) were implemented in the Anaconda Python environment. Figure 1a–e shows experimental results calculated by the proposed method for the TCS data set.

Figure 2 shows the proposed method's experimental results for the Microsoft stock exchange dataset.

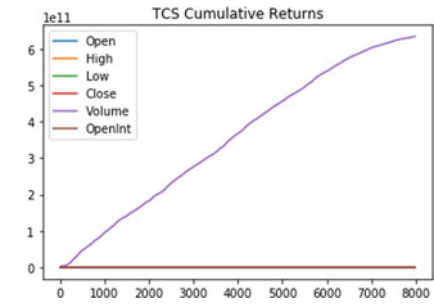
The experimental result from Figs. 1, 2, 3, and 4 shows the simulation results for TCS and Microsoft share prediction data. The experimental results clearly show that the proposed WEML method shows better precision, f-measure, and recall results over existing methods Long Short-Term Memory Network (LSTM) Multiple Regression, and Support Vector Regression methods.

## 6 Conclusion

The paper describes a weighted ensemble method consisting of MRM, SVRM, and LMN. The model employs the periodic inventory system based on the accuracy collected from social stock time series forecasting. The proposed model is compared with various machine learning methods such as SVM, random forest, and decision tree. The proposed WEML and the existing Long Short-Term Memory Network (LSTM), Multiple Regression, and Support Vector Regression methods are implemented on the NIFTY-50 online Kaggle dataset. An experimental result ensures that the ensemble learning (proposed WEML) shows 5% better results than many other base classifiers. The proposed method can be applied to solve several other categorization difficulties on a real-time dataset in future work.



(a) TCS share close data



(b) TCS Cumulative return



(c) Share TCS Prices



(d) TCS price prediction by a Proposed method



(e) TCS price prediction Final results using the proposed method

**Fig. 1** Experimental results of the proposed method for TCS share dataset

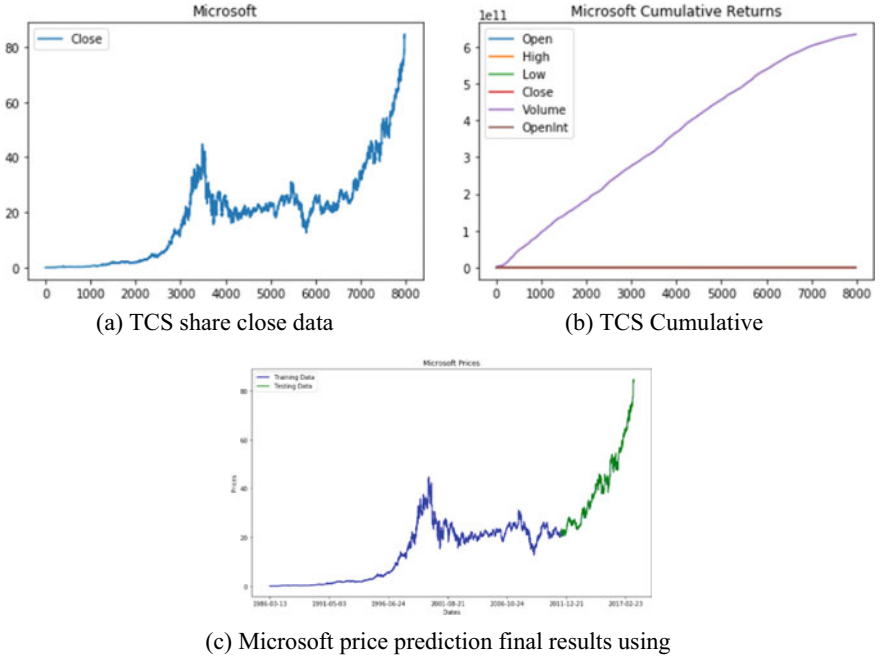


Fig. 2 Experimental results for the Microsoft stock exchange dataset

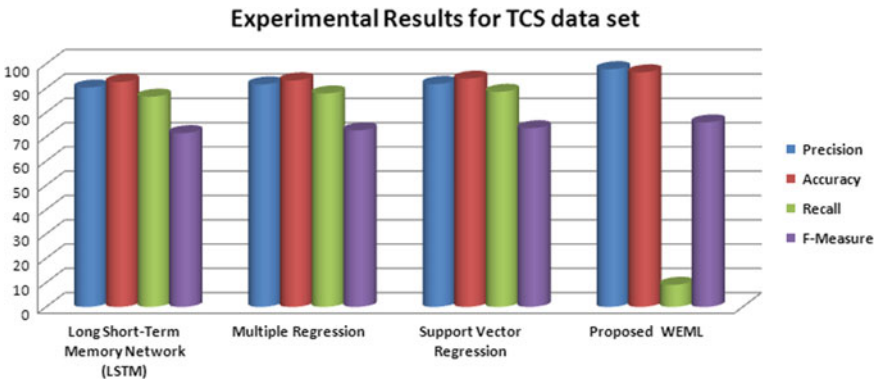
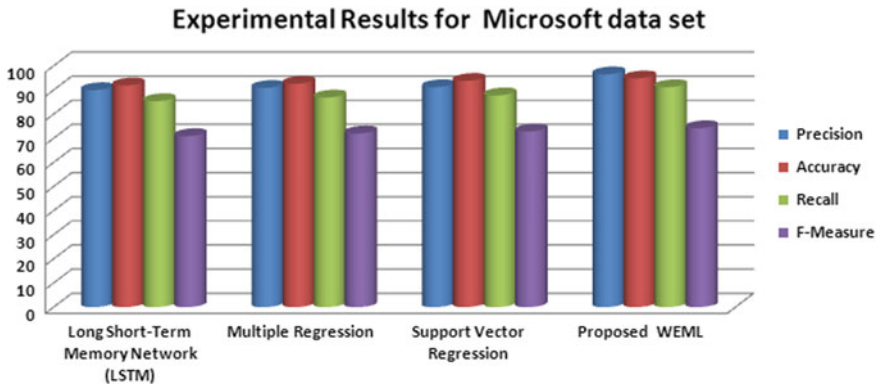


Fig. 3 Experimental results for TCS data set





**Fig. 4** Experimental results for Microsoft data set

**Conflict of Interest** All the authors declare no conflict of interest.

## References

1. Khattak MA, Ali M, Rizvi SAR (2021) Predicting the European stock market during COVID-19: a machine learning approach. *MethodsX* 8:101198
2. Liu X (2020) Analyzing the impact of user-generated content on B2B Firms' stock performance: big data analysis with machine learning methods. *Ind Mark Manag* 86:30–39
3. Saifan R, Sharif K, Abu-Ghazaleh M, Abdel-Majeed M (2020) Investigating algorithmic stock market trading using ensemble machine learning methods. *Informatica (Ljubljana)* 44. <https://doi.org/10.31449/inf.v44i3.2904>
4. Cervelló-Royo R, Guijarro F (2020) Forecasting stock market trend: a comparison of machine learning algorithms. *FMV* 6:37–49
5. Gupta R, Nel J, Pierdzioch C (2021) Investor confidence and forecastability of U.S. stock market realized volatility: evidence from machine learning. *J Behav Financ* 1–12
6. Sen J, Mehtab S (2020) A time series analysis-based stock price prediction using machine learning and deep learning models. *Int J Bus Forecast Mark Intell* 6:272
7. Mehtab S, Sen J (2020) A time series analysis-based stock price prediction using machine learning and deep learning models. *Int J Bus Forecast Mark Intell* 6:272
8. Mokhtari S, Yen KK, Liu J (2021) Effectiveness of artificial intelligence in stock market prediction based on machine learning. <http://arxiv.org/abs/2107.01031>
9. Ayala J, García-Torres M, Noguera JLV, Gómez-Vela F, Divina F (2021) Technical analysis strategy optimization using a machine learning approach in stock market indices. *Knowl Based Syst* 225:107119
10. Das S, Sahu TP, Janghel RR, Sahu BK (2021) Effective forecasting of stock market price by using extreme learning machine optimized by PSO-based group-oriented crow search algorithm. *Neural Comput Appl*. <https://doi.org/10.1007/s00521-021-06403-x>
11. Patil DST (2021) Machine learning model for stock market prediction. *Int J Res Appl Sci Eng Technol* 9:4057–4062
12. Mohamed ZE, El-Mesalamy E-AKE-D (2021) Analysis and prediction of stock market mining using machine learning clustering technique. *Int J Comput Appl* 183:39–44
13. Likhite R, Mahajan G, Padulkar S, Kakani S, Suradkar PT (2021) Future stock market prediction using machine learning. *Int J Adv Res Sci, Commun, Technol* 479–487

14. Rohatgi S, Kumar Singh K, Jasuja D (2021) Comparative analysis of machine learning algorithm to forecast Indian stock market. In: 2021 international conference on advanced computing and innovative technologies in engineering (ICACITE). IEEE
15. Ashfaq N, Nawaz Z, Ilyas M (2021) A comparative study of different machine learning regressors for stock market prediction. <http://arxiv.org/abs/2104.07469>
16. Likhite R, Mahajan G, Padulkar S, Kakani S, Suradkar PT (2021) Future stock market prediction using machine learning. *Int J Adv Res Sci, Commun, Technol* 553–557
17. Dinesh S, Rao N, Anusha SP, Samhitha (2021) Prediction of trends in the stock market using moving averages and machine learning. In: 2021 6th international conference for convergence in technology (I2CT). IEEE
18. Dichtl H, Drobetz W, Otto T (2021) Forecasting market crashes via machine learning: evidence from European stock markets. *SSRN Electron J*. <https://doi.org/10.2139/ssrn.3843319>
19. Trivedi NK, Simaiya S, Lilhore UK, Sharma SK (2021) COVID-19 pandemic: role of machine learning & deep learning methods in diagnosis. *Int J Curr Res Rev* 150–155
20. Lilhore UK, Simaiya S, Prasad D, Guleria K (2020) A hybrid tumour detection and classification based on machine learning. *J Comput Theor Nanosci* 17:2539–2544
21. Guleria K, Sharma A, Lilhore UK, Prasad D (2020) Breast cancer prediction and classification using supervised learning techniques. *J Comput Theor Nanosci* 17:2519–2522
22. Lilhore UK, Simaiya S, Guleria K, Prasad D (2020) An efficient load balancing method by using machine learning-based V.M. distribution and dynamic resource mapping. *J Comput Theor Nanosci* 17:2545–2551
23. Vanchurin V (2021) Toward a theory of machine learning. *Mach Learn: Sci Technol* 2:035012
24. Liu T, Barnard AS (2021) Fast derivation of Shapley based feature importances through feature extraction methods for nanoinformatics. *Mach Learn: Sci Technol* 2:035034
25. Knijff L, Zhang C (2021) Machine learning inference of molecular dipole moment in liquid water. *Mach Learn: Sci Technol* 2:03LT03
26. Nifty-50 dataset from “NIFTY-50 stock market data (2000–2021)”. <https://www.kaggle.com/rohanrao/nifty50-stock-market-data>. Accessed 15 July 2021

# Graph-Based Mechanism to Prevent Structural Attack Over Social Media



Jitendra Patel and Ravi Kumar Singh Pippal

**Abstract** Social media sites contain the personal information of the users, which entice the attackers. The attacker performs different types of attacks on the social media site to get the user's sensitive information. User privacy may be breached as the other kind of passive and active attacks are performed on social media sites; to prevent such a scenario, the network operator releases the data in an anonymized form. Social media operators fetch and store data from social media users to share among various third-party consumers. As the fetched information often contains sensitive data, the network operator releases the complete graph in anonymised and sanitised versions. But it does not provide a full guarantee of user privacy. This paper proposed a solution that provides a neighbourhood adjacency matrix based anonymisation process for the social network graph. This anonymisation process may be used to counter the neighbourhood attack over the social network graph. The proposed anonymisation process increases the number of isomorphic neighbourhood networks by adding dummy edges in the social network graph. Therefore, a user may not be re-identified in a social network graph based on its unique neighbourhood network.

**Keywords** Social media · Social media mining · Graph anonymisation · Neighbourhood attack · Graph-based attack · Adjacency matrix

## 1 Introduction

On the social network, end-users use the social media platform for sharing views, knowledge, information and mutual interact. End-users willingly provide their personal and private information for profile creation over social media sites including phone number, profile picture, relationship status, email, etc.

User interaction on social media generates a rich amount of data that contains sensitive information, e.g., users' attitudes towards any local, global, political, clinical [1, 2], environmental, critical and inflammable issue. Social media operators

---

J. Patel (✉) · R. K. S. Pippal  
Department of Computer Science, RKDF University, Bhopal, India  
e-mail: [jitendra.jp12@gmail.com](mailto:jitendra.jp12@gmail.com)

share user-generated content with third party for research and data analytics, such as advertisement companies, political parties, and manufacturing companies for revenue collection. As user-generated content contains sensitive information, social media operator shares anonymised graphical information [3, 4].

Suppose the adversary is having background knowledge of the social media structure. In that case, the anonymised graphical information of social media may be vulnerable in the sense of structure-based attacks such as subgraph and neighbourhood attacks [5, 6]. With structure-based attacks, an adversary can re-identify the targeted user node from the published social network [7].

Recently, neighbourhood anonymisation of the social network restricts the adversary with background knowledge of neighbourhood structure to prevent structure-based attacks [8–10]. This paper presents the anonymisation of social network data and preserves user privacy against neighbourhood attacks [11, 12]. The neighbourhood attack based on a user and its neighbour's information identifies the isomorphic structure [13–15]. If two or more neighbourhood networks are isomorphic in the social network graph, then the adversary cannot place a unique vertex neighbourhood sub-network [16, 17]. The proposed methodology increases the isomorphic neighbourhood network in the social network graph by adding established imitation relationship edges [18–20].

In this paper, a counter solution to prevent neighbourhood attacks has been presented. It has been found that if two or more neighbourhood networks are isomorphic in a social network graph then it is difficult for an attacker to re-identify a specified vertex from the released anonymous graph. Also simply adding dummy edges in the anonymous graph to avoid vertex re-identification will lead to greater information loss. So the proposed solution is based on increasing the isomorphic neighbourhood network in the social network by adding a minimum number of dummy edges. The detail of social large network data and implementation details of 1-neighbourhood adjacency matrix on big data environment is present in the second and third sections.

## 2 Proposed Method

The user's data provided by social networking sites is anonymised, which means that a meaningless identifier replaces the user's critical information. Recently, anonymisation procedures have been used to add dummy edges and corners to protect social networks' graphic data. However, the inclusion of false corners and edges increases the noise level in anonymous data.

This paper focuses on reducing the loss of information by lowering the noise level added for anonymisation. Minimise the degree of noise by optimising the number of false edges. An optimised artificial edge can maintain isomorphic groups and reduce information loss. Therefore, data from social networks remained helpful to researchers and data analysts.

In this paper, a neighbourhood adjacency matrix based anonymisation (NAMA) process has been proposed for an extensive social media network, as shown in Fig. 1. In NAMA, an optimised number of dummy edges are added to a unique sub-graph; however, the number of vertices remains the same. Unique sub-graph and vertices are highly vulnerable to structure-based attacks. NAMA approach has different modules for anonymised uniqueness of sub-graph and vertices.

For anonymised uniqueness of vertices, it initially extracts a set of vertex degrees in increasing order as follows:

$$v = \{v_i^d, v_j^d, \dots, v_m^d\} \tag{1}$$

where  $u_v$  is the set of vertices with superscript their degree and  $v_i^d$  represent vertices  $v_i$  having degree  $d$ . However, if degree ( $d$ ) of any  $v_i$  is unique for all  $v_i$  in  $v$ , then  $v_i$  is vulnerable vertices  $v_v$ .

$$d(v_i) = \begin{cases} \text{unique } \forall \text{ vulnerable } v_i \\ \text{common } \forall \text{ safe } v_i \end{cases} \tag{2}$$

After that, NUMA approach calculates identical degree vertices count (idvc), i.e., summarised all the vulnerable vertices ( $v_v$ ) as

$$idvc = \sum_{i=1}^n v_i \tag{3}$$

If idvc of  $v_v$  is greater than one, then the NUMA approach adds a dummy edge among  $v_v$ , else if idvc of  $v_v$  is one then add the dummy edge with vertices having more than two idvc until  $d(v_v)$  does not become common.

However, for anonymisation of the uniqueness of the sub-graph, NUMA approach extract vertices having an identical degree as vertices belong to vulnerable sub-graph  $v_{sg}$ . Next, extract the 1-neighbourhood networks and create adjacency matrices. Now sort the adjacency matrices. From the sorted vertices, vertices having a higher adjacency matrix have been selected. Subtract the lower adjacency matrix from the higher adjacency matrix. The resultant adjacency matrix shows the required dummy edges for anonymisation. Thus, dummy edges are added to that vertex neighbourhood network that has a lower adjacency matrix. After adding dummy edges, the vertices neighbourhood networks become isomorphic, as shown in Fig. 2.

**Algorithm 1 [Stepwise explanation of proposed solution]**

1. Start
2. Select the un-anonymised social network graph  $G (V, E)$  from the database, where  $V = \{v_1, v_2, v_3, \dots, V_n\}$  is a set of  $n$  vertices and  $E = \{e_{ij} = (v_i, v_j) | v_i, v_j \text{ in } V, i \text{ not equal to } j\}$  is the set of edges between the  $n$  vertices of graph  $G$ . The degree of a node is decided by the number of edges connected to with it. The set of degree  $D = \{d_1, d_2, d_n\}$  is the available degrees in  $G$ .

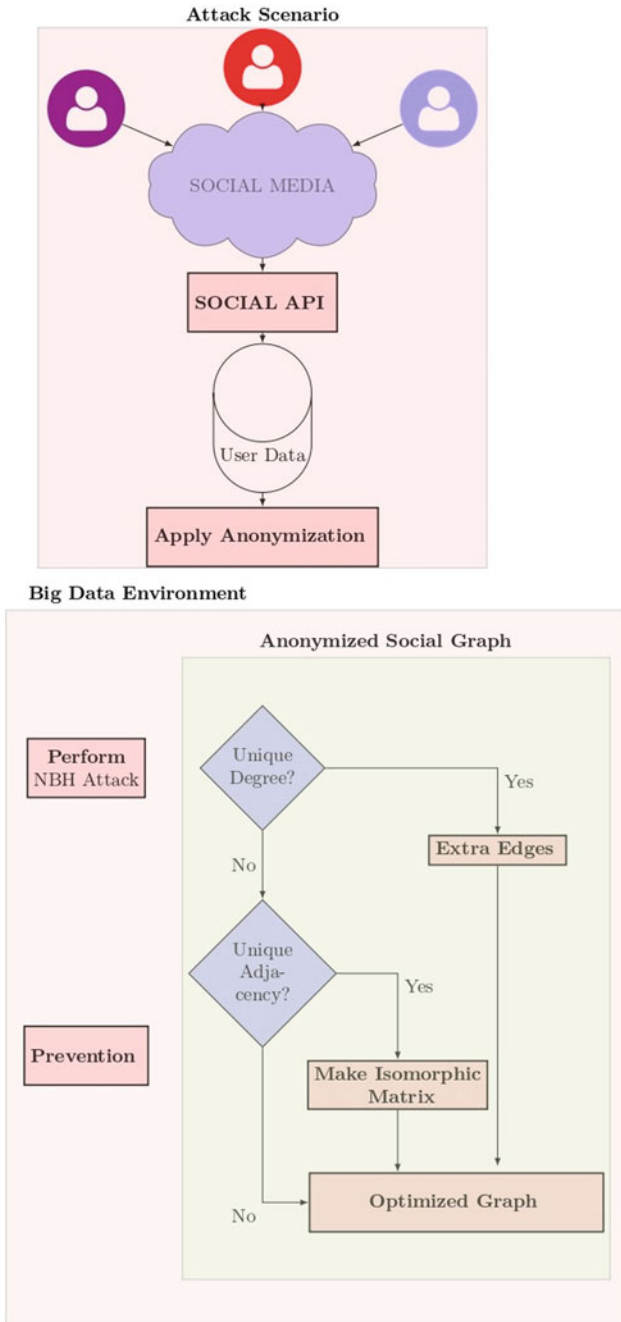
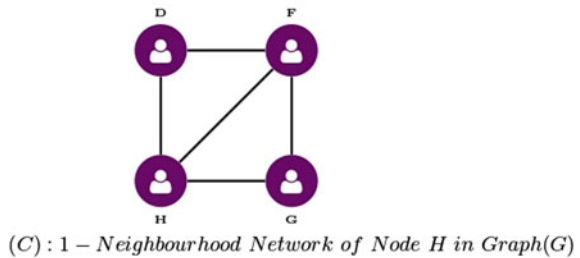
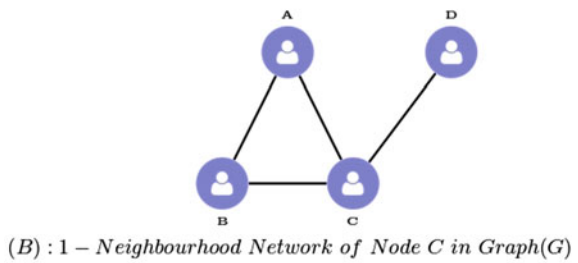
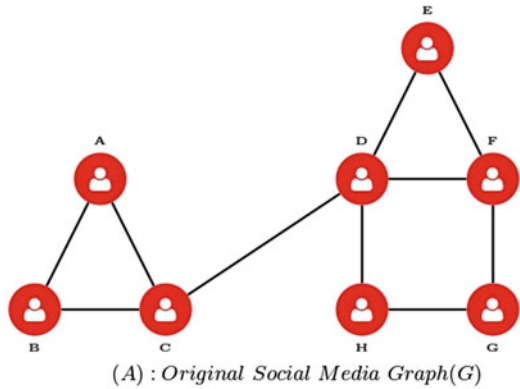


Fig. 1 Proposed framework to identify neighbourhood attack

**Fig. 2** Comparing of neighbourhood networks before anonymisation



3. Maintain a list of degree  $d_i$  in that all of the degrees from  $G$ , those appear once in  $G$ .
4. A vertex-degree list is created; in this list, all of the vertices  $V$  and their degrees are present.
5. Extract the neighbourhood networks  $N(v_i)$  from  $G$ . Here,  $N(v_i)$  is the set of neighbourhood networks  $N(v_i) = \{v_1, v_2, \dots, v_s\}$  in the graph  $G$ .
6. The adjacency matrices  $A_v$  for all  $N(v_i)$  are created. It maintains a list of adjacency matrices.
7. Now select a degree from degree list  $d_i$ , created in step 2.
8. Select the vertices from the vertex-degree list those degree equal to the  $d_i$ , vertex-degree list created in step 3.
9. Select the adjacency matrices of all vertices with degree  $d_i$ .

10. Check the cost of all adjacency matrices and sort it according to the increasing order of cost.
11. Now calculate the difference between all alternate adjacency matrices one after the other and store the result in AR, resulting in adjacency matrices.
12. Extract the edges from AR, update the set of dummy edges ED by inserting edge information, and go back to step 4 for the next degree.
13. Remove the duplicate edges from the set of dummy edges ED and union the set of edges E of G and the set of dummy edges ED and generate a new anonymised set of edges  $E' = \{E \text{ union } ED\}$ .
14. Update set of edges E with anonymised set of edges  $E'$  in the social network graph G.
15. Now the social network graph G becomes anonymised social network graph  $G'$ .
16. End.

The original social network is shown in Fig. 2a. The 1-neighbourhood networks of vertex C and vertex H are shown in Fig. 2b, c, respectively. Vertex C and vertex H have the same degree but their neighbourhood networks are different.

First, select a degree in social network graph G. Here, degree 3 is selected, the vertex C and vertex H has degree 3. Now, the 1-neighbour subgraph for both vertices is created. In the first subgraph vertex A, vertex B, and vertex D are connected with node C and also nodes A, B are also connected with each other. The same as in second subgraph, vertex D, vertex F, and vertex G are connected with node H. Vertex F, vertex G are connected with each other and also there is a connection between vertex D and vertex F. Both nodes C, H have the same number of neighbours but neighbourhood networks are not isomorphic. To find out the difference between neighbourhood subgraphs, the comparison can be done through adjacency matrices as shown in Fig. 3.

To find the difference between networks with one neighbourhood of node C, H, adjacent matrices are created and sorted in descending order. Here the node C adjacency matrix is higher than the node H adjacency matrix. Thus, the vertex C adjacency matrix is subtracted from the vertex H adjacency matrix. The resulting adjacent matrix  $C'$  shows the required dummy edge between vertex D and vertex A to anonymise the neighbourhood network of node M 1-neighbourhood network. After adding the dummy edges to the social network graph G, it changes to the anonymised social network G. The 1-neighbourhood network of M node and N node is isomorphic in the anonymised social network  $G'$ .

The anonymised social network graph  $G'$  is shown in Fig. 4a. After adding dummy edges to the original social network chart, the similarity between the vertical C and the vertex H1 neighbourhood networks is shown in Fig. 4b, c.



	A	B	C	D
A	-	1	-	1
B	1	-	1	-
C	1	1	-	1
D	-	-	1	-

(A):Adjacency Matrix of 1-Neighbourhood C

	F	D	H	G
F	-	1	1	1
D	1	-	1	-
H	1	1	-	1
G	1	-	1	-

(B):Adjacency Matrix of 1-Neighbourhood H

	F	D	H	G
F	-	-	-	1
D	-	-	-	-
H	-	-	-	-
G	1	-	-	-

(C):Adjacency Matrix of 1-Neighbourhood H-C

Fig. 3 Adjacency matrix

### 3 Experimental Setup

In the social big data network anonymisation, all the experiments were conducted on a system running the Ubuntu 16.04 operating system, with a 2.3 GHz Core(TM) i3 CPU, 3.0 GB RAM, and a 320 GB hard drive with open-source software Neo4j (version 3.3.5) and R (version 3.3.0). The program is implemented in the R programming language. This paper illustrates the comparative analysis of social network graph anonymisation on different sets of two different datasets, i.e., real-time Twitter dataset, and Gnutella Peer-to-Peer Network Dataset.

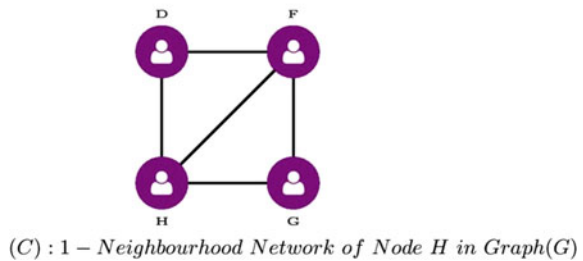
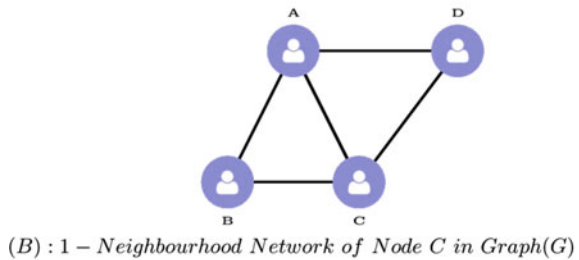
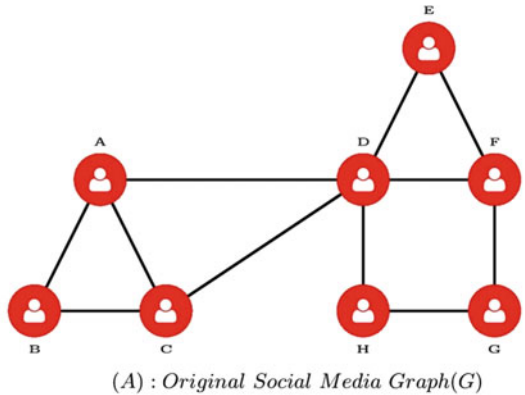
#### Real-Time Twitter Dataset:

Three sets of real-time data are extracted/crawled from twitter. Firstly, 100 tweets are crawled, the network has 315 vertices and 435 edges. The data is collected from the followers of Modi. Each node represents the followers and each edge represents a connection between a pair of hosts. Second, 1000 tweets have been selected, the network has 2313 nodes and 4110 edges. Third, 10,000 tweets are extracted, and the network has 20,350 vertex and 43,500 edges.

#### Gnutella Peer-to-Peer Network Dataset:

The Gnutella peer-to-peer network dataset represents a directed graph where the edge represents a connection between a pair of Gnutella hosts. It has 22,687 nodes and 54,705 edges.

**Fig. 4** Comparing of neighbourhood networks after anonymisation



### 4 Result Evaluation

The anonymisation process tampers the originality by pouring some noise over the network, i.e., adding or removing the edges in published data. However, the level of anonymisation depended upon the degree of noise added to the networks. The analysis is carried out on common parameters including the number of nodes, the number of actual edges, the number of dummy edges, and the average degree of the graph. In this paper, social network anonymisation experiments were performed on two datasets: real-time Twitter and Gnutella peer-to-peer network datasets, as shown in Table 1.

The graphical description of the anonymised social network graph, i.e., the number of vertices, the minimal number of dummy edges required to be added, the total

**Table 1** Parameters of unanonymised social network graph

Dataset	# Tweets	# Vertices	# Edges	# Average degree
Twitter dataset	100	315	435	3
	200	554	1085	4
	1000	2313	4110	4
Gnutella dataset	501	501	710	3
	1001	1001	2068	4
	5001	5001	15,798	6

number of resultant edges, and the average degree are shown in Table 2. The change in network information before anonymisation and anonymisation is visible after comparing Tables 1 and 2. The number of vertices remained unchanged. Whereas after incorporating the dummy edge, network density is increased which reflects both in the total number of edge counts and average degrees (Table 3).

Evaluation of the NUMA anonymisation approach is carried out by using the following evaluation metrics:

- a. **Anonymisation Reflection of Average Degree ( $\gamma^{ad}$ ):**  $\gamma^{ad}$  is the reflection of a change in average degree before and after the anonymisation through NUMA, as shown in Eq. 4.

**Table 2** Parameters of anonymised social network graph

Dataset	# Tweets	# Vertices	# Dummy edges	# Total edges	# Average degree
Twitter dataset	100	315	101	536	4
	200	554	251	1336	6
	1000	2313	2513	6623	5
Gnutella dataset	501	501	189	899	4
	1001	1001	958	3026	5
	5001	5001	4568	20,366	8

**Table 3** Evaluation of anonymised social network graph

	Tweet	RRAD	RRAE	Noise
Twitter dataset	100	0.33	0.23	7.5
	200	0.50	0.23	4.4
	1000	0.25	0.61	1.3
Gnutella dataset	501	0.33	0.27	4.9
	1001	0.25	0.46	2.3
	5001	0.33	0.29	0.2

$$\gamma^{ad} = \frac{\sum_{i=1}^n d(v_i) \in G' - \sum_{i=1}^n d(v_i) \in G}{\sum_{i=1}^n d(v_i) \in G} \tag{4}$$

where  $G$  and  $G'$  represent the graph before and after the anonymisation.

- b. **Anonymisation Reflection of Edge ( $\gamma^e$ ):**  $\gamma^e$  is the reflection of the change in the total number of edges after anonymisation, as shown in Eq. 5.

$$\gamma^e = \frac{\sum_{i=1}^n e_i \in G' - \sum_{i=1}^n e_i \in G}{\sum_{i=1}^n e_i \in G} \tag{5}$$

Noise Level ( $\alpha^l$ ): Noise in the network may be created due to a change in network parameters, i.e., number of vertices, edge, and average degree.  $\alpha^l$  is measured as the total reflection in network information after anonymisation as shown in Eq. 6.

$$\alpha^l = \frac{\gamma^v + \gamma^e + \gamma^{ad}}{|v| + |e| + |ad|} \tag{6}$$

Average degree of the graph before and after anonymisation is shown in Fig. 5 and the relative ratio of the average degree is shown in Fig. 6. The relative ratio of the average degree over the Twitter dataset is measured between 1.65 and 1.75 and the relative ratio of average degree in the Gnutella peer-to-peer network is measured between 1.3 and 1.75. The changes of a few edges or vertices using adjacency-based anonymisation still have a small effect on the average degree.

Edge change of the graph before and after anonymisation is shown in Fig. 7 and the relative ratio of edge change is shown in Fig. 8. The relative ratio of edge change over the Twitter dataset is measured between 0.05 and 0.2 and the relative ratio of

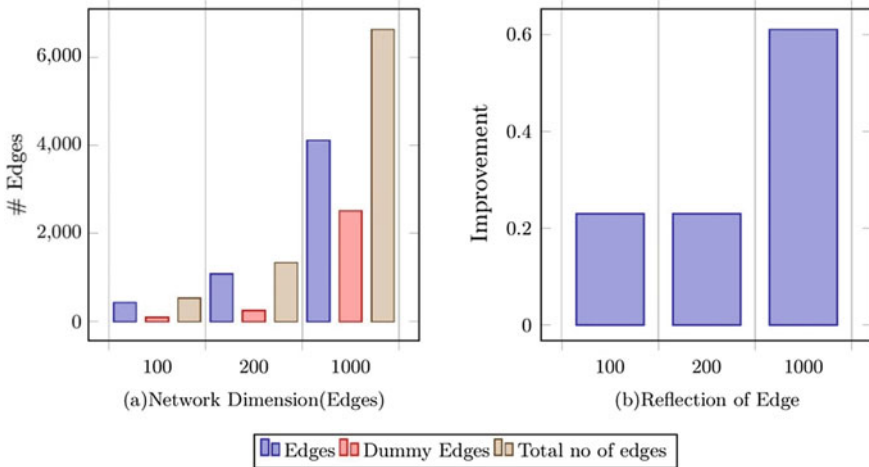


Fig. 5 Reflection of edges after anonymisation over Twitter dataset

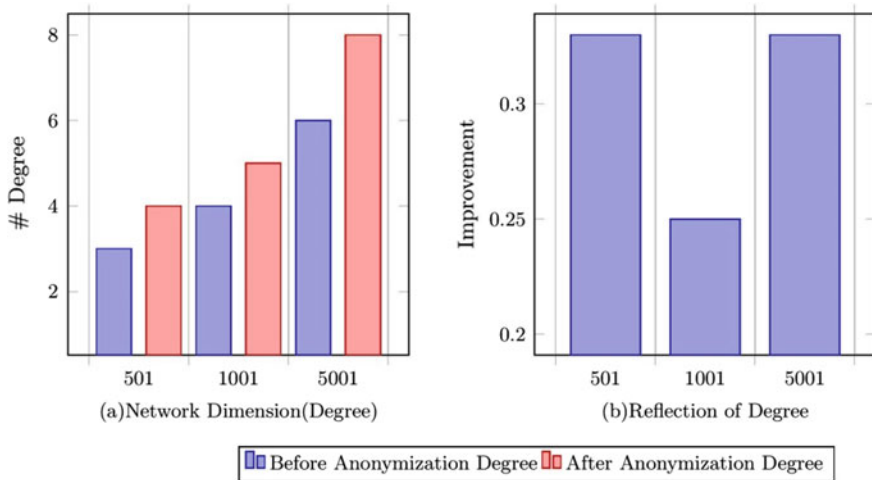


Fig. 6 Reflection of degree after anonymisation over Twitter dataset

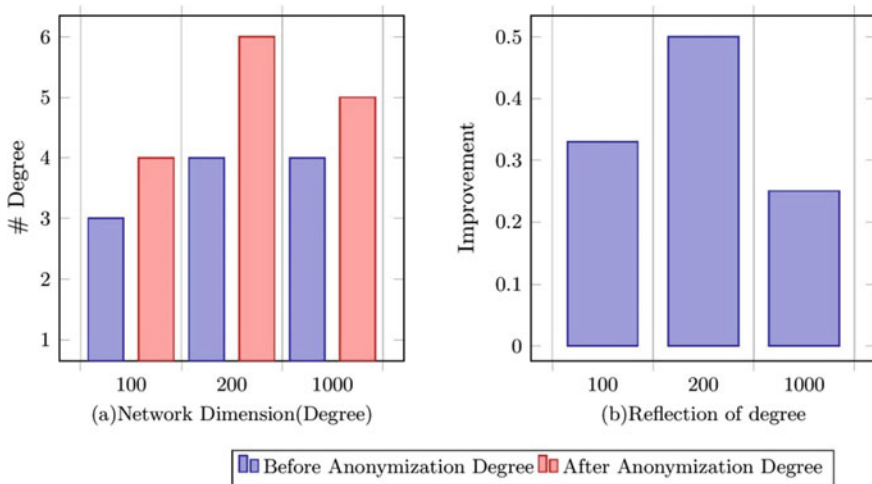


Fig. 7 Reflection of degree after anonymisation over Gnutella dataset

edge change in the Gnutella peer-to-peer network is measured between 0.3 and 0.85. In big social networks, this adjacency matrix based anonymisation changes a small portion of vertices and edges without significantly affecting the neighbourhood.

The noise level rise after anonymisation in the social network’s graph is shown in Figs. 9 and 10. The noise level is measured between 0.05–0.75 and 0.08–0.3, respectively, in the anonymised Twitter and Gnutella peer-to-peer network dataset.

The proposed approach decreases the noise level over social networks. Hence, it reduces information loss. The addition of a dummy edge in the social network

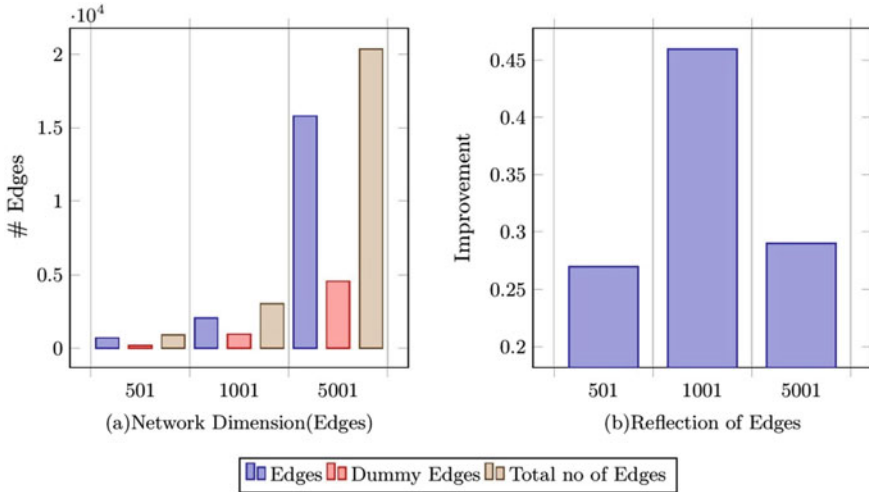
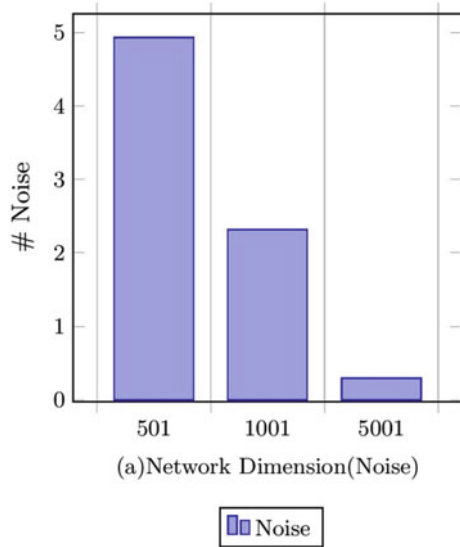


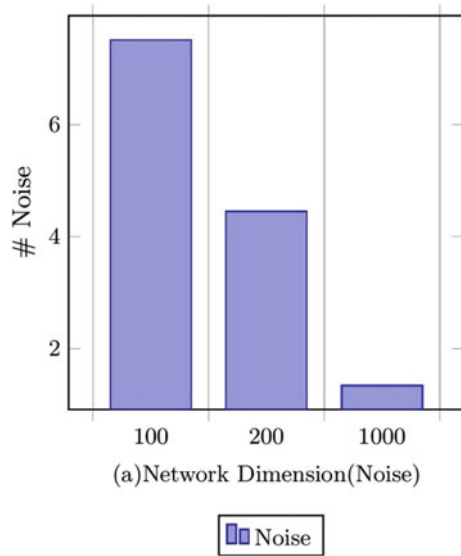
Fig. 8 Reflection of edge after anonymisation over Gnutella dataset

Fig. 9 Reflection of noise after anonymisation over Gnutella dataset



graph increases the social network’s isomorphic nature. It leads to an increase in the privacy-preserving level of social network data. But simultaneously, these dummy edges create some noise over social networks. Noise over social network data affects the results of those experiments which are conducted on social network data. The presented anonymisation process acquires a marginal noise level, i.e., approximately 0.05%–0.8% by adding and subtracting the optimised dummy edge.

**Fig. 10** Reflection of noise after anonymisation over the Twitter dataset



## 5 Conclusion

Social media becomes a significant part of human day-to-day life. Social media provides an easy medium to connect with family and friends for communicating and sharing. People use social media environment to interact with an old companion, maintain a relationship, or even met new friends, thus strengthening the overall connectivity among social media user. Social media sites contain the personal information of the users, which entice the attackers. The attacker performs different types of attacks on the social media site to get the user's sensitive information. Users' privacy may be breached as different types of passive and active attacks are performed on social media sites. In order to prevent such a scenario, network operator releases the data in an anonymised form. Recent anonymisation process preserves the use of social network graph data by adding dummy edges and vertices. The addition of dummy corners and edges increases the noise level, which can cause information loss. The amount of information loss is directly proportional to the number of dummy edges and the vertex added. Information loss is still a problem in social network anonymity. The inclusion of dummy corners and edges in social network data can change the social network graph's originality and increase the noise level. If excessive noise levels are presented in anonymous social network data, researchers and data analysts may receive an unfair result. This paper presents a neighbourhood adjacency matrix based anonymisation process to counter the neighbourhood attack on the social media data. NUMA increases the number of isomorphic neighbourhood networks by adding dummy edges. Any user may not be re-identified in a social network graph based on its unique neighbourhood network.

## References

1. Abawajy JH, Ninggal MIH, Herawan T (2016) Privacy preserving social network data publication. *IEEE Commun Surv Tutor* 18(3):1974–1997. <https://doi.org/10.1109/COMST.2016.2533668>
2. Jamil A, Asif K, Ghulam Z, Nazir MK, Mudassar Alam S, Ashraf R (2018) Mpmpa: a mitigation and prevention model for social engineering based phishing attacks on facebook. In: 2018 IEEE international conference on big data (big data). pp 5040–5048
3. Ji S, Li W, Gong NZ, Mittal P, Beyah R (2016) Seed-based de-anonymizability quantification of social networks. *IEEE Trans Inf Forensics Secur* 11(7):1398–1411
4. Ji S, Li W, Srivatsa M, Beyah R (2016) Structural data de-anonymization: theory and practice. *IEEE/ACM Trans Netw* 24(6):3523–3536
5. Ji S, Mittal P, Beyah R (2017) Graph data anonymization, de-anonymization attacks, and de-anonymizability quantification: a survey. *IEEE Commun Surv Tutor* 19(2):1305–1326
6. Kergl D (2015) Enhancing network security by software vulnerability detection using social media analysis extended abstract. In: 2015 IEEE international conference on data mining workshop (ICDMW). pp 1532–1533
7. Liu G, Wang C, Peng K, Huang H, Li Y, Cheng W (2019) Socinf: membership inference attacks on social media health data with machine learning. *IEEE Trans Comput Soc Syst* 6(5):907–921
8. Ninggal MIH, Abawajy J (2011) Attack vector analysis and privacy-preserving social network data publishing. In: 2011 IEEE 10th international conference on trust, security and privacy in computing and communications. pp 847–852
9. Orabi M, Mouheb D, Al Aghbari Z, Kamel I (2020) Detection of bots in social media: a systematic review. *Inf Process Manage* 57(4):102250
10. Patil NA, Manekar AS (2015) A novel approach to prevent personal data on a social network using graph theory. In: 2015 international conference on computing communication control and automation. pp 186–189
11. Rekha HS, Prakash C, Kavitha G (2014) Understanding trust and privacy of big data in social networks - a brief review. In: 2014 3rd international conference on eco-friendly computing and communication systems. pp 138–143
12. Reza KJ, Islam MZ, Estivill-Castro V (2017) Social media users' privacy against malicious data miners. In: 2017 12th international conference on intelligent systems and knowledge engineering (ISKE). pp 1–8
13. Sharma VD, Yadav SK, Yadav SK, Singh KN, Sharma S (2021) An effective approach to protect social media account from spam mail – a machine learning approach. *Mater Today: Proc*
14. Sushama C, Sunil Kumar M, Neelima P (2021) Privacy and security issues in the future: a social media. *Mater Today: Proc*
15. Tian W, Mao J, Jiang J, He Z, Zhou Z, Liu J (2018) Deeply understanding structure-based social network de-anonymization. *Procedia Comput Sci* 129:52–58
16. Wang B, Jia J, Zhang L, Gong NZ (2019) Structure-based sybil detection in social networks via local rule-based propagation. *IEEE Trans Netw Sci Eng* 6(3):523–537
17. Yang D, Qu B, Cudré-Mauroux P (2019) Privacy-preserving social media data publishing for personalized ranking-based recommendation. *IEEE Trans Knowl Data Eng* 31(3):507–520
18. Zhang J, Sun J, Zhang R, Zhang Y, Hu X (2018) Privacy-preserving social media data outsourcing. In: IEEE INFOCOM 2018 - IEEE conference on computer communications. pp 1106–1114



# Image Forgery Detection Using Supervised Learning Algorithm



R. Cristin, T. Daniya, and S. Divyatej

**Abstract** The pictures on the Internet and the photos-sharing platform are extremely possible to be edited with other universal alterations like compression, resizing, or screening. So, image forgery detection algorithms are used to detect such manipulations on images. A framework is presented to improve the reliability of image forgery detection. The main stage in the structure is to take into consideration the image performance of the software selected. Hence, utilize a camcorder model recognition based on CNN. JPEG is an image compressor that enables us to investigate the alteration and is the most popular sort of deliberate and involuntary concealment of picture forgeries. Consequently, a variety of two qualities compressed and uncompressed images is provided to the qualified CNN. To better decipher our trained CNN, we initially recommended thorough monitoring of the layers but also an experimental evaluation of the impact of the learning features. This examination will take us to a stronger and more precise approach. Furthermore, in an image forgery detection scenario, we will use this upgraded detection algorithm and show impressive outcomes in the identification and demonstration of the locations where the image has been forged or manipulated.

**Keywords** Forgery detection · Convolutional neural network · Camera model identification · Image pre-processing · Compression

## 1 Introduction

Nowadays, the ubiquitous use of intelligent devices including cameras and smartphones, and the accessibility of Internet sharing have made images a comprehensive part of our lives. The amount of image processing software grows in conjunction with reality. They are so approachable that everyone may quickly change and distribute

---

R. Cristin (✉) · T. Daniya · S. Divyatej  
GMR Institute of Technology, Rajam, Andhra Pradesh, India  
e-mail: [cristin.r@gmrit.edu.in](mailto:cristin.r@gmrit.edu.in)

T. Daniya  
e-mail: [daniya.t@gmrit.edu.in](mailto:daniya.t@gmrit.edu.in)

photographs online. At the same time, picture manipulation has become a way of seriously damaging our community. Numerous ways have existed: the most frequently used modifications are splicing, copy-moving, and removal.

**Kinds of Forgeries:** Generally, forgeries include splicing, copy-move, removal, image retouching, and morphing.

- **Splicing:** Splicing is a process of alteration that involves copying and pasting one or perhaps more image segments into a single image. This may be utilized to contribute a further aspect to a scenario.
- **Copy-move:** Copy-move may be utilized to add erroneous data or to hide data.
- **Removal:** Removal is the replacement of some sections of an image. This can be used to remove things for hidden information purposes.
- **Image Retouching:** With this forgery, image quality can be enhanced or degraded by some image alterations. The criteria that can be changed may include picture color, change image lights, etc. Major pixel-level alterations are performed through various movables such as a Gaussian filter, a smoothing filter that hides actual image truth and improves the appearance of the image. The goal here is to determine whether the image is smoothened. Steganalysis could be one of the ways to employ image retouching recognition.
- **Morphing:** Morphing is a distinctive effect in movies and animations, which changes (or moves) an image or forms into a different image through a smooth change.

For non-experts, some of the changes are hard to notice. In addition, some of the modified photographs try to communicate deceptive, socially threatening information (e.g., massive manipulation, cyber-crime, tamper-evident or removal of court evidence, and so on). The criminal research field, therefore, concentrated in the recent ten years on the development of instruments that evaluate the integrity of the image. Many ways have been presented to recognize the validity of the photos and to evaluate their integrity. Some of them concentrated their interest in forensic matters. Pictures are taken in many instances in which they can have high repercussions to determine their integrity and origins. In criminal investigations or for news coverage, for example, it is crucial. Thus, it is one of the most significant duties of the imaging forensic community to establish the image source and its legitimacy. The information collected from the image and from the accompanying multimedia content could reveal some discrepancies that demonstrate suspected image or document fabrication.

Deep learning is an artificial intelligence function (AI), which imitates the functioning of the human brain in the gathering of information. Deep learning is the subset of artificial intelligence learning with networks that can learn from unchecked or unstructured material without being monitored. Deep neural learning is also known as a deep neural network. Deep learning uses hierarchical neural networks to analyze the data as a subset of machine learning. In these, the neural coding identical to the human intellect is interconnected. The hierarchical structure of deep learning, unlike other typical linear programs, permits a nonlinear method, which is used to treat data over a number of layers, integrating succeeding levels of more information.

Computer programs that use deep learning are used to determine the dog by much the same procedure as the infant. In each algorithm, in the hierarchy, a nonlinear transformation is applied to the input of the algorithm and the results are what they learn. The changes continue up to an acceptable level of accuracy of the output. What motivated this label deeply was the number of levels through which data should transit. The learning process is supervised in classical machine learning, and the developer has to be exceedingly detailed when instructing the computer on what kind of item it should choose whether a picture contains a dog or not. The process of extraction of functions is lengthy and the success rate of a computer is totally dependent on the skill of the programmer to precisely define a dog's functionality. Deep learning has the advantage of allowing software to develop its own qualities without the need for human intervention. Learning that is not regulated, is usually more precise, and not merely faster.

Different techniques can be employed to develop strong models of deep learning. These strategies include decay, transfer, scratch, and drop-out education. Decline in learning rate. A hyperparameter is the rate of learning—a quantity that defines the system or sets its operating settings in advance of the learning process—which checks how much the model experiences change when the predicted error is revised. Too high levels of learning might lead to unstable training or a sub-optimal group of weights. Too little learning rates can lead to a long training process that can get stopped.

The method of decline in learning rates is also termed rinsing or adaptive learning rates by adjusting the study rate to improve performance and decrease training periods. The easiest and most popular adjustments to the training rate are approaches for decreasing the level of learning over time.

There are different varieties of neural networks, including recurring neural networks, convolutions neural networks, artificial neural networks, and feedforward neural networks. They all function identically, via inputting information and allowing the prototype to check if it has correctly understood or decided on a particular fact aspect.

Neural networks entail the process of trials and fault and hence require large quantities of data to train. It was only after several companies embraced Big Data Analytics and gathered vast data vaults that neural networks became prominent. Due to the slightly informed assumptions of the content of the image or sections of a speech in the initial few iterations of the model, the information used during the training phase should be labeled to allow the model to see whether its guess is accurate. This makes unstructured data less useful even though many businesses that utilize massive data have large amounts of data. Unstructured information can only be analyzed by a deep learning model if the model is trained and achieves an adequate degree of precision.

Deep learning refers to a subset of deep neural networks, which are most commonly employed in spectral evaluation to analyze visual images. We now think about matrix multiplication, but that's not the case with ConvNet. If you think about

a neural network, it employs convolution, a specific method. In mathematical convolution, two functions are a mathematical function that creates a third function, which demonstrates how one's form is altered by the other.

Multilayer perceptrons are CNN variations that have been regularized. Multilayer perceptrons are sometimes referred to as completely linked networks, in which every neural node in one layer is connected to all neural nodes in the following layer. These networks are “fully connected”, and therefore susceptible to data overfitting. Typical regularization or prevention methods involve punishing (such as weight decline) or cutting connection parameters throughout training (skipped connections, dropout, etc.) The CNNs adopt various approaches to regularity: employing smaller and more simple patterns embodied in their filters, and they employ a hierarchical data structure to create increasingly complicated patterns. CNNs also are at the lower extremity in the level of connection and complexity.

Artificial neurons are numerical values that estimate the summation of numerous input data and output data, as well as the initiation value, in a rudimentary replication of their natural correlatives. Each layer creates many activation functions when you enter an image in a ConvNet, which are transmitted to the next layer.

Recently, deeper neural networks such as the Deep Belief Network (DNN), the Deep Auto Encoder, and the CNN have been able to retrieve and effectively learn the hierarchical representations and computer vision (CV) tasks from complex statistical dependencies, including image classification tasks. In recent times, the deep learning approach found applications in the field of passive picture forensics.

The layout of this project is organized as follows:

- Initially, a short overview of the image forgery techniques published by employing convolutionary neural networks is presented.
- Then, the universal framework includes: pre-processing of images, CNN monitoring, and image forgery analyses.
- Then, we describe the need of taking compression modification into consideration, then assess our suggested structure to make CNN better understood.
- Finally, we emphasize the efficacy of our method for detecting forged images.
- Finally, we determine and depict various promising results of detecting the image, i.e., regardless of whether the image is immaculate or fake and show the performance analysis of our model.

## 2 Literature Survey

In paper [1], the authors proposed a picture recognition and location technique that leverages the characteristics of the footers left by various camera models in the photos. The suggested approach uses a CNN to analyze aspects from the image patch for the characteristic camera model. Those characteristics are then evaluated by iterative clustering techniques to find out if a picture is fabricated and to locate the alien region.

In this paper [2], they discussed copy-driving forgery, seam carving, splicing, and re-compressing popular image tampering techniques. Deep learning algorithms are currently utilized to detect picture tampering and to provide better outcomes. Drawback: The technique is not reliable for tampering detection where a post-processing procedure has not been performed.

In paper [3], the authors described Custom CNN for camera model identification architecture. They use methods like batch normalization and abandonment and do not use pre-trained networks to reduce the time of training. When compared to the previously proposed CNN-driven SVM classification technique, the efficiency of the suggested CNN-based strategy is improved in the form of enhanced diagnosis and fewer losses, based on a relatively small public repository. Drawback: The present framework is educated on pictures not resilient to any post-processing.

In paper [4], the authors proposed a generic platform that can resolve the identification model of the camera for altered still photos. The alteration factors and dedicated function extractor templates are proposed for the evaluation of the source camera using the CNN Architecture. Multiplexers are used to change the input images to the output of the CNN between the specific models. Drawback: The solution proposed tackles the scalability problem, as more modifications are predicted. More models mean that a product from the sector cannot be provided.

The aim of the work [5] is to identify the manipulated image by applying some algorithms and technical devices. CNN is employed for classifying the image as segmented type or copy-move type as they are common approaches for image forgery to give a better performance of the model. The simulation values depict the better performance of the presented model.

In this work [6], a new approach based on convolutionary neural network is suggested to identify forgery copy-move. The suggested method uses the already trained large-database model like ImageNet, and with short training samples the adaptation somewhat to the net structure. The drawback of the proposed solution is not resilient to the real scenario of copy-move images. The author presented a study on image forgery detection in this paper [7, 8] in a systematic way. The reflection [9, 10] and shadow [10, 11] based forgery detection are proposed. The various Machine Learning Applications like Agriculture [12, 13], Regression [14], Classification [14, 15] are popular in the research field.

### 3 Methodology

We discuss our methodology in this section. Its main aim is to enable the detection of picture forgery with a robust and effective framework. Figure 1 illustrates our proposal's national strategy. There are four particular pieces in this framework. Part one relates to all the processing phases of the image. In this part, we also demonstrate the significance to increase resilience by taking into consideration the quality of the incoming data. Then we discuss the strategy of classification utilizing the convolutional neural network. Next section, the analysis of our CNN is thorough and enables

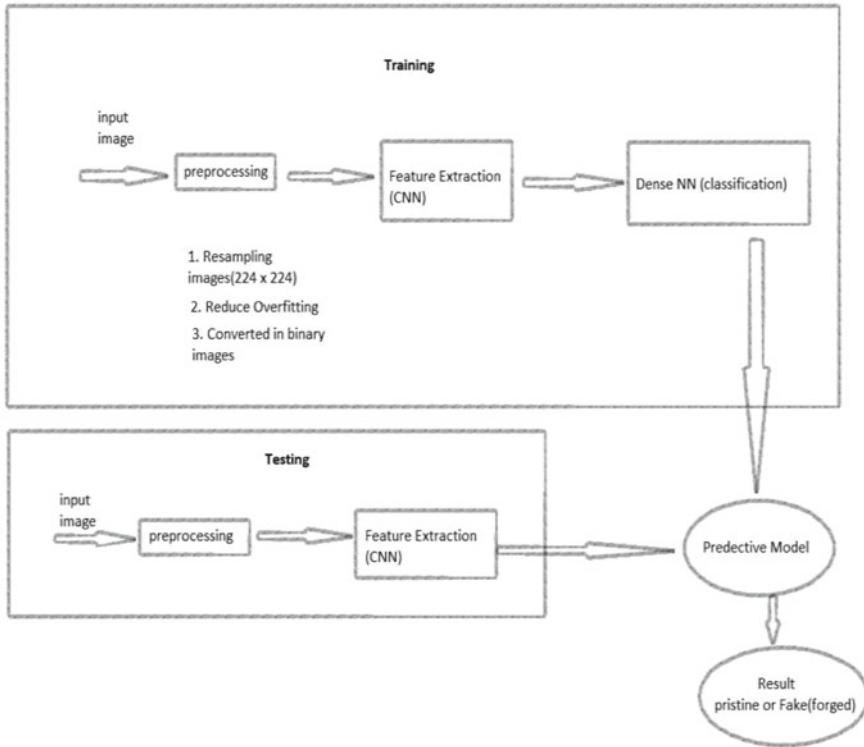


Fig. 1 Architecture of our proposed system model

us to grasp our methodology more clearly and enhance it. Finally, we are testing our suggested forgery detection framework application. The proposed system is a framework designed for pre-processing the dataset and then training the model using a Convolutional Neural Network and then testing the model by feeding the test set images to the trained Convolutional Neural Network and predicting the output result whether the given image is pristine (original) or forged (fake).

Initially, an input image is taken from the image dataset. Pre-processing techniques such as Grayscale, Resize, and Normalization are applied to the image. Then we use Convolutional Neural Network (CNN) for image segmentation such as pristine or fake. Now, our training model can predict the image whether it is pristine or fake. When the test set is fed into our training model as an input, then our predictive model predicts whether it is a pristine image or a fake image and gives us the result in binary form.

**Pre-processing:**

Data pre-processing is a process for the preparation and adaptation of raw information for a system of machine learning. It is the first and most important step in the development of a model of master learning. It is not always the case that we come

across clean and prepared data when establishing a machine learning project. And it is necessary to clean it and put it in the format while executing any function with data. We employ pre-processing data for this purpose. Generally, real-world data comprises noises, missing values, and perhaps in an unsuitable format that cannot be used straight for machine learning models. Data pre-processing is necessary in order to clean the information to make it acceptable for a machine-learning model, which also makes the machine-learning model more accurate and efficient.

### Gray scaling:

Gray scaling is the method for transforming an image to gray tones from different colour spaces such as RGB, CMYK, etc. It ranges from full black to full white. Image conversion to grayscale is done for simplicity and data reduction. Simplicity: Many image processing operations work on one plane of an image (i.e., single color) at a time. So, if you have an RGB image you have to apply operations of three image planes and combine the results. So gray scaling is done. Data reduction allows the algorithm to run in a reasonable amount of time.

### Resize:

It refers to changing the size of an image. Resizing can be done by scaling, cropping, and padding. In scaling, when we re-dimension an image, a new image with higher or lower pixels is generated. Here, we resized the image to a lower number of pixels.

### Normalization:

Normalization is a technique that adjusts the pixel intensity range. Normalization is a crucial step to assure a comparable data distribution for each input parameter (in this case, pixel). Photos in the range 0–1 are legitimate and images can be seen normally. All pixels can be divided up by the highest pixel value; that's 255 (Fig. 2).

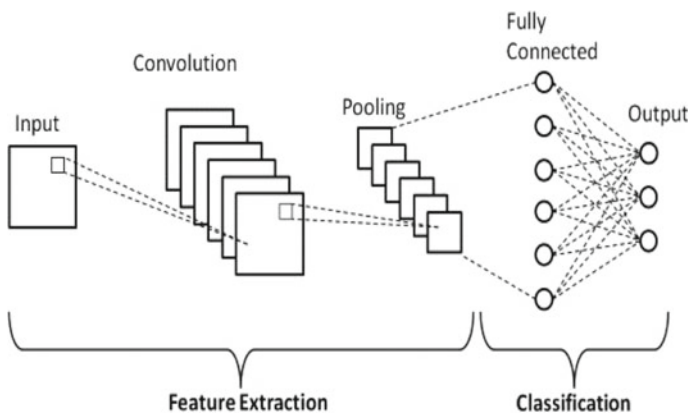


Fig. 2 Architecture of convolutional neural network

### **Convolutional Neural Network:**

Convolutional Neural Networks (CNNs) are one of the most important criteria for image recognition, picture categorization, object detection, and face recognition, among other things. Input images are processed and classified by CNN image classification (e.g., Dog, Cat, Tiger, Lion). Image input is used by computers as a pixelated array, and the quality of an image is determined by that. The resolution shows  $h \times w \times d$  ( $h$  = height,  $w$  = width,  $d$  = dimension) depending on the quality of the picture. Each input image will travel through a series of layers of deep CNN models for training and testing. Finally, use SoftMax to categorize an item using deterministic probabilities ranging from zero to one.

In the 1980s, CNNs were initially created and used. The most a CNN could do at that time was to distinguish manuscript numbers. In the postal sector, it was mostly used for reading zip codes, pin codes, etc. Any deep learning model is crucial to note that it needs a huge number of data to be trained and a lot of computer resources as well. This represented a severe disadvantage for the CNNs during that time, which led to CNN being restricted only to postal services and failing to reach the machine learning field.

LeCun et al. trained the handwritten letters in the late 1980s. In 1998, this CNN ancestor matured and its higher rating power was seen in the handwritten number basis of the MNIST. CNN was subsequently employed with good results in several applications in the 1990s. But with the introduction of other approaches like SVM's and Bayesian Networks (BNs), they quickly became obsolete.

In 2012, Alex Krizhevsky discovered that the field of deep learning that uses multilayered neural networks was time to come back. The availability of enormous data sets with millions of labelings and numerous computer resources made it possible for researchers to rekindle CNNs. The ImageNet dataset was also more particular. CNN has become a common tool for dealing with picture detection counterfeiting. In the range of computer vision and image processing applications, they have obtained an accuracy performance to address detection and ranking difficulties.

In 2012, CNN returned and became a leader in image classification techniques literature. Two primary elements made this event feasible. First of all, high-performance GPU devices became economical for everyone. So even on a laptop computer, training of CNN was now achievable with a tiny dataset. The second key element was the enormous number of photos on the Internet, especially with the Image Net Large Scale Visual Recognition Challenge (ILSVRC). This challenge in Image Net has now become a showdown benchmark for large firms such as Google, Microsoft, IBM, and many other deep learning models like Google Net, which are typically utilized in extracting deep characteristics that contain a particular application grading layer.

Convolutional neural networks have also been a fairly common method for fake imaging problems in recent years. The convolutionary layer has been built especially for the detection of the characteristics that have been manipulated. The technology is trained on a wide array of unmodified photos and datasets of various changes such as scaling.



In a neural network, the activation feature transforms the weighted summary input from the node into the node activation or output for that input. The revised linear function or ReLU for short is the linear function which, if positive, outputs the input directly, else outputs zero. It has become the standard objective factor for various types of neural networks since the model that uses it is comparatively faster and frequently improves performance.

Leaky ReLU is an enhanced ReLU Activation function variant. The descent for all input values is lower than zero, which would disable neurons in this region and create a ReLU death problem, which is 0 for the ReLU activation function. To solve this issue, Leaky ReLU is introduced. We describe it as an incredibly small linear component  $x$  rather than the ReLU Activation Function being 0 for negative inputs( $x$ ).

If the input is positive, this method returns  $x$ , but if the input is negative, it returns that even less number, 0.01 times as a result  $x$ . Thus, it gives an output for negative values as well. The gradient of the left-hand side of the graph appears to have a non-zero value by making this tiny change. Therefore, in this region, we no longer would meet dead neurons.

Adaptive Timing Estimation is a gradient descent optimization algorithm. The strategy for dealing with large issues including many information or criteria is highly efficient. It needs less storage and is efficient. The coupling of 'gradient descent with momentum' with the 'RMSP' method is intuitive. Adam optimizer includes two methods of gradient descent.

The most frequent approach to optimizing deep learning networks is gradient descent. In the 1950s, the technique was first developed to modify each of the system parameters, examine how a change affects the objective function, select a path that reduces the error rate, and continue to iterate until the target function reaches a minimum. SGD is a variation of gradient descent. The SGD only calculates small subsets or a random selection of data instances instead of computing the complete dataset, which is unnecessary and ineffective. If the learning rates are modest, SGD delivers the same performance as standard gradient descent.

### **Convolutional Layer:**

Convolutionary layer is sometimes termed the extractor layer, as the picture characteristics of this layer are extracted. First of all, the convolutional layer connects a portion of the image to the converting operating system as we had seen before, and computes the dot product between the receptive field and the filter. The output impedance is the result of the procedure. After that, we drag the filter to a Stride and repeat the process on the next reception sector of a certain input picture. Till we pass through the complete picture, we restart that process time and again. The output is the next layer's input. Convolution is the initial layer that analyzes characteristics from an inputted picture. The use of compact levels of data in the learning of visual characteristics maintains the link between frames. Two parameters, such as the Image Matrix and a filter or kernel, are required for the math operation.

### **Non-Linearity (ReLU):**

Rectified linear units rapidly disseminate gradient and hence lower the probability that deep neural architectures have an acute gradient problem. Rectified linear units have a negative threshold value of zero, so cancellation problems are solved and their volume is significantly slower. For several aspects, the sparsity is helpful but provides mostly resilience to tiny input changes such as clamor. The rectified linear units contain easy calculation functioning (mostly comparisons), which makes their application in convolutionary neural networks significantly more feasible. ReLU stands for a nonlinear operation as a rectified linear unit. The result is  $f(x) = \max(0, x)$ . The aim of ReLU is to incorporate in our CNN non-linearity. Since real-world data are non-negative linear numbers, our CNN would desire to learn. Other nonlinear functions, such as tanh or sigmoid, are also available in place of reliability. But use ReLU here because ReLU's effectiveness is superior to the two others. As a consequence of the benefits and efficiency of most of the newer layouts of convolutionary neural networks, rather than typical nonlinearity and rectification layers, rectified linear unit nonlinearity layers are utilized.

### **Pooling Layer:**

The main theory of the pooling layer is interpretation invariance, as characteristic detection and is largely compared to the directed position, especially in image recognition tasks. The pooling method, therefore, seeks to safeguard the characteristics found in a lesser representation, with less important data being thrown out at the expense of spatial resolution.

The pooling process includes the selection of a pooling procedure, similar to a functional map filter. The quantity of the pooling operation or filters, specifically 22 pixels applied with a step of 2 frames, will be less than the extent of the featured vector.

The pooling of map sizes is therefore always reduced by a ratio of 2. For example, each size is halved, decreasing to 1/4 of the size the number of frames or measures on each map. An outputted pooled map of  $3 \times 3$  (9 pixels) is produced, for instance, in a pooling layer provided to a map of  $6 \times 6$  (36 pixels). Pooling would decrease parameters if the pictures are too big. Spatial pooling is also referred to as subsamples or samples that lower the dimensionality of every map yet maintains significant information. From the corrected feature map, Max pooling chooses the largest element. It can alternatively take the average pooling of the greatest element. All the feature map items are summarized as sum pooling.

The largest element in Max Pooling is drawn from the function map. In a specified sized image segment average pools compute the expected elements. In Sum Pooling, the total number of components is calculated in the preset area. The layer of swimming usually acts as a bridge between the convolutionary layer and the FC layer. There is no pooling layer parameter, but the filter (P) and stride (R) are two hyperparameters. In general, if we have input dimension  $W1 \times H1 \times D1$ , then  $W2 = (W1 - P)/R + 1$   $H2 = (H1 - P)/R + 1$   $D2 = D1$ , where  $W2$ ,  $H2$  and  $D2$  are the output's width, height, and depth.

**Fully Connected Layer:**

Fully linked layer includes weights, distortions, and neurons. Each layer's neurons are linked to neurons in some other layer. It is used to identify pictures by training between different categories. The last layer of CNN is Softmax or Logistic Layer. It is located at the bottom of the FC. For simple binary grading, logistics is used, while Softmax is used for multi-segmentation grading. We have flattened the layer called the FC layer into a vector and transferred it into a completely connected layer. The feature matrix will be transformed as a vector ( $x_1, x_2, x_3, \dots$ ). These features were combined to create a technique with fully connected layers. Finally, the Softmax or sigmoid initiation model is used for classifying outputs.

**Approach for Modified CNN:**

Step 1: Loading Image Dataset consisting of pristine and fake images.

Step 2: Apply Pre-processing Techniques.

Step 3: Apply the CNN model.

Step 4: CNN using different activation functions, optimizers, and learning rates.

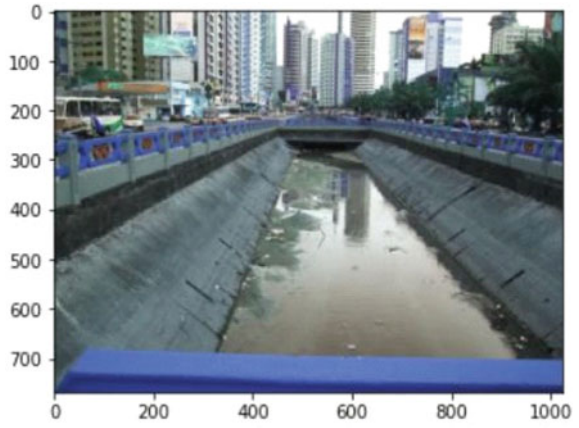
Step 5: Repeat step 2, 3, and 4 and find the results that the image is pristine or fake.

Step 6: Stop the procedure.

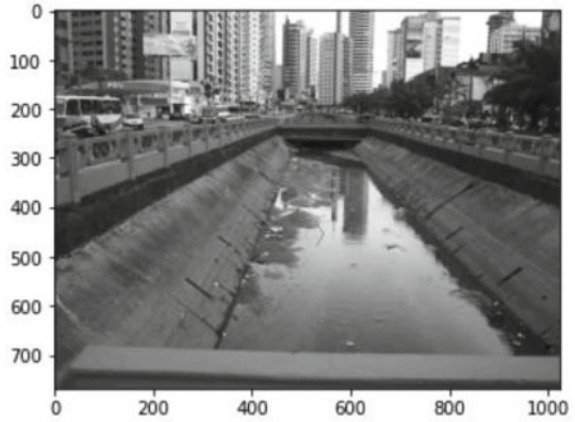
## 4 Results and Discussions

Gray scaling is a technique for converting a picture from other colors to grayscale, e.g., RGB, CMYK, etc. to shades of gray. It varies between complete black and complete white. Image converting to grayscale is done for simplicity and data reduction. Simplicity: Many Image processing operations work on one plane of an image (i.e., single color) at a time. So, if you have an RGB image you have to apply operations of three image planes and combine the results. So gray scaling is done. Data Reduction allows the algorithm to run in a reasonable amount of time. It refers to changing the size of an image. Resizing can be done by scaling, cropping, and padding. In scaling, when we resize a picture, we generate a new picture with greater or lesser pixel resolution. Here, we resized the image to a lower number of pixels (Figs. 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15) (Tables 1, 2, 3, 4, 5, and 6).

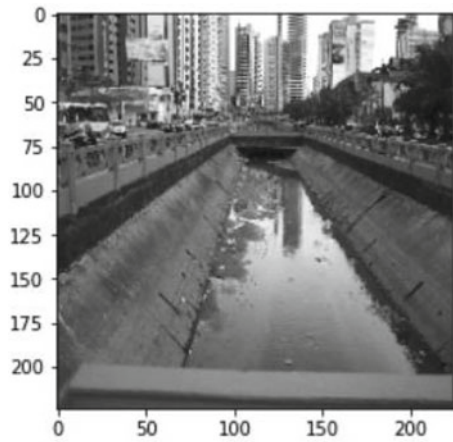
**Fig. 3** Input image



**Fig. 4** Grayscale image



**Fig. 5** Resize image



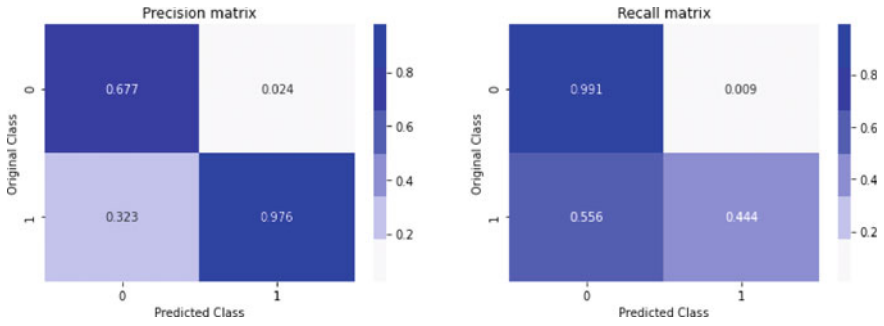


Fig. 6 Summary of CNN using ReLU activation function confusion matrix

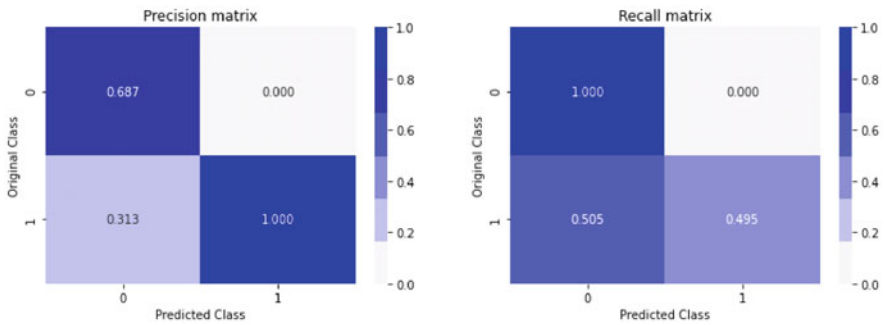


Fig. 7 Summary of CNN using Leaky ReLU activation function confusion matrix

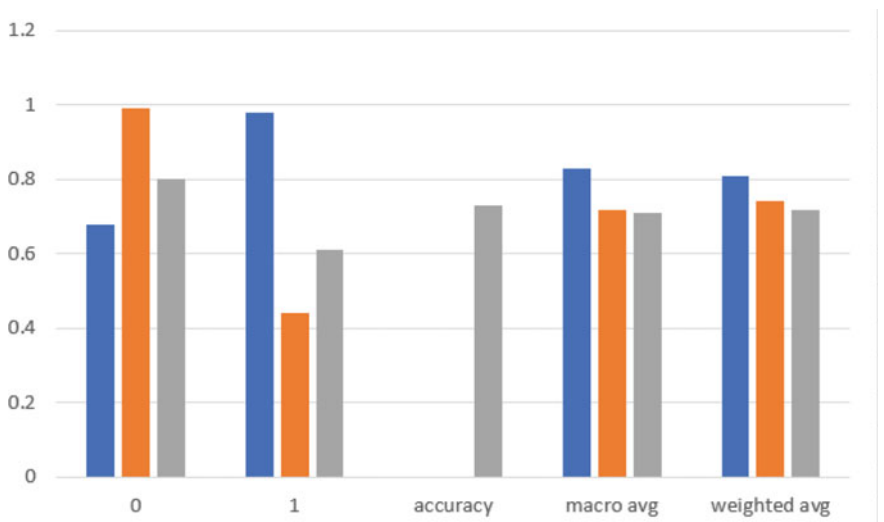


Fig. 8 Graph for performance using ReLU and Adam with 0.001 lr

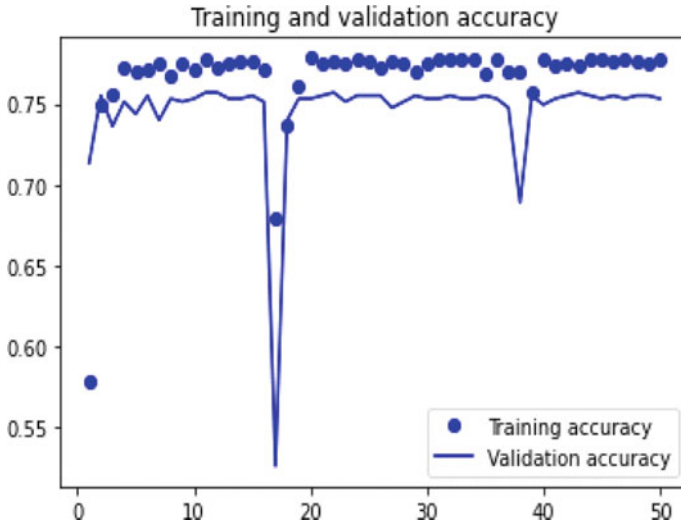


Fig. 9 Training and validation accuracy graph

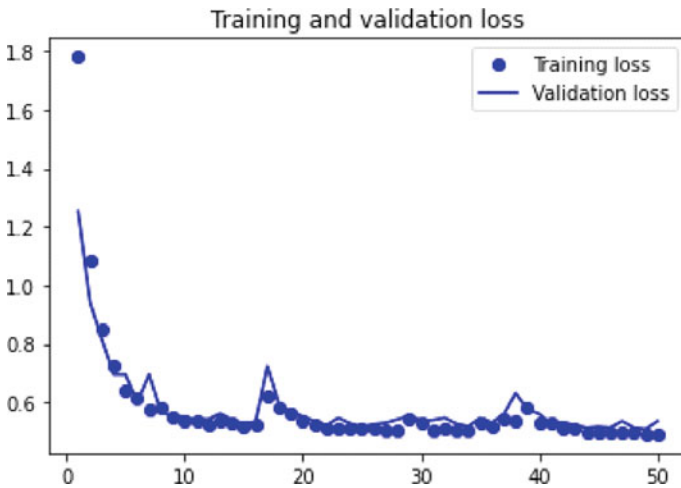


Fig. 10 Training and validation loss graph

## 5 Conclusion

Online sharing of images has a high chance of being modified or manipulated. Hence, we proposed a robust framework for Image forgery detection using image forgery detection algorithms. Here, we used CNN algorithm which is a deep learning

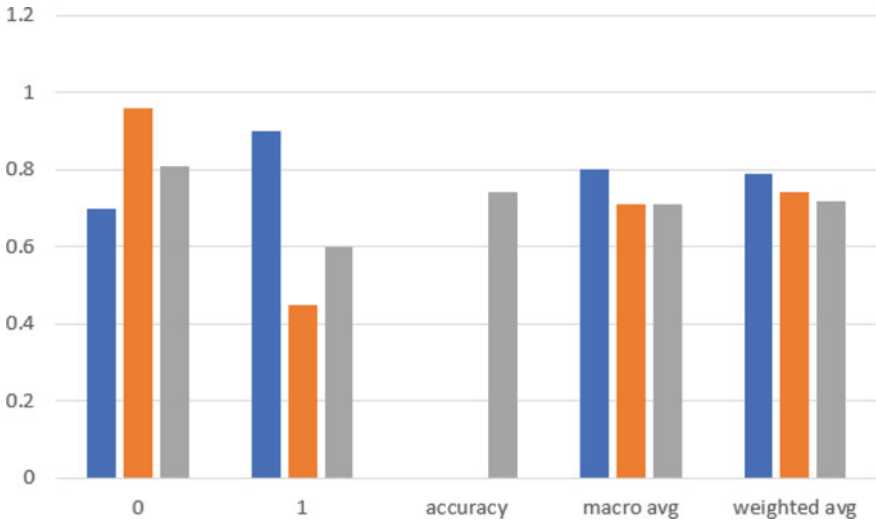


Fig. 11 Graph for performance using ReLU and Adam with 0.0001 lr

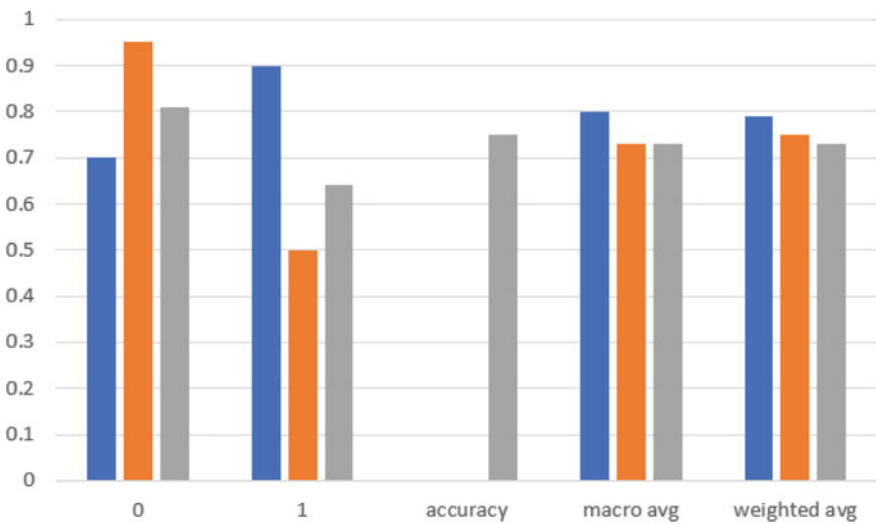


Fig. 12 Graph for performance using ReLU and Adam with 0.00001 lr

algorithm. Initially, pre-processing of the dataset images is done using various pre-processing techniques. Then, training of the CNN model is done using the training set and when we feed the CNN with the test dataset it predicts the results whether the image is pristine or fake. This task necessitates the use of a variety of stimulation mechanisms and optimizers in order to improve efficacy over a wide range of image changes. The future work includes increasing the layers of Convolutional

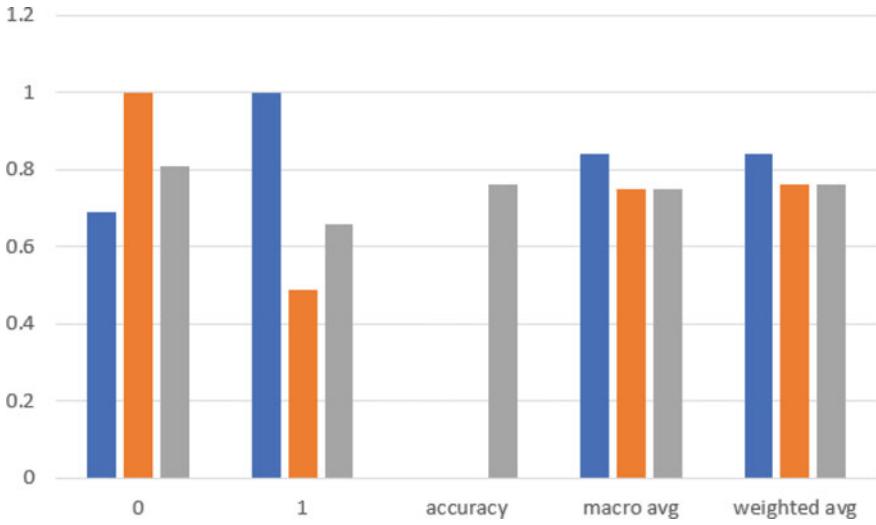


Fig. 13 Graph for performance using Leaky ReLU and Adam with 0.001 lr

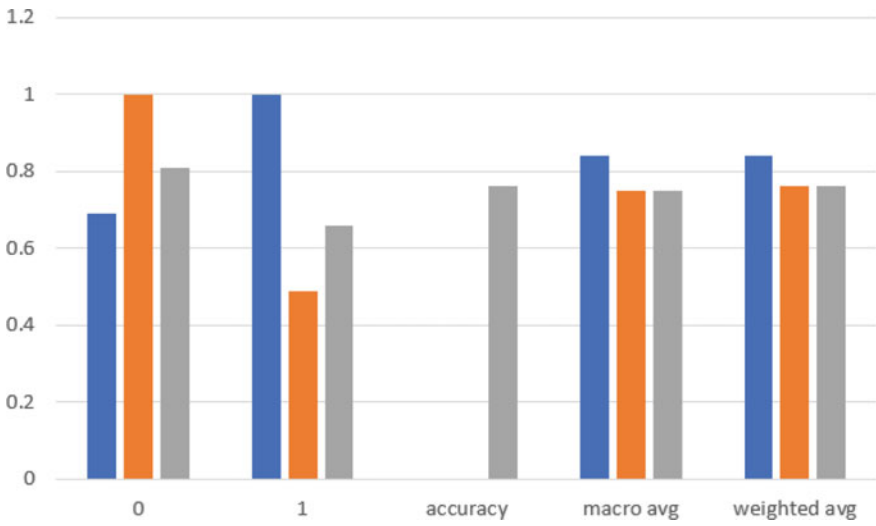
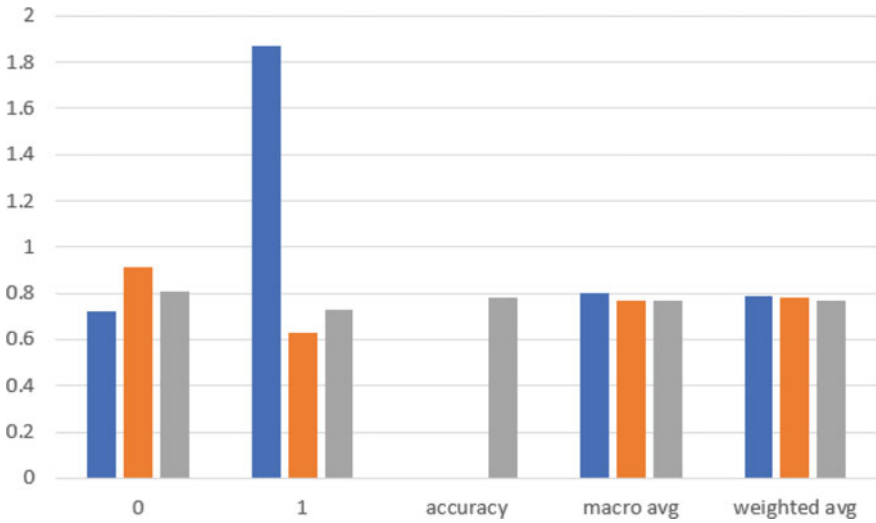


Fig. 14 Graph for performance using Leaky ReLU and Adam with 0.0001 lr

neural networks using different activation functions and optimizers and the use of other techniques like data augmentation and error level analysis (ELA) using CNN are used.





**Fig. 15** Graph for performance using Leaky ReLU and Adam with 0.00001 lr

**Table 1** Performance analysis of CNN with ReLU using 0.001 lr

	Precision	Recall	F1-score
0	0.68	0.99	0.80
1	0.98	0.44	0.61
Accuracy			0.73
Macro avg	0.83	0.72	0.71
Weighted avg	0.81	0.74	0.72

**Table 2** Performance analysis of CNN with ReLU using 0.0001 lr

	Precision	Recall	F1-score
0	0.70	0.96	0.81
1	0.90	0.45	0.60
Accuracy			0.74
Macro avg	0.80	0.71	0.71
Weighted avg	0.79	0.74	0.72

**Table 3** Performance analysis of CNN with ReLU using 0.00001 lr

	Precision	Recall	F1-score
0	0.70	0.95	0.81
1	0.90	0.50	0.64
Accuracy			0.75
Macro avg	0.80	0.73	0.73
Weighted avg	0.79	0.75	0.73

**Table 4** Performance analysis of CNN with Leaky ReLU using 0.001 lr

	Precision	Recall	F1-score
0	0.69	1.00	0.81
1	1.00	0.49	0.66
Accuracy			0.76
Macro avg	0.84	0.75	0.75
Weighted avg	0.84	0.76	0.76

**Table 5** Performance analysis of CNN with Leaky ReLU using 0.0001 lr

	Precision	Recall	F1-score
0	0.69	1.00	0.81
1	1.00	0.49	0.66
Accuracy			0.76
Macro avg	0.84	0.75	0.75
Weighted avg	0.84	0.76	0.76

**Table 6** Performance analysis of CNN with Leaky ReLU using 0.00001 lr

	Precision	Recall	F1-score
0	0.72	0.91	0.81
1	1.87	0.63	0.73
Accuracy			0.78
Macro avg	0.80	0.77	0.77
Weighted avg	0.79	0.78	0.77

## References

1. Diallo B, Urruty T, Bourdon P, Fernandez-Maloigne C (2020) Robust forgery detection for compressed images using CNN supervision. Elsevier B.V.
2. Padmanabhan R, Damodaran S, Batra VN, Gurugopinath S (2020) A convolutional neural network architecture for camera model identification with small datasets. IEEE
3. Thakur R, Rohilla R (2019) Copy-move forgery detection using residuals and convolutional neural network framework: a novel approach. In: IEEE 2nd international conference
4. Hema Rajini N (2019) Image forgery identification using convolutional neural network. Int J Recent Technol Eng (IJRTE)
5. El-Yamany A, Fouad H, Raffat Y (2018) A generic approach CNN-based camera identification for manipulated images. IEEE
6. Kuzin A, Fattakhov A, Kibardin I, Iglovikov VI, Dautov R (2018) Camera model identification using convolutional neural networks. IEEE
7. Santhosh Kumar B, Karthi S, Karthika K, Cristin R (2018) A systematic study of image forgery detection. J Comput Theor Nanosci (Scopus Indexed) 15(8):1–4
8. Santhosh Kumar B, Cristin R, Karthick K, Daniya T (2019) Study of shadow and reflection based image forgery detection. In: IEEE international conference on computer communication and informatics (ICCCI)
9. Cristin R, Gladiss Merlin NR, Daniya T (2020) Geometrical based technique for reflection based image forgery detection in digital images. Int J Sci Technol Res 9(1):2654–2659

10. Cristin R, Cyril Raj V (2017) Consistency features and fuzzy based segmentation for shadow and reflection detection in digital image forgery. *Sci China Inf Sci* 65(1):43–66
11. Merlin NRG, Santhosh Kumar B, Cristin R, Arun Sekar R (2019) Improved segmentation techniques for image manipulation with shadow based digital forgery using vanishing point computation. *Int J Sci Technol Res*
12. Daniya T, Vigneshwari S (2021) Deep neural network for disease detection in rice plant using the texture and deep features. *Comput J*. <https://doi.org/10.1093/comjnl/bxab022>
13. Daniya T, Vigneshwari S (2020) A review on machine learning techniques for rice plant disease detection in agricultural research. *Int J Adv Sci Technol* 8(13)
14. Daniya T, Geetha M, Santhosh Kumar B, Cristin R (2020) Least square estimation of parameters for linear regression. *Int J Control Autom* 13(2)
15. Daniya T, Geetha M, Kumar KS (2020) Classification and regression trees with Gini index. *Adv Math: Sci J* 9:8237–8247

# A Machine Learning Method for Customer Sentiment Analysis on Social Media



Chetan Agrawal, Anjana Pandey, and Sachin Goyal

**Abstract** Customer Data analysis is a significant part of different ventures utilizing figuring applications, for example, E-business and online shopping. Enormous information is utilized for advancing items which gives better availability among retailers and customers. These days, individuals consistently utilize online advancements to think about the best shops for purchasing better items. This shopping experience furthermore, assessment of the customer's shop can be seen by the client experience shared across online media stages. Another client while looking through a shop needs data about manufacturing date and manufacturing price, offers, quality, and ideas which must be given by the past client experience. The MRP and MRD are as of now accessible on the item cover or mark. A few methodologies have been utilized for anticipating the item subtleties however not giving precise data. This paper is persuaded toward applying Machine Learning algorithms for picking up, breaking down, and arranging the item data and the shop data dependent on the client experience. The accuracy of the proposed method is 99.4%.

**Keywords** Machine learning · Social media · Sentiment analysis · Twitter · Classifier

## 1 Introduction

Sentiment analysis is a method of studying the connotation of the text. A text can be written in various sentiments. It can be to encourage or to demean the very subject. In today's world, social media is gaining pace and each individual has somewhere or

---

C. Agrawal (✉)  
Department of CSE, UIT-RGPV, Bhopal, India  
e-mail: [chetan.agrawal12@gmail.com](mailto:chetan.agrawal12@gmail.com)

A. Pandey · S. Goyal  
Department of IT, UIT-RGPV, Bhopal, India  
e-mail: [anjanapandey@rgtu.net](mailto:anjanapandey@rgtu.net)

S. Goyal  
e-mail: [sachingoyal@rgtu.net](mailto:sachingoyal@rgtu.net)

the other expressed their interests, opinions, likes, and dislikes on any social media platform. The form of expressing views can be either explicit or implicit. But it is noteworthy that with this boom in online opinion sharing, there is immense scope to understand the view of the individual. When it comes to businesses, getting to know what is the perception of customers toward their brand, product, or service can be bliss. This intent of the customers can be obtained with the help of Sentiment Analysis. Consider the below examples:

- I. "It's a very nice old place since very authentic, peaceful place, a good place to dine. It's a simple place don't expect too much like having music and all. It's on 2 floors not too much parking space but you will get the paid parking space."

This is the review of a customer about a restaurant. We can sense that the customer is satisfied with the place. He has emotional connectivity as well with the restaurant. Here is another review of some other customers in the same restaurant.

- II. "I went to have a cup of coffee and grilled sandwich. The experience was painful. We were served a cold sandwich. It didn't appear fresh. Coffee didn't taste even as good as home. The toilet was worse than a public toilet. Flush didn't work. It wasn't mopped and waiters were wearing a dirty white dress. Wouldn't recommend."

Duh! This customer had a miserable experience. The restaurant will have to do miles to appease this customer. Now, for a moment think if you have 1 million such reviews and you have to read each of them to understand the customer's experience with your product or service. Just imagining it makes me tired. Also, we have to be cognizant that this is for only 1 brand 1 product. There are similar reviews for multiple brands and products of different regions. The good thing is that we do not need to dwell on this manually; we have sentiment analysis to our rescue. Sentiment analysis can extract the sentiments of the reviews. For example, review [1] will be given a score based on the intent of the customer.

**Score : {Pos – 0.28, Neu – 0.7, Neg – 0.02}**

This means that the review is 28% positive, 70% neutral, and 2% negative. Thus, at an overall level, we can say that the review is neutral-to-positive. Similarly, sentiment analysis is a quick utility method that can provide a general view of customers about a brand that can provide us with the overall perception of the market. The objective of our work is to build analytics capability in the system to gauge the overall sentiment of customers toward various social media.

The rest of the paper is organized as follows: in Sect. 2 different sentiment analysis techniques are defined, in Sect. 3 various works done previously by researchers are described, in Sect. 4 detailed proposed method is given which has six determined steps to apply the proposed method to machine learning, in Sect. 5 results are evaluated and discussed with the help of graphs and charts, and finally in Sect. 6 we concluded our work with future research direction of work; last, we listed all the papers we cited in our work.

## 2 Sentiment Analysis Methods

Opinion Mining and Sentiment Analysis are significant for deciding feelings on brands and administrations or understanding buyers' perspectives. Given the steady course of data on the Internet, somewhat recently the field of consequently extricating assessments has arisen, being impractical to stay aware of the progression of new data by manual techniques [1]. There is an enormous assortment of work on Opinion Mining for English, not intended for Italian, via programmed implies [2, 3]. All around the world, two strategies are utilized: Supervised Machine Learning [3, 4] and Unsupervised techniques, which utilize a vocabulary with words scored for extremity esteems as impartial, positive, or negative [5]. Managed strategies require a preparation set of writings with physically allowed extremity esteems and, from these models, they get familiar with the elements (for example, words) that relate with the worth. Chaovalit and Zhou [6] assessed normal executions for the two strategies on film surveys and presumed that Supervised procedures perform with about 85% exactness, though Unsupervised techniques perform about 77%. Regulated methods have the weakness that they require excellent preparation information for each sort of report, for every area and every language. Unaided frameworks are more vigorous across various sorts of writings and spaces and, when the lexical and semantic assets are created, can be conveyed all the more without any problem.

Other than the computational procedure that is utilized for Opinion Mining, there is an entire range of issues that assume a part in the quality and convenience of the assessment extraction. Most importantly, assessment mining can be applied to various degrees of text: words, phrases, sentences, passages, or records. Words, as the littlest units, can have various polarities in various implications (for example, 'star' which can be level-headed as an eminent body or positive) as well as in various areas (for example, 'eccentric' is useful for a film plot and awful for a vehicle). This requires word sense disambiguation of words in setting and space or theme location as earlier preparing [7]. Moreover, extremity communicated by a word might be switched inside an expression through refutation. Additionally, portions of a report might communicate various polarities. Truth be told, suppositions can be identified with themes (what's going on with the assessment) or related to various assessment holders.

**VADER sentiment analysis**—Valence Aware Dictionary and Sentiment Reasoner (VADER) is a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is used for sentiment analysis of text which has both polarities, i.e. positive/negative.

**Sentiment Analyzer (Naïve Bayes)**—A Sentiment Analyzer is a tool to implement and facilitate Sentiment Analysis tasks using NLTK features and classifiers, especially for teaching and demonstrative purposes.

### 3 Literature Review

With the online social media platforms, for example, Twitter, Facebook, Instagram, and WhatsApp, overwhelming the correspondence world, it has gotten basic that the information dwelling across these online media stages will pass on shrewd data about the assessment, state of mind, and conclusion of individuals over any item, thought, or strategy. A few works have been performed before to investigate Twitter substance and perform assessment mining over Twitter information. The creators of [8] have proposed a methodology that utilizes a profound convolution neural organization to dissect the Twitter channel. The list of capabilities was incorporated into profound CNN for preparing and foreseeing assessments by investigating Twitter information. With regards to Twitter information examination, text pre-preparing plays a huge part to play. A few book pre-preparing strategies have been investigated and looked at as a component of [9]. Since assessment investigation includes dissecting and deciphering various sorts of suppositions, multiclass conclusion examination is exceptionally fundamental and is talked about in [10]. Creators of [11] have created a device named SENTA that utilizes an example examination way to deal with and perform multiclass slant investigation. As examined, [12] is to comprehend the developing changes of assessment mining, to the unique circumstance and the powerful occasions that happen during the Twitter discussions.

The use of Twitter has aroused more exploration run after understanding the conclusions utilizing Twitter information. One such work talked about in [13] utilizes a crossover structure that utilizes a hereditary calculation-based way to deal with performing supposition examination. The focal point of this work was to upgrade the framework according to an adaptability viewpoint. Creators of [14] have investigated the varieties of public feeling over a given subject utilizing a numerical model to distinguish the closer viewpoints and advance the compelling positioning of competitors. An intriguing business related to wistful examination is to order feeling dependent on the subject of conversation.

Creators of [15] have given a viable theme and versatile estimation arrangement over tweets. The unique challenges presented while performing multiclass assumption investigation have been talked about in [16], and the creators have additionally fostered a clever model that utilizes multiclass supposition examination over Twitter information. Creators of [17] have proposed a compelling content-to-discourse transformation dependent on the sentences given in Twitter information. The creators of [18] have presented a clever strategy for various leveled themes displaying utilizing Twitter information to perform Online Analytical Processing (OLAP). To improve the adequacy of questions identified with the examination, a huge OLAP information base, for example, Vertica [19], Greenplum [20], and Teradata DB [14], has been recommended. Vertica [19] uses projection to improve the presentation of the question. Rather than creating records that are customary on segments, it holds the insights concerning min/max ranges, prompting higher idleness from lower pruning which is less proficient. Greenplum [20] and Teradata DB [21] endorse store section astute and permit determination of ordering in a segment by clients. Additionally, there

exist two downsides: at first, alteration of segment files in composes way which is exorbitant for each segment records; also, it requires more arbitrary I/Os for inquiries identifying with point-query.

From the above works, it tends to be effortlessly perceived that the substance of Twitter information gives valuable bits of knowledge about any point being talked about and passes on the assessment of individuals required over the specific theme. The fundamental benefit and the exceptional usefulness of AI calculations are extricating the embodiment of the given issue. It gives total data about the information and makes the designer ponder picking the proper calculation which can be utilized for the issue. A portion of the normal classes of AI issues are bunching, characterization, relapse, and rule extraction.

## 4 Proposed Method

We proposed a machine learning-based method for customer sentiment analysis of social media data for which we collect Customer review data from the Google Play Store of WhatsApp and Twitter with the help of various Python scripts. The most relevant 3000 reviews will be used for the analysis. This Data contains

1. Username—the name of the customer who has given a review.
2. Content—the review of the customer on Social media.
3. Score—the rating given by the customer to Social media.
4. thumbsUpCount—number of likes on this review by other customers.
5. At—time of sharing the review.

The process flow of our method is as shown in Fig. 1.

### 4.1 Data Extraction

Data is fetched from the Google Play Store with the help of Google play scrapper. Below is the sample data table.

Features:

- (i) The app is 'com. whatsapp, com.Twitter'.
- (ii) Sorted by Most relevant.
- (iii) 3000 reviews extracted.
- (iv) Language of review is 'English'.
- (v) Country of review is 'US'.



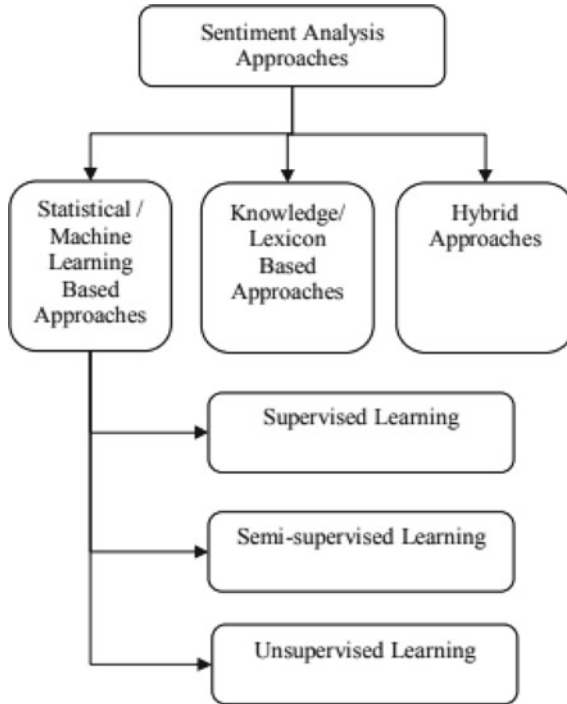


Fig. 1 Various approaches of sentiment analysis



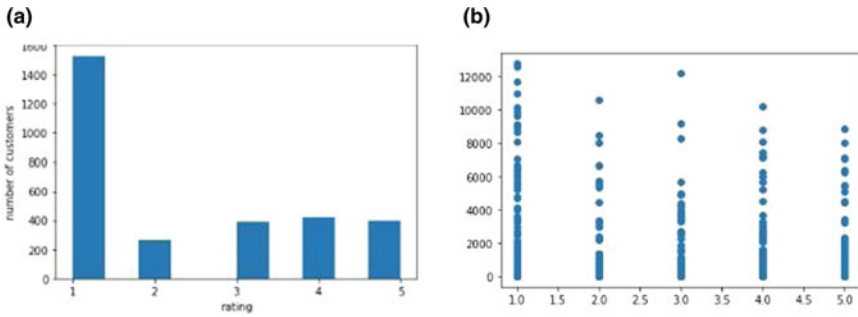
Fig. 2 Process flow of the proposed method

### 4.2 Data Preparation

The columns that are not required have been removed. Also, nulls are removed. Below is the sample output table with all necessary columns (Fig. 2).

### 4.3 Label Determination

To build a classification model on WhatsApp reviews, we will require predefined review labels such as ‘positive’, ‘negative’, or ‘neutral’. Since we do not directly get the tags, we will use ‘score’ and ‘thumbsUpCount’ to determine the labels.



**Fig. 3** a Distribution of ratings. b Distribution of thumbsups by ratings

Since most customers give a high rating to indicate a good service, an average rating to indicate satisfactory service, and less rating to indicate poor service, ‘rating’ will be an appropriate measure to label reviews. Also, more thumbs up on a review indicate a shared belief of other customers toward this review, hence we will use this measure to give weightage to ratings.

Checking Distribution of Ratings and ThumbsUps

From the above distributions, we observe that more customers have given a low rating to WhatsApp and a significant number of low ratings have approvals in the form of ‘thumbsup’ from other customers as well. Thus, to determine the label for classifying reviews as ‘positive’, ‘negative’, and ‘neutral’, we will do the following (Fig. 3):

1. All the reviews with a rating equal to 1 will be given a ‘negative’ label.
2. All the reviews with a rating equal to 3 will be given a ‘neutral’ label.
3. All the reviews with a rating equal to 5 will be given a ‘positive’ label.
4. For all the reviews with a rating equal to 2 or 4, we will give
  - a. ‘neutral’ label if a rating is 2 and thumbsup is 0.
  - b. ‘negative’ label if a rating is 2 and thumbsup is non-0.
  - c. ‘positive’ label if a rating is 4 and thumbsup is non-0.
  - d. ‘neutral’ label if a rating is 4 and thumbsup is 0.

### 4.4 Text Preprocessing

The following text preprocessing was performed on the reviews to make it a more digestible form so that machine learning algorithms can perform better.

1. Lower case conversion.
2. Remove stopwords and pronouns.
3. Special character removal.
4. Tokenization.

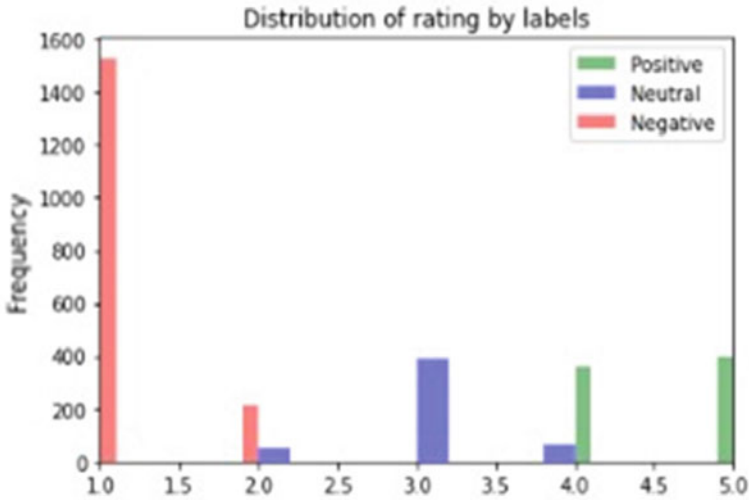


Fig. 4 Distribution of final labels across ratings

Mark Negation was also performed using sentiment analysis. ‘NEG’ subscript was added where the word had a negative connotation in the sentence of the review.

[‘please’, ‘add’, ‘feature’, ‘please’, ‘wish’, ‘guys’, ‘could’, ‘add’, ‘feature’, ‘phone’, ‘number’, ‘sent’, ‘could’, ‘click’, ‘call’, ‘it’, ‘like’, ‘ios’, ‘type’, ‘frustrating’, ‘im’, ‘point’, ‘im’, ‘considering’, ‘uninstalling’, ‘app’, ‘feature’, ‘isnt’, ‘app\_NEG’, ‘think\_NEG’] (Fig. 4).

### 4.5 Model Building

- **Train-Test split**

Train Data was divided into 70% training set and 30% test set.

- **Feature selection**

Training and Test reviews were divided into unigram feature sets where each word was a feature.

- **Model Fitting**

To classify the reviews as ‘positive’, ‘negative’, or ‘neutral’, we have used the Naïve Bayes classifier on the sentiment analyzer. It will use the unigram tokens to train the model.

## 5 Results and Discussion

The trained Naïve Bayes sentiment classifier model was tested on the test set. The model was able to successfully classify accurately 99.4% of the test set. The precision of negative reviews was much higher than the positive reviews (Figs. 5 and 6) (Tables 1 and 2).

Below are the 20 most informative words of this model. Words like ‘uninstall’, ‘worst’, ‘signal’, and ‘scary’ are more indicative of a negative review, whereas words like ‘individual’, ‘admin’, ‘suggestions’, and ‘helps’ are indicative of positive reviews. Similarly, neutral sentiment words are ‘improved’, ‘implement’, ‘emojis’, and ‘custom’ (Fig. 7).

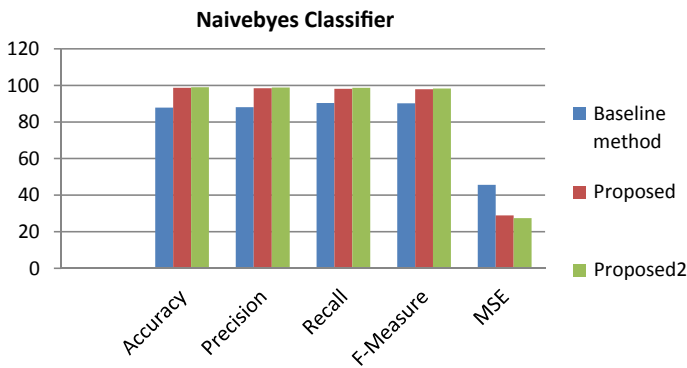


Fig. 5 Results of Naive Bayes classifiers

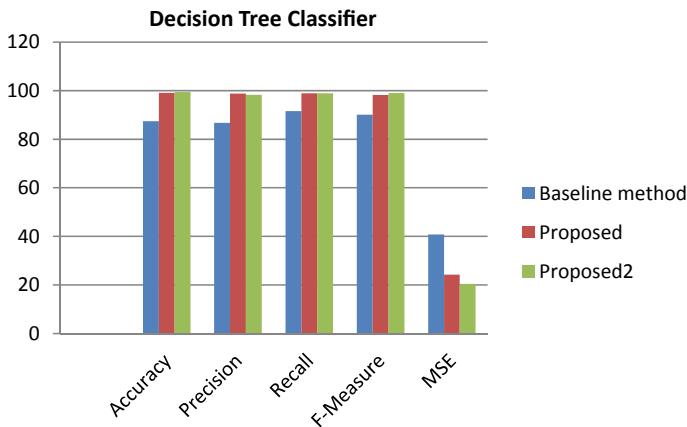


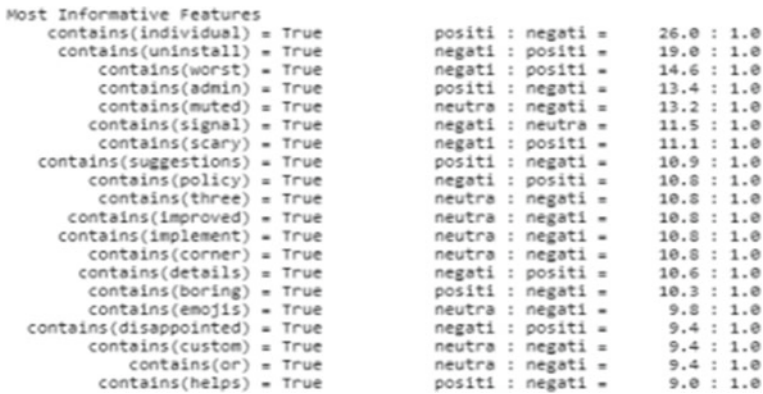
Fig. 6 Results of decision tree classifier

**Table 1** Evaluation results of Naive Bayes classifier

	Baseline method	Proposed (for WhatsApp)	Proposed (for Twitter)
Accuracy	87.9	98.7	99.01
Precision	88.1	98.4	98.9
Recall	90.4	98.1	98.7
F-Measure	90.2	97.9	98.25
MSE	45.6	28.9	27.4

**Table 2** Evaluation results of decision tree

	Baseline method	Proposed (for WhatsApp)	Proposed (for Twitter)
Accuracy	87.4	99.1	99.4
Precision	86.7	98.8	98.3
Recall	91.6	98.9	98.9
F-Measure	90.1	98.2	99.01
MSE	40.78	24.2	20.4



**Fig. 7** Top 20 informative words

## 6 Conclusion

We have successfully built a sentiment classifier model to determine the sentiment of customer reviews on WhatsApp with 99.4% accuracy. Most importantly, we have achieved 80% precision in classifying the negative reviews that require immediate attention from the business so that they can act quicker.

## References

1. Sentiment analysis symposium 2011, New York, 12 Apr 2011
2. Pang B, Lee L (2008) Opinion mining and sentiment analysis. *Found Trends Inf Retr* 2(1–2):1–135
3. Pang B, Lee L, Vaithyanathan S (2002) Thumbs up? sentiment classification using machine learning techniques. In: *Proceedings of the ACL-02 conference on empirical methods in natural language processing*, vol 10, pp 79–86
4. Socher R et al Semi-supervised recursive autoencoders for predicting sentiment distributions. In: *Proceedings of EMNLP '11 - the conference on empirical methods in natural language processing*. ISBN: 978-1-937284-11-4, pp 151–161
5. Taboada M, Brooke J, Tofiloski M, Voll K, Stede M (2011) Lexicon-based methods for sentiment analysis. *Comput Linguist* 37(2):267–307
6. Chaovalit P, Zhou L (2005) Movie review mining: a comparison between supervised and unsupervised classification approaches. In: *Proceedings of the Hawaii international conference on system sciences (HICSS)*
7. Esuli A, Sebastiani F (2006) SentiWordNet: a publicly available lexical resource for opinion mining. In: *Proceedings of LREC-2006*, Genova, Italy
8. Jianqiang Z, Xiaolin G, Xuejun Z (2018) Deep convolution neural networks for Twitter sentiment analysis. *IEEE Access* 6:23253–23260. <https://doi.org/10.1109/ACCESS.2017.2776930>
9. Jianqiang Z, Xiaolin G (2017) Comparison research on text preprocessing methods on Twitter sentiment analysis. *IEEE Access* 5:2870–2879. <https://doi.org/10.1109/ACCESS.2017.2672677>
10. Bouazizi M, Ohtsuki T (2018) Multi-class sentiment analysis in Twitter: what if classification is not the answer. *IEEE Access* 6:64486–64502. <https://doi.org/10.1109/ACCESS.2018.2876674>
11. Bouazizi M, Ohtsuki T (2017) A pattern-based approach for multiclass sentiment analysis in Twitter. *IEEE Access* 5:20617–20639. <https://doi.org/10.1109/ACCESS.2017.2740982>
12. Ebrahimi M, Yazdavar AH, Sheth A (2017) Challenges of sentiment analysis for dynamic events. *IEEE Intell Syst* 32(5):70–75. <https://doi.org/10.1109/MIS.2017.3711649>
13. Iqbal F et al (2019) A hybrid framework for sentiment analysis using genetic algorithm based feature reduction. *IEEE Access* 7:14637–14652. <https://doi.org/10.1109/ACCESS.2019.2892852>
14. Tan S et al (2014) Interpreting the public sentiment variations on Twitter. *IEEE Trans Knowl Data Eng* 26(5):1158–1170. <https://doi.org/10.1109/TKDE.2013.116>
15. Liu S, Cheng X, Li F, Li F (2015) TASC:topic-adaptive sentiment classification on dynamic tweets. *IEEE Trans Knowl Data Eng* 27(6):1696–1709. <https://doi.org/10.1109/TKDE.2014.2382600>
16. Bouazizi M, Ohtsuki T (2019) Multi-class sentiment analysis on twitter: classification performance and challenges. *Big Data Min Anal* 2(3):181–194. <https://doi.org/10.26599/BDMA.2019.9020002>
17. Trilla A, Alias F (2013) Sentence-based sentiment analysis for expressive text-to-speech. *IEEE Trans Audio Speech Lang Process* 21(2):223–233. <https://doi.org/10.1109/TASL.2012.2217129>
18. Yu D, Xu D, Wang D, Ni Z (2019) Hierarchical topic modeling of Twitter data for online analytical processing. *IEEE Access* 7:12373–12385. <https://doi.org/10.1109/ACCESS.2019.2891902>
19. Lamb AF, Varadarajan M, Tran R, Vandier N, Doshi BL, Bear C (2012) The vertica analytic database: C-store 7 years later. arXiv:1208.4173
20. Greenplum Database. [greenplum.org/](http://greenplum.org/)
21. Solutions TW (2002) Teradata Database technical overview, pp 1–7. <http://www.teradata.com/brochures/Teradata-Solution-Technical-Overview-eb3025>

# A Novel Hybrid Approach for the Designing and Implementation of Dogri Spell Checker



Shubhnandan S. Jamwal and Parul Gupta

**Abstract** A spell checker is a software program that highlights words in a text that may be incorrect. A spell checker is an essential feature of every word processor, regardless of the language. The spell checker examines the written text for misspellings and provides the best right suggestions for such misspellings. In this paper, first ever attempt has been made for the designing and implementation of the Dogri spell checker using a hybrid approach. The principal components of this system are error detection, error rectification by producing suggestions, and replacement by manually or automatically using the proposed methodology. The algorithm correctly suggests 74.72% of erroneous words and finds 80.79% of misspelled words.

**Keywords** Error correction · Error detection · Dogri spell checker · Minimum edit distance · Morphology · Dogri

## 1 Introduction

Most of the researchers working on the Dogri language find difficulty in text processing application of NLP because of the non-availability of tools. The development of the spell check can help the researchers and other people working on the Dogri language who might be unable to write effortlessly and correctly. The computerization of Dogri language processing has been initiated recently, and it is still in the novice stage and suffers from inadequacy of resources; therefore, a spell checker can be a very important tool which can be used for finding and rectifying several spelling errors. It is also proposed that in Natural Language Processing (NLP), finding and rectifying errors in the words has been extensively used in

---

S. S. Jamwal · P. Gupta (✉)

Department of Computer Sciences and IT, University of Jammu, Jammu, India  
e-mail: [parul.gupta@jammuuniversity.ac.in](mailto:parul.gupta@jammuuniversity.ac.in)

S. S. Jamwal

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

normalizing data. Spell error finding and rectifying can be very useful if it is executed by machines because humans may not be able to find spelling errors and then rectify them by themselves.

Several NLP tasks neglect the challenge of text normalization in a code-mixed social media text, and many academics propose approaches for detecting spelling problems in social media code-mixed text. Most of the netizens on social media are using English words with Romanized transliteration of the Dogri language. Dogri language has no longer any full-fledged spell checker which is available for regular use like many different Indian and foreign languages. There are different types of researches which are available on spell checkers for Indian languages and many other European languages. However, low-resourced languages like Dogri have limited research in developing spell checkers may be because of their highly inflected and morphologically rich nature.

## 2 Literature Review

Hasan et al. [1] proposed Bengali speech facts via way of means of using ‘Deep Speech’, which produces a neural network for spotting audio documents in Bengali and later remodel the speech into its textual content format. It has been found that integrating the aforesaid method with a Bengali spell corrector improves its performance and correctness. Roy and Ali [2] utilized cosine similarity to choose the best choices for misspelled words and used an unsupervised contextualized spelling correction and detection method for conventional Bangla text. Fast-text embedding concept of using character n-gram embedding has been used for the experimental setup. Unknown words are generated by using the character n-gram embedding. Kaur et al. [3] together with others proposed a hybrid method to develop a Punjabi spell checker. Traditional methods for error detection, such as dictionary lookup algorithms and minimum edit distance for error correction, have also been discovered in current spell checkers. They suggested a tree data structure for storing local language words in a dictionary, and then utilised this tree-based method, in conjunction with n-gram analysis, to detect misspelled words. The use of recurrent neural networks and a hybrid method based on the rule and minimum edit distance have been used to correct misspelled words. Zaky and Romadhony [4] proposed an LSTM model in which the input word is encoded at the character level, with the help of Parts of Speech tag context surroundings as features. The experiment was conducted using Indonesian Wikipedia articles as an artificial dataset. These errors are generated by simulating some artificial spelling errors at the character level and testing them on a real dataset, with an accuracy rate of 83.76%. Ismail and Rahman [5] developed Bangla word clusters by using an unsupervised ML. These words are based on their semantic and contextual similarity using the N-gram language model. Based on the successful application of the N-gram model to English word clustering, the work of Bangla is taken into account. Dutta et al. [6] developed a CRF-based machine learning method for word level identification task and spell



checking tasks in English, reaching an accuracy of 90.5%, while the spelling checker reached 69.4% only. Uthayamoorthy et al. [7] developed the spell checker and suggestion generator and called it as DDSpell for the Tamil language. It was created utilizing a language-independent and data-driven methodology. During the project, 4.0 million Tamil words from diverse sources are utilised to check the spelling. To provide suggestions, bi-gram similarity matching at the character level, minimal edit distance metrics, and word frequencies were employed, and DDSpell performed better than other available tools. Many systems employ spell checkers, which include natural language processing, data mining, machine learning, and appropriate data processing techniques, Jo et al. [8]. In addition, Azmi et al. [9] suggested an Arabic spell checker to identify and repair error words. Along with machine learning, the technique suggested a word and stem n-gram  $n = 1-3$  language model. Grobbelaar and Kinyua [10] presented a multithreaded, spell checking, and correcting software for South Sotho typed text in the form of the windows application program called eSpellingPro. Joshi et al. [11] have developed a tool that converts Romanized Hindi to regular Hindi text. Regarding the research on the Dogri language, a hybrid model generated by the Dogri language has been proposed [12, 13].

### 3 Errors Classification in Dogri

The grammar of a language and its syntax are determined by the words that are joined in meaningful ways. The semantics of a language determines the real meaning of words and their combinations. Spell checking can be defined as the method of detecting misspelled words in written text and providing correct suggestions. Optical character recognition (OCR), text editors, word processors, search engines, and other word processing tools all have spell correction as one of their primary functions. The spell checker is composed of error detection, suggestion generator, and error correction module.

The initial step is to discover mistakes in the input text. This section employs a language model that takes into consideration the words that can be used in the language. The second part deals with generating possible suggestions of possible correct candidates for the wrong words. The final step is to correct any spelling errors made by the user either manually by selecting the appropriate candidate word from the suggestion list or automatically by the system. To discover alternative replacements for misspelled words, an error model is required. This part also includes the ranking of replacement candidates. Ranking can be based on edit distance, string similarity metric, phonetic metric, or word frequency.

In the realm of language processing, error correction is a critical issue. Many studies have been conducted in this field throughout the years. It's important to consider how errors in spelling happen before understanding error detection and correction. Error detection and error correction techniques are proposed based on the type of error words.

Based on existing studies, spelling errors are bifurcated into two classes: Non-word errors and real-word errors [3]. Non-word errors are misspellings that aren't in the dictionary. Example i.e. पुच्छेअ for पुच्छेआ (Ask). Non-word errors, such as typographical errors [13], occur when we know the proper spelling of a word, but the term is typed wrongly. Incorrect typing are the major cause of these problems. For example, typing दसामी for दसमी (Tenth). The term "real-word errors" refers to words that are permissible in the dictionary but are wrong in the context of the phrase. For example, कलम देश दा राष्ट्रीय फूल है (incorrect) for कमल देश दा राष्ट्रीय फूल है (correct), the term कलम is a correct word in the dictionary, however it appears as a misspelling of कमल in the preceding phrases.

## 4 Proposed Architecture

There is some work in progress in the area of Dogri spell detection and correction, and identifying errors in Dogri text is not easy. Figure 2 depicts the blueprint of the proposed spell checker system. Three modules make up the whole system: one is the preprocessing module, which is used to prepare the resources required by the system; the other is the error detection module, which marks the misspelled words in a given document; and the last is the error correction module, which generates, ranks the candidates' suggestions by wrongly spelled words and automatically replacing the incorrect words.

Our system takes Dogri's text as input, marks the text, and recognizes errors when looking up that specific word in the Dogri dictionary. It generates a list of candidate suggestions for that misspelled word. The user will get to choose a suggestion from the list and modify the error appropriately. Automatic replacement of the error word requires the ranking of the suggestions that use n-gram probabilities which in return the top k corrections. Every n-gram ( $n = 2, 3, 4$ ) of the sentence is checked with every k suggestion and based upon their frequency it is replaced accordingly. The end result is a corrected text with no spelling errors.

### 4.1 Preprocessing Module

The preprocessing module is in charge of generating the system's resources. It serves two key purposes: To begin, the system tokenized the phrase, segregating words from punctuation marks. For the Dogri language, we use information from the writing system and language recognition charts to create a list of supported characters. In this writing system, we include capital and lowercase letters (if appropriate) as well as numerals while omitting any punctuation. Any character that does not appear on this list is assumed to be a foreign character in the language and will be noted as a separate mark. Using regex expression rules, we extract all consecutive character sequences in the support list. The other function is that the

module will also use a built-in stemming system [14] that is particularly designed to handle Dogri linguistic phenomena while creating the language model. Stemming reduces the influence of affixes, because of the language's high morphology. Considering the extremely dynamic nature of Dogri, the module is capable of retraining the system's resources. This ensures that the system can use the most up-to-date lexicon for the language.

## ***4.2 Error Detection Module***

Initially the approach will convert the sentence into a word, separates the word from the punctuation mark, and then sends it to the error detection module. This module uses a dictionary look-up to determine if the tokenized word exists in the language. The dictionary lookup approach will detect the error in the text entered by the user by checking every word of the input in the Dogri lexicon. It is a correct word if the word is found; otherwise, it is considered a wrong word. In the event that the dictionary search is failed to find a word, the word will further undergo stemming [14] and the dictionary look-up method will be executed again.

If the dictionary lookup and stemming approaches fail to locate the word, the module's n-gm lookup technique will be utilized. The word stem's character trigram sequences are created using this method. This module checks if every trigram of the root word exists in the corpus. And, when the trigram does not exist, mark the wrong word.

For the system to be more convenient, we only consider wrongly spelled tokens of length more than two. In a manual analysis of the Dogri misspelling dataset, we found that misspellings of lengths one and two are meaningless, so it is illogical to calculate suggestions and rank them. To deal with real word errors, we created n-gram ( $n = 2, 3$ ) and their frequency. Corpus is used to check real word errors for every n-gram (where  $n = 2, 3, 4$ ) of the sentence. The possibility of error raises if the frequency of grams is less.

## ***4.3 Error Correction Module***

This module is used to create candidate suggestions for every wrongly spelled word that is detected using a Levenshtein's Minimum edit distance (LMED) computation. We build a list of all known words with an LMED of two and name them candidate suggestions, when given an unknown token. On an error word, LMED is used to produce plausible alternatives for that term. LMED transforms an erroneous word into one of the available right words during the fundamental editing operations of Insertion, Substitution, and Deletion. The distance between the error word and the dictionary words is calculated. The word in the dictionary that is closest to

**Table 1** Shows the suggestions and minimum edit distances

Misspelled word	Corresponding Suggestions	No. of Operations performed(LMED's)
अगड़	<ol style="list-style-type: none"> <li>1. अगड़ा</li> <li>2. अगर</li> <li>3. अगला</li> </ol>	Insertion(ा) (LMED =1) Substitution(र) (LMED =1) Substitute(ल),Insertion(ा) (LMED=2)
आ-गेई	<ol style="list-style-type: none"> <li>1. आई-गेई</li> </ol>	Insertion(ई) (LMED=1)

the misspelled word appears first in the list of suggestions. Table 1 depicts the possible suggestions and LMED of some error words.

Once all the feasible candidate suggestions are generated, the proposed module sorts the most likely suggestions and the corresponding candidate suggestions according to the LMED information of the misspelled words. We sort these candidate words in increasing order of edit distance in order to generate a rank. Words with the same edit distance are sorted according to the frequency of occurrence.

To prioritize suggestions with the same minimal edit distance, the statistical machine translation (SMT) approach is used. SMT is suitable for recommendations to search the most meaningful words in the suggestion list. For this method to work, at least three words are required as input. SMT compares the text entered with paragraphs stored in the Dogri corpus to determine the best appropriate recommendation in our proposed system. Prioritization is determined by substituting erroneous words for the suggestions depending on the preceding and subsequent terms. If the database paragraph has the right combination of them, it is advised that you use the recommendation as the most appropriate term and change it appropriately. For example, महाराज गुलाब सहि जम्मू और कश्मीर दे राजा हे. Here, गुलाब is an error word. It may be plausible that the right word will be गुलाब or गुलाम. By combining words like [गुलाब सहि जम्मू और कश्मीर दे] and [गुलाम सहि जम्मू और कश्मीर दे], SMT will priorities recommendation. That suggestion will be a suitable recommendation based on the word combination that occurs in the database. Suggestions with the greatest scores will be given a priority as the candidate correction for the wrongly spelled word.

**Algorithm**

1. **Procedure** Err\_Detection(dictionary, tokenization, stemmer, extraction, Dogri\_corpus)
2. **Variables:**
3. Txfl: Textfile.txt
4. Snt  $\leftarrow$  **extraction**(Txfl)
5. Tkn  $\leftarrow$  **tokenization**(Txfl)
6. stmTkn  $\leftarrow$  **stemmer**(Tkn)
7. FltrTkn  $\leftarrow$  **regex**(Tkn)
8. Result1, Result2, Result3: String
9. **Begin:**
10. Snt [j]  $\leftarrow$  **extraction**(Txfl)
11. While Snt[j]  $\neq$  EOF do
12.     Tkn[i]  $\leftarrow$  **tokenization**(Snt[j])
13.     FltrTkn[k]  $\leftarrow$  **regex**(Tkn[i])
14.     While k < FltrTkn.len do
15.         Result1= FltrTkn[k].**Dictionary\_lookup**(dictionary, Binary\_Search)
16.         If (**Not\_found**(Result1))
17.             stmTkn[k]  $\leftarrow$  **stemmer**(FltrTkn[k])
18.             Result2=stmTkn[k].**Dictionary\_lookup**(dictionary, Binary\_Search)
19.             If (**Not\_found**(Result2))
20.                 Result3=stmTkn[k].**n-gram\_lookup**(Dogri\_corpus)
21.                 If (**Not\_found**(Result3))
22.                     Err\_tkn[q]=**Mark\_as\_error**(stmTkn[k])
23.             Else
24.                 Txfl.**Update\_Dictionary\_Context**()
25. **end:**

**Algorithm**

1. **Procedure** Err\_Correction(dictionary, extraction, Dogri\_corpus, Err\_tkn[q])
2. **Variables:**
3.     Min\_distance  $\leftarrow$  2
4.     n-gram\_value  $\leftarrow$  3
5.     Err\_tkn[q]  $\leftarrow$  **extraction**(Err\_tkn[q])
6.     Cnd\_sugg[p]  $\leftarrow$  Err\_tkn[q].**LMED\_Calculation**(dictionary, Min\_distance)
7.     Rnk\_Cnd\_sugg[p]  $\leftarrow$  **Ranking**(Cnd\_sugg[p])
8.     Replc[q]  $\leftarrow$  Err\_tkn[q].**n-gram\_approach**(Rnk\_Cnd\_sugg[p], n-gram\_value)
9.     Result1, Result2, Result3: String

**10. Begin:**

11. Err\_tkn[q] ← **extraction**(Err\_tkn[q])
12. **While** Err\_tkn[q] ≠ '\0' **do**
13. Cnd\_sugg[p] ← Err\_tkn[q].**LMED\_Calculation**(dictionary  
Min\_distance)
14. Rnk\_Cnd\_sugg[p] ← **Ranking**(Cnd\_sugg[p])
15. **If**( Option\_choose.**Automatic**(YES))
16. Replc[q] ← Err\_tkn[q].**n-gram\_approach**(Rnk\_Cnd\_sugg[p], n  
gram\_value)
17. **Else**
18. Sugg\_list.**Display**(Rnk\_Cnd\_sugg[p] )
19. p ← p+1
20. q ← q+1
21. **end:**

## 5 Implementation

The proposed spell checker for the Dogri language is shown in Fig. 1. The proposed architecture is implemented in node js using JSON as the backend. The user gives the input in the form of Dogri text in the textbox or by uploading the text file. The next step is to click the spell check button in order to identify the error words in the input text. The proposed Dogri spell checker retrieves tokens from the input text one by one and compares it with a dictionary using binary search. The occurrence of a word in the dictionary depicts its correction, otherwise it is considered as misspelled word. Error words are added to the error list and also displayed in red color within the text in the input textbox as shown in the figure. When a user clicks an error word from the error word list, possible suggestions are displayed in the suggestion list. If the most expected suggestion is discovered in the suggestion list, the user will replace the error word by picking the most desired suggestion from the list. Or the other option is also available, the error word will automatically replace by the candidate suggestion based on the ranking by n-gram lookup approach using the surrounding context. Similarly users can perform tasks like reset, upload, download, add to the dictionary, undo, redo, etc. (Fig. 2).

## 6 Results and Discussions

The quality test dataset was the most difficult aspect of the spell checker evaluation. The majority of publicly accessible datasets are for English and a few of Indian languages, but the Dogri dataset is still not available anywhere on the web. To

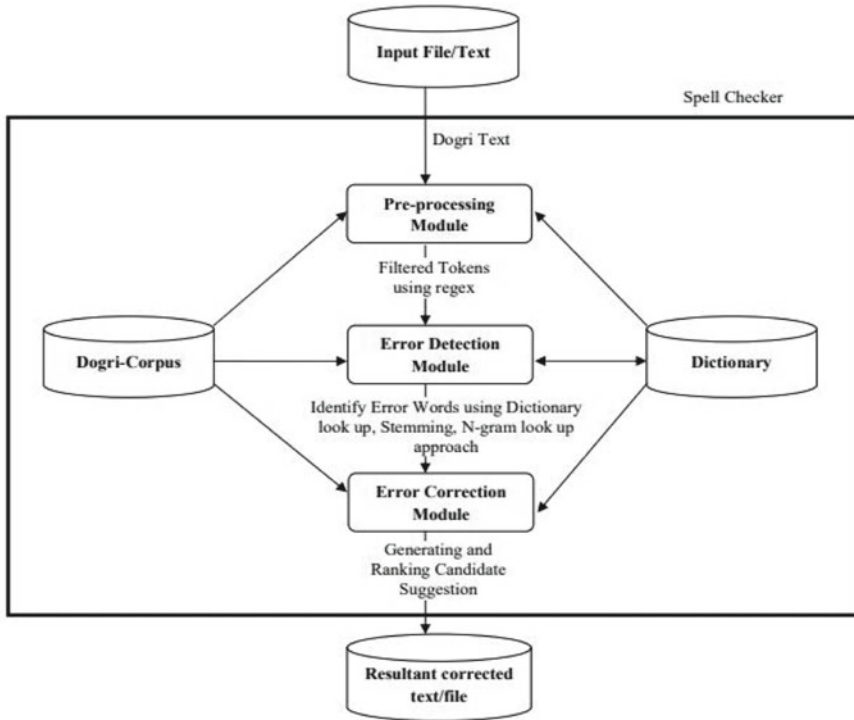


Fig. 1 Data flow of propose Dogri spell checker

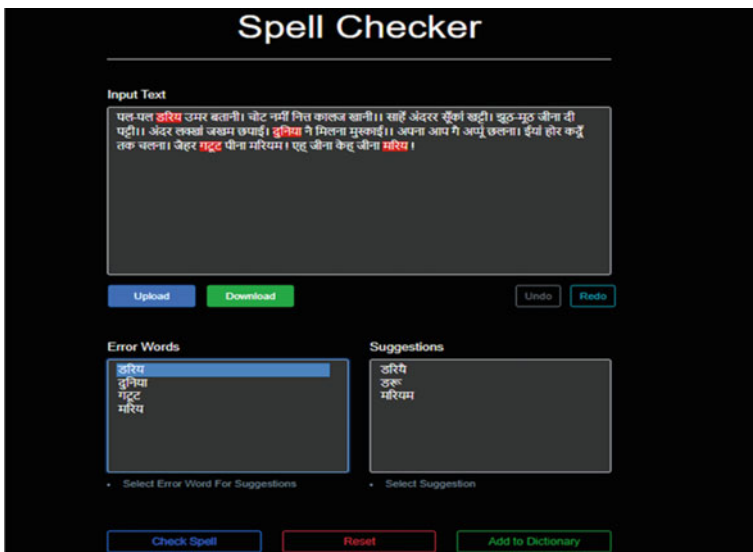


Fig. 2 Graphical user interface of proposed spell checker

**Table 2** Shows the experimental result

Dataset	Total error words	Detected error words	Intended word in suggestion list	Detection rate (%)	Correction rate (%)
Test-set-1	102	81	75	79.42	73.52
Test-set-2	256	206	191	80.46	74.61
Test-set-3	343	279	257	81.34	74.92
Test-set-4	389	312	289	80.21	74.3
Test-set-5	480	396	366	82.5	76.25
Average	314	255	236	80.79	74.72

illustrate noisy channels, we propose three techniques for introducing typographical errors in the right words. We took all of the sentences in which we couldn't detect a spelling mistake and added more than one error per sentence. To test the algorithm, we manually created 1570 misspell terms from data acquired from books, research theses, and individuals, among other sources. There are 1274 words recognized as incorrect words in the result analysis, while the algorithm gives accurate suggestions for 1178 words. The average detection rate of the system is nearly 80.79% and its correction rate is around 74.72%. The table shows the results of different data sets (Table 2).

Accuracy is determined by the number of editing operations required to convert an incorrect word to a correct word, as well as the length of the letters. It is observed from the table above that there are slight variations in the results; this is due to the domain of the test data collected. Another aspect is that the system's ability to capture morphological phenomena is dependent on the stemmer and resources employed. Finally, the system's candidate suggestions for the misspelled word are based on linguistic resources.

## 7 Conclusion

This paper introduces the Dogri spell checker, which is not included in any application program like word processor, text editor, etc. or website. The first attempt at developing a spell checker for a low-resource morphologically rich language like Dogri using hybrid ML methods has been performed. Using 1570 misspelled words manually made from the exercise documents of Dogri research scholars, the spell checker system has an average detection rate of 80.79% and a correction rate of 74.72% and the algorithm was able to properly identify a wide range of frequent errors in Dogri. As there is a very small amount of digitized data of Dogri Language and machine learning algorithms available till now, the proposed spell checker will be helpful for them in developing various natural language processing systems. In the future, better use of the larger corpus would contribute much to improve the spell checker accuracy.



## References

1. Hasan HMM, Islam MA, Hasan MT, Hasan MA, Rumman SI, Shakib MN (2020) A spell-checker integrated machine learning based solution for speech to text conversion. In: 2020 third international conference on smart systems and inventive technology (ICSSIT), pp 1124–1130. <https://doi.org/10.1109/ICSSIT48917.2020.9214205>
2. Roy S, Ali FB (2019) Unsupervised context-sensitive Bangla spelling correction with character N-gram. In: 2019 22nd international conference on computer and information technology (ICCIT), pp 1–6. <https://doi.org/10.1109/ICCIT48885.2019.9038604>
3. Kaur G, Kaur K, Singh P (2019) Spell checker for Punjabi language using deep neural network. In: 2019 5th international conference on advanced computing & communication systems (ICACCS), pp 147–151. <https://doi.org/10.1109/ICACCS.2019.8728369>
4. Zaky D, Romadhony A (2019) An LSTM-based spell checker for Indonesian text. In: 2019 international conference of advanced informatics: concepts, theory and applications (ICAICTA), pp 1–6. <https://doi.org/10.1109/ICAICTA.2019.8904218>
5. Ismail S, Rahman MS (2014) Bangla word clustering based on N-gram language model. In: 2014 international conference on electrical engineering and information & communication technology, pp 1–5. <https://doi.org/10.1109/ICEEICT.2014.6919083>
6. Dutta S, Saha T, Banerjee S, Naskar SK (2015) Text normalization in code-mixed social media text. In: 2015 IEEE 2nd international conference on recent trends in information systems (ReTIS), pp 378–382. <https://doi.org/10.1109/ReTIS.2015.7232908>
7. Uthayamoorthy K, Kanthasamy K, Senthalaan T, Sarveswaran K, Dias G (2019) DDSpell - a data driven spell checker and suggestion generator for the Tamil language. In: 2019 19th international conference on advances in ICT for emerging regions (ICTer), pp 1–6. <https://doi.org/10.1109/ICTer48817.2019.9023698>
8. Jo S, Trummer I, Yu W, Wang X, Yu C, Liu D, Mehta N (2019) AggChecker: a fact-checking system for text summaries of relational data sets. *Proc VLDB Endow* 12(12):1938–1941. <https://doi.org/10.14778/3352063.3352104>
9. Azmi AM, Almutery MN, Aboalsamh HA (2019) Real-word errors in Arabic texts: a better algorithm for detection and correction. *IEEE/ACM Trans Audio, Speech Lang Proc* 27(8):1308–1320. <https://doi.org/10.1109/TASLP.2019.2918404>
10. Grobelaar LA, Kinyua JDM (2009) A spell checker and corrector for the native South African language, South Sotho. In: Proceedings of the 2009 annual conference of the Southern African computer lecturers' association (SACLA '09). Association for Computing Machinery, New York, NY, USA, pp 50–59. <https://doi.org/10.1145/1562741.1562747>
11. Joshi N, Mathur I, Mathur S (2010) Frequency based predictive input system for Hindi. In: Proceedings of the international conference and workshop on emerging trends in technology (ICWET '10). Association for Computing Machinery, New York, NY, USA, pp 690–693. <https://doi.org/10.1145/1741906.1742065>
12. Jamwal SS, Gupta P, Sen VS (2021) Hybrid model for generation of verbs of Dogri language. In: Singh TP, Tomar R, Choudhury T, Perumal T, Mahdi HF (eds) *Data driven approach towards disruptive technologies. Studies in autonomic, data-driven and industrial computing*. Springer, Singapore. [https://doi.org/10.1007/978-981-15-9873-9\\_39](https://doi.org/10.1007/978-981-15-9873-9_39)
13. Jamwal SS, Gupta P, Sen VS (2021) A novel approach for identification and classification of verbs in Dogri language. *Int J Intell Eng Inform* 9(4):412–423
14. Gupta P, Jamwal S (2021) Designing and development of stemmer of Dogri using unsupervised learning. In: Marriwala N, Tripathi CC, Jain S, Mathapathi S (eds) *Soft computing for intelligent systems. Algorithms for intelligent systems*. Springer, Singapore, pp 47–156